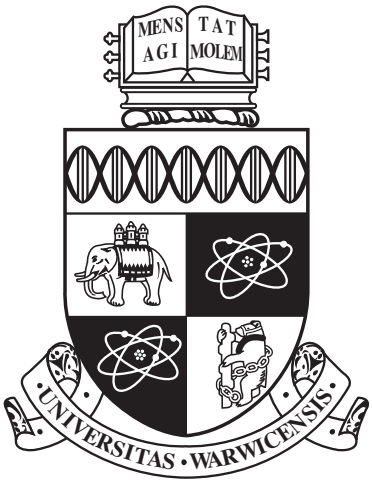

Bayesian inference for nonparametric hidden Markov models with applications to physiological data

by

Sida Chen

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

October 2021

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I owe a lot to my mum and dad, who supported me financially and mentally over my eight years of study at Warwick, including my PhD journey, without which this experience would not have been possible. I am incredibly grateful to them for always having full faith in me and encouraging me to pursue my interest and further my study. I am also indebted to my grandparents and other family members for their love and care over the years.

I would like to thank my PhD supervisor Prof Bärbel Finkenstädt Rand for offering me the opportunity to pursue a PhD and for her patience, constant encouragement and exceptional guidance over the past four years. It has been a pleasant and fruitful journey to work under her mentorship, where I was given so much freedom to discover and explore my research project and taught how to be a good researcher. I want to thank my personal tutor Dr Ric Crossman for his help and support thought my study and my PhD panel Prof Wilfrid Kendall and Dr Ritabrata Dutta, for their valuable feedback and suggestions regarding my research. A special thank you to my Master's supervisor Prof David Firth who inspired and supported me to further my research in Statistics. I also want to extend my gratitude to all other staff at the Warwick statistics department for creating such a vibrant and supportive environment from which I benefitted a lot.

I am also grateful to the Association of British Chinese Professors for offering me the bursary fund and to Prof Jihong Wang and Prof Qing Wang for their support on my application. I wish to acknowledge Dr Qi Huang, Sandra Komarzynski and Prof Francis Lévi for their support on the analysis of the human accelerometer data. Thanks also to Prof Roland Langrock for helpful research discussions and

feedback. Finally, I would like to thank my many friends and colleagues for making the research journey more enjoyable. I choose not to list everyone as they all count.

Declarations

I hereby confirm that this thesis is based on my original work unless stated otherwise. This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. Part of the work in chapters 2 and 3 formed the manuscript for "Bayesian inference for spline-based hidden Markov models", co-authored with my supervisor Prof Bärbel Finkenstädt Rand, and is currently under review. The oceanic whitetip shark data analysed in chapter 2 are kindly provided by Prof Roland Langrock, Dr Yannis Papastamatiou and Dr Yuuki Watanabe. The dichotomy $I < O$ and rhythm indices used in chapter 3 are kindly provided by Dr Qi Huang.

Abstract

This thesis develops new nonparametric Bayesian hidden Markov models (HMM) and estimation methods that address some of the challenges and limitations of existing nonparametric approaches. In chapter 2, we introduce for the first time a fully Bayesian method for inference in spline-based HMMs where the number of states may be unknown along with other model parameters including the knot configuration of the B-splines. Regarding the latter, we propose the use of a trans-dimensional Markov chain Monte Carlo (MCMC) algorithm, while model selection regarding the number of states can be achieved based on the estimated marginal likelihood. Our methodology compares favourably with existing competing methods in terms of estimation accuracy, stability and efficiency. We then extend the spline-based HMM proposed in chapter 2 to develop a novel hierarchical conditional HMM approach, which allows us to analyse the specific state of an HMM at a finer level with another sub-HMM, achieving inferences that are otherwise not possible with a single HMM. We apply the proposed method to human activity data from wearable devices where we can jointly identify and characterise sleep periods, an area of interest to sleep and circadian biology research. In the last part of the thesis, we exploit the strength of the hierarchical Dirichlet process and a suitable integration with HMMs to develop new Bayesian nonparametic multivariate HMMs. The resulting models allow for flexible yet parsimonious modelling of the emission distributions and automatic learning of the state cardinality, generalising existing models to offer greater modelling flexibility. We develop novel MCMC methods which combine the slice sampling technique and a dynamic programming algorithm for exact and efficient posterior inference. Finally, we apply our proposed models to motion and heart rate data collected from the Apple watch for learning human sleep dynamics in an unsupervised context.

Abbreviations

- AC: Acceleration
- AIC: Akaike information criterion
- BIC: Bayesian information criterion
- BNP: Bayesian nonparametric
- CDLL: Complete data log-likelihood
- CTS: Circadian timing system
- CRF: Chinese restaurant franchise
- CRP: Chinese restaurant process
- DIC: Deviance information criterion
- DP: Dirichlet process
- DPMM: Dirichlet process mixture model
- d-sHDP: Disentangled sticky HDP
- EM: Expectation-Maximization
- FFBS: Forward filtering backward sampling
- FPR: Forecast pseudo residuals
- GEM: Griffiths, Engen, and McCloskey

- HDP: Hierarchical Dirichlet process
- HDPM: HDP mixture model
- HMM: Hidden Markov model
- HHMM: Hierarchical HMM
- HPD: Highest posterior density
- HR: Heart rate
- HSMM: Hidden semi-Markov model
- iHMM: Infinite HMM
- IW: Intermittent wake
- KLD: Kullback-Leibler divergence
- LIDS: Locomotor Inactivity During Sleep
- MCMC: Markov chain Monte Carlo
- MESA: Multi-Ethnic Study of Atherosclerosis
- MFM: Mixture of finite mixture
- MH: Metropolis-Hastings
- ML: Maximum likelihood
- ODBA: Overall dynamic body acceleration
- lODBA: Log-transformed ODBA
- OPR: Ordinary pseudo residuals
- PA: Physical activity
- PSG: Polysomnography

- P-splines: Penalized B-splines
- REM: Rapid eye movement
- RJMCMC: Reversible jump MCMC
- SCN: Suprachiasmatic nucleus
- TST: Total sleep time
- WASO: Wake after sleep onset

Chapter 1

Introduction to hidden Markov models

Hidden Markov models (HMMs) are arguably one of the most important and popular class of time series models for extracting information from sequential data, which are central to many modern statistical and machine learning problems. Instead of directly modelling the relationship between consecutive observations, they explain the patterns in the data by introducing an additional latent structure, with which complex dependencies between the observations may be handled while retaining a relatively simple and interpretable modelling framework. Since the successful application in speech recognition [Rabiner, 1989], they have been useful in areas throughout applied sciences; a few examples are economics [Hamilton, 1989; Kim, 1994], finance [Rydén et al., 1998; Langrock et al., 2012b], pattern recognition [Epaillard and Bouguila, 2016; Nguyen et al., 2005], genomics [Yau et al., 2011], biophysics [Chen et al., 2016], medical sciences [Li et al., 2013] and ecology [McClintock et al., 2020]. We refer to Cappé et al. [2005], Dymarski [2011] and Zucchini et al. [2016] for illustrations of various successful applications of HMMs and an extended bibliography. This chapter is not intended to give an extensive review of HMMs but to introduce the core ideas and basic inference and learning algorithms which will lay the foundation for the development of the thesis.

1.1 The basic framework

We begin with a introduction to the HMM in its most basic form whereas extensions of the basic model will be introduced in later sections. Suppose we have a sequence of random variables $y^{(T)} = (y_1, \dots, y_T)$, where y_t may be discrete or

continuous and take values in an observation space Ω. The basic N-state HMM for the observed process $y^{(T)}$ introduces an unobserved finite state space Markov chain $x^{(T)} = (x_1, \ldots, x_T)$ taking values in $S = \{1, \ldots, N\}$. The Markov chain is parametrized by the initial distribution δ and a homogeneous transition probability matrix $\Gamma = (\gamma_{i,j})_{i,j=1,\ldots,N}$ such that

$$\begin{aligned} P(x_1 = i) &= \delta_i, \quad i = 1, \ldots, N, \\ P(x_t | x^{(t-1)}, \Gamma) &= P(x_t | x_{t-1}, \Gamma) = \gamma_{x_{t-1}, x_t}, \quad t = 2, \ldots, T. \end{aligned} \tag{1.1}$$

Therefore the joint probability of $x^{(T)}$ is given by

$$P(x^{(T)} | \delta, \Gamma) = \delta_{x_1} \prod_{t=2}^{T} P(x_t | x_{t-1}, \Gamma).$$

We assume further that the distribution of an observed data point y_t, $t = 1, \ldots, T$, given all other observed and latent variables, depends only on the current state x_t

$$f(y_t | y^{(-t)}, x^{(T)}, \phi) = f(y_t | x_t, \phi) = f(y_t | \phi_{x_t}) \tag{1.2}$$

where $y^{(-t)} = (y_1, \ldots, y_{t-1}, y_{t+1}, \ldots, y_T)$, $\phi = (\phi_1, \ldots \phi_N)$ is a stacked vector of parameters associated with the distributions in (1.2) and here, and throughout this chapter, we shall use $f(\cdot|\cdot)$ as a generic notation to represent conditional densities as specified by their arguments. The conditional distribution in (1.2) is referred to as the state-dependent distribution or the emission distribution in the HMM literature (denoted as $f_{x_t}(y_t)$ for short) and is usually assumed to belong to some parametric family of distributions, such as the normal or gamma family. Equations (1.1) and (1.2) together form the two basic modelling assumptions in HMMs and they can be compactly represented by a direct acyclic graph (DAG) as shown in Figure 1.1. A basic HMM is hence fully specified by the triplet $\theta = (\delta, \Gamma, \phi)$. The joint density of the observed and hidden variables (also known as the complete data likelihood) takes a simple form thanks to the Markov and conditional independence assumptions

$$f(y^{(T)}, x^{(T)} | \theta) = f(x^{(T)} | \theta) f(y^{(T)} | x^{(T)}, \theta) = \delta_{x_1} \prod_{t=2}^{T} P(x_t | x_{t-1}, \Gamma) \prod_{t=1}^{T} f(y_t | \phi_{x_t}). \tag{1.3}$$

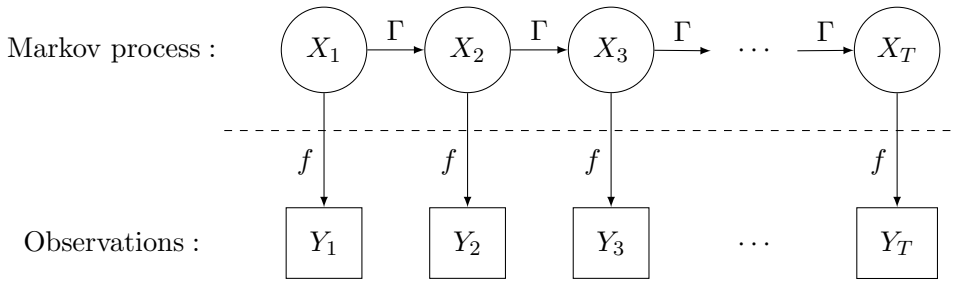

Figure 1.1: Graphical representation of a basic hidden Markov model. The unobserved variables, i.e. the hidden states, are shown in circles while the observed variables are shown in boxes.

The marginal likelihood of the observed data with hidden states integrated out (known as the observed data likelihood) can be obtained as

$$L_T = f(y^{(T)}|\theta) = \sum_{x_1,\ldots,x_T} f(y^{(T)}, x^{(T)}|\theta). \tag{1.4}$$

Note that direct evaluation of (1.4) requires $O(N^T)$ steps so the computation would become infeasible as T grows large (even with N held fixed). An efficient recursive algorithm to evaluate this likelihood will be introduced in the next section.

1.2 Inference in HMMs

In this section we introduce the main inference problems in the HMM context (with parameter θ assumed given for now), namely the filtering, smoothing, prediction and decoding problems. Importantly, the conditional independence assumptions implied by equations (1.1) and (1.2) permit dynamic programming techniques for solving these tasks efficiently, regardless of the particular parametric forms of the emission distributions. For convenience of notation, the dependent parameters θ will be dropped from the expressions of the (conditional) probability densities from now onward. The notations and terminology used here closely follow the ones used in Zucchini et al. [2016].

1.2.1 The filtering problem

The filtering problem concerns inferring the state at time $t \leq T$ given all observations up to time t, i.e. solving for $P(x_t|y^{(t)})$. To facilitate computation, we introduce what is known as the forward probability vector $\boldsymbol{\alpha}_t = (\alpha_t(1), \ldots, \alpha_t(N))$, $t = 1, \ldots, T$,

where

$$\alpha_t(i) = f(y^{(t)}, x_t = i), \quad i = 1, \ldots, N,$$

which is the unnormalized filtering distribution at time t. Note that for $t = 2, \ldots, T$, $i = 1, \ldots, N$, a "forward" recursion for $\alpha_t(i)$ can be derived as

$$\begin{aligned}
\alpha_t(i) &= \sum_{k \in S} f(y^{(t-1)}, y_t, x_t = i, x_{t-1} = k) \\
&= \sum_{k \in S} f(y^{(t-1)}, x_{t-1} = k) f(x_t = i | x_{t-1} = k, y^{(t-1)}) f(y_t | x_t = i, x_{t-1} = k, y^{(t-1)}) \\
&= \sum_{k \in S} f(y^{(t-1)}, x_{t-1} = k) P(x_t = i | x_{t-1} = k) f(y_t | x_t = i) \\
&= f_i(y_t) \sum_{k \in S} \alpha_{t-1}(k) \gamma_{k,i},
\end{aligned}$$

where $\alpha_1(i) = f(y_1, x_1 = i) = \delta_i f_i(y_1)$. The whole procedure therefore has a computational complexity of $O(TN^2)$ and is summarized in Algorithm 1 using equivalent matrix expressions, where $\mathbf{P}(y_t)$ is defined as the diagonal matrix with i-th diagonal element $f_i(y_t)$. By Bayes theorem, the HMM filter is obtained as

$$P(x_t = i | y^{(t)}) = \frac{f(x_t = i, y^{(t)})}{\sum_{i=1}^{N} f(x_t = i, y^{(t)})} = \frac{\alpha_t(i)}{\sum_{k \in S} \alpha_t(k)}, \quad t = 1, \ldots, T, \quad i = 1, \ldots, N.$$

Another by-product from the forward algorithm is the marginal likelihood which can now be efficiently computed in linear time in sample size T

$$L_T = f(y^{(T)}) = \sum_{i \in S} \alpha_T(i) = \delta \mathbf{P}(y_1) \Gamma \mathbf{P}(y_2) \cdots \Gamma \mathbf{P}(y_T) \mathbf{1}',$$

where $\mathbf{1}$ is a row vector of ones of dimension N.

Algorithm 1: The forward algorithm

- Initialize $\boldsymbol{\alpha}_1 = \delta \mathbf{P}(y_1)$
- For $t = 2, \ldots, T$, set $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} \Gamma \mathbf{P}(y_t)$

1.2.2 The smoothing problem

For the smoothing problem we are interested in computing $P(x_t | y^{(T)})$ for $t < T$, that is, the marginal probability of the state at a past time given the entire observations

$y^{(T)}$. To this end, we shall introduce another basic recursive scheme for the HMM, namely the backward algorithm. We define the vector of "backward" probabilities $\boldsymbol{\beta}_t = (\beta_t(1), \ldots, \beta_t(N))$, $t = 1, \ldots, T$, by

$$\beta_t(i) = f(y_{t+1}^T | x_t = i), \quad t = 1, \ldots, T-1, \quad i = 1, \ldots, N,$$
$$\beta_T(i) = 1, \quad i = 1, \ldots, N,$$

where $y_{t+1}^T = (y_{t+1}, \ldots, y_T)$. Then a backward recursion for $\beta_t(i)$ can be derived as follows

$$\begin{aligned}
\beta_t(i) &= f(y_{t+1}^T | x_t = i) \\
&= \sum_{k \in S} f(y_{t+2}^T, y_{t+1}, x_{t+1} = k | x_t = i) \\
&= \sum_{k \in S} P(x_{t+1} = k | x_t = i) f(y_{t+1} | x_{t+1} = k, x_t = i) f(y_{t+2}^T | y_{t+1}, x_{t+1} = k, x_t = i) \\
&= \sum_{k \in S} P(x_{t+1} = k | x_t = i) f(y_{t+1} | x_{t+1} = k) f(y_{t+2}^T | x_{t+1} = k) \\
&= \sum_{k \in S} \gamma_{i,k} f_k(y_{t+1}) \beta_{t+1}(k), \quad t = T-1, \ldots, 1, \quad i = 1, \ldots, N.
\end{aligned}$$

The whole procedure has a computational cost of $O(TN^2)$ can be executed independently of the forward recursion, and is summarized in Algorithm 2 using equivalent matrix expressions. The HMM smoother can be obtained as a function of the for-

Algorithm 2: The backward algorithm

- Initialize $\boldsymbol{\beta}_T = \mathbf{1}$
- For $t = T-1, \ldots, 1$, set $\boldsymbol{\beta}_t = \Gamma \mathbf{P}(y_{t+1}) \boldsymbol{\beta}_{t+1}$

ward and backward probabilities

$$P(x_t = i | y^{(T)}) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{k \in S} \alpha_t(k)\beta_t(k)}. \tag{1.5}$$

To verify (1.5), note that

$$f(x_t = i | y^{(T)}) = \frac{f(x_t = i, y^{(T)})}{\sum_{i \in S} f(x_t = i, y^{(T)})}$$

and

$$\begin{aligned} f(x_t = i, y^{(T)}) &= f(x_t = i, y^{(t)}, y_{t+1}^T) \\ &= f(x_t = i, y^{(t)}) f(y_{t+1}^T | x_t = i, y^{(t)}) \\ &= f(x_t = i, y^{(t)}) f(y_{t+1}^T | x_t = i) \\ &= \alpha_t(i)\beta_t(i). \end{aligned}$$

Therefore we have another expression for the marginal likelihood $L_T = \sum_{k \in S}(\alpha_t(k)$ $\beta_t(k))$ for any given t. Another quantity of interest in the smoothing context is the pairwise marginal probability $P(x_t = i, x_{t+1} = j | y^{(T)})$, which is used in applying the Expectation-Maximization (EM) [Dempster et al., 1977] algorithm for learning HMM parameters. It can be easily evaluated given the forward and backward probabilities as

$$\begin{aligned} P(x_t = i, x_{t+1} = j | y^{(T)}) &= \frac{f(x_t = i, x_{t+1} = j, y^{(T)})}{L_T} \\ &= \frac{f(x_t = i, x_{t+1} = j, y^{(t)}, y_{t+1}, y_{t+2}^T)}{L_T} \\ &= \frac{f(x_t = i, y^{(t)}) P(x_{t+1} = j | x_t = i) f(y_{t+1} | x_{t+1} = j) f(y_{t+2}^T | x_{t+1} = j)}{L_T} \\ &= \frac{\alpha_t(i)\gamma_{i,j} f_j(y_{t+1})\beta_{t+1}(j)}{L_T}. \end{aligned} \tag{1.6}$$

1.2.3 The prediction problem

The prediction problem involves predicting the hidden state and the data at a future time point $T + h$, where the integer $h > 0$ is known as the forecast horizon, given observations up to time T. Therefore there are two conditional distributions of interest, namely $P(x_{T+h} | y^{(T)})$ and $f(y_{T+h} | y^{(T)})$, each of which is easy to compute given the forward probabilities $\boldsymbol{\alpha_T}$. Starting from state prediction, we have

$$\begin{aligned} P(x_{T+h} = i | y^{(T)}) &= \frac{f(x_{T+h} = i, y^{(T)})}{f(y^{(T)})} \\ &= \frac{\sum_{k \in S} f(x_{T+h} = i, x_T = k, y^{(T)})}{L_T} \\ &= \frac{\sum_{k \in S} \alpha_T(k) P(x_{T+h} = i | x_T = k)}{L_T} \\ &= \frac{\boldsymbol{\alpha}_T \Gamma^h e_i^{'}}{L_T}, \quad i = 1, \ldots, N, \end{aligned} \tag{1.7}$$

where Γ^h is the h-step transition matrix for the Markov chain and $e_i = (0, \ldots, 1, \ldots, 0)$ is a row vector of dimension N that has a one in the i-th entry, with the rest being zero. Note that the standard Markov chain theory indicates that under the regularity conditions, as the horizon $h \to \infty$,

$$\frac{\boldsymbol{\alpha}_T}{L_T}\Gamma^h \to \boldsymbol{\pi} = (\pi_1, \ldots, \pi_N),$$

where $\boldsymbol{\pi}$ is the stationary distribution of the Markov chain and thus $P(x_{T+h} = i|y^{(T)}) \to \pi_i$. Built on (1.7), the forecast distribution $f(y_{T+h}|y^{(T)})$ can then be derived as

$$\begin{aligned} f(y_{T+h}|y^{(T)}) &= \sum_{k\in S} f(y_{T+h}, x_{T+h} = k|y^{(T)}) \\ &= \sum_{k\in S} f(y_{T+h}|x_{T+h} = k)P(x_{T+h} = k|y^{(T)}) \\ &= \frac{\sum_{k\in S} f_k(y_{T+h})\boldsymbol{\alpha}_T\Gamma^h e_i^{'}}{L_T} \\ &= \frac{\boldsymbol{\alpha}_T\Gamma^h\mathbf{P}(y_{T+h})\mathbf{1}^{'}}{L_T}. \end{aligned} \tag{1.8}$$

Note that the last row of (1.8) can be rewritten as a finite mixture of the emission distributions $\sum_{i=1}^N w_i^h f_i(y_{T+h})$, where the weight w_i^h is given by the i-th entry of $\boldsymbol{\alpha}_T\Gamma^h/L_T$. As $h \to \infty$, the weights will tend to the stationary probabilities and thus the limiting forecast distribution is given by the marginal distribution of a stationary HMM.

1.2.4 The decoding problem

Decoding refers to the process of inferring the hidden state sequence given observations and is often of central interest in many applied problems. Generally speaking, there are two different strategies for solving this problem. The first approach, which is referred to as local decoding, selects the state at each time point by maximizing its marginal state probability, that is, to find

$$\hat{x}_t = \underset{i=1,\ldots,N}{\arg\max} P(x_t = i|y^{(T)}),$$

where $P(x_t = i|y^{(T)})$ is the HMM smoother as given by (1.5). The resulting estimated state sequence is known as the maximum accuracy path. Alternatively, one can perform global decoding which aims at finding the the most likely sequence of the hidden states, i.e. the state path $(x_1, \ldots, x_T)$ that maximizes the conditional

probability $P(x^{(T)}|y^{(T)})$ or equivalently, the joint density $P(x^{(T)}, y^{(T)})$

$$\hat{x}^{(T)} = \arg\max_{x^{(T)}} P(x^{(T)}|y^{(T)}) = \arg\max_{x^{(T)}} f(x^{(T)}, y^{(T)}).$$

Clearly, direct optimization by comparing all possible state paths is not a feasible solution as there would be a total of N^T paths to be considered. The Viterbi algorithm [Viterbi, 1967] turns out to be an efficient method of solving this problem, with computational complexity of the same order as that for the forward and backward algorithms. Let $\mathbf{V}_t = (V_t(1), \ldots, V_t(N))$, $t = 1, \ldots, T$, where $V_1(i) = f(x_1 = i, y_1)$ and $V_t(i) = \max_{x_1,\ldots,x_{t-1}} f(x^{(t-1)}, x_t = i, y^{(t)})$, $i = 1, \ldots, N$. Then $\mathbf{V}_2, \ldots, \mathbf{V}_T$ can be computed in a recursive manner since

$$\begin{aligned} V_t(i) &= \max_{x_1,\ldots,x_{t-2}} \max_{x_{t-1}} f(x^{(t-2)}, x_{t-1}, x_t = i, y^{(t-1)}, y_t) \\ &= \max_{x_1,\ldots,x_{t-2}} \max_{x_{t-1}} f(x^{(t-2)}, x_{t-1}, y^{(t-1)}) f(x_t = i|x^{(t-2)}, x_{t-1}, y^{(t-1)}) \\ &\qquad\qquad f(y_t|x_t = i, x^{(t-2)}, x_{t-1}, y^{(t-1)}) \\ &= \max_{x_{t-1}} V_{t-1}(x_{t-1}) \gamma_{x_{t-1},i} f_i(y_t). \end{aligned}$$

It then becomes clear that we can reconstruct the most likely state path in a backward manner by first identifying the optimal state at time T as

$$\hat{x}_T = \arg\max_{i=1,\ldots,N} V_T(i),$$

and then recursively recovering the remaining states from

$$\hat{x}_t = \arg\max_{i=1,\ldots,N} V_t(i) \gamma_{i,\hat{x}_{t+1}} f_{\hat{x}_{t+1}}(y_{t+1}) = \arg\max_{i=1,\ldots,N} V_t(i) \gamma_{i,\hat{x}_{t+1}}, \quad t = T-1, \ldots, 1.$$

The whole procedure is summarized in Algorithm 3.

Algorithm 3: The Viterbi algorithm

- For $i = 1, \ldots, N$, set $V_1(i) = f(x_1 = i, y_1)$
- For $t = 2, \ldots, T$, $i = 1, \ldots, N$, set $V_t(i) = \max_{x_{t-1}} V_{t-1}(x_{t-1}) \gamma_{x_{t-1},i} f_i(y_t)$
- Set $\hat{x}_T = \arg\max_{i=1,\ldots,N} V_T(i)$
- For $t = T-1, \ldots, 1$, set $\hat{x}_t = \arg\max_{i=1,\ldots,N} V_t(i) \gamma_{i,\hat{x}_{t+1}}$

1.3 Learning for HMMs

Parametric estimation theory for HMMs is well established and here we briefly introduce the two basic inferential frameworks for estimating the unknown parameters $\theta = (\delta, \Gamma, \phi)$ of an HMM, the maximum likelihood and Bayesian estimation methods. To simplify discussion we shall assume that the cardinality N is known throughout this section and will discuss its selection as a model selection problem in the next section.

1.3.1 Maximum likelihood estimation

In this framework the parameters θ are treated as fixed quantities, although unknown, and we want to find the parameters that maximize the observed data likelihood, i.e. solving the constrained optimization problem

$$\hat{\theta} = \arg\max_{\theta \in \Theta} f(y^{(T)}|\theta), \tag{1.9}$$

where Θ is the joint parameter space. Consistency and asymptotic normality of the maximum likelihood (ML) estimator for general HMMs were established in Leroux [1992] and Bickel et al. [1998], respectively. Although conceptually simple and theoretically attractive, its implementation in practice is generally a challenging problem due to the complicated structure of $f(y^{(T)}|\theta)$ induced by the latent process and the fact that no closed form solution exist. In the literature we may identify two different strategies to perform ML estimation, each of which has its relative merits. The first uses numerical maximisation techniques to directly find solutions to (1.9) and usually requires little programming effort (see e.g. Langrock et al. [2012a]), thanks to the ease of evaluating the likelihood with forward algorithm and the existence of many optimization routines available in many software packages. However, the approach can suffer from numerical issues or poor convergence properties, especially for increasing cardinality N. The second method relies on the EM algorithm (known as the Baum-Welch algorithm [Baum et al., 1970] in the context of HMMs), a general yet powerful method for finding ML estimates for models with missing data. It works with the complete data log-likelihood (CDLL), which takes a much simpler form

$$\log(f(y^{(T)}, x^{(T)})) = \log(\delta_{x_1}) + \sum_{t=2}^{T} \log(\gamma_{x_{t-1},x_t}) + \sum_{t=1}^{T} \log(f_{x_t}(y_t)).$$

Let us further define $I_i(t) = 1$ if $x_t = i$, $t = 1, \ldots, T$, and $I_{ij}(t) = 1$ if $x_{t-1} = i$ and $x_t = j$, $t = 2, \ldots, T$, the CDLL can then be rewritten as:

$$\log(f(y^{(T)}, x^{(T)})) = \sum_{i=1}^{N} I_i(1) \log(\delta_i) + \sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} I_{ij}(t) \log(\gamma_{i,j}) + \sum_{t=1}^{T} \sum_{i=1}^{N} I_i(t) \log(f_i(y_t)).$$

Given the starting values for the parameter vector $\theta^{(0)}$, the EM algorithm proceeds to create a sequence of $\theta^{(i)}$ according to:

$$\begin{aligned}
\theta^{(i+1)} &= \arg\max_{\theta} \mathbb{E}[\log(f(y^{(T)}, s^{(T)}))|\theta^{(i)}, y^{(T)}] \\
&= \arg\max_{\theta} \Big(\sum_{i=1}^{N} \hat{I}_i(1) \log(\delta_i) + \sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{I_{ij}}(t) \log(\gamma_{i,j}) + \sum_{t=1}^{T} \sum_{i=1}^{N} \hat{I}_i(t) \log(f_i(y_t)) \Big)
\end{aligned}$$

where the conditional expectation is taken with respect to $x^{(T)}$, $\theta^{(i)}$ is the current parameter estimate, $\hat{I}_i(t) = \mathbb{E}[I_i(t)|\theta^{(i)}, y^{(T)}] = P(x_t = i|\theta^{(i)}, y^{(T)})$ and $\widehat{I_{ij}}(t) = \mathbb{E}[I_{ij}(t)|\theta^{(i)}, y^{(T)}] = P(x_{t-1} = i, x_t = j|\theta^{(i)}, y^{(T)})$ are given by the HMM smoother (1.5) and (1.6), respectively, computed conditional on $\theta^{(i)}$. It can be shown that the observed data log-likelihood is nondecreasing at each iteration of the algorithm

$$\log f(y^{(T)}|\theta^{(i)}) \leq \log f(y^{(T)}|\theta^{(i+1)})$$

and will converge at least to a local maximum (proof is omitted here). Compared with the direct maximization approach, the EM algorithm enjoys better theoretical guarantees but can suffer from slow convergence and usually requires higher programming effort and thus can be more costly to implement. We refer to Altman and Petkau [2005], Bulla and Berzel [2008] and Cappé et al. [2005] (chapter 10) for more comparative discussions on the use of the EM and direct numerical approach for learning HMMs. It should be pointed out that the initialization of the parameters is a subtle and challenging issue for both approaches and can have a significant impact on the final results, see Maruotti and Punzo [2021] and references therein for a more detailed discussion and related strategies. For such ML based methods, standard errors (and approximate confidence intervals) for the parameter estimates are usually estimated either from the approximate covariance matrix for the parameter estimates or by parametric bootstrap technique. We note however that various issues exist with these approaches. Particularly, with respect to the former approach, the estimated standard error can be unreliable when some of the parameters are close to (or on) the boundary of their parameter space. The latter approach avoids relying on asymptotics, but the computations are usually very

time-consuming [Zucchini et al., 2016].

1.3.2 Bayesian estimation

In a Bayesian paradigm, the model parameters $\theta = (\delta, \Gamma.\phi)$ are treated as random variables, for which prior distributions are introduced to express our beliefs before seeing the data. After obtaining the observations we update our beliefs on the parameters via Bayes theorem and our inference on θ can be performed based on its posterior distribution. In the HMM context we are typically interested in the posterior over the parameters and hidden variables

$$f(\theta, x^{(T)}|y^{(T)}) \propto f(x^{(T)}, y^{(T)}|\theta)f(\theta) \tag{1.10}$$

where $f(\theta)$ is the joint prior distribution. Direct inference on this joint posterior is, however, intractable in general and we need to resort to approximate inference techniques. One option is to use Bayesian variational method, an optimization-based deterministic approach that is originally introduced to HMMs in MacKay [1997]. The basic idea is to approximate the target posterior in by a tractable family of distributions q that is usually assumed fully factorizable in its components, i.e. $q(\theta, x^{(T)}) = q_\delta(\delta)q_\Gamma(\Gamma)q_\phi(\phi)q_{\mathbf{x}}(x^{(T)})$ (known as the mean field approximation). The approximating density q can then be optimized (in the sense of minimizing the discrepancy between $q(\theta, x^{(T)})$ and $f(\theta, x^{(T)}|y^{(T)})$) in an iterative manner by defining and optimizing a variational free energy, using a so-called variational Bayes EM algorithm which is guaranteed to converge to a stationary point. We refer to MacKay [1997], Beal [2003] and McGrory and Titterington [2009] for more theoretical and implementational details. Another more popular solution is to use Markov chain Monte Carlo (MCMC) method; some pioneer works include, for instance, Robert et al. [1993], Chib [1996], Robert and Titterington [1998] and Scott [2002]. It is a simulation-based approach that yields a stochastic representation of a potentially complex posterior distribution. Various quantities of interests, such as the posterior means of the parameters, can be easily approximated using their sample-based averages. Uncertainty quantification of the parameter estimates can be achieved by studying the variability of their posterior samples. The most popular MCMC algorithm for HMMs is perhaps the block Gibbs sampler based on data augmentation [Frühwirth-Schnatter, 2006]. In outline, the sampler switches between the following two steps until "convergence"

- draw $x^{(T)} \sim P(x^{(T)}|\theta, y^{(T)})$

- draw $\theta \sim f(\theta|x^{(T)}, y^{(T)})$

An important fact is that we can efficiently simulate $x^{(T)}$ from its posterior using a dynamic programming algorithm, known as a forward filtering backward sampling (FFBS) algorithm (see e.g. Scott [2002]), at a cost of $O(TN^2)$ regardless of any particular parametric form for the emission distribution $f_{x_t}(y_t)$. The key insight here is that the joint posterior of the hidden states can be factorized as

$$\begin{aligned} P(x^{(T)}|y^{(T)}) &= P(x_T|y^{(T)}) \prod_{i=1}^{T-1} P(x_{T-i}|x_{T-i+1}^T, y^{(T)}) \\ &= P(x_T|y^{(T)}) \prod_{i=1}^{T-1} P(x_{T-i}|x_{T-i+1}, y^{(T)}), \end{aligned}$$

where the second equality follows from the basic assumptions (1.1) and (1.2). Therefore we can jointly sample $x^{(T)}$ by first sampling $x_T \sim P(x_T|y^{(T)})$, and then for $t = T-1, \ldots, 1$, sampling $x_t \sim P(x_t|x_{t+1}, y^{(T)}) \propto f(x_t, y^{(t)})P(x_{t+1}|x_t)$, where x_{t+1} is the most recent sampled state at $t+1$. Note that to implement this procedure we need to compute the forward probabilities $\boldsymbol{\alpha}_t$, $t = 1, \ldots, T$, in advance, which explains the first part of the name "forward filtering". The whole process is summarized in Algorithm 4. Conditional on the sampled state sequence and the observed

Algorithm 4: The forward filtering backward sampling algorithm

- For $t = 1, \ldots, T$, compute $\boldsymbol{\alpha}_t$ using the forward algorithm
- Sample $x_T \sim P(x_T = i|y^{(T)}) \propto \alpha_T(i)$
- For $t = T-1, \ldots, 1$, sample $x_t \sim P(x_t = i|x_{t+1}, y^{(T)}) \propto \alpha_t(i)\gamma_{i,x_{t+1}}$

data, model parameters $\theta = (\delta, \Gamma, \phi)$ can, depending on the model setting, be easily sampled in blocks using Gibbs sampling steps. We refer to Frühwirth-Schnatter [2006] and Rydén [2008] for more details and some examples. An alternative to the above Gibbs sampler is to use Metropolis-Hastings Algorithm, with the hidden state sequence integrated out via the forward algorithm, see Cappé et al. [2005] (chapter 13.1) for an example. Notably, MCMC methods are asymptotically exact as opposed to variational methods, in the sense that it will sample from the exact posterior of interest as the number of iterations goes to infinity, and the algorithm can be easily adapted and applied to almost arbitrary models. On the other hand, they can be more computationally intensive to implement and it may take excessive time for the sampler to converge that is also hard to justify. There is also an unidentifiability

issue we need to take care of - the posterior in (1.10) would be invariant to permutations of the state labels (i.e. has $N!$ symmetric modes) if the priors are invariant to relabelling of the states. This potential label switching problem means that the samples generated by MCMC cannot be directly used for state-specific inference as the state labels can permute during the MCMC iterations. A number of strategies has been proposed to deal with this issue, see for instance Marin et al. [2005] and Spezia [2009]. We will discuss this issue in more details in later chapters.

1.4 Model selection

Up to now we treated the number of states N as given. This is the case in certain scenarios where we have sufficient prior information regarding the underlying physical process that generated the data or we may fall into a classification task with pre-defined categories to be allocated. In many other applied problems, determining the cardinality N is of scientific interest in itself as it conveys important information regarding the underlying process, and thus it needs to be estimated from the data along with other model parameters. Informally speaking, we would like to select the cardinality N, neither overly high (i.e. danger of overfitting) nor overly low (i.e. danger of underfitting), that best explains the key features of the data and allow us to extract as much information as possible from the limited data at hand. We note that for the special case of finite-alphabet HMMs (observed data have a finite support), this order selection problem is well studied, see e.g. Rydén [1995] and Gassiat and Boucheron [2003]. However, for more general HMMs this remains a challenging task. Under the frequentist estimation framework, it has been shown that the likelihood ratio statistic for HMMs has a nonstandard behaviour and is unbounded even in some simple parametric cases [Gassiat and Keribin, 2000]. Penalized likelihood methods such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC) are popular in practice to compare HMMs with varying complexity [Rydén et al., 1998; Punzo and Maruotti, 2016], however, there is a lack of theoretical guarantees and they can be problematic, as for instance, the former may favour HMMs with an undesirably large number of states while the latter may over-penalize the larger models [Scott, 2002; Pohle et al., 2017; Li and Bolker, 2017]. Alternative methods based on the cross-validated likelihood are proposed in Celeux and Durand [2008] and were found to be strong competitors to the information criteria based approaches. However, no consensus has been made to date regarding the optimal selection criterion. For the special case of univariate Gaussian HMMs, Hung et al. [2013] developed a double penalized likelihood method for model selection and estab-

lished consistency and good finite sample performance. In the Bayesian framework, the cardinality N is treated as an additional model parameter and its value can be inferred according to its posterior distribution

$$P(N|y^{(T)}) \propto f(y^{(T)}|N)P(N),$$

where $P(N)$ expresses our prior belief on the number of states with support usually defined over a finite set $\{2, \ldots, M\}$, and

$$f(y^{(T)}|N = k) = \int f(y^{(T)}|, \theta_k, N = k) f(\theta_k|N = k) d\theta_k, \tag{1.11}$$

where θ_k denotes the parameter set for a k-state HMM. The quantity in (1.11) is known as the evidence or the integrated likelihood, which can be informally regarded as the averaged fit of the model to the data conditional on the cardinality N and it automatically penalizes models with larger number of parameters (principle of Occam's razor, see e.g. Jefferys and Berger [1992] and MacKay [1992]). The evidence (1.11) therefore plays an important role in Bayesian model selection, although it is generally difficult to estimate [Scott et al., 2005]. Various techniques (mostly Monte Carlo methods) have been proposed to approximate this evidence (or the ratios between two models), a few examples include the importance sampling of Geweke [1989], the annealed importance sampling of Neal [2001], the harmonic mean of Gelfand and Dey [1994], the serial methods of Chib [1995] and Chib and Jeliazkov [2001], the bridge sampling of Meng and Wong [1996] and Frühwirth-Schnatter [2004] and the variational method of Corduneanu and Bishop [2001]. We will continue our discussion on its estimation in chapter 2. Instead of directly targeting (1.11), we can estimate the cardinality N jointly with other unknown parameters using Transdimensional MCMC methods, see Frühwirth-Schnatter [2006] for an overview. A notable example is the use of the reversible jump MCMC of Green [1995] for inference in HMM with Gaussian emission distributions [Robert et al., 2000; Spezia, 2010; Spezia et al., 2011], where model parameters are updated via Gibbs sampling or Metropolis-Hastings algorithm and the dimension of the model changes in split/merge and birth/death moves. However, such algorithms can be computationally costly to implement and require careful algorithmic design depending on model settings, and thus such methods are not widely used [Boys and Henderson, 2001; Murphy, 2012]. Another possible solution to tackle this model selection problem is to specify the hidden state process nonparametrically using Bayesian nonparametric techiniques to allow for an unbounded number of states a-priori, and the cardinality N can be determined a-posterori in a fully data-driven manner [Beal et al., 2002;

Teh et al., 2006]. Of course, given the observed data only a finite number of states will be initiated (i.e. have at least one allocated observation). We will come back to this topic to investigate it further in chapter 4.

1.5 Model checking

Model checking is an important latter stage in the model development process, after one "best" model is selected using some criteria. We need to examine if the modelling assumptions are met and if the selected model can adequately explain the patterns observed in the data to assure that we can make trustful and meaningful inferences with the fitted model. For parametric HMMs estimated with ML-based approaches, the so-called pseudo residuals provide an effective way to examine the overall goodness-of-fit of the estimated model and detect possible outliers in the data [Zucchini et al., 2016]. They are motivated from the simple fact that if a random variable Y is distributed according to a distribution F_Y, i.e. $Y \sim F_Y$, then the random variable $Z = \Phi^{-1}(F_Y(Y)) \sim N(0,1)$, where Φ is the cumulative distribution for a standard normal distribution. For HMMs, two versions of pseudo residuals, namely ordinary pseudo residuals (OPR) and forecast pseudo residuals (FPR), are popularly used. Here we restrict ourselves to the case of continuous observations. For the discrete counterpart the OPR or FPR need to be modified accordingly, see Zucchini et al. [2016]. The OPR for a realisation $y_t = y_t^O$ is based on the conditional distribution of y_t given all other observations, and is defined as:

$$z_t = \Phi^{-1}\big(\int_{-\infty}^{y_t^O} f(y_t = y|y^{(-t)})dy\big), \quad t = 1, \dots, T,$$

where the integrand $f(y_t|y^{(-t)}) = f(y^{(T)})/f(y^{(-t)})$ and simple algebra shows that

$$\begin{aligned} f(y^{(-t)}) &= \sum_{x^{(-t)}} f(x^{(-t)}, y^{(-t)}) \\ &= \delta \mathbf{P}(y_1)\Gamma\mathbf{P}(y_2)\cdots\Gamma\mathbf{P}(y_{t-1})\Gamma^2\mathbf{P}(y_{t+1})\cdots\Gamma\mathbf{P}(y_T)\mathbf{1}', \end{aligned} \tag{1.12}$$

where $\mathbf{P}(y_t)$ denotes the diagonal matrix with i-th diagonal element $f_i(y_t)$. Note that (1.12) takes the same form as (1.4) except that $P(y_t)$ is replaced by the identity matrix. The FPR for a realisation $y_t = y_t^O$ is based on the conditional distribution of y_t given all preceding observations and is defined as:

$$z_t = \Phi^{-1}\big(\int_{-\infty}^{y_t^O} f(y_t = y|y^{(t-1)})dy\big), \quad t = 2, \dots, T,$$

where $f(y_t|y^{(t-1)}) = f(y^{(t)})/f(y^{(t-1)})$. Note that $f(y^{(t)})$ can be evaluated via the forward algorithm as in (1.4) (with T replaced by t). If the model is adequate, the computed pseudo residuals should have an approximate standard normal distribution, which can be easily checked using quantile-quantile plots or various normality tests. Extreme values in OPR or FPR would indicate that the corresponding observations are unlikely to occur given the fitted model and other observations. More details regarding the theory, construction and application of these pseudo residuals can be found in Zucchini and MacDonald [1999] and Zucchini and MacDonald [2009]. An alternative strategy to model checking is to simulated data conditional on the point estimate of model parameters, and to check if they can reproduce the key features presented in the empirical data, for instance in terms of the marginal distribution and the correlation structure of the data [Langrock et al., 2014; Touron et al., 2018; Adam et al., 2019b]. This is less formal than the residual-based method but has proved to be useful in identifying lack of fit and may be able to provide useful insights into potential ways to improve the model. We also note the recent work of Buckby et al. [2020] who proposed new residual-based model checking methods following results from point process models, which have advantages over previous methods in checking more complicated extensions of the basic HMMs. We omit the details here and refer the reader to their paper and references therein for additional information. In a Bayesian estimation framework, model checking is typically performed based on the posterior samples (e.g. obtained via MCMC) for the parameters and the hidden state sequence For instance, we can check the Markov assumption of the latent state process by analyzing the transition/waiting patterns at each state from the posterior samples of $x^{(T)}$ [Chen et al., 2016]. The predictive distribution of the data, $f(y|y^{(T)})$, provides another valuable diagnostic for model adequacy [Scott, 2002]. Note that

$$f(y|y^{(T)}) = \int f(y|\theta) f(\theta|y^{(T)}) d\theta,$$

therefore, we can simulate from $f(y|y^{(T)})$ by simulating data from the marginal distribution of the HMM conditional on each $\theta^{(i)}$ with $\theta^{(i)} \sim f(\theta|y^{(T)})$. More detailed posterior predictive check can be performed by simulating the entire data set from the HMM conditional on each simulated parameter set $\theta^{(i)}$, and compare summaries from the simulated data set to those from the empirical data [Gelman et al., 1995]. This is to be compared with the frequentist simulation-based checking approach described above where the assessment is based on a fixed point estimate of θ, whereas here the uncertainty regarding θ is appropriately taken into account.

1.6 Extensions of the basic HMM

Extensions of the basic HMMs have been broadly explored and applied to allow for more flexible and accurate modelling of the increasingly complex real data. They are typically achieved by relaxing the assumptions made in (1.1) and (1.2) and/or adding more structures to the basic modelling framework. For instance, the Markov assumption of the state process is mathematically convenient yet can be overly simple, as it implicitly assumed that the sojourn time at each state is geometrically distributed. One natural extension is to explicitly model the state dwell-time at each state with more arbitrary distributions while still keeping the Markovian structure, which leads to the so-called hidden semi-Markov models [Yu, 2015]. We can also construct HMMs that allow for additional serial dependence structure at the observation level. For instance, the distribution of y_t may depend both on the current state x_t as well as previous observations. Such an extension is known as the Markov switching autoregressive models which are found useful in financial time series modelling [Hamilton, 2020]. Covariates can also be incorporated into HMMs by allowing HMM parameters to depend on them through suitable "link" functions as in the generalized linear regression framework. For example, we can make the latent state process inhomogeneous by assuming that the transition probabilities between time t and $t+1$ (i.e. ${}_t\gamma_{i,j}$; $i,j = 1,\ldots,N$) are functions of other covariates. In this case the standard multinomial logistic link function can be used and by treating diagonal elements of ${}_t\Gamma$ as reference variables, the off-diagonal elements of ${}_t\Gamma$ can be expressed as

$$ {}_t\gamma_{i,j} = \frac{\exp(\beta_{\mathbf{ij}}\mathbf{c_t'})}{1+\sum_{k=1;k\neq i}^{N}\exp(\beta_{\mathbf{ik}}\mathbf{c_t'})}, \quad i,j = 1,\ldots,N, \quad i \neq j, $$

where $\mathbf{c_t}$ represents a row vector of covariates and $\beta_{\mathbf{ij}}$ is the transition-specific coefficient vector (see chapter 10 of Zucchini et al. [2016]). Emission distributions can also be covariate-dependent by using similar strategies, see e.g. Mccallum and Wang [2013]. We refer to Mor et al. [2021] for a more comprehensive review of various HMM extensions within the parametric framework and their applications. Another important direction for extending the basic HMM that has received an increasing amount of interests is to explore the use of non-parametric techniques for flexible modelling of the emission and state models, as is the main topic of this thesis. This is motivated by the fact that the emissions and the cardinality N can be hard to select and justify in practice, and their misspecification could lead to less accurate or even erroneous inferences in HMMs. The great potentials of using nonparametric

methods have already been demonstrated in various real-world applications, see Yau et al. [2011] and Langrock et al. [2015], and the theory, development and application of these methods for HMMs are still under exploration. We will continue our journey on nonparametric HMMs in the following chapters.

Chapter 2

Bayesian inference for spline-based hidden Markov models

2.1 Introduction

A basic N-state HMM consists of a discrete-time stochastic process (x_t, y_t) with $x_t \in \{1, \ldots, N\}$ and $y_t | x_t \sim f_{x_t}(y_t)$, where the process $\{x_t\}$ is unobserved and assumed to be distributed as an N-state time-homogeneous Markov chain, and, conditionally on $\{x_t\}$, the y_t's are independent with state-dependent so called *emission* distributions f_{x_t}. The latter are usually assumed to belong to some parametric family of distributions, such as the normal or gamma family. Despite its relatively simple model structure, the HMM has the ability to handle complex dependencies between the observations due to the assumption of the hidden state process. Estimation theory for parametric HMMs is well established, both in the frequentist and Bayesian framework, see Douc et al. [2004]; Mevel and Finesso [2004]; Douc et al. [2011] for the theory on maximum likelihood estimation and De Gunst and Shcherbakova [2008]; Gassiat et al. [2014]; Douc et al. [2020] for Bayesian inference methods.

There has been considerable effort in extending the structure of the basic HMM formulation to introduce more flexibility and to allow for a more realistic modelling in real data applications. In particular, it is recognized that simple parametric choices for the emission distributions are not always well justified where misspecification can lead to seriously erroneous inference on the number of hidden states and on the classification of the observations to the states [Yau et al., 2011;

Gassiat et al., 2016a; Pohle et al., 2017]. Nonparametric approaches offer much more flexibility and may serve as exploratory tools to investigate the suitability of a parametric family of emission distributions [Langrock et al., 2015]. Considerable effort has been invested in the use of semi- and nonparametric emission distributions such as proposed in Piccardi and Pérez [2007] for activity recognition in videos, Yau et al. [2011] for the analysis of genomic copy number variation, Langrock et al. [2015, 2018] for modelling animal movement data and Kang et al. [2019] for delineating the pathology of Alzheimer's disease, among many others. Theoretical properties for inference in such models have been studied in a number of recent papers. Alexandrovich et al. [2016] proved that model parameters as well as the order of the Markov chain are identifiable (up to permutations of the hidden states labels) if the transition probability matrix of $\{x_t\}$ has full rank and is ergodic, and if the emission distributions are all distinct. These conditions are fairly generic and in practice they will usually be satisfied. We also refer to Gassiat et al. [2016a,b] for other useful identifiability results in this context. Based on this building stone, theoretical results for different kinds of estimation procedures in this nonparametric setting have been developed in recent works, see for instance Alexandrovich et al. [2016] for consistency of a maximum likelihood estimator, Vernet et al. [2015] for posterior consistency of Bayesian procedures, and also Lehéricy [2018] and references therein for asymptotic results on spectral and least square estimators. We also note the work by De Castro et al. [2017] who provided theoretical guarantees for estimating the filtering and marginal smoothing distributions in nonparametric HMMs.

Unsurprisingly, the increased flexibility and modelling accuracy obtained by specifying the emission distributions in a non-parametric way comes at the cost of a higher computational complexity in terms of model estimation and inference. For instance, the computational cost of the standard HMM algorithms (e.g. the forward algorithm introduced in Rabiner [1989]) for kernel-based HMMs [Piccardi and Pérez, 2007] is subject to a quadratic growth with the data size n and thus can be prohibitive for long time series data. Also for mixtures of Dirichlet process (MDP) HMMs [Yau et al., 2011] the increased complexity of the model space poses challenges to the existing sampling methods [Hastie et al., 2015]. Here, we focus on spline-based HMMs which are attractive for real applications as they exploit the strengths of two powerful tools, namely the forward algorithm for efficient and exact likelihood evaluation, and the flexibility of B-splines for estimating the emission densities, while retaining a relatively simple model formulation. A frequentist approach for inference based on penalized B-splines (P-splines) was recently introduced by

Langrock et al. [2015, 2018]. Their methods, however, require pre-specification of the number and positions of knots which can strongly influence the computational costs and the convergence results. In practice, a large number of equidistant knots has to be used to ensure some flexibility, leading to high computational challenges (e.g. convergence to suboptimal local extrema of the likelihood) and cost. Furthermore, the selection of the state-specific smoothing parameters and the quantification of uncertainty associated with parameter estimates remain to date challenging inferential tasks in the frequentist framework. Current methods rely on cross-validation and parametric bootstrap techniques, which are extremely computationally intensive and can be numerically unstable especially for increasing cardinality N. Their approach is therefore only feasible for models with a small number of states which may severely limit its applicability.

We note that nonparametric methods such as those in Piccardi and Pérez [2007]; Yau et al. [2011] and Lehéricy [2018] require the number of states to be known or fixed in advance. This is the case for certain types of applications of HMMs, such as classification in a supervised learning context (e.g. speech recognition) where the states and their interpretation are predefined. In other scenarios, estimating N is often a question of scientific interest in itself and introduces an additional level of complexity to HMM inference. The problem of order estimation for parametric HMMs has been extensively studied in the literature. We refer to Celeux and Durand [2008]; Costa and De Angelis [2010] and Pohle et al. [2017] for discussions on various criterion-based methods, and to Barber et al. [2011] (chapter 15) for a review of relevant Bayesian methods. In contrast, few theoretical or practical results have been obtained for the nonparametric case. We note that Lehéricy et al. [2019] recently proposed two different estimators for N which are proved to be consistent in a fairly general setup. The first method uses model selection techniques that involves minimization of a penalized least square criterion and the second one relies on a thresholding method on the singular values of the estimated density of two consecutive observations. Although theoretically attractive, both estimators suffer from implementation difficulties. The former requires a separate estimator for the penalty term and the minimization problem is non-convex so there is a danger of getting stuck in local minima, while the latter needs custom heuristics to tune the threshold which is critical in this algorithm. An alternative strategy to tackle this model selection problem is to use Bayesian nonparametric techniques to allow for a potentially infinitely large state space, leading to the so-called infinite HMM (iHMM), see Barber et al. [2011] (chapter 15) for a review of the topic. In this chapter, however, we will consider HMMs with a finite and fixed state space whose

cardinality N may be unknown.

To the best of our knowledge, spline-based methods have only been used in a Bayesian HMM for modelling covariate effects [Song et al., 2018], but not for the purpose of density estimation. We therefore propose and develop a fully Bayesian methodology that jointly estimates the spline functions for the emission densities and other HMM parameters, and provides a consistent and principled framework for quantifying uncertainties associated with model parameters including the number of states. We develop an almost "tuning-free" reversible jump Markov chain Monte Carlo (RJMCMC) algorithm [Green, 1995] which exploits the use of a forward filtering backward sampling (FFBS) procedure for efficient simulation of the hidden state process, a stochastic approximation based adaptive MCMC scheme for automatic tuning, a reparametrization scheme for enhancing the sampling efficiency and an adaptive knot selection scheme that modifies and extends ideas considered in different modelling contexts, see for instance DiMatteo et al. [2001] for Bayesian curve fitting and more recently Sharef et al. [2010] for baseline hazard modelling. As shown later, the proposed adaptive spline based algorithm has significant advantages over its frequentist P-spline counterpart and also compares favourably to the Bayesian P-spline [Lang and Brezger, 2004] version which is investigated for the first time here for density estimation in HMMs. We also address the issue of model selection of N through a parallel sampling scheme which is straightforward to implement and computationally efficient as it only involves simultaneous and independent runs of the proposed algorithm for candidate values of N, from which quantities such as marginal likelihood for each model can be estimated. A further advantage of a Bayesian inference framework is that the modularity of its components can be used to perform inference rigorously in a more complex hierarchical HMM model (as will be demonstrated in chapter 3).

This chapter is structured as follows: Section 2.2 gives a brief introduction to B-splines and reviews existing Bayesian estimation strategies, Section 2.3 provides details of a Bayesian formulation of the spline-based HMM, Section 2.4 details the structure of the RJMCMC algorithm, Section 2.5 outlines the parallel sampling method for selecting the number of states and Section 2.6 examines the performance of the proposed methods in comparison to other related methods in various simulation settings. Section 2.7 illustrates our methods on animal activity data, while the last section provides a discussion and possible directions of further work.

2.2 B-splines

Splines have been extensively used for tasks such as interpolation and curve fitting as they have good approximation properties for a rich class of functions, see De Boor et al. [1978] and Schumaker [2007] for details on the related theoretical results. A spline function of order O is a piecewise polynomial function of degree $O-1$ where the polynomial pieces are connected at the so-called knot points [De Boor et al., 1978]. Provided that the knots are distinct, the derivatives of these piecewise polynomials are $(O-2)$-times continuously differentiable at the knots. B-splines (short for basis splines) of order O provide basis functions for representing spline functions of the same order defined over the same set of knots. In other words, any spline function can be uniquely constructed via a linear combination of B-splines [Prautzsch et al., 2002]. To set up the B-splines, let a and b be the two boundary knots which define the domain of interest over which the splines are evaluated. Let K be a positive integer indicating the number of interior knots with location given by the K-dimensional vector $R_K = (r_1, \ldots, r_K)$, with $a < r_1 < \ldots < r_K < b$ (for simplicity we do not consider knot duplication here). For computational reasons the knot sequence is usually augmented by introducing additional knots such that

$$\bar{R}_K = (r_{1-O}, \ldots, r_0, R_K, r_{K+1}, \ldots, r_{K+O}),$$

where $r_{1-O} \leq \ldots \leq r_0 \leq a$, $b \leq r_{K+1} \leq \ldots \leq r_{K+O}$ and O is the order of the spline function [Friedman et al., 2001]. The positions of the left and right external knots are usually arbitrary and for convenience we may set them equal to the boundary values a and b, respectively. Rewriting the augmented knot vector as $\bar{R}_K = (\tilde{r}_1, \ldots, \tilde{r}_{K+2O})$, where $\tilde{r}_i = r_{i-O}$, we can define the i-th B-spline basis function of order j ($j \leq O$) for $\bar{R}_K$, $B_{i,j}(y)$, using the Cox–de Boor recursion starting with

$$B_{i,1}(y) = \begin{cases} 1 & \tilde{r}_i \leq y < \tilde{r}_{i+1} \\ 0 & \text{else} \end{cases}, \quad i = 1, \ldots, K + 2O - 1.$$

Then,

$$B_{i,j}(y) = \frac{y - \tilde{r}_i}{\tilde{r}_{i+j-1} - \tilde{r}_i} B_{i,j-1}(y) + \frac{\tilde{r}_{i+j} - y}{\tilde{r}_{i+j} - \tilde{r}_{i+1}} B_{i+1,j-1}(y), \quad i = 1, \ldots, K + 2O - j.$$

The order O and the number and location of the knots thus fully specify the B-splines of that order. We can see from this construction that the B-spline basis functions are non-negative over the domain and have compact support (and so does any linear

combination of them): $B_{i,j}(y) > 0$ only when $y \in (\tilde{r}_i, \tilde{r}_{i+j})$. This local property has important computational consequences, making their computation numerically stable and efficient even for large values of K. These basis functions can be easily generalized to bivariate (or higher dimensional) scenarios via a tensor product of the univariate B-spline basis in each dimension as defined above. The computational advantages of the univariate B-splines directly carry over to the multivariate case.

The great flexibility and nice computational properties make B-splines (in the form of $\sum_{i=1}^{K+O} a_i B_{i,O}(y)$, where the a_i are spline coefficients to be estimated) a popular tool in semi-/nonparametric statistical modelling, especially in nonlinear regression analysis [Denison et al., 2002; Zanini et al., 2020; Michelot et al., 2016] and density estimation [Koo, 1996; Edwards et al., 2019; Maturana-Russel and Meyer, 2021]. Inference in spline-based models is available in both frequentist and Bayesian framework, and here we concentrate on the latter. Broadly speaking, there are two main Bayesian estimation strategies, depending on whether the number and positions of knots are treated as fixed or not. The Bayesian P-spline method uses a rather large number of evenly spaced knots over the domain. To balance against overfitting, suitable smoothness priors, which usually take the form of random walk priors with roughness parameters, are imposed to the adjacent spline coefficients [Lang and Brezger, 2004; Brezger and Lang, 2006]. Smooth functions and roughness parameters can be jointly estimated via MCMC or Laplace approximation techniques [Gressani and Lambert, 2018, 2021]. A key limitation with this estimation strategy is that the initial settings on the P-spline parameters (e.g. number of knots and priors on the roughness parameters) can have a strong impact on the estimation results. In addition, this method tends to have difficulty in capturing functions that have locally varying curvatures unless significant modifications are introduced to allow for spatially adaptive roughness parameters, which may greatly complicate the inference process [Yue et al., 2012]. The second possibility is the Bayesian adaptive (regression) spline method which takes the uncertainty regarding the number and/or positions of knots into account, treating them as unknown parameters to be inferred from the data along with other parameters. The unknown function can in principle be estimated in a locally adaptive fashion and additional regularization is usually not needed as it is indirectly achieved via model choice strategies [Jeong et al., 2020]. The final functional estimate may be obtained by model averaging over the functions sampled from a MCMC algorithm. Under this framework we may further distinguish between two estimation strategies. The first is based on the idea of Bayesian variable selection, where an adequate subset of basis functions are selected to be used in the model from a large prespecified number

of candidate basis functions [Smith and Kohn, 1996, 1997]. A potential drawback with this route is that the estimation results may rely on the initial choice of the candidate bases as is for the Bayesian P-splines, and the potentially large model and parameter spaces may pose computational challenges to MCMC algorithms (see chapter 20.2 of Gelman et al. [2013]). The alternative Bayesian free knot spline technique puts priors on both the number and location of the knots and allows for full modelling flexibility of the splines. RJMCMC algorithms are typically used to explore different knot configurations in a data-driven manner, following the seminal works of Denison et al. [1998], Biller [2000] and DiMatteo et al. [2001]. The challenges with this method lie in the design of efficient trans-dimensional moves of the MCMC which may be model dependent.

2.3 A Bayesian HMM with spline-based emissions

We assume that the emission densities $f_1, \ldots, f_N$ can be approximated by mixtures of standardized cubic B-spline basis functions (i.e. order $O = 4$) with knots located between boundary knots a and b (assumed fixed) [Langrock et al., 2015]. We use (K, R_K) to denote the interior knot configuration and set $r_{-3} = r_{-2} = r_{-1} = r_0 = a$ and $b = r_{K+1} = r_{K+2} = r_{K+3} = r_{K+4}$. Note that $K = k$ corresponds to the case of $k + 4$ B-spline basis functions, and we assume $K \geq 2$ for identifiability. Under these settings, f_i is formulated as:

$$f_i(y) = \sum_{k=1}^{K+4} a_{i,k} B_k(y), \qquad i = 1, \ldots, N, \tag{2.1}$$

where $B_k(y)$, $k = 1, \ldots, K + 4$, denotes the k-th normalized (such that it integrates to one) B-spline basis function of degree 3 for the augmented knot sequence $\bar{R}_K$ and the $a_{i,k}$ are the corresponding coefficients such that $\sum_{k=1}^{K+4} a_{i,k} = 1$ and $a_{i,k} \geq 0$, for all $k = 1, \ldots, K + 4$. In the time-homogeneous case, i.e. where the transition probabilities of the Markov chain are constant over time, the resulting class of HMMs is fully specified by the initial state distribution, $\delta = (\delta_1, \ldots, \delta_N)$, with $\delta_i = P(x_1 = i)$, the transition probability matrix, $\Gamma = (\gamma_{i,j})_{i,j=1,\ldots,N}$, with $\gamma_{i,j} = P(x_t = j | x_{t-1} = i)$, and the emission densities defined in (2.1). The joint (complete) likelihood of observations $\mathbf{y}^{(n)} = (y_1, \ldots, y_n)$ and the hidden states $\mathbf{x}^{(n)} = (x_1, \ldots, x_n)$ is

$$f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)} | K, R_K, \delta, A_K, \Gamma) = \delta_{x_1} \prod_{t=2}^{n} f(x_t | x_{t-1}, \Gamma) \prod_{t=1}^{n} f_{x_t}(y_t), \tag{2.2}$$

where here, and throughout this chapter, we shall use $f(\cdot|\cdot)$ as a generic notation to represent conditional densities as specified by their arguments and A_K denotes the set of spline coefficients $a_{i,k}$, $i = 1, \ldots, N, k = 1, \ldots, K+4$. The marginal likelihood integrating out the hidden states can be evaluated in $O(N^2 n)$ steps using the forward algorithm (in the form of Zucchini et al. [2016]), via the matrix product expression

$$\begin{aligned} f(\mathbf{y}^{(n)}|K, R_K, \delta, A_K, \Gamma) &= \int f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K, \delta, A_K, \Gamma) d\mathbf{x}^{(n)} \\ &= \delta P(y_1) \Gamma P(y_2) \cdots \Gamma P(y_n) \mathbf{1}, \end{aligned} \tag{2.3}$$

where $P(y_t)$ is a diagonal matrix with i-th diagonal entry given by $f_i(y_t)$, $\mathbf{1}$ is a column vector of ones of dimension N.

To complete the Bayesian formulation of the model, we assume the following factorization of the complete joint density

$$f(K, R_K, \delta, A_K, \Gamma, \mathbf{y}^{(n)}, \mathbf{x}^{(n)}) = f(K) f(\delta) f(\Gamma) f(R_K|K) f(A_K|K)$$

$$\times f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K, \delta, A_K, \Gamma).$$

The assumption that the parameters associated with the observed and hidden process is a-priori independent is commonly adopted in Bayesian HMMs. We use a uniform prior on $\{2, \ldots, K_{max}\}$ for K, with K_{max} fixed to 50 in our examples where a preliminary study suggested that this was large enough to cover the support of K. Clearly larger default values, including the sample size n, may be used instead and the estimation results are insensitive to its choice as long as it is large enough. For the knot positions, we propose that the r_k are taken to be the k-th order statistics of K independent uniform random variables on $[a, b]$, i.e. $f(R_K|K) = K!/(b-a)^K$. The state-specific spline coefficients $(a_{i,1}, \ldots, a_{i,K+4})$, $i = 1, \ldots, N$, are reparametrized as

$$a_{i,j} = \frac{\exp(\tilde{a}_{i,j})}{\sum_{l=1}^{K+4} \exp(\tilde{a}_{i,l})}, \qquad \tilde{a}_{i,j} \in \mathbb{R},$$

so that the positivity and unit sum constraints will not hinder the design of our reversible jump moves. The fact that the $\tilde{a}_{i,j}$ are not identifiable need not be a concern as we are only interested in the $a_{i,j}$, which remain identifiable, and in this way the mixing of the MCMC may also be improved [Cappé et al., 2003]. We choose to use a log-gamma prior with shape parameter ζ and rate parameter 1 on the $\tilde{a}_{i,j}$, i.e. $\exp(\tilde{a}_{i,j}) \sim \text{Gamma}(\zeta, 1)$, giving a symmetric Dirichlet, i.e. $\text{Dir}(\zeta, \ldots, \zeta)$ distribution on the corresponding $(a_{i,1}, \ldots, a_{i,K+4})$. We choose a vague Gamma$(1, 1)$

hyperprior on ζ to reflect our uncertainty on its value. For the transition probability matrix we assume that the rows are a-priori independent, each of which has a vague Dirichlet prior

$$(\gamma_{i,1}, \ldots, \gamma_{i,N}) \sim Dir(1, \ldots, 1), \quad i = 1, \ldots, N,$$

and we assume that the initial distribution is fixed and uniform on $\{1, \ldots, N\}$. Note that it is not possible to estimate it consistently as there is only one unobserved variable associated with it. Thus the complete joint density incorporating the reparametrization can be rewritten as

$$f(\zeta, K, R_K, \tilde{A}_K, \Gamma, \mathbf{y}^{(n)}, \mathbf{x}^{(n)}) = f(\zeta) f(K) f(\Gamma) f(R_K|K) f(\tilde{A}_K|K, \zeta)$$

$$\times f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K, \tilde{A}_K, \Gamma), \tag{2.4}$$

where $\tilde{A}_K$ represents the set of $\tilde{a}_{i,k}$ $(i = 1, \ldots, N, k = 1, \ldots, K + 4)$. Note that if we relabel the hidden states and rearrange the state specific parameters, the HMM likelihood and the joint prior for the parameters remain unchanged, and thus the joint posterior density corresponding to (2.4) is defined on $N!$ subspaces, one for each permutation of the labels of the hidden states.

2.4 The reversible jump MCMC algorithm

Our aim is to obtain realisations from the posterior distribution of $(K, R_K, A_K, \Gamma, \zeta)$, which can be achieved by simulating from the joint posterior density defined through (2.4). To allow for model searches between parameter subspaces of different dimensionality, we develop a RJMCMC algorithm which combines a Metropolis-within-Gibbs sampler with birth and death trans-dimensional moves of the knot points (and the associated spline coefficients). The structure of our algorithm is listed in Algorithm 5, where $b_K = \mathbf{I}(K = 2) + 0.5 \times \mathbf{I}(3 \leq K < K_{max})$ and $\mathbf{I}(\cdot)$ is the indicator function. Steps (a)-(e) propose moves within a dimension while the last step proposes a birth or death of a knot point which changes the model dimension. We now give the rules for each of the updating steps while further details regarding the validity and implementation of the algorithm are provided in the appendix of this chapter. Throughout this section we assume that the cardinality N is fixed noting that model selection will be addressed in section 2.5.

Algorithm 5: Reversible jump MCMC algorithm for spline-based HMMs

Initialize K, R_K, ζ, $\tilde{A}_K$, Γ ;
for *i=1, . . . , T* **do**
  (a) update the hidden state sequence $\mathbf{x}^{(n)}$;
  (b) update the transition probability matrix Γ;
  (c) update the knot location vector R_K;
  (d) update the set of B-spline coefficients A_K (via $\tilde{A}_K$);
  (e) update the hyperparameter ζ;
  draw $U \sim U(0, 1)$;
  if $U < b_K$ **then**
    consider the birth of a knot point in the B-spline representation in (2.1);
  else
    consider the death of a knot point in the B-spline representation in (2.1);
  end
end

2.4.1 Within-model moves

The moves in steps (a) and (b) are of Gibbs type whereas those in steps (c) to (e) are of Metropolis-Hastings (henceforth MH) type. In step (a), $\mathbf{x}^{(n)}$ can be simulated exactly and efficiently from its full conditional distribution, $f(\mathbf{x}^{(n)}|\mathbf{y}^{(n)}, \mathrm{K}, R_K, \tilde{A}_K, \Gamma)$, via a standard FFBS procedure as introduced in chapter 1, with transition matrix Γ and emission densities $f_i(y_t)$ given in (2.1). Step (b) is performed via a standard Gibbs step as used routinely for basic Bayesian HMMs. Conditional on $\mathbf{x}^{(n)}$, the rows of Γ are conditionally independent and are updated from their conjugate Dirichlet posterior

$$(\gamma_{i,1}, \ldots, \gamma_{i,N}) \sim Dir(1 + n_{i,1}, \ldots, 1 + n_{i,N}), \quad i = 1, \ldots, N,$$

where $n_{i,j}$ denotes the number of transitions from state i to j in $\mathbf{x}^{(n)}$. In step (c) a knot r_{k^*} is chosen uniformly from the set of existing knots $\{r_1, \ldots, r_K\}$ and proposed to be moved to a candidate point, r_c, which is generated from a normal distribution with mean r_{k^*} and standard deviation τ_1, truncated to $[a, b]$ [DiMatteo et al., 2001]. The proposal in step (d) is generated by a random walk on the reparametrized spline coefficients $\tilde{a}_{i,j}$ $(i = 1, \ldots, N; j = 1, \ldots, K + 4)$, i.e.

$$\tilde{a}'_{i,j} = \tilde{a}_{i,j} + \eta_{i,j},$$

where $\eta_{i,j} \sim \mathcal{N}(0, \tau_2^2)$. In step (e), we update ζ via a log-normal random walk

$$\log(\zeta^{'}) = \log(\zeta) + \nu,$$

where $\nu \sim \mathcal{N}(0, \tau_3^2)$. The variance parameters τ_1, τ_2 and τ_3 may be regarded as tuning parameters that need to be adjusted to achieve a satisfactory mixing of the chain. Here, we adopt a simple yet well-used adaptive MCMC scheme based on a stochastic approximation procedure to allow for automatic tuning during the burn-in period [Atchade et al., 2011], without incurring additional computational burden. More specifically, the scaling parameter τ_i is adapted from iteration $t-1$ to t as

$$\tau_i^{(t)} = \max\left(\tau_i^{(t-1)} + \epsilon(t)\text{sgn}\big(\frac{1}{T_a}\sum_{j=t-T_a+1}^{t}\rho_i^{(j)} - \rho_i^*\big),\ \epsilon_i\right)$$

where $\epsilon(t) = \min(0.01, 1/\sqrt{t})$ following suggestions in Roberts and Rosenthal [2009] and Rosenthal [2007], $\text{sgn}(\cdot)$ is the sign function, $\rho_i^{(j)}$ is the MH acceptance rate in iteration j, ρ_i^* is the targeted acceptance rate and ϵ_i is a sufficiently small positive number. That is, we adjust the scaling parameter at every iteration by adding or subtracting a factor $\epsilon(t)$ (whose magnitude is diminishing) if the averaged acceptance rate over the past T_a iterations is below or exceed the target ρ_i^*. We refer to Green et al. [2015] and references therein for an in-depth description of the theory and methods behind the adaptive MCMC approaches. In our context we set $\rho_1^* = \rho_3^* = 0.4$ and $\rho_2^* = 0.24$ based on the optimal scaling results for MH algorithms (see, e.g. Gelman et al. [1997] and Roberts and Rosenthal [2001]) and set $T_a = 10$ (inspired by results in Marshall and Roberts [2012]) and $\epsilon_i = 10^{-6}$ in our examples. In our examples, the lower bound ϵ_i is usually not reached as the algorithm stabilizes well.

2.4.2 Birth and death moves

The birth and death moves allow for increasing or decreasing the number of knots, or equivalently, the number of B-spline basis elements. Our design extends the ideas in DiMatteo et al. [2001] and Sharef et al. [2010] to the framework of HMMs defined in section 2.3. Suppose that the current model has knot configuration (K, R_K), we first make a random choice between birth and death with probabilities b_K and $d_K = 1 - b_K$, respectively. In the birth move, we select a knot, r_{b^*}, at random from the existing knots and create a candidate new knot, r_c, by drawing from a normal distribution (truncated to $[a, b]$) with mean r_{b^*} and standard deviation $\tau(R_K, b^*)$, where τ is chosen as a function having the form $(r_{b^*+1} - r_{b^*-1})^\alpha$ and α is a positive real constant. The intuition here is that a new knot is more likely to

be needed in locations where existing knots are relatively "dense". To complete the birth step we update the corresponding spline coefficients, which now has dimension $K+5$ for each state. Here, our design is guided by the deterministic knot insertion rule described in De Boor [2001] which allows a new knot to be inserted without changing the shape of the overall B-spline curve, noting that in our context this exact relationship becomes approximate as we are working with normalized basis functions. We extend the scheme by adding more degrees of freedom in order to meet the dimension matching condition required for the validity of the RJMCMC algorithm. More specifically, for the birth of a candidate knot point $r_c \in (r_{n^*}, r_{n^*+1})$, the associated spline parameters $\tilde{a}'_{i,j}$, for $i = 1, \ldots, N$, are created as

$$\tilde{a}'_{i,j} = \begin{cases} \tilde{a}_{i,j} & 1 \leq j \leq n^*+1 \\ c_j \tilde{a}_{i,j} + (1-c_j)\tilde{a}_{i,j-1} & n^*+1 < j < n^*+4 \\ u_i \tilde{a}_{i,j} + (1-u_i)\tilde{a}_{i,j-1} & j = n^*+4 \\ \tilde{a}_{i,j-1} & n^*+4 < j \leq K+5 \end{cases} \tag{2.5}$$

where $c_j = (r_c - r_{j-4})/(r_{j-1} - r_{j-4})$ and $u_i \overset{iid}{\sim} U(0,1)$. Here the $\tilde{a}'_{i,j}$ are generated using the deterministic rule in De Boor [2001], except for $\tilde{a}'_{i,n^*+4}$ we introduce one degree of freedom through u_i. This way of updating allows us to effectively use knowledge from current spline parameters, while also allowing for a possible improvement on the fit resulting from the introduction of a new knot point. Our design can also be related to the idea of "centering" reversible jump proposals proposed in Brooks et al. [2003] where current and proposed parameters produce similar likelihoods. The parameters associated with the state process are unchanged for this move.

Next, consider the death of a knot point from the current knot configuration (K, R_K). A knot, r_{d^*}, is chosen at random from the set of existing knots $\{r_1, \ldots, r_K\}$ and then deleted. The spline parameters associated with this move are updated according to the inverse transformation of (2.5):

$$\tilde{a}'_{i,j} = \begin{cases} \tilde{a}_{i,j} & 1 \leq j \leq d^* \\ \frac{\tilde{a}_{i,j} - (1-c_j)\tilde{a}'_{i,j-1}}{c_j} & d^* < j < d^*+3 \\ \tilde{a}_{i,j+1} & d^*+3 \leq j \leq K+3 \end{cases}$$

where $c_j = (r_{d^*} - r_{j-4})/(r_{j-1} - r_{j-4})$. The parameters for the state process remain unaltered in this move. We note the difference between our birth and death

proposals to those in Sharef et al. [2010] (see equation 3.1 therein), who propose a parameterization where the transformation acts on the exponentials of the spline coefficients (restricted to be positive). Such a scheme is problematic as the proposed parameters from the death step based on the deterministic rules are not guaranteed to be positive.

2.5 Bayesian model selection

Up to now we have taken the cardinality N. Next we address model selection. In principle we could extend the proposed RJMCMC algorithm, Algorithm 5, by introducing an additional reversible jump step on the number of states, or by working with a product space search algorithm to sample from the joint posterior of parameters from all competing models (e.g. Carlin and Chib [1995]). However, in the present HMM setting with spline-based emissions, efficient and computationally practical trans-dimensional algorithms are very difficult to design due to the potentially large and complex parameter space. Instead, we propose to perform Bayesian model selection based on the marginal likelihood (also known as the evidence)

$$f(\mathbf{y}^{(n)}|N=j) = \int f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_j, N=j) f(\boldsymbol{\theta}_j|N=j) d\boldsymbol{\theta}_j, \quad j=1,\dots,M, \tag{2.6}$$

where $\boldsymbol{\theta}_j$ is the parameter set associated with the j-state model, $f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_j, N=j)$ is the observed likelihood given in (2.3) and M denote the maximum number of states that we want to consider. Given prior model probabilities $p(N=j)$, the posterior model probabilities can be computed as

$$P(N=j|\mathbf{y}^{(n)}) = \frac{f(\mathbf{y}^{(n)}|N=j)p(N=j)}{\sum_{i=1}^{M} f(\mathbf{y}^{(n)}|N=i)p(N=i)}, \quad j=1,\dots,M, \tag{2.7}$$

and following Bayesian decision theory we can pick the model that gives the highest posterior probability, i.e. $N^* = \text{argmax}_{k=1,\dots,M} P(N=k|\mathbf{y}^{(n)})$. For most models of interest (including HMMs), however, the integral in (2.6) has no closed-form expression and needs to be approximated. We refer to Ardia et al. [2012], Friel and Wyse [2012] and Llorente et al. [2020] for some recent reviews of various Monte Carlo based approximation schemes for the evidence (or ratios of two evidences, i.e. Bayes factors).

We propose to approximate the evidence of a spline-based HMM by using a harmonic mean estimator originally proposed in Gelfand and Dey [1994], although modifications of other popular estimators such as the method by Chib and Jeliazkov

[2001] may also be applicable in this context. The main advantage of the former approach is that it allows direct estimation of the evidence using the simulation output and thus is straightforward to implement, while the latter requires additional simulation runs and calibrations during the estimation process, leading to more computational costs. Our chosen estimator relies on the simple fact that for any proper density function h, we have for the expectation

$$\mathbf{E}_{\boldsymbol{\theta}_j|\mathbf{y}^{(n)}}\left[\frac{h(\boldsymbol{\theta}_j)}{f(\boldsymbol{\theta}_j)f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_j)}\right] = \int \frac{h(\boldsymbol{\theta}_j)}{f(\boldsymbol{\theta}_j)f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_j)} f(\boldsymbol{\theta}_j|\mathbf{y}^{(n)})d\boldsymbol{\theta}_j = \frac{1}{\mathbf{M}_j},$$

where $\mathbf{M}_j = \int f(\boldsymbol{\theta}_j)f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j$. A Monte Carlo approximation of the evidence is thus obtained as

$$\hat{\mathbf{M}}_j = \left\{\frac{1}{T}\sum_{i=1}^{T}\frac{h(\boldsymbol{\theta}_j^{(i)})}{f(\boldsymbol{\theta}_j^{(i)})f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_j^{(i)})}\right\}^{-1}$$

where $\boldsymbol{\theta}_j^{(i)}$ is the i-th sample simulated from the posterior $f(\boldsymbol{\theta}_j|\mathbf{y}^{(n)})$. This estimator enjoys a finite variance if $\int h^2(\boldsymbol{\theta})/(f(\boldsymbol{\theta})f(\mathbf{y}^{(n)}|\boldsymbol{\theta}))d\boldsymbol{\theta} < \infty$, i.e. $h(\boldsymbol{\theta})$ must have lighter tails than $f(\boldsymbol{\theta})f(\mathbf{y}^{(n)}|\boldsymbol{\theta})$ [DiCiccio et al., 1997]. We follow Robert and Wraith [2009] and Marin and Robert [2009] to construct such an appropriate density h based on truncated highest posterior density (HPD) regions derived from the MCMC samples. The resulting estimator is known as a truncated harmonic mean estimator and has been successfully used in various model settings, see for instance Durmus et al. [2018] and Acerbi et al. [2018]. More specifically, we define a sample-based $100\beta\%$ HPD region as (omitting the dependence on the index of state j for clarity)

$$\tilde{\mathbf{H}}_\beta = \{\boldsymbol{\theta}^{(i)} : \ f(\boldsymbol{\theta}^{(i)})f(\mathbf{y}^{(n)}|\boldsymbol{\theta}^{(i)}) > \tilde{q}_\beta\},$$

where $\tilde{q}_\beta$ is the empirical upper β quantile of the $(f(\boldsymbol{\theta}^{(i)})f(\mathbf{y}^{(n)}|\boldsymbol{\theta}^{(i)}))$ produced in the output of the MCMC. We then construct the density h as

$$h(\boldsymbol{\theta}) = \frac{1}{V(\xi)\beta T}\sum_{j:\boldsymbol{\theta}^{(j)}\in\tilde{\mathbf{H}}_\beta, dim(\boldsymbol{\theta}^{(j)})=dim(\boldsymbol{\theta})} \mathbf{I}(d(\boldsymbol{\theta}^{(j)},\boldsymbol{\theta}) < \xi),$$

where $V(\xi)$ is the volume of a ball centered at $\boldsymbol{\theta}$ with radius ξ (small), $dim(\cdot)$ is the dimensionality of the argument and $d(\cdot,\cdot)$ is a suitable distance measure. It is easy to check that h is a proper density function and has a finite support, noting that the parameter space of $\boldsymbol{\theta} = (K, \zeta, R_K, \tilde{A}_K, \Gamma)$ is a union of subspaces of varying dimension. Our proposal h may be interpreted as a histogram-like nonparametric

estimator of the posterior $f(\boldsymbol{\theta}|\mathbf{y}^{(n)})$ based only on samples in the HPD regions. Notice also that $V(\xi)$ does not need to be computed as it will be cancelled out in (2.7), as long as we fix ξ across models.

We would like to comment further on two appealing alternatives in the literature which attempt to perform model selection using only independent MCMC outputs from each candidate model, thus also permitting a parallel sampling framework as for the harmonic mean estimator proposed here. The first one is Congdon's estimator [Congdon, 2006] which is inspired by the product space approach of Carlin and Chib [1995], and is advocated in popular HMM textbooks [Bartolucci et al., 2019; Zucchini et al., 2016] and few research papers such as Chen et al. [2011]. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ and $\boldsymbol{\Theta}^{(i)} = (\boldsymbol{\theta}_1^{(i)}, ..., \boldsymbol{\theta}_M^{(i)})$, $i = 1, ..., T$, represent the i-th parallel draw from $f(\boldsymbol{\theta}_k|y^{(n)})$, $k = 1, ..., M$. With the simplifying assumptions that (i) $f(\mathbf{y}^{(n)}|\boldsymbol{\Theta}, N = k) = f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_k, N = k)$, (ii) $f(\boldsymbol{\Theta}|N = k) = \prod_{i=1}^{M} f(\boldsymbol{\theta}_i|N = k)$ and (iii) $f(\boldsymbol{\theta}_{j\neq k}|N = k) \propto constant$, Congdon [2006] propose to estimate $P(N|\mathbf{y}^{(n)})$ by the ensemble average

$$\hat{P}(N = k|\mathbf{y}^{(n)}) = \frac{1}{T}\sum_{i=1}^{T} P(N = k|\mathbf{y}^{(n)}, \boldsymbol{\Theta}^{(i)}), \tag{2.8}$$

where $P(N = k|y^{(n)}, \boldsymbol{\Theta}^{(i)}) \propto f(y^{(n)}|\theta_k^{(i)}, N = k) f(\theta_k^{(i)}|N = k) P(N = k)$ (see Congdon [2006] for more details). However, as pointed out in Robert et al. [2008], the estimator in (2.8) is biased (and not valid in a strict sense) as the aggregated chain $\boldsymbol{\Theta}^{(i)}$ is essentially simulated based on $\prod_{i=1}^{M} f(\boldsymbol{\theta}_i|y^{(n)})$, and not the "correct" joint posterior $f(\boldsymbol{\Theta}|\mathbf{y}^{(n)}) \propto \sum_{i=1}^{M} p(N = i) f(\boldsymbol{\Theta}|N = i) f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_k, N = i)$. In fact, assumption (iii) makes $f(\boldsymbol{\Theta}|\mathbf{y}^{(n)})$ undefined noting that it can be expressed as $\sum_{i=1}^{M} p(N = i|\mathbf{y}^{(n)}) f(\boldsymbol{\theta}_i|y^{(n)})$ (see Robert et al. [2008] for further discussions). Although we found it to perform accurately when testing it in our simulation experiments, it should be used with caution as the theoretical underpinnings of the estimator are problematic.

The second approach, that we also tested in our simulations, is based on the deviance information criterion (DIC) which is originally developed in Spiegelhalter et al. [2002] and may be regarded as a Bayesian version of the Akaike information criterion (AIC). For latent variable models including HMMs, we may distinguish between the observed likelihood DIC and the conditional likelihood DIC, depending on whether the latent variables are integrated out or not [Celeux et al., 2006]. The latter version has several issues from both practical and theoretical viewpoints and the former is generally preferred as long as the observed likelihood can be easily

computed (see Li et al. [2020b] and references therein), which is indeed the case for HMMs. Adopting the form used in Chan and Grant [2016], the observed likelihood DIC is defined as

$$DIC(m) = \mathbf{E}_{\boldsymbol{\theta}_m}[D(\boldsymbol{\theta}_m)|\mathbf{y}^{(n)}] + P_D(m), \tag{2.9}$$

where $D(\boldsymbol{\theta}_m) = -2\log(f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_m))$ is the deviance for model m, $P_D(m) = \mathbf{E}_{\boldsymbol{\theta}_m}[D(\boldsymbol{\theta}_m)|\mathbf{y}^{(n)}] - D(\hat{\boldsymbol{\theta}}_m)$ is a measure of effective number of parameters and $\hat{\boldsymbol{\theta}}_m$ is the posterior mode of $\boldsymbol{\theta}_m$. A sample-based approximation of (2.9) is obtained as

$$\widehat{DIC}(m) = -\frac{4}{T}\sum_{i=1}^{T}\log(f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_m^{(i)})) + 2\log(f(\mathbf{y}^{(n)}|\hat{\boldsymbol{\theta}}_m),$$

where $\hat{\boldsymbol{\theta}}_m = \text{argmax}_{\boldsymbol{\theta}_m^{(i)}:i=1\ldots T} f(\mathbf{y}^{(n)}|\boldsymbol{\theta}_m^{(i)})f(\boldsymbol{\theta}_m^{(i)})$. The model with the lowest $\widehat{DIC}$ is favoured. In our simulation study described below, however, we observe consistent poor performance of the DIC criterion, with a tendency towards overfitting. In addition, there are general criticisms of the DIC in the context of Bayesian model comparison, including, for instance, the use of a plug-in predictive approach instead of a proper predictive distribution and the fact that it may not be invariant with respect to reparametrization [Maity et al., 2021; Spiegelhalter et al., 2014].

2.6 Simulation studies

We conduct four simulation experiments to thoroughly evaluate the feasibility of the proposed Bayesian methodology (adSP method), and to compare its performance mainly with alternative spline-based methods for the HMM emissions, including a Bayesian P-spline approach (bpSP) which is investigated for the first time in this context (see below for further descriptions) and the frequentist P-spline approach of Langrock et al. [2015] (fpSP). Our comparison is mainly based on the following two criteria

1. **Ability to recover the true emission distributions**: This is quantified by the average Kullback-Leibler divergence (KLD):

$$(\mathbf{KLD}(\hat{f}_i||f_i) + \mathbf{KLD}(f_i||\hat{f}_i))/2, \quad i = 1,\ldots,N,$$

where $\mathbf{KLD}(\hat{f}_i||f_i) = \int \hat{f}_i(y)\log(\hat{f}_i(y)/f_i(y))dy$ and $\hat{f}_i$ is the estimated emission density for state i. In Bayesian MCMC, used with adSP and bpSP methods, we estimate the unknown emission densities (pointwise) through the posterior expectation $\mathbf{E}[f_i(y)|\mathbf{y}^{(n)}]$, which is the Bayes estimator of $f_i(y)$ under

posterior mean squared error loss and can be approximated by the Monte Carlo average

$$\hat{f}_i(y) = \frac{1}{T}\sum_{j=1}^{T} f_i^{(j)}(y), \quad i = 1, \ldots, N,$$

where $f_i^{(j)}(y)$ is the emission density arising from the j-th MCMC sample (as a function of the knot configuration and spline coefficients) and y is a fixed point in the domain of the observed data. Density estimates for the fpSP method are constructed as in Langrock et al. [2015] where a single set of the penalized maximum likelihood estimates of the spline coefficients is plugged into equation (2.1). In our implementations we used the KLD function in the R package *LaplacesDemon* [Statisticat and LLC., 2021] for approximating the average KLD.

2. **Decoding accuracy**: Decoding is to infer the hidden state process $\mathbf{x}^{(n)}$ based on the observed $\mathbf{y}^{(n)}$ and is one of the key inference tasks in HMMs. We quantify the discrepancy/agreement between the estimated and true state sequences via the normalized Hamming distance/decoding accuracy as commonly used in the HMM context [Fox et al., 2011], which essentially measures the proportion of the incorrectly/correctly classified states. For the adSP and bpSP methods we estimate the states by first estimating the marginal state probability based on the MCMC samples of $\mathbf{x}^{(n)}$

$$\hat{P}(x_t = k|\mathbf{y}^{(n)}) = \frac{1}{T}\sum_{j=1}^{T} \mathbf{I}(x_t^{(j)} = k), \quad t = 1, \ldots, n, \quad k = 1, \ldots, N,$$

where $x_t^{(j)}$ is the value of x_t from the j-th MCMC draw. We then determine the value of x_t such that its posterior state probability is maximized (also referred to as local decoding in the HMM literature). Note that a more elaborate Rao–Blackwellized estimator may be used for $P(x_t = k|\mathbf{y}^{(n)})$, see Scott [2002]. However, it requires additional computational effort for running the forward-backward recursion at each MCMC iteration, and in our studies no accuracy gains have been found in terms of the final decoding performance. For the fpSP method we follow Langrock et al. [2015] to perform decoding via the Viterbi algorithm, conditional on the point estimates of the model parameters [Zucchini et al., 2016].

2.6.1 Description of experiments

We now describe in detail the 4 simulation models used for testing and comparing the estimation performance of the competing methods. Model 1 is a 2-state HMM (see Figure 2.1) originally considered in Langrock et al. [2015] with emission distributions:

$$y_t|x_t = 1 \sim \mathcal{N}(-15, 11^2),$$

$$y_t|x_t = 2 \sim 0.35\mathcal{N}(-5, 9^2) + 0.65\mathcal{N}(30, 10^2),$$

where the states of the underlying Markov chain were generated from $\delta = (1/2, 1.2)$ and $\gamma_{12} = \gamma_{21} = 0.1$ and $n = 800$. For Model 2, we consider a 3-state HMM with a unimodal positively skewed emission distribution in state 1, a bimodal distribution in state 2 and a unimodal negatively skewed distribution in state 3 (see Figure 2.2). We use B-splines to construct these densities and the details of the spline parameters are omitted here. The states were generated using $\delta = (1/3, 1/3, 1/3)$ and

$$\Gamma = \begin{pmatrix} 0.85 & 0.1 & 0.05 \\ 0.075 & 0.85 & 0.075 \\ 0.05 & 0.1 & 0.85 \end{pmatrix},$$

from which $n = 1500$ observations were simulated from the corresponding emission distribution using the inverse transform sampling scheme [Devroye, 1986]. Model 3 is motivated and modified from the *bimod* model considered in Yau et al. [2011]. We construct the emissions using a mixture of a Laplace and a generalized Student's t distribution (see Figure 2.3):

$$y_t|x_t = 1 \sim 0.5\mathbf{Laplace}(-1, 0.2) + 0.5\mathbf{t}_3(2, 2),$$

$$y_t|x_t = 2 \sim 0.5\mathbf{Laplace}(0.5, 0.2) + 0.5\mathbf{t}_3(3.5, 2),$$

where $\mathbf{Laplace}(\mu, \sigma)$ denotes a Laplace distribution with location parameter μ and scale parameter σ and $\mathbf{t}_\nu(\mu, \sigma)$ denotes a generalised t distribution with ν degrees of freedom (assume $\nu > 2$), mean μ and standard deviation σ. By construction one of the emission can be obtained by translating the other emission horizontally, and the emissions have varying degree of smoothness across the domain, which can be expected to pose challenges to P-spline based inference methods. For this model we set $\delta = (0.5, 0.5)$, $\gamma_{12} = \gamma_{21} = 0.05$ and $n = 2000$. The last model, Model 4, is a 2-state HMM considered in Yau et al. [2011] (the *trimod* case) with emissions

specified as a mixture of three well-separated normal distributions (see Figure 2.4):

$$y_t|x_t = 1 \sim \frac{1}{3}\mathcal{N}(-4,1) + \frac{1}{3}\mathcal{N}(0,1) + \frac{1}{3}\mathcal{N}(8,1),$$

$$y_t|x_t = 2 \sim \frac{1}{3}\mathcal{N}(-3,1) + \frac{1}{3}\mathcal{N}(1,1) + \frac{1}{3}\mathcal{N}(9,1),$$

and the same Markov chain parameters as in Model 3. For this model we consider a relatively large data set of length $n = 5000$. In each of our chosen simulation scenarios, the emissions exhibit a subset of the features such as multi-modality, skewness, heavy-tailedness and excess kurtosis, and the number or the structure of the emissions are not directly identifiable by visually inspecting the empirical marginal distributions (e.g. histogram). Models 3 and 4 pose the most serious computational challenges even when the correct number of states is assumed to be known as in Yau et al. [2011].

2.6.2 Settings and results

We first present the settings for implementing the methods mentioned above. For our proposed algorithm the unspecified constants are set to $a = \min(\mathbf{y}^{(n)}) - 10$, $b = \max(\mathbf{y}^{(n)}) + 10$ and $\alpha = 0.65$ in all scenarios (experiments suggest that our results are not very sensitive to these specific choices). For the fpSP method we select the number of equidistant knots K and the state-specific smoothing parameters $\lambda = (\lambda_1, \ldots, \lambda_N)$ either based on the original choices in Langrock et al. [2015] (for Model 1) or on pre-experiments of our own (for Models 2-4), and the details are summarized in Table 2.1. To set up the Bayesian P-spline model for the emissions, we use the same knot configuration (prefixed) as for the frequentist P-spline counterpart to facilitate comparison. We follow Lang and Brezger [2004] to use a second order random walk prior on the reparametrized spline coefficients, which is translated to a multivariate normal prior on the $\tilde{a}_{i,j}$

$$\tilde{\mathbf{a}}_i = (\tilde{a}_{i,1}, \ldots, \tilde{a}_{i,K+4})|\tilde{\tau}_i \sim \mathcal{N}_{K+4}(0, (\tilde{\tau}_i P)^{-1}), \quad i = 1, \ldots, N, \tag{2.10}$$

where $\tilde{\tau}_i$ is regarded as the state-specific roughness parameter, $P = D_2^T D_2 + \tilde{\epsilon}\mathcal{I}_{K+4}$ is the penalty matrix with $D_2 \in \mathbf{R}^{(K+2)\times(K+4)}$ the second order difference matrix defined as

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & & & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & & & 0 \\ \vdots & & \ddots & \ddots & \ddots & & & & \vdots \\ 0 & & & & & \cdots & 0 & 1 & -2 & 1 \end{pmatrix},$$

Table 2.1: Settings for the fpSP method.

Parameters	Model 1	Model 2	Model 3	Model 4
K	27	47	71	51
λ	$(2048, 1024)$	$(2800, 1600, 2400)$	$(1, 1)$	$(600, 600)$

$\tilde{\epsilon}$ is a small positive quantity and $\mathcal{I}_{K+4}$ is the identity matrix of dimension $K+4$. Note that the addition of $\tilde{\epsilon}\mathcal{I}_{K+4}$ makes P a full rank matrix and here we take $\tilde{\epsilon} = 10^{-6}$ as in Lambert and Bremhorst [2020] and Maturana-Russel and Meyer [2021]. For $\tilde{\tau}_i$ a conjugate Gamma prior is commonly used [Lang and Brezger, 2004]. Here, to address the impact of the prior parameters on the smoothness of the spline fit, we adopt a robust specification by introducing the following hyperpriors [Jullion and Lambert, 2007]

$$\tilde{\tau}_i|\tilde{\tau}^{'} \sim \mathbf{Gamma}(\alpha_{\tilde{\tau}}, \alpha_{\tilde{\tau}}\tilde{\tau}^{'}), \quad i = 1, \ldots, N,$$
$$\tilde{\tau}^{'} \sim \mathbf{Gamma}(\alpha_{\tilde{\tau}'}, \beta_{\tilde{\tau}'}).$$

We choose $\alpha_{\tilde{\tau}'} = \beta_{\tilde{\tau}'} = 10^{-3}$ following the suggestion in Jullion and Lambert [2007]. The choice of $\alpha_{\tilde{\tau}}$ is not influential and we set $\alpha_{\tilde{\tau}} = 1$ as in Bremhorst and Lambert [2016] and Maturana-Russel and Meyer [2021]. Details for the associated MCMC algorithm are provided in the appendix. Clearly more sophisticated Bayesian P-spline methods exist, such as those permitting spatially adaptive smoothing parameters, but our aim here is to compare with the "standard" approach as commonly used in practice.

To take care of the variability in the simulated data set, for each of the four simulation models 10 random replications of the data were generated and the three methods were implemented for each replications. For Models 1-3 our posterior samples (for the adSP and bpSP methods) are based on 25k iterations of the corresponding MCMC samplers after a burn-in of 25k iterations, and for Model 4 the burn-in period is increased to 35k iterations. We monitored convergence of the parameters as well as the likelihood of the models generated by the Markov chain and found that the chosen burn-in periods are generally sufficient to obtain reliable results. Further details regarding the performance of the proposed RJMCMC algorithm are discussed in the Appendix.

Model selection

For each simulation model, we first implement the marginal likelihood based approach described in section 2.5 to examine its performance in recovering the true

number of states, when applied in conjunction with the Bayesian MCMC algorithms for the spline-based HMM. In each repetition we collect MCMC samples using both the adSP and bpSP methods where we place a uniform prior on N over the candidate set $\{2, 3, 4, 5\}$. Throughout we set $\beta = 0.2$ and $\xi = 0.01$ and experiments suggest that the results are robust to these choices provided that they are chosen to be relatively small as specified here. For the adSP method, the correct number of states is identified in all repetitions of all simulation scenarios, with averaged posterior probability of the correct model equal to one (rounded to 3 decimal places). The bpSP method, however, suffered from an overestimation of the number of states in most repetitions. We hypothesize that this may be due to the specific structure of the prior in (2.10), which is "almost improper" (becomes improper as $\tilde{\epsilon} \to 0$). It is well known that improper priors can cause troubles in the evaluation of the marginal likelihood.

Comparison with fixed N

We now report results obtained for each of the three methods, conditional on the true value of N for each simulation model. Figure 2.1 (top and bottom left panels) shows the true as well as estimated emission densities obtained in the 10 repetitions for Model 1 and these agree reasonably well in general for all methods. $K = 6$ (i.e. 10 basis elements) or 7 are suggested by the adSP method as the posterior mode with an equal frequency of 50%. This is to be compared with the bpSP and fpSP methods where 31 basis elements are used for the estimation. We can see that fitted models obtained by our proposed method are considerably more parsimonious but still achieve comparable or even slightly better (for state 1) density fits in terms of the average KLD (see bottom right panel of Figure 2.1). In terms of decoding, the adSP, bpSP and fpSP methods yielded accuracy of 94.3%, 94% and 93.6% (averaged over the 10 repetitions), respectively, indicating a slightly better performance of the adSP method. For the second simulation model, $K = 9$ or 10 are suggested by the adSP method as the posterior mode with an equal frequency of 50%, and our algorithm gives reasonable appearing estimates for the emissions in all 10 repetitions (see top left panel of Figure 2.2). By contrast, due to inefficient knot placements, the two P-spline-based approaches use a much larger $K = 47$ to prevent underfitting yet still give generally poorer fits compared with our adSP (see bottom right panel of Figure 2.2), with the fpSP method performing worst in modelling both the smooth and non-smooth parts of the densities. The average decoding accuracy over the 10 replicates is 90.7%, 90.5% and 90.5% for models estimated using the adSP, bpSP and fpSP methods, respectively (see Figure 2.5 for more details) and thus relatively

similar for all 3 approaches.

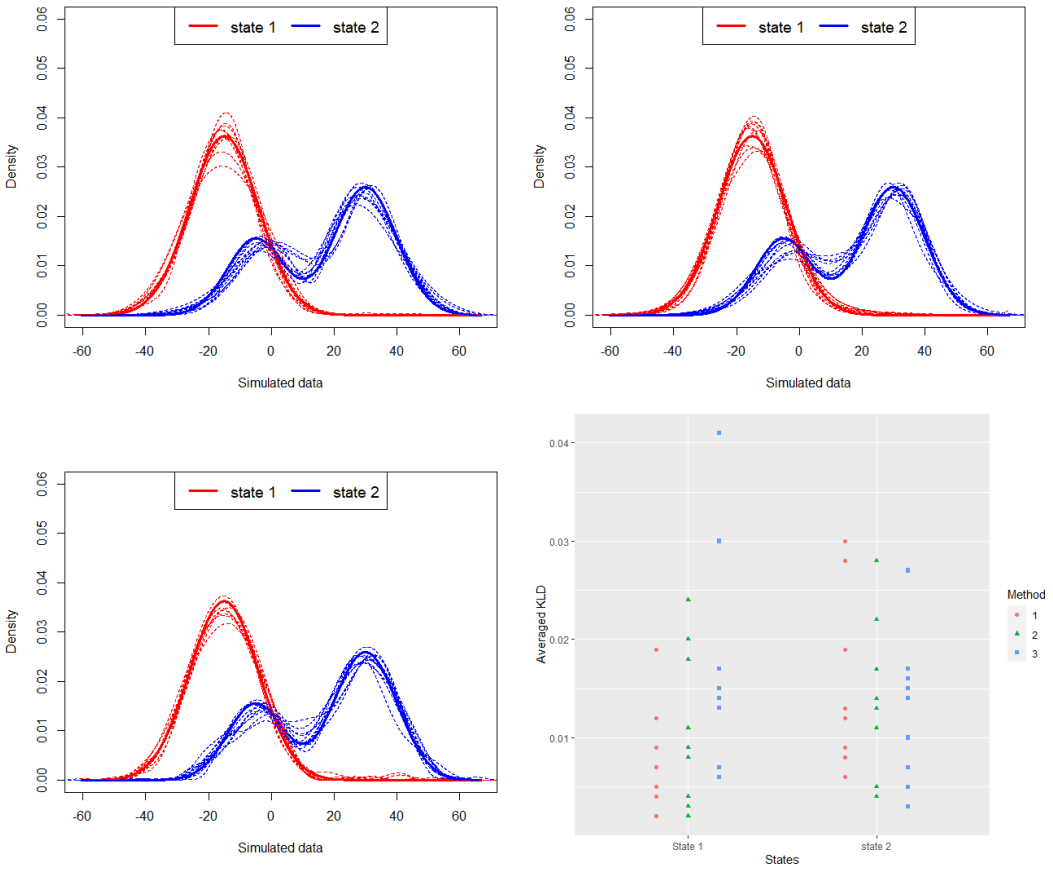

Figure 2.1: Estimation results for 10 simulations of Model 1. Top left, top right and bottom left panels show the true (solid curves) and estimated (dashed curves) densities of the emission distributions obtained in each replication using the adSP, bpSP and fpSP methods, respectively. Bottom right panel shows the average KLD for each state obtained in each replication with the three methods (methods 1, 2 and 3 correspond to the adSP, bpSP and fpSP methods, respectively).

Models 3 and 4 are, by design, more challenging simulation scenarios, where we see the adSP performs well in both cases but the other two methods suffer from numerical and convergence issues. For model 3, $K = 13$ or 14 is suggested by our algorithm with an equal frequency of 50%, and we can see from Figure 2.3 (top left panel) that the estimated emissions capture both the sharp peak and the smooth parts reasonably well. The fpSP method, however, even with careful selection of the initial parameter values, failed to converge (or converged to sub-optimal solutions) in 50% of the simulation runs. Within the convergent repetitions, both bpSP and fpSP

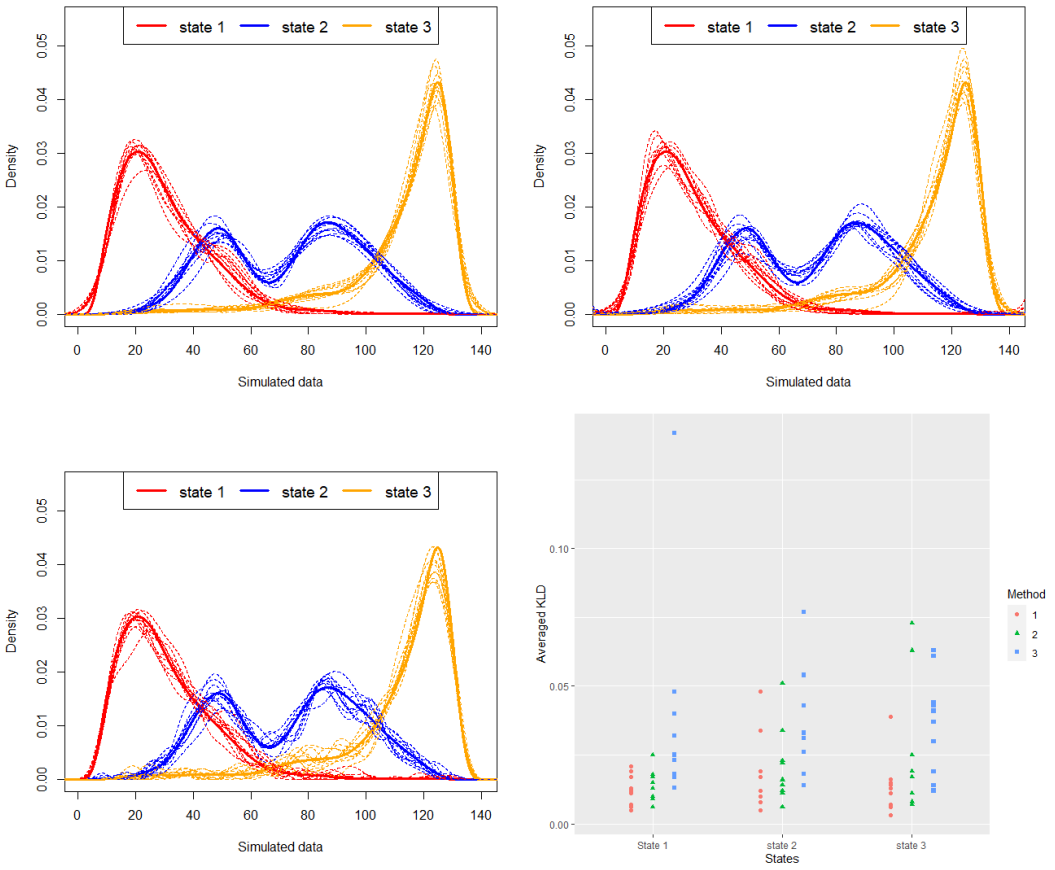

Figure 2.2: Estimation results for 10 simulations of Model 2. The settings are the same as in Figure 2.1.

methods lead to significantly larger average KLDs for the emissions (see bottom right panel of Figure 2.3). These results are not very surprising since both methods assume global smoothing parameters whereas the true densities are subject to significantly differing degree of smoothness over the domain. Regarding decoding, the adSP and bpSP methods perform roughly equally well, whereas results from the fpSP method exhibits much higher variability (see Figure 2.5).

For model 4, the tri-modal nature of the emissions is successfully identified with the adSP method in all 10 replicates despite the vague prior information, where $K = 16$ is suggested by our algorithm in 60% of the simulation runs. The bpSP and fpSP methods, however, fail to converge for 2 and 3 of the repetitions, respectively. Restricting to the convergent repetitions, we see (Figures 2.4 and 2.5) a slight advantage of the bpSP method over the adSP method, which is in some sense

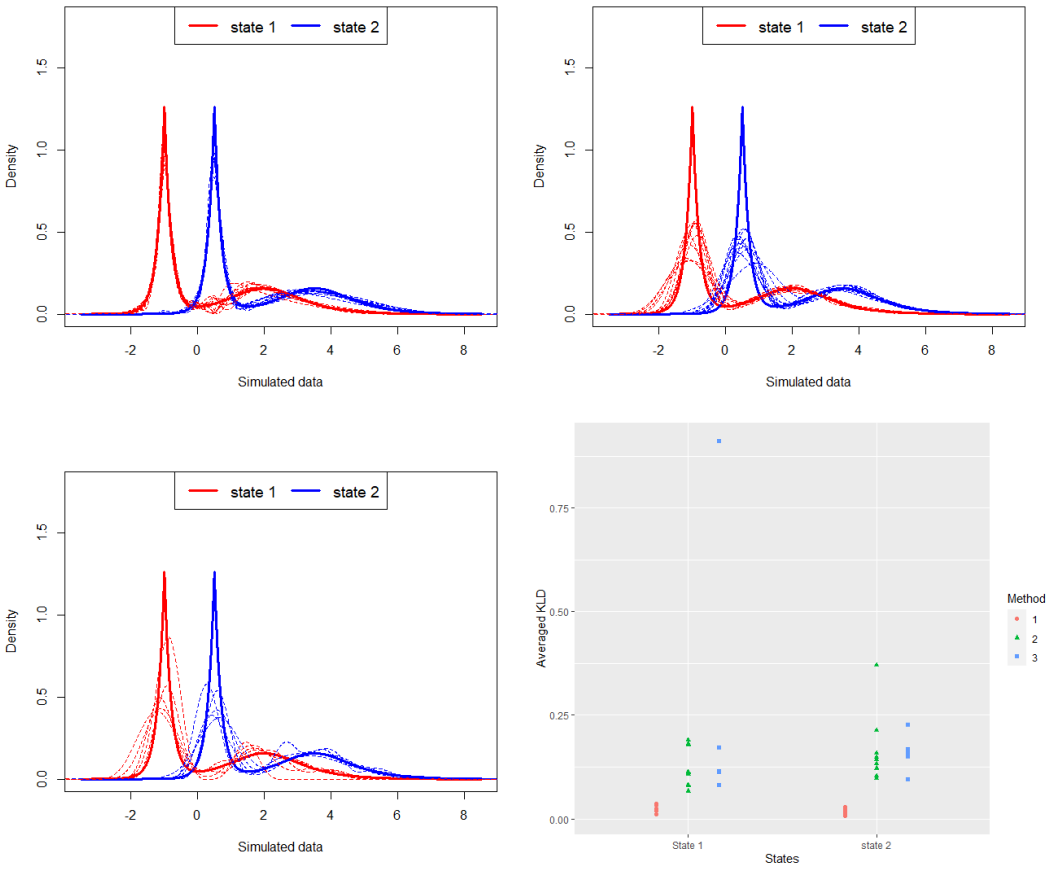

Figure 2.3: Estimation results for 10 simulations of Model 3. The settings are the same as in Figure 2.1 (for the fpSP method only the 5 convergent cases are included).

understandable given the smooth nature of the emissions and the fact that a much larger number of basis elements ($K = 51$) are employed in the P-splines. In this scenario, we additionally compare our adSP method with the Bayesian nonparametric approach of Yau et al. [2011] on retrieving the transition dynamics of the hidden Markov chain, which is one of the main foci in their work. The average posterior means (± 1 standard deviation) of the transition probabilities obtained from our method are $\gamma_{1,2} = 0.056$ (± 0.009) and $\gamma_{2,1} = 0.057$ (± 0.01), which is consistent with the true value $\gamma_{1,2} = \gamma_{2,1} = 0.05$, and is comparable to those reported in Table 2 (case when T=5000) of Yau et al. [2011]. However, it should be pointed out that in contrast to Yau et al. [2011] who assume the translation nature of the emission and that $\gamma_{1,2} = \gamma_{2,1}$ is known a-priori, here we estimate the model structure from the data.

Table 2.2: Average computational time (in seconds) for generating 10k MCMC samples for the adSP and bpSP methods.

Method	Model 1	Model 2	Model 3	Model 4
adSP	113	230	294	679
bpSP	109	227	274	619

Overall, we may conclude that the adSP and bpSP methods have roughly comparable performance in more "regular" settings (see e.g. model 1) and the advantages of using the adSP method are more significant in more complicated scenarios (see, e.g. model 3), where both bpSP and fpSP methods become very sensitive to the initial parameter settings and may suffer from poor density fits and/or convergence issues. In general, fpSP is the least accurate and reliable method. In terms of computational efficiency, the adSP and the bpSP approaches are roughly comparable, see Table 2.2 for a comparison of the computational time for each model (based on a PC computer having Intel(R) Core(TM) i7-6700 CPU, at 3.4 GHz and 16 GB RAM). The computational cost required for the fpSP method is more difficult to quantify as it is hugely influenced by the initial knot settings and the grid search strategy adopted for the smoothing parameters, not to mention that non-ignorable (and often significant) additional computational effort is needed for performing uncertainty quantification for the parameters (which we did not perform here).

2.7 Analysis of oceanic whitetip shark acceleration data

HMMs provide a useful tool for modelling animal movement metrics to study the dynamical patterns of an animal's behavioural states (e.g., resting, foraging or travelling) in ecology [Patterson et al., 2009; Langrock et al., 2012a, 2018]. Here we consider a time series of the overall dynamic body acceleration (ODBA) collected from an oceanic whitetip shark at a rate of 16 Hz over a time span of 24 hours noting that a larger replicate data set was analyzed in Langrock et al. [2018]. For our analysis, the raw ODBA values are averaged over non-overlapping windows of length 15 seconds and log transformed (lODBA), resulting in a total of 5760 observations. The marginal distribution of the transformed data is illustrated in Figure 2.6.

We first model the lODBA values using our proposed spline-based HMM with N fixed to three states as in Langrock et al. [2018], who present potential biological interpretations of these three states. The constants used in our algorithm

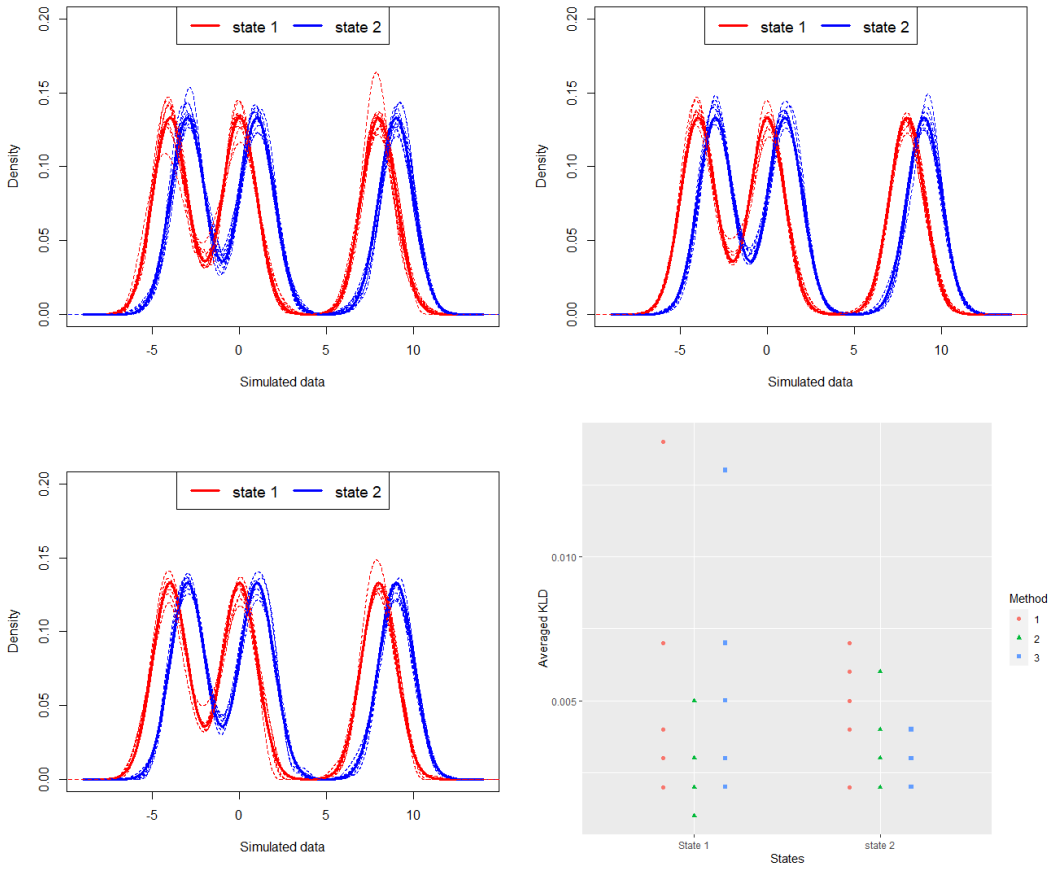

Figure 2.4: Estimation results for 10 simulations of Model 4. The settings are the same as in Figure 2.1 (for the bpSP and fpSP methods only 8 and 7 convergent cases are included).

are specified as $a = -5.5$, $b = -1$ and $\alpha = 2$. Our posterior inference was based on 25k sweeps of the RJMCMC algorithm after a burn-in of 25k sweeps. The whole sampling process took about 70 minutes in R (for the same hardware specification as above). Figure 2.6 (top left panel) shows the estimated emission densities. The posterior modal number of knots is 10, with $\hat{P}(K = 10|data) = 0.741$, followed by $K = 11$, with $\hat{P}(K = 11|data) = 0.245$. Our posterior summaries for the entries of the transition probability matrix are

$$\hat{\Gamma} = \begin{pmatrix} 0.941_{(0.006)} & 0.059_{(0.006)} & 0.001_{(0.001)} \\ 0.031_{(0.003)} & 0.96_{(0.004)} & 0.009_{(0.002)} \\ 0.006_{(0.004)} & 0.048_{(0.009)} & 0.946_{(0.01)} \end{pmatrix},$$

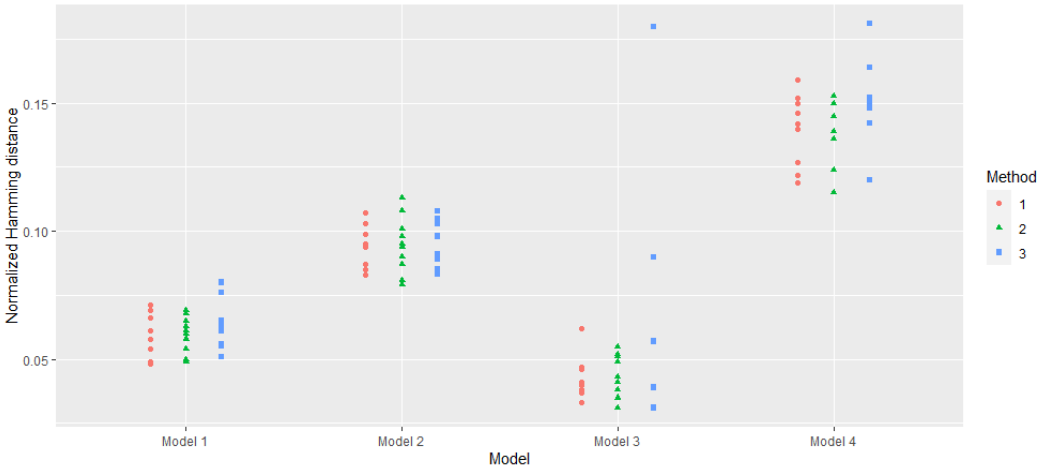

Figure 2.5: Summary of decoding results obtained for each replication of each model, where methods 1, 2 and 3 correspond to the adSP, bpSP and fpSP methods, respectively (for the bpSP and fpSP methods only convergent cases are included).

where the point estimates are the posterior means and the associated standard deviations are shown in brackets. In particular, the diagonal entries of $\hat{\Gamma}$ are all close to 1 highlighting the serial dependence. The estimated state sequence (indicated in colours) obtained via local decoding is shown in the top right panel of Figure 2.6. Note that, as was pointed out in Langrock et al. [2018], there could be a potential lack of fit if one chooses to model this type of data by some common parametric HMMs. To illustrate this, we fit a Gaussian HMM with $N = 3$ to the lODBA data using the standard MLE approach and the resulting estimates of the emissions are shown in the bottom left panel of Figure 2.6, which exhibits a certain level of underfitting (e.g. fails to accurately capture the spikiness and the slight right tail of the emission of the middle state). The corresponding diagonal entries of the transition matrix are estimated as $\hat{\gamma}_{1,1} = 0.941$, $\hat{\gamma}_{2,2} = 0.952$ and $\hat{\gamma}_{3,3} = 0.915$, indicating an underestimation of the state persistency in states 2 and 3. To further demonstrate the advantage of our approach, we also fitted a B-spline HMM using Langrock et al.'s (2015) method, where we have set $K = 39$ to ensure enough flexibility and selected $\lambda = (300, 1, 400)$ for the smoothing parameters based on our experiments. While the resulting transition probability estimates and the density fits seem to be comparable to our results (see bottom right panel of Figure 2.6), their fitted 3-state HMM uses a total of 129 parameters for estimating the emissions, for which we experienced numerical stability issues in the process of estimation and the results are found to be very sensitive to the initial parameter setting (as we observed in the simulation studies). It can thus

be expected that their bootstrap-based uncertainty quantification approach would be challenging and costly to implement as well. In contrast, our method used 52 parameters if conditional on the posterior modal number of knots and we obtained posterior uncertainties for the parameters at no extra computational cost.

To verify if the shark data support a model with $N = 3$, we also proceed to performing model selection using the marginal likelihood based approach as described in section 2.5. Our posterior estimates are based on 30k draws after a burn-in of 30k iterations. Interestingly, with a discrete uniform prior over $\{2, \ldots, 9\}$, the posterior modal number of states is estimated to be $N = 8$, with a posterior probability of 1, thus strongly indicating that the data support a considerably larger number of states than originally assumed in Langrock et al. [2018]. This is perhaps not surprising given the expected complexity of shark's behavior in reality and the potential rich information contained in this high-resolution signal [Bres, 1993]. Figure 2.7 displays the estimated emission densities (left panel) and the corresponding decoded times series (right panel). We can see that the estimated hidden states roughly correspond to 8 different levels of activity which resolves the multimodality seen in the 3-state model (state 1) into a mixture of unimodal emission densities. The posterior means of the transition probabilities are

$$\hat{\Gamma}_8 = \begin{pmatrix} 0.861 & 0.121 & 0.006 & 0.003 & 0.003 & 0.003 & 0.002 & 0.002 \\ 0.071 & 0.778 & 0.126 & 0.008 & 0.005 & 0.01 & 0.002 & 0.001 \\ 0.013 & 0.118 & 0.682 & 0.144 & 0.02 & 0.018 & 0.003 & 0.001 \\ 0.001 & 0.004 & 0.061 & 0.785 & 0.136 & 0.009 & 0.005 & 0.001 \\ 0.002 & 0.004 & 0.012 & 0.103 & 0.831 & 0.041 & 0.007 & 0.001 \\ 0.004 & 0.03 & 0.046 & 0.041 & 0.062 & 0.723 & 0.087 & 0.006 \\ 0.007 & 0.009 & 0.03 & 0.026 & 0.032 & 0.256 & 0.494 & 0.147 \\ 0.002 & 0.002 & 0.003 & 0.003 & 0.003 & 0.01 & 0.05 & 0.927 \end{pmatrix}.$$

Almost all of the estimated states are persistent in the sense that there usually is a large probability of staying in the current state (see diagonal entries of $\hat{\Gamma}_8$). In comparison to the 3-state model, the diagonal entries are naturally lower as the shark's movement is now sub-divided into more states. The fitted model with 8 states indicates, for example, that the lower and middle activity mode of the 3-state model can each be associated to several separate states, while the higher activity mode mainly corresponds to a single state, state 8, whose onset typically occurs following state 7. To assign biologically meaningful interpretations to these states, however, additional ecological information is required and we shall not pursue it further here. Nevertheless, our analysis demonstrates the ability of our method

to deal with model selection in an HMM with a relatively large number of states, while Langrock et al.'s approach can be expected to be severely challenged with increasing N. Our nonparametric approach could also be used in an explorative way to identify a suitable parametric model. We can see from Figure 2.7 that most of the emission densities of the 8-state model have a relatively regular shape. An interesting further question would be whether the 3-state nonparametric model can be replaced by a 8-state parametric approach, assuming for instance a mixture of Gaussian emission densities, which can capture a higher degree of multimodality in the marginal distribution.

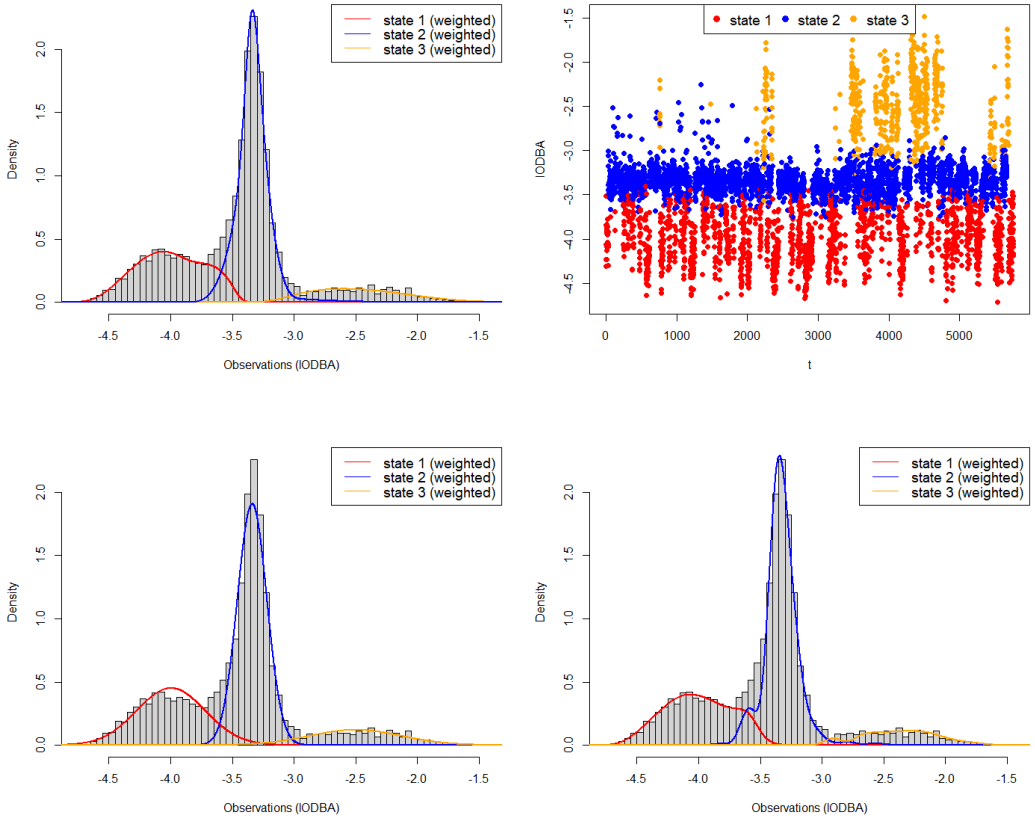

Figure 2.6: Histogram of 15s-averaged lODBA values along with the estimated emission densities (weighted according to their proportion in the stationary distribution of the estimated Markov chain) obtained from our method (top left), the Gaussian HMM (bottom left) and Langrock et al's method (bottom right); top right panel: 15s-averaged lODBA series, where colour indicates the locally decoded state at each time obtained with our method. Here the state labels are sorted according to their mean lODBA levels.

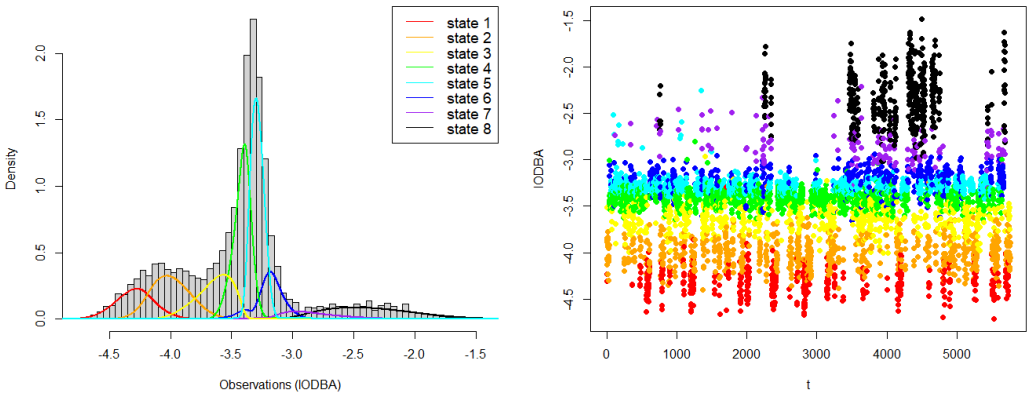

Figure 2.7: Left panel: histogram of 15s-averaged lODBA values along with the eight estimated emission densities (weighted according to their proportion in the stationary distribution of the estimated Markov chain) obtained from our method; right panel: the corresponding locally decoded time series of the 8-state model. Here the state labels are sorted according to their mean lODBA levels.

2.8 Discussion

In this chapter we focus on spline-based HMMs which are attractive in comparison to alternative nonparametric HMMs due to their simplicity in model interpretation and their modelling flexibility. We propose and develop a Bayesian methodology for inference in B-spline-based HMMs where the number of hidden states, N, may be unknown along with all other model parameters including the spline knot configuration. With N fixed, we introduce a RJMCMC algorithm that allows for a parsimonious and efficient positioning of the spline knots as we were able to demonstrate in simulations and the case study. It further appears that the implied computational efficiency allows us to realistically conduct model selection on N which is a notoriously difficult problem due to challenging convergence problems even in - or perhaps because of - parametric approaches and when dealing with more states and larger data sets. Here we propose to estimate the marginal likelihood of a spline-based HMM via a truncated harmonic mean estimator, under a parallel sampling scheme. Our simulation studies demonstrate the resulting effectiveness of the approach in selecting the correct model.

We have demonstrated that within the spline-based modelling framework, our proposed method has significant advantages over the frequentist P-spline-based approach proposed by Langrock et al. [2015], and compares favourably to a Bayesian P-spline approach which is investigated for the first time. Our method circumvents

the challenging problems such as the selection of smoothing parameters and the quantification of uncertainty on model parameters, and allows for more stable and also more parsimonious estimation of the emission densities while achieving comparable or even better performance. These advantages also mean that we are able to estimate an HMM where N may potentially be higher. Langrock et al. [2018] didn't approach model selection on N but we were able to estimate a B-spline based HMM where N could be, and was indeed estimated to be, considerably higher for the same kind of animal movement data. This comparison highlights the advantages gained from being able to address model selection on N and the use of a nonparametric approach for explorative data analysis. Comparing with the Bayesian nonparametric model developed in Yau et al. [2011], our approach does not assume the translation nature of the emissions and thus can be applied for more general data sets. Even restricted to the class of translation HMMs our method is still a strong competitor as it enjoys a relatively simple model formulation and the ability to address the case when N is unknown.

The modelling framework may be extended in other ways, with appropriate modifications of the proposed algorithm. For instance, the homogeneous assumption on the hidden Markov chain of our spline-based HMM can be relaxed to allow for a covariate dependent transition probability matrix. One way to achieve this is to employ the standard multinomial logistic link function for the transition probabilities [Zucchini et al., 2016] to reparametrize Γ in terms of the covariates. Efficient MCMC inference can be achieved by incorporating the Polya-Gamma data augmentation approach of Polson et al. [2013] into the present modelling framework, as was successfully applied in Holsclaw et al. [2017] for parametric nonhomogeneous HMMs. A generalization to a multivariate observation process also is straightforward - at least in principle. In particular, assuming contemporaneous conditional independence among the M observed variables, i.e. $f_{x_t}(y_1, \ldots, y_M) = \prod_{i=1}^{M} f_{x_t,i}(y_i)$, one can model the state-dependent joint density by assuming univariate B-splines used here for the corresponding marginal densities. In this case multiple birth and death moves are required for the RJMCMC to update the knot configuration for each component of $(y_1, \ldots, y_M)$ in a deterministic or random manner. However, designing an efficient MCMC methodology for spline-based HMMs with more general multivariate distributions is beyond the scope of this chapter.

2.A Further details of the reversible jump MCMC algorithm

In this section, we give further computational and implementational details related to the reversible jump MCMC algorithm presented in Section 2.4 and establish the validity of the algorithm.

2.A.1 Acceptance probabilities for the Metropolis-Hastings moves

Moves (c) and (d) in Algorithm 5 are standard Metropolis-Hastings updates. For (c), the acceptance probability for relocating the knot r_{k^*} to the candidate point r_c is:

$$\min\big(1, \frac{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K', \tilde{A}_K, \Gamma) f_{N,[a,b]}(r_{k^*}|r_c, \tau_1^2)}{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K, \tilde{A}_K, \Gamma) f_{N,[a,b]}(r_c|r_{k^*}, \tau_1^2)}\big),$$

where R_K' differs from R_K only in the replacement of r_{k^*} by r_c, and $f_{N,[a,b]}(\cdot|\mu, \sigma^2)$ denotes the density of the truncated normal distribution with mean μ, standard deviation σ and bounded on $[a, b]$. For move (d) since the proposal is symmetric the acceptance probability is:

$$\min\big(1, \frac{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K, \tilde{A}_K', \Gamma) f(\tilde{A}_K'|K, \zeta)}{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K, \tilde{A}_K, \Gamma) f(\tilde{A}_K|K, \zeta)}\big),$$

where the set of $\tilde{a}_{i,j}'$ is denoted by $\tilde{A}_K'$. In step (e) we used a log-normal random walk to update the parameter due to the positivity constraint, and the corresponding acceptance probability after adjusting for the log-transformation is

$$\min\big(1, \frac{f(\tilde{A}_K|K, \zeta') f(\zeta')\zeta'}{f(\tilde{A}_K|K, \zeta) f(\zeta)\zeta}\big).$$

2.A.2 Acceptance probabilities for the birth and death moves

Using the notation of Green [1995], the birth move regarding spline parameters is accepted with probability $\min(1, A)$, where A could be expressed in the form

$$\text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio} \times \text{Jacobian}.$$

In our context the likelihood ratio is:

$$\frac{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K+1, R_{K+1}, \tilde{A}_{K+1}', \Gamma)}{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K, \tilde{A}_K, \Gamma)},$$

where $\tilde{A}'_{K+1}$ stands for the set of proposed $\tilde{a}'_{i,j}$ and the complete data likelihood $f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|\cdot)$ is given by equation (2.2). The prior ratio is given by the product of the ratio of the priors on each block of parameters that are involved in this update:

$$\frac{f_U(K+1)}{f_U(K)} \frac{\frac{(K+1)!}{(b-a)^{K+1}}}{\frac{K!}{(b-a)^K}} \frac{\prod_{i=1}^{N} \prod_{j=1}^{K+5} f_{LG}(\tilde{a}'_{i,j}|\zeta, 1)}{\prod_{i=1}^{N} \prod_{j=1}^{K+4} f_{LG}(\tilde{a}_{i,j}|\zeta, 1)},$$

where $f_U(\cdot)$ denotes the probability mass function of a uniformly distributed random variable on $\{2, \ldots, K_{\max}\}$, and $f_{LG}(\cdot|\zeta, 1)$ stands for the log-gamma density with shape parameter ζ and rate parameter 1. The proposal ratio is given by

$$\frac{d_{K+1}}{(K+1)} \left\{ b_K \frac{\sum_{i=1}^{K} f_{N,[a,b]}(r_c|r_i, \tau(R_K, i)^2)}{K} \right\}^{-1},$$

where $f_{N,[a,b]}(\cdot|\mu, \sigma^2)$ denotes the density of a truncated normal distribution with mean μ, standard deviation σ and bounded on $[a, b]$. Lastly, the Jacobian corresponding to the transformation from $(R_K, \tilde{A}_K, \Gamma, r_c, u_1, \ldots, u_N)$ to $(R_{K+1}, \tilde{A}'_{K+1}, \Gamma)$ is

$$|(r_{n^*+2} r_{n^*+3})^N \prod_{i=1}^{N} (\tilde{a}_{i,n^*+4} - \tilde{a}_{i,n^*+3})|.$$

After simplification A is thus given by

$$A = \frac{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K+1, R_{K+1}, \tilde{A}'_{K+1}, \Gamma)}{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|K, R_K, \tilde{A}_K, \Gamma)} \frac{1}{(b-a)} \frac{\prod_{i=1}^{N} \prod_{j=n^*+2}^{n^*+4} f_{LG}(\tilde{a}'_{i,j}|\zeta, 1)}{\prod_{i=1}^{N} \prod_{j=n^*+2}^{n^*+3} f_{LG}(\tilde{a}_{i,j}|\zeta, 1)}$$

$$\times \frac{d_{K+1}}{b_K} \left\{ \frac{\sum_{i=1}^{K} f_{N,[a,b]}(r_c|r_i, \tau(R_K, i)^2)}{K} \right\}^{-1} |(r_{n^*+2} r_{n^*+3})^N \prod_{i=1}^{N} (\tilde{a}_{i,n^*+4} - \tilde{a}_{i,n^*+3})|.$$

Since the birth and death moves are defined in a symmetric way, the acceptance probability for this death move is $\min(1, A^{-1})$, where K is replaced by $K-1$ and $n^* = d^* - 1$.

2.A.3 Tackling label switching

A practical consequence of the properties of the model and its prior is that samples generated by the reversible jump MCMC algorithm are subject to the label switching problem, i.e. the state labels can permute during the MCMC iterations without changing the posterior density [Scott, 2002]. As a result, the MCMC output cannot be directly used for inference about the state specific parameters. To tackle this issue we choose to use the Kullback–Leibler relabelling algorithm developed in Stephens

[2000], which has been successfully applied for both parametric HMMs [Rodríguez and Walker, 2014] and nonparametric HMMs [Hadj-Amar et al., 2020a]. The basic ideas are as follows. For each of the T collected MCMC samples, we first construct a $n \times N$ dimensional classification probability matrix whose (i,k)-th entry in our context is given by

$$P_{i,k}^{(t)} = \frac{\pi_k^{(t)} f_k^{(t)}(y_i)}{\sum_{j=1}^{N} \pi_j^{(t)} f_j^{(t)}(y_i)}, \quad t = 1, \ldots, T,$$

where $f_k^{(t)}(\cdot)$ is the emission density for state k constructed from the t-th MCMC sample and $(\pi_1^{(t)}, \ldots, \pi_N^{(t)})$ is the stationary distribution associated with the transition matrix $\Gamma^{(t)}$. The algorithm then involves iteratively searching a specific permutation of state labels to minimize the KLD between classification probabilities averaged over MCMC iteration, $q_{i,k} = (\sum_{t=1}^{B} P_{i,k}^{(t)})/B$, and the classification probabilities obtained in each MCMC iteration. In other words, we make the state labels associated with each MCMC draw agree on the classification probabilities $[P_{i,k}^{(t)}]$. The "optimized" permutation searched for each MCMC sample can then be used to relabel the samples to achieve a consistent ordering of the labels. We refer to Stephens [2000] and Rodríguez and Walker [2014] for more details of the algorithm. In our implementations, we use the R package *Label.switching* of Papastamoulis [2016b] to perform this minimization procedure.

2.A.4 Validity of the algorithm

When the adaptive tuning scheme ceases after a period of burn in, the validity of the proposed reversible jump MCMC algorithm can be established following standard Markov chain theory as presented in Tierney [1994] and Robert and Casella [2013]. First note that the Markov transition kernel for each of the move steps admits the target posterior distribution, f (defined through equation (2.4)), as invariant distribution, so a concatenation of these kernels also admits f as invariant distribution. Irreducibility of the constructed chain can be deduced as the chain can move from one value of K to any other possible value by increasing or decreasing its value by one at a time, with positive probability. In step (a) all possible state allocations have positive probability, In steps (b), (d) and (e) the full conditional distribution/proposal density is positive on the natural parameter space and the same holds true for step (c) if we consider several consecutive sweeps. The chain is also aperiodic as there is a strictly positive probability that the chain remains in a neighbourhood of the current state after one sweep of the MCMC procedure. With

the above properties the chain is guaranteed to converge to the posterior distribution from almost all initial states (except for a set of posterior probability zero). To replace "almost all" by "all" we require a stronger condition called Harris recurrence, which is generally difficult to verify in the trans-dimensional MCMC set-up [Roberts and Rosenthal, 2006; Hastie and Green, 2012]. However in practice we could tackle this issue by drawing the initial state using a continuous distribution centered around the posterior mode (or other approximations to that such as the maximum likelihood estimate). This strategy is employed in our initialization process. To accelerate the convergence of the chain we initialize the knot points at the empirical quantiles of the data so that more knots are initially placed at data-rich regions. For the remaining parameters the initial values are drawn from appropriate truncated normal distributions centered at their respective maximum likelihood estimates computed given the initial knot configuration.

2.B Further details for the simulation study

In this section we present the MCMC algorithm used for inference in the Bayesian P-spline-based HMM (the bpSP method) described in section 2.6 and discuss in more detail the performance of the proposed RJMCMC algorithm in the simulation study.

2.B.1 MCMC details for the Bayesian P-spline-based model

For the Bayesian P-spline-based model the knot configuration (K, R_K) is prefixed, and the parameter set is $(\tilde{A}, \Gamma, \tilde{\boldsymbol{\tau}}, \tilde{\tau}^{'})$, where $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, \ldots, \tilde{\tau}_N)$ and $\tilde{A} = (\tilde{a}_1, \ldots, \tilde{a}_N)$. We assume that the joint posterior distribution of the state sequence and model parameters takes the form

$$f(\mathbf{x}^{(n)}, \Gamma, \tilde{A}, \tilde{\boldsymbol{\tau}}, \tilde{\tau}^{'} | \mathbf{y}^{(n)}) \propto f(\Gamma) f(\tilde{\tau}^{'}) f(\tilde{\boldsymbol{\tau}} | \tilde{\tau}^{'}) f(\tilde{A} | \tilde{\boldsymbol{\tau}}) f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)} | \tilde{A}, \Gamma),$$

where $f(\tilde{\boldsymbol{\tau}} | \tilde{\tau}^{'}) = \prod_{i=1}^{N} f(\tilde{\tau}_i | \tilde{\tau}^{'})$, $f(\tilde{A} | \tilde{\boldsymbol{\tau}}) = \prod_{i=1}^{N} f(\tilde{a}_i | \tilde{\tau}_i)$ and the complete likelihood (last term) is given by (2.2) with spline coefficients derived from $\tilde{A}$. We use the same prior distribution for Γ as for our proposed spline-based HMM and the priors $f(\tilde{\tau}^{'})$, $f(\tilde{\tau}_i | \tilde{\tau}^{'})$ and $f(\tilde{a}_i | \tilde{\tau}_i)$ are specified as described in section 2.6.2. Posterior simulation for the resulting model can be achieved using a Metropolis-within-Gibbs sampler as outlined in Algorithm 6. Steps (a) and (b) can be performed exactly as in our RJMCMC algorithm (see steps (a) and (b) of Algorithm 5), and the details are omitted here. For step (c), we update the state-specific spline coefficients $\tilde{\mathbf{a}}_i$ in

Algorithm 6: MCMC algorithm for Bayesian P-spline-based HMMs

Initialize $\tilde{A}_K$, Γ, $\tilde{\boldsymbol{\tau}}$, $\tilde{\tau}'$; **for** *i=1, ..., T* **do**
- (a) update the hidden state sequence $\mathbf{x}^{(n)}$;
- (b) update the transition probability matrix Γ;
- (c) update the set of reparametrized B-spline coefficients $\tilde{A}$;
- (d) update the roughness parameters $\tilde{\boldsymbol{\tau}}$;
- (e) update the hyperparameter $\tilde{\tau}'$

end

blocks via a random walk MH step as used in step (d) of Algorithm 5 due to the potential high-dimensionality of $\tilde{A}$. For each state a separate variance parameter is used for the MH update and they are tuned adaptively as described in section 2.4. For step (d), note that the full conditional distribution of $\tilde{\tau}_i$ is

$$f(\tilde{\tau}_i|rest) \propto \tilde{\tau}_i^{\alpha_{\tilde{\tau}}-1} \exp(-\alpha_{\tilde{\tau}}\tilde{\tau}'\tilde{\tau}_i)\tilde{\tau}_i^{\frac{K+4}{2}} \exp(-\frac{\tilde{\tau}_i}{2}\tilde{\mathbf{a}}_i^T P\tilde{\mathbf{a}}_i), \quad i = 1, \ldots, N.$$

Therefore we update $\tilde{\tau}_i$ by drawing from

$$\tilde{\tau}_i|rest \sim \textbf{Gamma}(\frac{K+4}{2} + \alpha_{\tilde{\tau}}, \alpha_{\tilde{\tau}}\tilde{\tau}' + \frac{1}{2}\tilde{\mathbf{a}}_i^T P\tilde{\mathbf{a}}_i), \quad i = 1, \ldots, N.$$

Step (e) is also a Gibbs step. The full conditional distribution of $\tilde{\tau}'$ is

$$f(\tilde{\tau}'|rest) \propto (\tilde{\tau}')^{\alpha_{\tilde{\tau}'}-1} \exp(-\beta_{\tilde{\tau}'}\tilde{\tau}') \prod_{i=1}^{N} (\tilde{\tau}')^{\alpha_{\tilde{\tau}}} \exp(-\alpha_{\tilde{\tau}}\tilde{\tau}_i\tilde{\tau}'),$$

and thus we draw:

$$\tilde{\tau}'|rest \sim \textbf{Gamma}(N\alpha_{\tilde{\tau}} + \alpha_{\tilde{\tau}'}, \alpha_{\tilde{\tau}} \sum_{i=1}^{N} \tilde{\tau}_i + \beta_{\tilde{\tau}'}).$$

For the same reason stated before, here MCMC inference is subject to the label switching problem, which can be tackled using the Kullback-Leibler relabelling algorithm described above.

2.B.2 Performance details of the proposed RJMCMC algorithm

Here we give more details and discuss the performance of the proposed algorithm for our simulation models presented in Section 2.6. Some diagnostic plots that are related to the mixing and convergence of the sampler are shown in Figure 2.8, where the results for each model are obtained for a randomly selected replication of the

simulated data set.

For adaptive MCMC steps (steps (c)-(e) of Algorithm 5), the empirical acceptance rates are closed to the pre-specified desired levels. We analyzed the traces and running averages for selective parameters (including the tuning variance parameters) across MCMC iterations and found acceptable mixing patterns in most cases. Step (c) may mix slower than MH steps due to the relatively high dimension of the spline coefficients vector. To further improve sampling efficiency, a Metropolis adjusted Langevin algorithm (MALA) [Roberts and Tweedie, 1996] may be employed, which is a specific class of MH algorithms where the proposal distribution exploits the local information of the target

$$q(\tilde{A}_K'|\tilde{A}_K) = \mathcal{N}_{K+4}(\tilde{A}_K + \frac{h}{2}\Sigma\nabla_{\tilde{A}_K}\log f, h\Sigma),$$

where h is a positive real number, Σ is a symmetric positive definite matrix and $\nabla_{\tilde{A}_K}\log f = (\frac{\partial}{\partial\tilde{a}_{i,j}}\log f)_{i=1,....N;j=1,...,K+4}$, with

$$\frac{\partial}{\partial\tilde{a}_{i,j}}\log f = \sum_{t:x_t=i}\frac{\partial}{\partial\tilde{a}_{i,j}}\log f_i(y_t) + \frac{\partial}{\partial\tilde{a}_{i,j}}\log f_{LG}(\tilde{a}_{i,j}|\zeta,1).$$

Both h and Σ are tuning parameters that need to be selected using pilot runs or tuned on-the-fly using adaptive MALA techniques, see, e.g. Atchadé [2006]. On the other hand, the potential gain in efficiency is counterbalanced by an increase in the computational cost in evaluating the gradients at each iteration and the tuning is more subtle yet important than that for the random walk MH algorithm [Christensen et al., 2005]. In our experiments, the MALA algorithm did not offer noticeable advantage over the random walk MH in terms of mixing and the final estimation accuracy.

For the dimension changing moves, the averaged acceptance rates for models 1-4 are 0.32%, 0.17%, 0.37% and 0.38%, respectively. While it is lower than desired in our simulation cases, we did not detect any apparent convergence issues (see Figure 2.8). For Model 2 the acceptance rate can be expected to be lower than for the others as it has a larger number of states whose emission distributions have quite different characteristics. In general, a new knot is more likely to be accepted if it contributes to the fit of all the emission densities. On the other hand, the degree of precision in the posterior distribution of K is limiting the achievable acceptance rates for the dimension changing moves. For instance, as n increases, the posterior for K is expected to be more concentrated, leading to a generally lower acceptance rate. From the algorithmic perspective this rate may be mildly affected

by the standard deviation τ of the truncated normal distribution used to generate the new knot or the proposal distribution for u_i used in the birth move. In our experiments, with the functional form of τ fixed, the results are not very sensitive to the value of α, provided that it is chosen in a reasonable way. For instance we prefer to set $0 < \alpha < 1$ when the averaged distance between knots is much larger than 1 and vice versa. The use of other potential proposal distributions for u_i within the Beta family was also investigated, but no clear evidence was found in terms of their superiority over the noninformative choice $U(0, 1)$. We also compare our algorithm with the one that integrates out the latent state sequence $\mathbf{x}^{(n)}$ (via the forward algorithm in (2.3)), and draw $\mathbf{x}^{(n)}$ afterwards at each iteration via FFBS conditional on the simulated model parameters. Although the latter strategy gives a slightly higher acceptance rate, which is not unexpected as the dimensionality of the parameter space is greatly reduced, there is no noticeable gain in terms of the overall computational cost and the estimation accuracy. A higher acceptance rate for the dimension changing move may be achieved by modifying the current model or the proposal mechanism. For instance, one might consider using a separate knot configuration for each emission density and propose state-specific jump moves, or generating the random variables u_i used in equation (2.5) from a more informative proposal distribution such as a truncated normal distribution centred at its maximum likelihood estimate (approximated via some numerical optimization routines). However, in either scenario the model complexity or the computational effort may increase significantly, which may prevent a successful application of the algorithm in some settings. Though not presented here, we also investigated the estimation performance of the algorithm with other simulation settings. Our preliminary findings indicate that the performance generally improves as serial correlation and/or sample size increases, while it declines as the number of states and/or the overlap of the emission distributions increases.

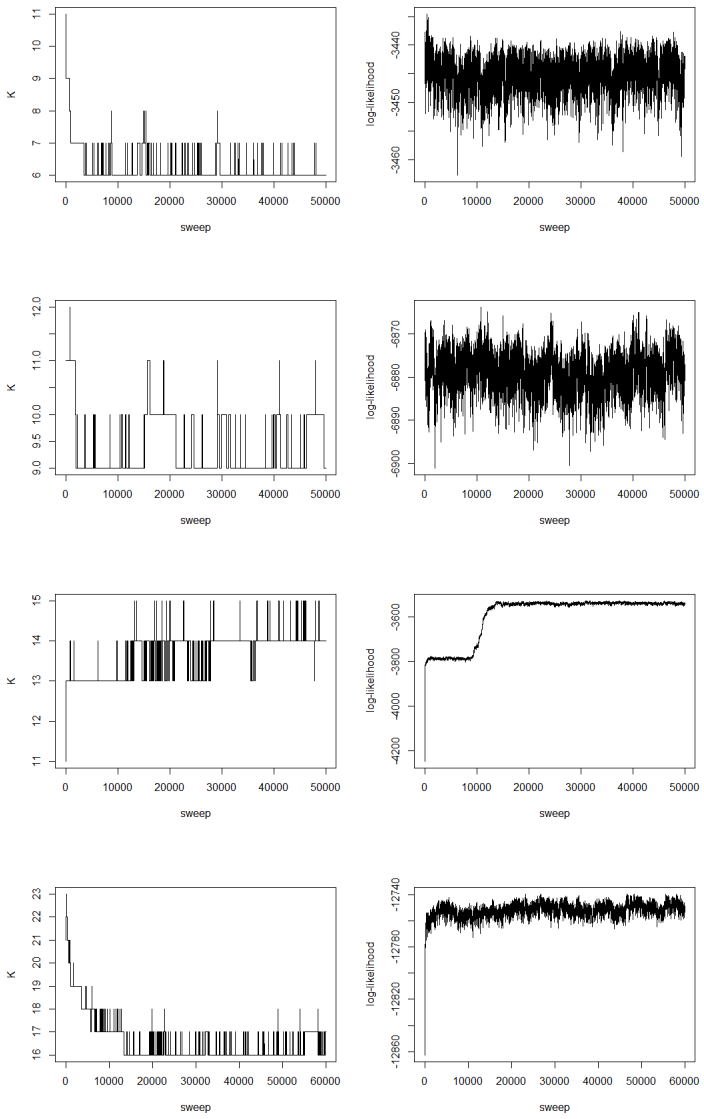

Figure 2.8: Selected simulation outputs for the four simulation models (conditioned on the true value of N). Panels 1-4 show trace of K for the complete MCMC run including burn-in (left) and the corresponding observed data log-likelihood of the model generated by the algorithm (right) for models 1-4, respectively.

.

Chapter 3

A conditional hidden Markov model for inferring circadian and sleep patterns

3.1 Introduction

The human circadian (approximately 24 hours) timing system (CTS) is an innate clock formed by a complex hierarchical network of molecular clocks, which are governed by specific clock genes in most cells of the body [Zhang and Kay, 2010; Partch et al., 2014]. The suprachiasmatic nucleus (SCN), the central pacemaker of the CTS, is primarily entrained by external light-dark cycles via visual afferents as well as inputs from other molecular oscillators, and it regulates and synchronizes downstream cellular clocks in peripheral tissues and organs via signals such as hormone secretion. The CTS as a whole plays an important role in coordinating diverse physiological processes including core body temperature, heart rate, blood pressure, among many others, and behavioural processes such as sleep/wake, with the geophysical time [Ruben et al., 2019; Dunlap et al., 2004]. Recent research illustrated the close association between the function or the status of our CTS and our physical and mental health. In particular, it is revealed that disruptions and perturbations of the CTS are reliably linked to mental health related problems, greater risk of developing numerous diseases such as cancer and psychiatric disorder and the worsening of pre-existing pathologies [Roenneberg and Merrow, 2016; Ortiz-Tudela et al., 2010; Evans and Davidson, 2013; Ortiz-Tudela et al., 2016; Lyall et al., 2018]. Regarding cancer it has also been shown that optimal circadian timing of the administration of medication based on an individual's circadian rhythm, termed chronotherapy,

could enhance tolerance and efficacy to some extent in both experimental and clinical situations [Lévi et al., 2010; Lévi and Okyar, 2011; Ortiz-Tudela et al., 2013, 2014]. There is now a great deal of interest in assessing and monitoring an individual's CTS, with the aim for timely interventions and more effective personalized treatments.

Measurements of circadian rhythms can be achieved by recording and analyzing the rhythmicity of the "circadian biomarkers", which serve as indicators of the CTS [Hofstra and de Weerd, 2008; Abdullah et al., 2017]. Among the most commonly recognized biomarkers are core body temperature and certain hormones (e.g. melatonin, cortisol), but also physical activity (PA), with the last one receiving a lot of attention in recent research on chronobiology and chronotherapeutic healthcare. Measuring the CTS with PA relies on the fact that there is increased movement during wake periods and reduced movement during sleep periods, and that it can be easily and objectively measured from wearable computing devices (e.g. accelerometers) in a non-invasive way under normal living conditions. The reliability and validity of this approach have been established in numerous studies, see, e.g. Jean-Louis et al. [1996] and Hofstra and de Weerd [2008]. However, extracting clinically relevant and interpretable summaries from high volume and complex PA data for the purpose of long-term monitoring and assessing the circadian rhythmicity of an individual is a challenging task. The most widely used summary statistics, which attempt to evaluate the function of an individual's CTS from different perspectives, are often open to different suggestions for estimators and, moreover, their uncertainties, which are important information for medical decisions, are often challenging or impossible to estimate. For instance, to calculate the *dichotomy index* $I < O$ used in the oncological literature [Minors et al., 1996], which measures the proportion of epochs during time in bed when activity is lower than the median activity during time out of bed, we usually need to subjectively determine the sleep-period time for the target subject.

More sophisticated statistical approaches have been used to develop more insights from PA data. Functional and smoothing based approaches allow for extracting further parameters of interest and direct modelling of the effects of external covariates on PA [Xiao et al., 2015; Morris et al., 2006]. However, these methods usually don't provide explicit classification of activity levels/modes and struggle to capture stochastic and abrupt transitions in activity levels (e.g. from inactive to active states; see Huang et al. [2018] for further discussion). Here, HMMs offer an attractive probabilistic modelling framework as they naturally account for the temporal dependence and variability in the data by employing a discrete la-

tent variable, distributed according to a Markov chain. Using simulated data sets, Witowski et al. [2014] established the advantages of the HMM method over traditional threshold-based methods in terms of activity classification. Huang et al. [2018] investigated the PA using a nonhomogeneous HMM and proposed a novel model-derived circadian parameter for monitoring and quantifying a subject's circadian rhythm. de Chaumaray et al. [2020] use a mixed HMM with discrete random effects for characterizing the activity pattern of a subject in a longitudinal setting. A potential limitation with the current HMM based methods is that they usually rely on parametric assumptions of the distribution of the PA data which may not always be unproblematic.

While modelling the sleep-wake cycles is of important interest, the sleep itself, as a vital physiological process that recharges our energy and rejuvenates the body, deserves additional attention. Our current knowledge indicates that sleep homeostat impacts the sleep-wake cycle together with the circadian clock, and sleep is strongly affecting a person's physical and mental well-being and plays a crucial role in adverse health conditions such as, for example, diabetes, cardiovascular disease and depression [Foster, 2020]. Monitoring sleep could be essential to gaining insight into a subject's circadian rhythm and health status. Conventionally, sleep is monitored and evaluated under laboratory settings using the polysomnography (PSG), a multi-sensor approach that collects multiple physiological signals with regards to brain activity, muscle and eye movements, respiratory and cardiac activity, from which the sleep stage in each epoch (typically in 30-s intervals) of the monitoring period can be evaluated [Berry et al., 2012]. Whilst being considered as the gold standard for measuring sleep, it is of very limited use to investigate daily sleep rhythms in practice due to its high cost, cumbersomeness and intrusiveness of measurement settings. Instead, actigraphy has become increasingly popular in (large scale) sleep research for similar reasons stated above and received extensive validation against the PSG [Ancoli-Israel et al., 2003, 2015; Quante et al., 2018]. We note however that statistical methods for analysing accelerometer data, such as functional data analysis and HMM based methods, mainly focused on studying physical activity and/or the sleep-wake cycle of a subject, with little attention being paid to analysing the sleep periods. We noticed the works of Li et al. [2020a] and Lüdtke et al. [2021] who developed parametric HMMs for sleep analysis and demonstrate the superiority of HMM methods over alternative state-of-the-art algorithms when restricting to 2-stage sleep/wake identification. In another work, Winnebeck et al. [2018] demonstrated the potential of accelerometer recordings taken at the wrist for extracting more detailed patterns of sleep physiology by analyzing the "*Locomotor*

Inactivity During Sleep" (LIDS), a simple ad-hoc inverse transformation of the PA which enhances non-movement during sleep. They found that LIDS oscillate in phase (low activity) with markers of "deeper" sleep, and out of phase (high activity) with markers of "lighter" sleep as well as rapid eye movement (REM) sleep, and they were able to establish some systematic relationships between PSG sleep parameters and LIDS-derived parameters. While their findings are very promising, their approach did not allow them to systematically quantify the sleep periods or analyse the stochastic dynamics of the sleep process on an individual basis.

Here, we focus on retrospectively analyzing the PA data collected over multiple days/weeks with a two-fold objective: (i) characterize the sleep-wake patterns in the entire activity data set and (ii) analyze overnight sleep patterns of an individual. We would like to have a probabilistic approach that jointly achieves these two tasks in a principled and coherent manner, with as little human intervention into tuning the estimation process as possible. To this end, we develop a hierarchical conditional hidden Markov modelling approach building on the spline-based inference method for HMMs proposed in chapter 2. More specifically, we assume that sleep periods are contained within the state (State 1) that is associated with the lowest activity level of a HMM on the entire PA data (main-HMM). We then insert a second or "sub-HMM" which is invoked conditional on State 1 of the main-HMM and aims at refining the activity behaviour within State 1. Both main and sub models are specified nonparametrically by using spline-based emission densities. The strength of our approach comparing to existing methods will be that, firstly, it gives a very flexible stochastic model from which we can systematically estimate many important parameters that characterize an individual's circadian rhythm, such as duration and center of rest times, amount and regularity of activity etc. (see Huang et al. [2018]) and, secondly, it allows us to model the individual stochastic dynamics of the rest state activity which does not appear strictly periodic as seen on the aggregate level in Winnebeck et al. [2018]. By conditioning on the rest state alone there is no need to apply any ad-hoc transformation to the data and also our likelihood for the sub-model will not be influenced by the relatively large values and variability of the activity observed during the day. Another feature of our method is that the whole algorithm operates in an unsupervised manner, i.e. it does not require PSG labels for learning the model, which is desirable in applied settings as these labels are very costly or even impossible to acquire [Li et al., 2020a].

Note that the method developed here can be applied in much more general settings where we may be interested in analyzing specific state(s) of an HMM at a finer level with separate hidden Markov process(es), achieving inferences that are

otherwise not possible with a single HMM. Also, our conditional HMM approach should not be confused with what is called the hierarchical HMM (HHMM) [Fine et al., 1998], which is originally developed in the field of machine learning for pattern recognition tasks and has later been used in different contexts such as animal movement modelling [Leos-Barajas et al., 2017; Adam et al., 2019a; Sacchi and Swallow, 2021]. The fundamental difference is that in the HHMM, a joint model is formulated for multiple observed processes at different temporal resolutions, each of which is modelled via a hidden Markov process and the process at the coarser level determines the onset of a specific finer level process for each epoch (see Adam et al. [2019a] for more details). In contrast, our method by default operates on a single time scale and our focus is to refine specific states of an HMM. The term HHMM may also refers to the scenario where a hierarchical prior model is employed on top of multiple HMMs (see e.g. Chen et al. [2016]), which is different to our context as well.

This chapter is organised as follows. Section 3.2 gives a brief introduction and description of the sleep dataset from the Multi-Ethnic Study of Atherosclerosis (MESA) and the cohort used in our analysis. Section 3.3 introduces our proposed conditional HMM modelling approach in a general set-up and outlines the associated Bayesian inference procedure. In section 3.4 we apply our method to the selected MESA cohort and present relevant results and we close this chapter with a brief discussion.

3.2 MESA data description

MESA is a multisite collaborative longitudinal study aimed at investigating the progression of subclinical to clinical cardiovascular disease [Chen et al., 2015; Zhang et al., 2018]. In the initial stage (2000-2002), a total of 6814 individuals free of clinically apparent cardiovascular diseases, aged from 45 to 84 and with diverse ethnic backgrounds, participated in the study. Between 2010-2012 about one third of the participants were enrolled in the MESA sleep study (MESA Exam 5) where each participant wore an actigraph (Actiwatch Spectrum) on the non-dominant wrist for one week and underwent a full unattended home-based PSG session for one night. In this sleep study the activity is measured in each 30-s epoch by counting the number of times movement intensity crosses a threshold and its value reflects the overall activity intensity in that epoch. Figure 3.1 gives an example of the raw PA data collected for an example subject over the monitoring period of 7 days. In addition, wake and four sleep stages, namely N1, N2, N3 and REM, are identified for

every 30-s epoch from PSG recordings for one night using the criteria set out by the American Academy of Sleep Medicine. Among sleep stages, N1 and N3 correspond to light and deep sleep, respectively, while N2 is the intermediate stage. N1, N2 and N3 are collectively referred to as the non-REM stages of sleep [Berry et al., 2012]. The REM stage is associated with dreaming and is physiologically very different from the other stages of sleep [Stein and Pu, 2012].

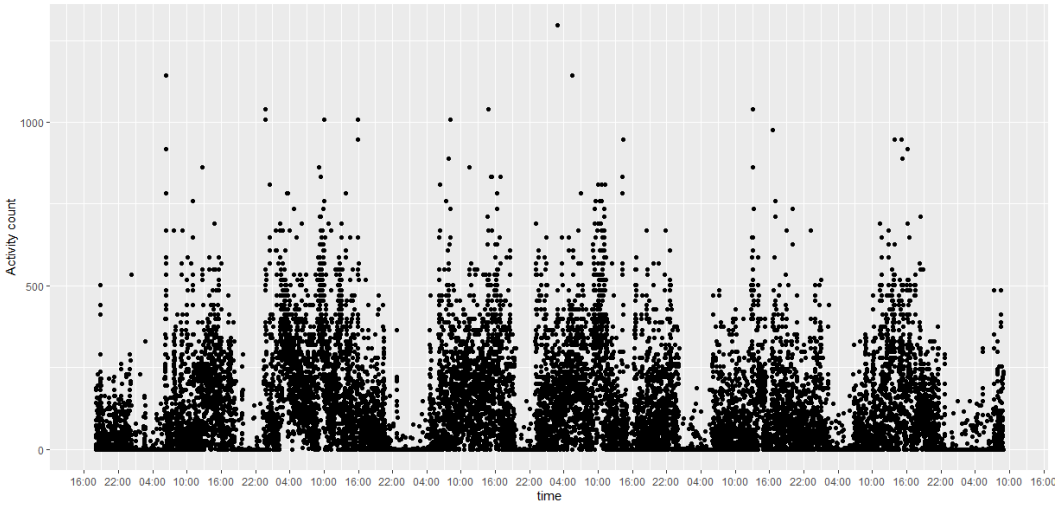

Figure 3.1: Example of raw PA data (MESA subject 2243). Activity counts are recorded per 30-s over 7 days.

In our analysis we considered a sub-cohort of 44 subjects, which are randomly selected from the overall cohort conditioned on having an equal number of males and females, a spread of age covering a span from 55 to 82 years, and top quality scores for both PSG and PA data, i.e. excellent or outstanding actigraphy quality (scored as 6 or 7) and outstanding PSG quality (scored as 7). More details including the demographic characteristics of the selected subjects are shown in Table 3.1. Note that most subjects in our cohort do not have sleep related medical conditions according to their self reports. Table 3.2 gives a summary of some parameters related to the circadian rhythm and sleep for our cohort. All quantities are obtained from the MESA database except for the dichotomy $I < O$ and rhythm indices, which are parameters associated with the rest-activity rhythm and are computed based on the actimetry data using the methods developed in Huang et al. [2018]. The chronotype score is a summary score based on the modified Horne-Ostberg Morningness-Eveningness Questionnaire (MEQ) which reflects the sleep-wake behaviour of an individual [Horne and Östberg, 1976]. A higher value indicates a stronger tendency towards the morning type. The proportions of the

sleep stages are evaluated based on the PSG data, conditional on the PSG-derived sleep period (from sleep onset to offset). We refer to the online MESA website [1] for more information on these variables. It is worth noting the big differences in the duration of the sleep stages during the sleep period (identified based on PSG) in that stage N2 accounts on average for nearly half of an individual's total sleep time, while stages N1 and N3 on average only account for about 10% of the sleep time (see Table 3.2 and Figure 3.2). Note that the PSG also recorded wake epochs that occurred during the PSG session, which can happen, for instance, during periods of sleep interruptions. We will refer to these as intermittent wake (IW). Some results on the relationships between activity and PSG-derived sleep stages are shown in the bottom panel of Figure 3.2. We can see that, as expected, the mean PA levels conditional on the PSG stages tend to decrease with increasing sleep depth, with the deep sleep stage $N3$ having the overall lowest activity level. But on the other hand, the empirical distributions of PA conditional on each PSG stage (including IW) are all highly skewed to the right, indicating that all stages spent a high proportion of time in very low activity levels, including zero counts. For instance, as shown in bottom right panel of Figure 3.2, the proportions of zero activity counts are high for all sleep and also the IW stages. Therefore it can be hypothesised that PA alone may be insufficient for distinguishing between all four sleep stages, which is in agreement with findings from previous studies [Zhai et al., 2020; Boe et al., 2019]. However, we can identify a general trend in terms of averaged PA levels for deeper and lighter sleep (and wake) as found in Winnebeck et al. [2018]. The temporal dependence in the sleep stages is moderate to high, with the intermittent wake and N2 states being the most stable stages followed by REM. This can be derived from the empirical transition probabilities of the four sleep and IW stages for the PSG data of all individuals (details not shown here). We also investigated the correlations between sex/age and the circadian and sleep parameters listed in Table 3.2, and the results are shown in Table 3.3. Regarding sleep, we found that compared to males, females tend to have a higher sleep duration (p=0.005), larger proportions of N3 (p=0.044) and REM sleep (p=0.022) and a smaller proportion of intermittent wake (p=0.041). We did not find significant differences in gender between the circadian-related parameters (i.e. the chronotype score, the dichotomy $I < O$ and rhythm indices) as the p-values are all greater than 0.5 (see Table 3.3), indicating that females and males in our cohort may have roughly similar rest-activity rhythms. In terms of age, elderly subjects tend to have weaker circadian rhythms in activity, i.e., they have smaller values in the dichotomy $I < O$ and rhythm indices

[1] https://sleepdata.org/datasets/mesa

Table 3.1: Characteristics of the MESA cohort

Variables	Subjects ($D = 44$)
Gender, counts (proportion)	
Male	22 (50%)
Female	22 (50%)
Race/ethnicity, counts (proportion)	
White, Caucasian	19 (43.2%)
Chinese American	5 (11.4%)
Black, African-American	10 (22.7%)
Hispanic	10 (22.7%)
Sleep-related conditions, counts (proportion)	
Sleep apnea	2 (4.5%)
Insomnia	1 (2.3%)
Restless legs syndrome	1 (2.3%)
Age, mean (SD)	67.5 (8.7)

Table 3.2: Circadian and sleep statistics for the MESA cohort. For each subject, the choronotype score is obtained from their sleep questionnaire, the dichotomy $I < O$ and rhythm indices are computed from the PA data using the method of Huang et al. [2018] and the remaining sleep related parameters are computed from the PSG data.

Variables	mean $\pm$ SD
Chronotype score	17.81 ± 3.91
Dichotomy index $I < O$	$98.7\% \pm 1.4$
Rhythm index	0.687 ± 0.129
Sleep efficiency	$79.3\% \pm 11.6$
Total sleep time (TST)	$380.4 \text{ min} \pm 69.5$
Wake after sleep onset (WASO)	$75.3 \text{ min} \pm 58.7$
Wake proportion	$16.1\% \pm 11.2$
N1 proportion	$11.1\% \pm 6.3$
N2 proportion	$45.6\% \pm 10.3$
N3 proportion	$11.0\% \pm 7.0$
REM proportion	$16.4\% \pm 5.8$

(both significant at 10% level), which is consistent with the fact that elderly people have lower sleep efficiency (p=0.032), a larger percentage of IW (p=0.01) and light sleep N1 (p=0.031) during their sleep. The chronotype is also age-related to some extent where elderly people are more likely to be morning-type (p=0.054).

3.3 The conditional HMM methodology

In this section we introduce the proposed conditional HMM approach under a general set-up and then present the sampling scheme for performing Bayesian inference of the resulting model. Additional details regarding the MCMC algorithm are provided

Table 3.3: Gender and age effects in the circadian and sleep parameters. t represents the t statistics in the Welch two-sample t-test (between female and male) and r denotes the Spearman correlation coefficient (corresponding p-values are indicated in the bracket). Significant effects (at 10% level) are highlighted in red.

Variables	Gender (Female/Male)	Age
Chronotype score	$t = -0.07$ (0.944)	**r = 0.296 (0.054)**
Dichotomy index $I < O$	$t = 0.569$ (0.572)	**r = −0.448 (0.002)**
Rhythm index	$t = 0.047$ (0.963)	**r = −0.264 (0.084)**
Sleep efficiency	$t = 1.62$ (0.113)	**r = −0.324 (0.032)**
Total sleep time (TST)	**t = 2.98 (0.005)**	r=−0.244 (0.111)
Wake after sleep onset (WASO)	$t = -1.62$ (0.114)	**r = 0.358 (0.017)**
Wake proportion	**t = −2.12 (0.041)**	**r = 0.383 (0.01)**
N1 proportion	$t = -0.728$ (0.471)	**r = 0.326 (0.031)**
N2 proportion	$t = 0.069$ (0.945)	**r = −0.298 (0.049)**
N3 proportion	**t = 2.08 (0.044)**	$r = -0.169$ (0.273)
REM proportion	**t = 2.37 (0.022)**	**r = −0.397 (0.008)**

in the appendix.

3.3.1 The Bayesian model

Let $\mathbf{y}^{(n)} = (y_1, \ldots, y_n)$ be the observed data of interest, here activity counts. The main model employs a N-state spline-based HMM introduced in Chapter 2 of the thesis for characterizing the general pattern of the data at a relatively coarse level which takes the whole series into account, where the cardinality N may be estimated via the marginal likelihood based approach as described in chapter 2. Details for setting up the main-HMM are omitted here (see Chapter 2). Here we shall focus on the sub-HMM, which is introduced for characterizing a specific state i of the main model in more detail by assuming N_S sub-states of state i (without loss of generality, we set $i = 1$ and for clarity omit this subscript in what follows). We assume that

$$f(\boldsymbol{\theta}^S|\mathbf{y}^{(n)}) = \int f(\boldsymbol{\theta}^S|\mathbf{x}^{(n)}, \mathbf{y}^{(n)}) f(\mathbf{x}^{(n)}|\mathbf{y}^{(n)}) d\mathbf{x}^{(n)}, \tag{3.1}$$

where $\boldsymbol{\theta}^S$ is the parameter set for a N_S-state sub-HMM (with hidden state variables integrated out), $\mathbf{x}^{(n)}$ is the hidden state sequence associated with the main-HMM and

$$f(\boldsymbol{\theta}^S|\mathbf{x}^{(n)}, \mathbf{y}^{(n)}) \propto f(\boldsymbol{\theta}^S) f(\mathbf{y}^{(n)}|\boldsymbol{\theta}^S, \mathbf{x}^{(n)}). \tag{3.2}$$

We refer to the second term in (3.2) as the "conditional likelihood" for the sub-HMM. We further assume that by conditioning on state 1 of the main-HMM, only observations that are associated with $\{t : x_t = 1\}$ will contribute to this likelihood,

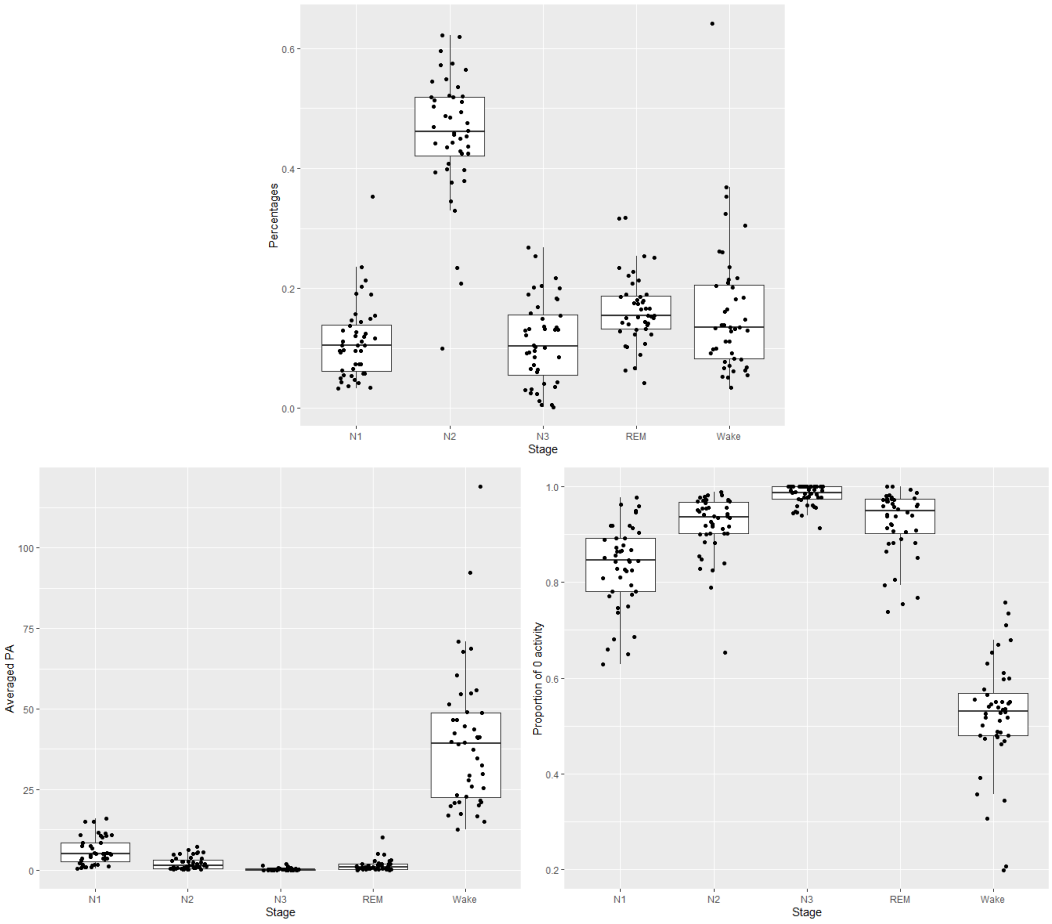

Figure 3.2: Characteristics of sleep and IW stages during PSG sessions for 44 subjects. Top panel shows the proportions of the time spent in each stage; bottom left panel shows mean PA levels conditional on each PSG stage and bottom right panel shows the percentage of zero activity in each stage. All panels use the boxplots, where the dot superimposed represents the corresponding value computed for each subject.

whereas the remaining observations $\{y_t : x_t \neq 1\}$ are treated as "missing data". The advantage of such a conditioning concept is that the resulting conditional likelihood can be easily handled in the HMM framework. More specifically, let $(t_1, \ldots, t_{T_1})$ be the collection of time points in ascending order such that $x_{t_j} = 1$, $j = 1, \ldots, T_1$.

Then, using the notation of chapter 2, we have

$$\begin{aligned} f(\mathbf{y}^{(n)}|\boldsymbol{\theta}^S, \mathbf{x}^{(n)}) &= f(y_{t_1}, \ldots, y_{t_{T_1}}|\boldsymbol{\theta}^S) \\ &= \sum_{x_{t_1}, \ldots, x_{t_{T_1}}} f(y_{t_1}, \ldots, y_{t_{T_1}}, x_{t_1}, \ldots, x_{t_{T_1}}|\boldsymbol{\theta}^S), \\ &= \delta P(y_1) \Gamma P(y_2) \cdots \Gamma P(y_n) \mathbf{1} \end{aligned} \tag{3.3}$$

where $P(y_t) = I_{N_S}$, the identity matrix of dimension N_S, for $t \neq t_1, \ldots, t_{T_1}$. Note that the last row of (3.3) takes exactly the same form as a marginal likelihood of a standard HMM and thus the standard forward algorithm applies for efficient evaluation of the conditional likelihood [Zucchini et al., 2016]. To facilitate illustrating the distinctive roles that the main and sub-HMMs play, let's consider a specific modelling context as follows. Supposing that the overall trajectory of a physiological process is well described by a 3-state HMM and that the sub-process corresponding to a specific state, say state 1 (in our application this state is related to the sleep process), is of particular interest where we would like to analyze the sub-dynamics associated with this process with 2 sub-states. In this context, the main HMM would have 3 states, aiming at capturing the main patterns for the overall trajectory while identifying the observations that are relevant to the sub-HMM (i.e. observations allocated to state 1) based on the posterior samples of the state sequence of the main HMM. A 2-state sub-HMM is introduced focusing explicitly on the observations related to state 1 by conditioning on the underlying state process of the main HMM. In particular, given a realisation of the state process, the conditional likelihood described in (3.3) effectively removes the influence of irrelevant observations (assigned to states 2 or 3) on the sub-HMM while automatically respecting the temporal patterns of the observations that are of interest (bypass the issue of manually handling the small (e.g. interruptions during sleep) or large (e.g. wake period during the day) time gaps between these observations). To appropriately take the uncertainty of the state classification into account, we further integrated out the state sequence to obtain the marginal posterior for the sub model as defined in (3.1), on which our inference for the sub-HMM will be based. It is important to note that fitting the main HMM with 4 states will not necessarily split state 1 of the original model into 2 states as desired, whereas in our framework we have direct control over this due to the aforementioned conditioning set up. Note also that while the marginal distribution of the data associated with the sub-HMM (defined through a mixture of the emission densities for the sub-states) is expected to be similar to the emission distribution of state 1, there is no explicit analytical relationship between the two

as the former and latter are inferred based on partial and full data, respectively.

When modelling accelerometer data, a large number of zeros are typically observed during a rest or sleep state [Ae Lee and Gill, 2018], causing problems to the estimation of continuous valued emission densities (irrespective of whether they are parametric or spline-based). To address this issue we assume a zero inflation of the spline-based emission distributions at both HMM levels as follows

$$f_{x_t}(y_t) = w_{x_t,1}\delta_0 + w_{x_t,2}f^B_{x_t}(y_t),$$

where x_t indicates the underlying state at time t, $w_{x_t,1}$ represents the state-specific zero weight such that $0 \leq w_{x_t,1} \leq 1$ and $w_{x_t,1} + w_{x_t,2} = 1$, δ_0 is the Dirac delta distribution and $f^B_{x_t}(y_t)$ is a spline-based emission density as defined in chapter 2. Following Gassiat et al. [2016a] we can establish identifiability of the resulting HMM provided that at most one $w_{x_t,1}$ is equal to one and that $\{\delta_0, f^B_1, \ldots, f^B_N\}$ are linearly independent. In our analysis these conditions are always satisfied. The proposed emission model can easily be adapted and used in wider applications where the data may be modelled as a mixture between a point mass (or multiple point passes, if necessary) and a continuous distribution. The priors for parameters involved in the sub-HMM are specified as follows. For the knot configuration and spline coefficients, we employ the same priors as in chapter 2. Note that there is no conjugate prior for Γ for the sub-HMM as the associated hidden state process is not simulated. Following the reparametrization scheme used in chapter 2, here we reparametrize each row of the transition probability matrix as $\gamma_{i,j} = \tilde{\gamma}_{i,j}/\sum_{l=1}^{N}\tilde{\gamma}_{i,l}$, $\tilde{\gamma}_{i,j} > 0$, and place a vague gamma prior on the $\tilde{\gamma}_{i,j}$, i.e. $f(\tilde{\gamma}_{i,j})$=gamma$(1,1)$, which gives a Dir$(1,\ldots,1)$ distribution on $(\gamma_{i,1},\ldots,\gamma_{i,N})$. Similarly, we reparameterize the weights as $w_{i,j} = \tilde{w}_{i,j}/(\tilde{w}_{i,1} + \tilde{w}_{i,2})$, $\tilde{w}_{i,j} > 0$, $i = 1,\ldots N_S$, $j = 1,2$. We choose a vague gamma prior on the $\tilde{w}_{i,j}$, i.e. $f(\tilde{w}_{i,j})$=gamma$(1,1)$, leading to a Dir$(1,1)$ distribution on $(w_{i,1}, w_{i,2})$. Using the notation of chapter 2, we assume the following factorization of the joint distribution in (3.2) with all the reparametrizations

$$f(\tilde{\boldsymbol{\theta}}^S|\mathbf{y}^{(n)},\mathbf{x}^{(n)}) \propto f(\zeta)f(K)f(\tilde{W})f(\tilde{\Gamma})f(R_K|K)f(\tilde{A}_K|K,\zeta)f(\mathbf{y}^{(n)}|\boldsymbol{\theta}^S,\mathbf{x}^{(n)}) \quad (3.4)$$

where the reparamatrized parameter vector $\tilde{\boldsymbol{\theta}}^S = (\zeta, K, \tilde{W}, \tilde{\Gamma}, R_K, \tilde{A}_K)$, $\tilde{\Gamma} = (\tilde{\gamma}_{i,j})_{i,j=1,\ldots,N_S}$, $\tilde{W} = (\tilde{w}_{i,k})_{i=1,\ldots,N_S;\ k=1,2}$, and the $\tilde{w}_{i,k}$ and $\tilde{\gamma}_{i,j}$ are assumed to be a-priori independent.

3.3.2 Markov chain Monte Carlo methodology

Posterior inference for the main and sub-HMMs can be achieved by sequentially sampling from $f(\boldsymbol{\theta}, \mathbf{x}^{(n)}|\mathbf{y}^{(n)})$ and $f(\boldsymbol{\theta}^S|\mathbf{y}^{(n)})$, where $\boldsymbol{\theta}$ represent the parameter set of the main-HMM. The MCMC methodology described in chapter 2 (Algorithm 5) can be used to simulate from $f(\boldsymbol{\theta}|\mathbf{y}^{(n)})$. In cases where the emission are specified as a mixture of point masses and splines (e.g. zero-inflated emissions), an additional MH step is needed to update the state-specific weights for the point mass and the details for this update are given in the appendix. For the sub-HMM, we simulate from $f(\boldsymbol{\theta}^S|\mathbf{y}^{(n)})$ according to (3.1) by first generating samples from $f(\mathbf{x}^{(n)}|\mathbf{y}^{(n)})$, which are obtained as a by-product from the posterior simulation for the main-HMM. Conditional on each realisation of $\mathbf{x}^{(n)}$, we then simulate from $f(\boldsymbol{\theta}^S|\mathbf{x}^{(n)}, \mathbf{y}^{(n)})$ by drawing a sample for $\boldsymbol{\theta}^S$ from the joint density defined in (3.4) using essentially the same sampling scheme as used for the main-HMM, where the RJMCMC updates are run for several iterations and the last sample is kept. Tuning of the MH scaling parameters was achieved via a separate pilot run using the same adaptive procedure as described in chapter 2, where the conditioning variable $x^{(n)}$ may be fixed at a specific realisation from its posterior distribution or the local decoding result obtained from the simulation output for the main-HMM. Posterior simulation regarding the hidden state process associated with the sub-HMM for a given segment(s) of the time series can be achieved by running a standard FFBS algorithm, conditional on each simulated parameter set $\boldsymbol{\theta}^S$. Note that data points that were assigned to the higher active states (states 2 or 3) by the local decoding result of the main-HMM are treated as "missing" when implementing the forward procedure. Mathematically this corresponds to replacing the emission densities by the constant of one for the associated time points. In running the backward simulation procedure, the whole sub-state sequence associated with the given series will be simulated. However, only sub-states that correspond to the "non-missing" data points (assigned by state 1 of the main-HMM) are of interest.

3.4 Application to the MESA cohort

In this section we present our results for analysing the MESA data set introduced in section 3.2. For the main-HMM, the number of states N is fixed at 3 as in Huang et al. [2018] which is found to achieve a good balance between model complexity and interpretability for all individuals in the cohort, with the lowest activity state (State 1) corresponding to rest/sleep periods that mostly occurred during night-time. To save computational time inference for the main-HMM is based on the transformed

PA data obtained by first averaging PA over 5-min windows as in Huang et al. [2018] (this resolution was found to be useful in identifying sleep-wake patterns), followed by a log transformation as used in Li et al. [2020a], i.e. $\log(1 + PA)$ to handle high variability observed in the MESA data. For the emission model we put a second point mass at $\log(1.1)$ (the second smallest possible value for 5-min averaged PA after the log-transformation) in addition to the point mass on zero to handle its high occurrence. For the sub-HMM we assume 2 sub-states, 1.1 and 1.2, of State 1 to potentially capture the ultradian oscillations between higher and lower intensity of movement during sleep such as identified by Winnebeck et al. [2018] who alluded to the possibility that these oscillations might be due to the approximately 120-min periodic transitions between the Non-REM and REM stages of sleep. Inference for the sub-HMM is based on the raw 30-s PA counts rather than the 5-min aggregates to better focus on the detail of activity during sleep, and we introduced point masses at the first four lowest values including 0 in specifying the emissions for the sub-states. Throughout we use two example subjects from the cohort (subjects 921 and 3439), both free from diagnosed sleep related diseases, to facilitate illustrating our proposed method. Some additional estimation results are postponed to the appendix.

3.4.1 Results for the main-HMM

Our results for the main-HMM was based on 25k iterations of the proposed algorithm, of which the first 25k were discarded as burn-in. Figures 3.3 and 3.4 (top panel) depict the 5-min averaged PA data along with the locally decoded states (indicated by colors) and, in the panel underneath, the cumulative probabilities of the three states at each time point conditioned on the set of all observations (i.e. $P(x_t \leq i|\boldsymbol{\theta}, \mathbf{y}^{(n)}); i = 1, 2, 3$) are plotted for subjects 921 and 3439, respectively. It is apparent that for both subjects, State 1 (in blue) of the fitted main-HMMs is characterized by relatively long periods of immobility which typically occurred at night time. Other states (in pink and red shades) usually correspond to day-time activities of varying intensity which will depend on the subject's lifestyle, but also to potential interruptions of sleep such as seen for subject 3439. Thus the main-HMM suggests that in comparison to subject 3439, subject 921 appears to have an overall more active lifestyle and a more regular sleep-wake routine with no significant sleep disruptions during the monitoring period, whereas subject 3439 seems to suffer from a more disturbed circadian rhythm. Our visual impressions are backed by the lower dichotomy $I < O$ and rhythm indices for subject 3439, which are 96.4% and 0.553, respectively, than those for subject 921, which are 99.4% and 0.774. They are also in

line with their sleep questionnaires where subject 3439 reported to have a generally restless sleep and sometimes have trouble falling asleep. To obtain more evidence in support of our interpretations on the states, we compare our sleep (State 1)/wake (states 2 and 3) decoding results with the corresponding PSG-derived sleep/wake labels (available during the first night) in an epoch-by-epoch manner, where the PSG stages in each 5-min epoch are summarized by the most frequent stage of the corresponding ten 30-s bins in the raw PSG labels. The sleep/wake classification performance of the main-HMM in terms of overall accuracy, sensitivity for sleep (proportion of true sleep epochs identified correctly) and specificity for wake (proportion of true wake epochs identified correctly) are 88.2%, 100%, 70.7% and 89.1%, 93.3%, 79.5% for subjects 921 and 3439, respectively, indicating reasonably good agreement for both subjects. The relatively lower accuracy for detecting wake is understandable as there are usually the in-bed time before falling asleep or gentle sleep interruptions, which are characterized by low/no activity (see also bottom right panel of Figure 3.2). Our main-HMM also provides useful quantitative summaries of an individual's rest-activity profile. Our posterior summaries for the entries of the transition probability matrix for subjects 921 and 3439 are

$$\hat{\Gamma} = \begin{pmatrix} 0.98_{(0.006)} & 0.009_{(0.005)} & 0.011_{(0.005)} \\ 0.012_{(0.006)} & 0.914_{(0.014)} & 0.074_{(0.014)} \\ 0.008_{(0.004)} & 0.077_{(0.013)} & 0.915_{(0.014)} \end{pmatrix},$$

and

$$\hat{\Gamma} = \begin{pmatrix} 0.913_{(0.016)} & 0.08_{(0.016)} & 0.006_{(0.004)} \\ 0.088_{(0.017)} & 0.862_{(0.019)} & 0.05_{(0.01)} \\ 0.004_{(0.004)} & 0.087_{(0.016)} & 0.908_{(0.016)} \end{pmatrix},$$

respectively, where the point estimates are the posterior means and the associated standard deviations are shown in brackets. A particular transition of interest is $\hat{\gamma}_{1,1}$ where low values may be indicative of a higher tendency of transiting from sleep to wake, and thus a more interrupted sleep and a more disrupted circadian rhythm [Huang et al., 2018]. We can see that as expected, subject 3439 has a lower value for $\hat{\gamma}_{1,1}$. The time spent at the three different activity levels can be estimated according to the stationary distribution associated with the estimated transition matrix, which are $(0.328, 0.336, 0.336)$ and $(0.392, 0.375, 0.232)$ for states $(1, 2, 3)$ for subjects 921 and 3439, respectively. We can see that subject 3439 tends to have a higher proportion of time spent in states of lower activity levels and thus a more sedentary lifestyle, although noting that the interpretations of the active states (states 2 and 3) may not always be directly comparable across subjects due

to the individualized nature of the model [de Chaumaray et al., 2020]. We can nevertheless always compare activity level associated with each state based on the estimated emission distributions for each subject.

We now summarize some of our estimation results for the entire cohort of 44 subjects. Figure 3.5 shows the performance of our main-HMM in terms of sleep/wake classification with the PSG labels as reference. We see an overall high sensitivity for sleep, i.e. almost all sleep epochs are correctly decoded as State 1, with a mean percentage of 98.5%. By contrast, the specificity for wake is generally lower (with a mean of 56.9%) and exhibits high inter-subject variability, ranging from 10% to 100%. This can be understood from the fact that small or no activity is not synonymous to sleep. Revealed by a Spearman correlation analysis, we found that the specificity decreases with wake after sleep onset (WASO) (p=0.068) and the proportions of IW (p=0.086) and N1 (p=0.085) during sleep, indicating that subjects with a lighter and more disturbed sleep are more likely to have undetected wake epochs in the main-HMM. The overall classification accuracy has a mean of 86% and a standard deviation of 9.8%. Altogether, our main-HMM achieves the state-of-the-art performance in comparison to related HMM-based studies in terms of sleep/wake identification [Li et al., 2020a; Lüdtke et al., 2021]. To extract further understanding of the potentially useful parameter estimate $\gamma_{1,1}$, we assessed its association with other circadian and sleep parameters discussed above using the Spearman correlation. Significant correlations were found between $\gamma_{1,1}$ and the dichotomy ($r = 0.33,\ p = 0.027$) and rhythm indices ($r = 0.64,\ p < 0.01$), both of which are in agreement with the findings of Huang et al. [2018] and therefore suggest its potential in providing insights into an individual's circadian rhythm. However, we did not detect significant and interpretable associations between $\gamma_{1,1}$ and the PSG-derived sleep related parameters (as defined in Table 3.2), which motivates the need of a sub-HMM as will be discussed below. We found a significant age effect in $\gamma_{1,1}$ that is consistent with our previous findings, in that elderly people tend to have a lower value of $\gamma_{1,1}$, while there was no discernible gender effect within the cohort.

3.4.2 Results for the sub-HMM

For inference in the sub-HMM, our proposed algorithm was run for 25k updates (based on the last 25k posterior samples of $\mathbf{x}^{(n)}$ obtained from the main-HMM analysis), 10k of which are discarded as burn-in. The bottom panel of Figures 3.3 and 3.4 show the locally decoded time series of the 30-s PA data during the PSG monitoring period along with the cumulative probability of the two sub-states at each time point for subjects 921 and 3439, respectively. The graphs clearly show

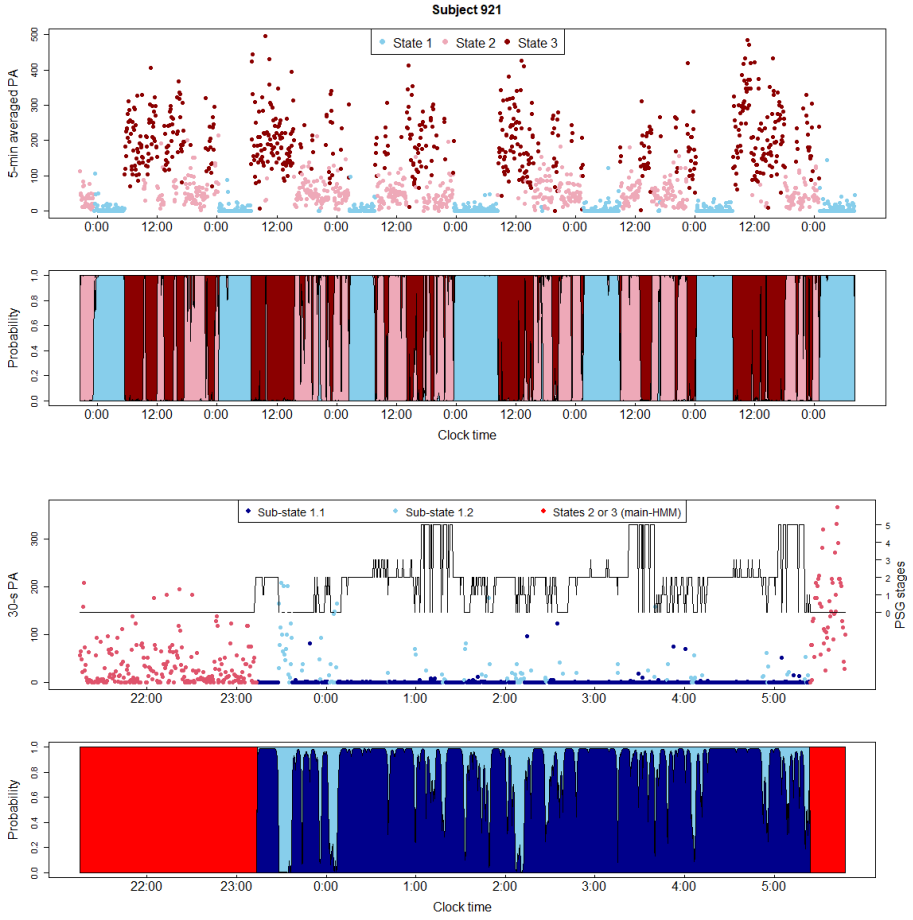

Figure 3.3: Results for subject 921. Top row upper panel: 5-min averaged PA data over a monitoring period of 7 days, where colour indicates the locally decoded state at each time under the estimated main-HMM; top row lower panel: the cumulative posterior probability of the state at each time under the estimated main-HMM (i.e. $P(x_t \leq i|\boldsymbol{\theta}, \mathbf{y}^{(n)}); i = 1, 2, 3$). The Bottom row displays the 30-s PA data during the PSG monitoring period of 1 night, with colours indicating the locally decoded state at each time (top) and the corresponding cumulative probability of each sub-state at each time (bottom) under the estimated sub-HMM for sleep bout identified by the main-HMM (i.e. State 1). Data in red are those that are assigned to more active states outside state 1 by the local decoding result for the main-HMM.

that for both subjects the transitions between and the times spent in the sub-states are subject to stochasticity. State 1.1 is characterized by a large probability of observing zero, with posterior mean of zero weight $\hat{w}_{1.1,1}$ of 0.913 and 0.962 for subjects 921 and 3439, while State 1.2 corresponds to a moderately higher level of activity where the posterior mean $\hat{w}_{1.2,1}$ for the two subjects are 0.36 and 0.474,

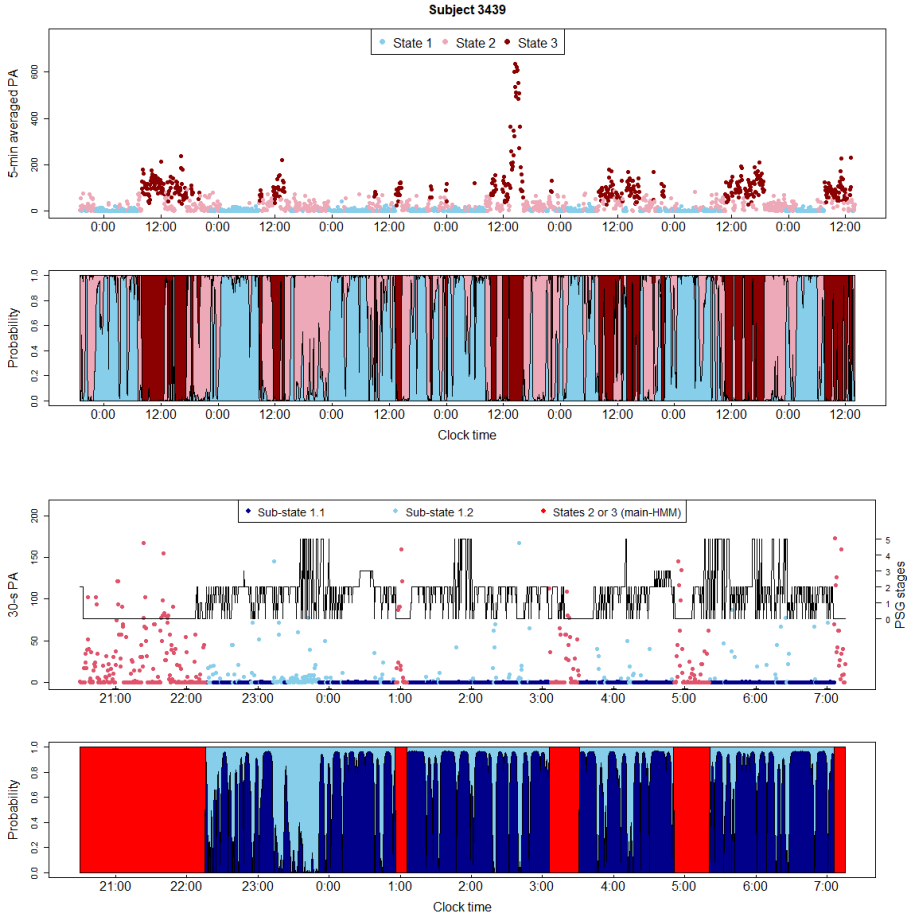

Figure 3.4: Results for subject 3439. The settings are the same as in Figure 3.3.

respectively. To obtain a clearer interpretation of the two sub-states, for each subject we compute the proportion of the five PSG stages, namely wake, N1, N2, N3 and REM, conditional on each sub-state, and the results are shown in Table 3.4. We can see that state 1.1 is highly mixed with respect to all sleep stages including wake, which is unsurprising as this sub-state is dominated by zero activity, which in turn accounts for a moderate to high proportion in all five stages. This can also be seen by looking at the percentage of the PSG stages to be decoded as State 1.1, which are $(22.6\%, 74.4\%, 94.8\%, 100\%, 91.8\%)$ for subject 921 and $(26\%, 67.4\%, 77\%, 100\%, 67.9\%)$ for subject 3439 for (wake,N1,N2,N3,REM). On the other hand, State 1.2 has a relatively clearer tendency to be associated with lighter sleep stages as well as disruptions into wake which were not identifiable by the main-HMM. We therefore expect it to provide additional useful information regarding the sleep quality of a subject that can not be extracted from the main-HMM.

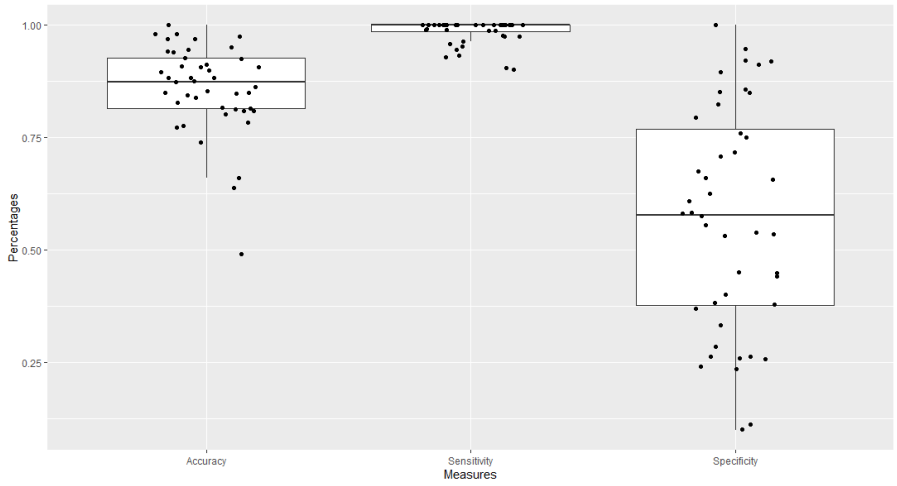

Figure 3.5: Boxplot showing sleep/wake classification performance of the main-HMM in terms of the overall accuracy, sensitivity for sleep and specificity for wake. The dot superimposed represents the corresponding value computed for each subject.

An analysis of the estimated parameters, in particular the transition probabilities of the fitted sub-HMM will provide a systematic quantitative summary which could be used, for example, to compare sleep behaviour between subjects. For subject 921, the posterior means (± 1 standard deviation) of the diagonal entries of Γ are $\hat{\gamma}_{1.1,1.1} = 0.961$ (± 0.01) and $\hat{\gamma}_{1.2,1.2} = 0.668$ (± 0.053), and those for subject 3439 are $\hat{\gamma}_{1.1,1.1} = 0.909$ (± 0.009) and $\hat{\gamma}_{1.2,1.2} = 0.733$ (± 0.046). The latter individual has a lower value of $\hat{\gamma}_{1.1,1.1}$ and higher value of $\hat{\gamma}_{1.2,1.2}$, meaning that the subject has a higher probability of leaving state 1.1 and a larger expected staying time in state 1.2 which may be associated with poorer sleep quality during the monitoring period. Indeed, subject 3439 has a slightly lower sleep efficiency compared to the other subject, which is 63.15% and 66.37%, respectively. Our results are also in accordance with what we see in table 3.5, which shows that subject 3439 spent a larger proportion of sleep time in wake and N1 stages while having a lower proportion of time in the deep and REM stages. The results of decoding and state probabilities for the sub-HMM further allow us to investigate the variation within and between courses of a sleep bout. For instance, it appears that subject 921 seems to experience more interruptions and/or lighter sleep during the initial period of the sleep bout (defined by state 1), whereas subject 3439 suffers from more frequent sleep interruptions/transitions to lighter sleep throughout the night as we see a more fragmented blue region in the state probability plot. These observations are in line with their own reports in the sleep questionnaire and match reasonably well with the PSG recordings. It is important to note that these are detailed patterns that allow us

Table 3.4: Composition of the states of the sub-HMM with respect to PSG stages

Subject	sub-state	Wake	N1	N2	N3	REM
Subject 921	1.1	0.151	0.091	0.576	0.039	0.142
	1.2	0.516	0.198	0.198	0	0.088
Subject 3439	1.1	0.174	0.168	0.529	0.051	0.078
	1.2	0.262	0.222	0.401	0	0.114

Table 3.5: Proportions of time spent in different PSG stages during sleep for the example subjects

Subject	Wake	N1	N2	N3	REM
Subject 921	0.21	0.105	0.519	0.035	0.132
Subject 3439	0.369	0.143	0.394	0.03	0.066

to focus on studying rest or sleep periods which are not (fully) discernible from the main-HMM. The hierarchical modelling approach by means of a conditional HMM is also justified noting that an unconditional HMM with number of states fixed to four is rarely likely to assign two sleep-related states as for most subjects the higher values and variability of day-time activity will dominate the likelihood and the assignment of states.

To further support our findings for the example subjects, we computed the composition of the two sub-states with respect to the PSG stages for each subject in the cohort (see Figure 3.6). It can be seen that in general state 1.1 is dominated by the intermediate N2 stage, followed by REM, wake and N3 stages, with the latter two accounting for similar proportions. State 1.2, by contrast, tends to have a relatively higher proportion of wake, followed by N2, with the deep sleep N3 occupying the least proportion. We can also see that, as expected, the ranking observed in the bottom panel of Figure 3.6 is consistent with the ordering in activity levels/percentage of zero PA as seen in Figure 3.2 (bottom panel), where the percentages of state 1.1 decoding conditional on each PSG stage increase with the sleep depth. Table 3.6 examines the correlations between the key parameters of the sub-HMM, namely $\gamma_{1.1,1.1}$, $\gamma_{1.2,1.2}$, $w_{1.1,1}$ and $w_{1.2,1}$, and the circadian and sleep parameters. It is interesting to note that $\gamma_{1.2,1.2}$ is positively correlated with WASO, wake and N1 proportions, while negatively correlated with REM proportion, the dichotomy index and sleep efficiency. The zero weight for state 1.1, $w_{1.1,1}$, is significantly positively associated with the dichotomy index and REM proportion and negatively associated with wake and N1 proportion, which are of opposite sign to those for $\gamma_{1.2,1.2}$. All these significant correlations are comprehensible and in line with our previous findings. For $\gamma_{1.1,1.1}$ and $w_{1.2,1}$, however, no significant correlations were found for this cohort. We also found an age effect on $w_{1.1,1}$ with $r = -0.396$

and a p-value of $p = 0.008$, which is consistent with elder people experiencing more lighter sleep and sleep interruptions. We did not detect any significant gender effect on the sub-HMM parameters.

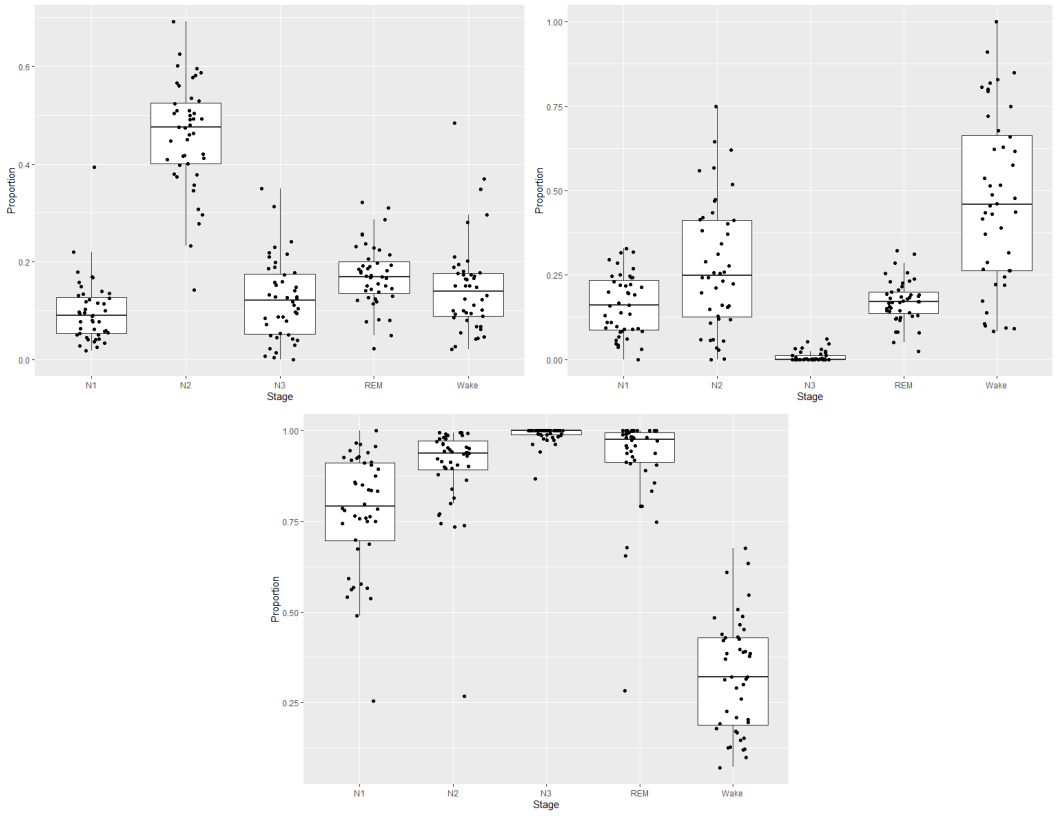

Figure 3.6: Top panel: composition of the five PSG stages conditional on state 1.1 (left) and 1.2 (right); bottom panel: percentage of the PSG stages to be decoded as state 1.1. All panels use the boxplots, where the dot superimposed represents the corresponding value computed for each subject.

3.5 Discussion

In this chapter, we extend the spline-based nonparametric HMM introduced in chapter 2 to develop a Bayesian conditional HMM modelling approach where a sub-HMM (or multiple sub-HMMs) can be introduced within an overall HMM for more detailed characterizations of the dynamics within states of the main model. We illustrate the potential usefulness of the proposed method by analysing a cohort from the MESA data set which has simultaneous recordings of the accelerometer and the PSG data. The main novelty and advantage of our modelling approach lies in the hierarchi-

Table 3.6: Spearman correlation between parameters of sub-HMM and circadian and sleep parameters. P-values are indicated in brackets and significant correlations (at 10% level) are highlighted in red.

Variables	$\gamma_{1.1,1.1}$	$\gamma_{1.2,1.2}$	$w_{1.1,1}$	$w_{1.2,1}$
Chronotype score	0.031 (0.845)	−0.202 (0.194)	−0.056 (0.723)	−0.127 (0.416)
Dichotomy index	−0.173 (0.263)	**−0.287 (0.059)**	**0.395 (0.008)**	0.123 (0.427)
Rhythm index	0.035 (0.82)	−0.048 (0.758)	−0.029 (0.852)	−0.202 (0.189)
Sleep efficiency	0.042 (0.787)	**−0.255 (0.095)**	0.169 (0.272)	0.019 (0.904)
TST	−0.024 (0.875)	0.012 (0.936)	0.233 (0.128)	0.164 (0.287)
WASO	0.064 (0.679)	**0.33 (0.028)**	−0.245 (0.108)	0.05 (0.749)
Wake proportion	0.061 (0.696)	**0.295 (0.052)**	**−0.267 (0.079)**	0.029 (0.849)
N1 proportion	0.169 (0.274)	**0.273 (0.073)**	**−0.377 (0.012)**	−0.053 (0.733)
N2 proportion	−0.133 (0.389)	−0.222 (0.147)	0.097 (0.53)	−0.093 (0.548)
N3 proportion	0.128 (0.408)	−0.015 (0.925)	0.173 (0.261)	0.121 (0.434)
REM proportion	−0.03 (0.848)	**−0.267 (0.079)**	**0.447 (0.002)**	−0.034 (0.827)

cal framework that allow us to analyse retrospectively the time-varying features of a person's sleep–wake cycle and quantify the sleep periods in a coherent and systematic way, and to model directly the PA during sleep which would otherwise be problematic with a standard HMM as the emission distributions would be highly positively skewed. What's more, as supported by an analysis with the PSG data, our method allow us to systematically quantify an individual's stochastic dynamic behaviour of transitions between, and sojourn times within, sub-states that may be associated with deeper and lighter or interrupted sleep stages. We also found interesting associations between parameters derived from the main and sub-HMMs and key PSG parameters and circadian parameters. The method developed here is thus of high interest to sleep and circadian biology research.

We recognize that there are still limitations with the current study. For instance, the cohort we consider here is still relatively small in size, and our model only considers a univariate time series (PA) and assumes a homogeneous Markov chain. In future, it would be of interest to exploring the potential benefits of integrating other sleep related biomarkers such as heart rate and skin temperature (e.g. by considering multivariate emissions), and to incorporate the possible covariate information (e.g. introducing a generalized linear model on the transition probabilities) into our HMM-based modelling approach for a finer sleep stage analysis. Another interesting perspective is considering a longitudinal extension of the current modelling framework that jointly analyses the heterogeneity and homogeneity among

subjects to allow for comparison between subjects and information pooling across subjects when making predictions. However, developing computationally feasible Bayesian inference methods for these tasks is non-trivial and is beyond the scope of this chapter.

3.A Further details of the MCMC algorithm

In this section we give additional computational details for updating the zero weights in the zero-inflated emission distributions. For cases where multiple point masses are used (as in our MESA application) the scheme described here can be easily adjusted in a similar fashion. More specifically, we update the reparametrized state-specific zero weights $\tilde{w}_{i,j}$, $i = 1, \dots N$, $j = 1, 2$, via a log-normal random walk

$$\log(\tilde{w}_{i,j}^{'}) = \log(\tilde{w}_{i,j}) + \phi_{i,j},$$

where $\phi_{i,j} \sim \mathcal{N}(0, \tau_w^2)$. The acceptance probabilities of this move for the main and sub-HMM are

$$\min\left(1, \frac{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|\boldsymbol{\theta}^{'})f(\tilde{W}^{'})}{f(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}|\boldsymbol{\theta})f(\tilde{W})} \prod_{i=1}^{N}\prod_{j=1}^{2} \frac{\tilde{w}_{i,j}^{'}}{\tilde{w}_{i,j}}\right)$$

and

$$\min\left(1, \frac{f(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \boldsymbol{\theta}^{S'})f(\tilde{W}^{'})}{f(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, \boldsymbol{\theta}^{S})f(\tilde{W})} \prod_{i=1}^{N}\prod_{j=1}^{2} \frac{\tilde{w}_{i,j}^{'}}{\tilde{w}_{i,j}}\right),$$

respectively, where $\tilde{W}^{'}$ denotes the vector of proposed $\tilde{w}_{i,j}^{'}$ and $\boldsymbol{\theta}^{'}$ and $\boldsymbol{\theta}^{S'}$ denote the corresponding updated parameter set for the main and sub-HMM, respectively.

3.B Further details of the MESA application

In this section we present additional implementation details and estimation results for applying the proposed conditional HMM approach on the example subjects. For the main-HMM we have chosen $a = 0.1$, $b = \max(\log(1 + PA)) + 3$ and $\alpha = 0.65$ (defined as in chapter 2). Figure 3.7 (left panel) displays the estimated emission densities (on the positive domain) obtained by averaging over the emissions generated across MCMC iterations for subject 921 and 3439, respectively. The posterior modal number of knots is 8 and 10 for subject 921 and 3439, respectively. The posterior means for the state specific weights of the point masses at 0 ($w_{i,1}$) and $\log(1.1)$ ($w_{i,2}$) are shown in table 3.7. For the sub-HMM we set $a = 4.5$, $b =$ where is the 30-s PA data corresponding to state 1 of the main-HMM and $\alpha = 0.65$. The

Table 3.7: Posterior means for the state specific weights of the point masses at 0 ($w_{i,1}$) and log(1.1) ($w_{i,2}$)

Subject	$w_{1,1}$	$w_{1,2}$	$w_{2,1}$	$w_{2,2}$	$w_{3,1}$	$w_{3,2}$
Subject 921	0.314	0.155	0.002	0.002	0.002	0.001
Subject 3439	0.317	0.133	0.009	0.012	0.003	0.002

estimated emission densities (on the positive domain) for the two subjects are displayed in the right panel of Figure 3.7. We can clearly see that these emissions are highly skewed to the right.

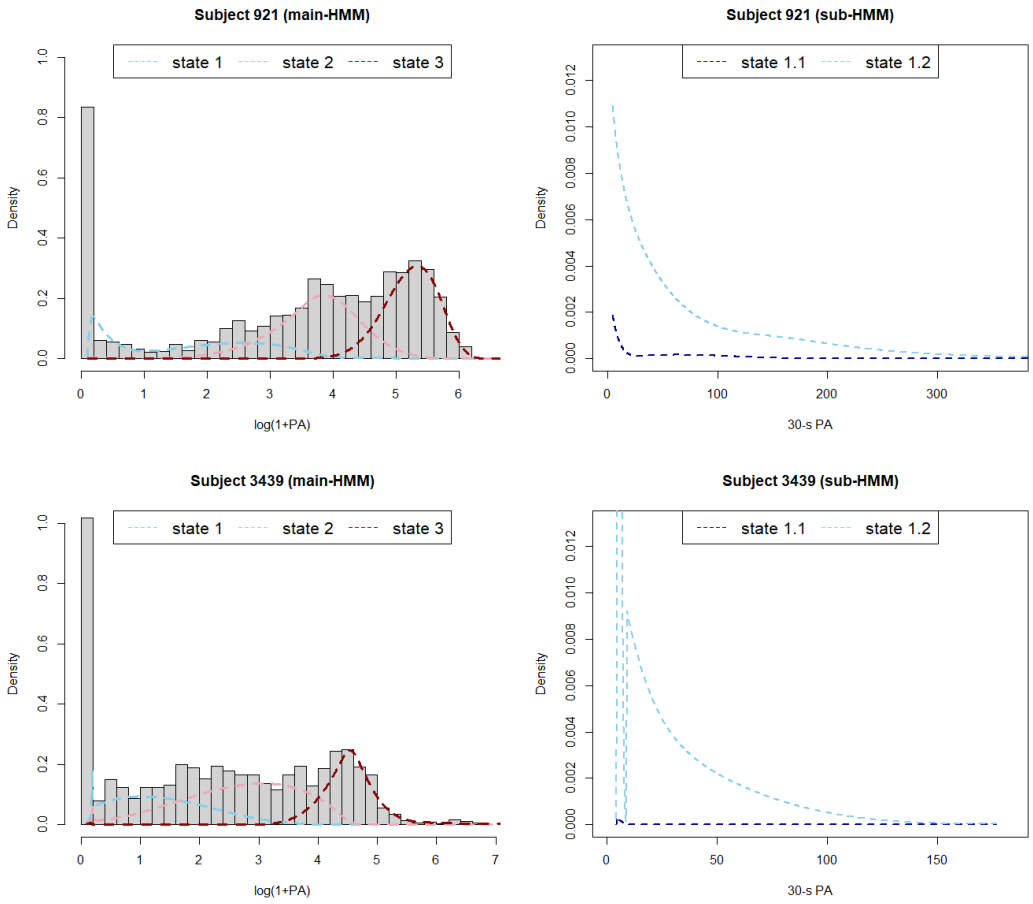

Figure 3.7: Left panel: histogram of 5-min transformed PA data along with the estimated emission densities (weighted according to their proportion in the stationary distribution of the estimated Markov chain) for the main-HMM; right panel: estimated emission densities for the sub-HMM. The weights for the point masses are not shown in the graph.

Chapter 4

Bayesian inference for nonparametric hidden Markov models with hierarchical Dirichlet process priors

4.1 Introduction

Bayesian nonparametric (BNP) models are playing an increasingly important role in modern statistical learning due to their great representation power and modelling flexibility, and also the development of relatively efficient learning algorithms. Equipped with an infinite dimensional parameter space, such models allow the sophistication of a stochastic system to scale automatically with the complexity of the data in a Bayesian framework, eliminating the need of performing tedious model selection. We refer to Xuan et al. [2019] for a state-of-the-art review of different variants and extensions of the BNP models and their applications. In this chapter, our interests lie in the use of an important class of BNP models, namely the hierarchical Dirichlet process (HDP) [Teh et al., 2006], for constructing multivariate hidden Markov models (HMM) that permit flexible emission distributions and a potentially unbounded number of hidden states. HDP has been successfully used to define nonparametric priors in probabilistic graphical models for a wide range of modelling tasks. For instance, it is used in mixture models for learning the latent cluster structures among groups of data [Sohn et al., 2009; Savage et al., 2010], and in HMMs and partially observable Markov decision processes, for automatic learning of the number of hidden states as well as the corresponding transition dynamics

[Hines et al., 2015; Doshi-Velez et al., 2013]. Here, by exploiting the strengths of the HDP and a suitable integration with HMMs, we will develop a new Bayesian hidden Markov modelling framework that generalize existing nonparametric Bayesian HMMs to offer greater modelling flexibility.

As discussed in earlier chapters, in many scenarios choosing appropriate emission distributions for an HMM is important yet challenging. This may particularly be the case for multivariate HMMs where the observed variable is multi-dimensional (we restrict our focus on the continuous case here). A convenient and widely used choice in the parametric multivariate setting is the multivariate normal distribution [Phillips et al., 2015; Maruotti et al., 2017]. Alternatively, conditional independence of the observed variables given the underlying state is often assumed and thus the joint emission distributions can be specified based on the corresponding marginal distributions [Choo-Wosoba et al., 2020; DeRuiter et al., 2017]. Unsurprisingly, these modeling assumptions can be overly simplistic and inadequate in some cases. More flexible parametric models, such as the multivariate t distribution [Scott et al., 2005], Gaussian mixture models [Volant et al., 2014] and copulas [Härdle et al., 2015] have been introduced to HMMs to address distributional features like heavy-tailedness, multi-modality and non-linearity within state dependence. However, estimation in such models usually involves non-trivial model selection problems (e.g. selection of the number of mixture components and choice of the copula) and due to limited flexibility, a particular model may only work for certain types of data. More recently, a few nonparametric estimation procedures have been developed, see for instance Yau et al. [2011] for a Bayesian nonparametric method, Alexandrovich et al. [2016] and Gassiat et al. [2016a] for various maximum likelihood based methods and Lehéricy [2018] for spectral and least squares estimators. While these methods seem to offer promising practical or theoretical results, their implementation can be quite challenging in practice and they are of very limited use.

In this chapter we first investigate the use of HDP-based mixture models for flexible yet parsimonious modelling of the emission distributions in a multivariate HMM with finite state space. We propose to specify the emission distributions via infinite mixture models, where the mixing measure associated with each state is induced and coupled via the HDP. The Dirichlet process (DP) mixture models, which may be regarded as a special case of HDP mixture models, have been identified as an attractive nonparametric approach to density estimation since the seminal work of Escobar and West [1995]. They provide flexible priors that have dense support over the entire class of continuous distributions, and strong asymptotic results exist on the posterior consistency and convergence rates in both univariate

and multivariate settings (see Ghosal and Van Der Vaart [2001]; Tokdar [2006]; Shen et al. [2013]; Canale et al. [2017] and references therein). In our context it is natural to consider a collection of DP mixture models, one for each hidden state. To achieve a parsimonious model representation and higher efficiency in parameter estimation, we propose to introduce a further hierarchy on top of these DP mixtures to encourage sharing of mixture components and therefore data points across states. Our model is in contrast to the approach in Yau et al. [2011] where the focus is restricted to finite translation HMMs [Gassiat et al., 2016b] and the emission distributions are specified based on a single DP mixture model. We also note the work of Torbati and Picone [2015] who considered using HDP mixture models for the emission distributions; however, operationally they are approximated by hierarchical finite mixture models with a pre-fixed number of mixture components. We develop a novel Markov chain Monte Carlo (MCMC) methodology for asymptotically exact posterior simulation in such HMMs, without resorting to finite truncated approximations of HDP prior. The algorithm effectively combines an efficient dynamic programming algorithm for HMMs, which jointly samples the state sequence conditional on the observations, with the slice sampling technique [Neal, 2003], which efficiently samples from the HDP mixture model. With minor adaptations the proposed algorithm can also be used for inference in HMMs where each emission distribution is modelled separately by a DP mixture model (if more appropriate), extending Yau et al. [2011] to a more general HMM framework.

Our second goal of this chapter is to build a fully nonparametric HMM that allows the number of hidden states to be automatically inferred from data when this information is unavailable a-priori. In our HMM setting where the emission distributions are specified via HDP mixtures, commonly used information criteria and the parallel sampling approach of Congdon [2006] that we described in Chapter 2 are not applicable here for model selection since the parameter space is now infinitely dimensional and exact evaluation of the joint posterior density is not possible (unless using some sort of truncated approximations). The marginal likelihood with respect to the number of states would also be difficult to estimate accurately and computationally intensive simulations are often required for approximating integrals involved. Therefore, instead of trying to pick a single "best" value for the number of states, we propose to make further use of the HDP to model the transition matrix nonparametrically. In this direction, the use of the HDP was originally introduced by Teh et al. [2006] (known as the HDP-HMM) to define a prior on infinite dimensional transition matrices and allow the number of states to be unbounded a-priori. An improvement of the HDP-HMM, known as the sticky HDP-HMM, was proposed

in Fox et al. [2011] in the framework of an application to speaker diarization. This model introduced an additional "sticky" parameter to encourage more self transitions in the state process and thus provide a more realistic modelling of the temporal persistence, guarding against the generation of redundant states and unrealistically rapid state transitions. Since then, the sticky HDP-HMM is almost regarded as the default version of the Bayesian nonparamtric HMM and has enjoyed varying degrees of success in a wide range of real world tasks, such as speech recognition [Torbati and Picone, 2015], motion pattern learning [Hu et al., 2017], financial time series modelling [Song, 2014], and physiological data modelling [Hadj-Amar et al., 2020a], among many others. However, as pointed out in Zhou et al. [2021], the sharing of the sticky parameter across states induces an undesirable coupling effect between the prior on transition probabilities and that on self persistence probabilities. For instance, when the transition probabilities are expected to be similar across states, so are the self persistence probabilities, which limits the expressiveness of the prior. To address this issue, we propose to reformulate the prior on the transition matrix as in Zhou et al. [2021] (termed as the disentangled sticky HDP model) to disentangle this intrinsic correlation, thus offering extra flexibility to capture more complex transition dynamics in real data. The new model admits the original sticky HDP model as a special case. However, our model further extends over Zhou et al. [2021] as we also model the emission distributions nonparametrically in a general multivariate setting. Furthermore, we develop a MCMC methodology for fully Bayesian inference in the resulting nonparametric HMM. The algorithm exploits the beam sampling technique developed in Van Gael et al. [2008] for efficient simulation of the state sequence while avoiding the truncated approximation to the HDP prior for the transition matrix (see e.g. Fox et al. [2011]; Zhou et al. [2021]), which can lead to misleading posterior samples (when truncation level is small) or significant computational burden (when truncation level is large). To the best of our knowledge, this is the first Bayesian methodology developed for HMMs with fully nonparametric HDP priors without relying on such approximations, and our proposed modelling framework is still computationally accessible despite being remarkably flexible.

The structure of this chapter is organized as follows. In section 4.2 we will introduce the building blocks for constructing the Bayesian nonparametric HMMs, the DP and HDP. In section 4.3 we introduce the use of HDP mixture models for nonparametrically modelling the emission distributions in a multivariate HMM with fixed number of states and develop the associated MCMC methodology for posterior inference. In section 4.4 we extend the model developed in section 4.3 to allow automatic learning of the number of states and we develop the simulation strategy

for the resulting model. In section 4.5 we apply our proposed model to motion and heart rate data collected from Apple watch [Walch, 2019] for unsupervised learning of sleep macrostructure. Lastly, in section 4.6, we conclude with a brief discussion.

4.2 Dirichlet processes and hierarchical Dirichlet processes

In this section we introduce the DP and its extension HDP that will lay the foundation of our nonparametric inference methods. We will also review sampling based algorithms available for inference in such models.

4.2.1 Dirichlet processes

We begin by giving a formal definition of DP that is due to Ferguson [1973]. Let Θ be a measurable parameter space, H be a probability distribution defined over Θ and α be a positive real number. We say a random measure G is distributed according to a DP with base measure H and concentration parameter α, denoted as $DP(\alpha, H)$, if for any finite partition $\{A_1, \ldots, A_K\}$ of Θ

$$(G(A_1), \ldots, G(A_K)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_K)), \tag{4.1}$$

where $Dir(\cdot)$ denotes the Dirichlet distribution. Equivalently speaking, given any finite partition of Θ, the measure of a random probability distribution $G \sim DP(\alpha, H)$ on this partition set is Dirichlet distributed according to (4.1). From the properties of the Dirichlet distribution we see that $E[G(A_i)] = H(A_i)$, which is independent of α, and $Var(G(A_i)) = H(A_i)(1 - H(A_i))/(\alpha + 1)$. Therefore the base measure H can be understood as the "mean" of the DP while the concentration parameter α controls the variability (or level of concentration) of a random draw G from the DP around its mean. In particular, as $\alpha \to \infty$, $Var(G(A_i)) \to 0$ and thus G converges weakly to H. On the other hand when $\alpha \to 0$, $Var(G(A_i)) \to H(A_i)(1 - H(A_i))$, which indicates that $G(A_i)$ becomes a Bernoulli random variable and G assigns either full mass or no mass in A_i.

We now turn to a more direct and constructive definition of the DP, known as the stick-breaking construction [Sethuraman, 1994], which gives us a way to draw

a single distribution from it. Let

$$\begin{aligned} \beta_i^{'} &\sim Be(1,\alpha), \quad i = 1,\ldots,\infty, \\ \beta_1 = \beta_1^{'}, \quad \beta_i &= \beta_i^{'}\prod_{j=1}^{i-1}(1-\beta_j^{'}), \quad i = 2,\ldots,\infty, \\ \theta_i &\sim H, \quad i = 1,\ldots,\infty, \\ G &= \sum_{k=1}^{\infty}\beta_k\delta_{\theta_k}, \end{aligned} \tag{4.2}$$

where Be denotes the Beta distribution, H is the base measure as defined above and δ_{θ_i} denotes the Dirac delta measure centred at θ_i. Then it is shown that with probability 1, $\sum_{k=1}^{\infty}\beta_k = 1$ and $G \sim DP(\alpha, H)$ [Sethuraman, 1994]. The construction of the weights $\beta = \{\beta_i\}_{i=1}^{\infty}$ via the first two rows of (4.2) is often denoted by $\beta \sim GEM(\alpha)$ for convenience (GEM is short for Griffiths, Engen, and McCloskey; see e.g. Pitman [2002]), where the expectation of β_k decreases exponentially in k. An important observation from this definition of the DP is that a random draw from a DP is discrete (with probability one), even if H is a continuous distribution. We can also recognize the role α plays in the DP as controlling the relative magnitude of the weights $\{\beta_i\}_{i=1}^{\infty}$. As $\alpha \to 0$, $E[\beta_i^{'}] \to 1$ and G will reduce to a point mass, while as $\alpha \to \infty$, $E[\beta_i^{'}] \to 0$ and the infinite weights $\{\beta_i\}_{i=1}^{\infty}$ in G tends to be evenly distributed.

The polya urn representation proposed in Blackwell et al. [1973] provides another defining property of the DP that makes inference with DP priors computationally tractable and it has been used to establish the existence of the DP. More specifically, if $G \sim DP(\alpha, H)$, where α and H are defined as above, and let $\theta_1, \theta_2, \ldots$ be a sequence of random variables that are independently distributed according to G, i.e. $\theta_1, \theta_2, \ldots | G \overset{iid}{\sim} G$. Then the predictive distribution of θ_{n+1} given the preceding draws $\theta_1, \ldots, \theta_n$ (with G integrated out) can be expressed as:

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{1}{\alpha+n}(\alpha H + \sum_{i=1}^{n}\delta_{\theta_i}) = \frac{1}{\alpha+n}(\alpha H + \sum_{j=1}^{K_n} n_j\delta_{\theta_j^*}) \tag{4.3}$$

where $\{\theta_1^*, \ldots, \theta_{K_n}^*\}$ represent unique values of $\{\theta_1, \ldots, \theta_n\}$ and $n_j := \sum_{i=1}^{n}\mathbf{I}(\theta_i = \theta_j^*)$, with $\mathbf{I}(\cdot)$ denoting the indicator function that assumes a value of one when the argument is true. Therefore (4.3) provides a way to draw observations (the θ_i) directly from a DP prior without first constructing the infinite dimensional probability measure G in (4.2). If we additionally introduce a cluster assignment variable S_i

such that $S_i = k$ if $\theta_i = \theta_k^*$, then it can be shown that the process described by (4.3) can also be characterized by the predictive distribution for the S_i:

$$P(S_{n+1} = k|S_1, \ldots, S_n) = \frac{1}{\alpha + n}(\alpha \mathbf{I}(k = K_n + 1) + \sum_{j=1}^{K_n} n_j \mathbf{I}(k = j)), \tag{4.4}$$

where $K_n + 1$ represents an empty new cluster. A key observation from both (4.3) and (4.4) is the clustering property of the DP, which lays the foundation of using DP for constructing mixture models. That is, we tend to see observations that we have seen before, and they are more likely to join existing larger clusters (the θ_i^* with larger n_i). This is the so-called "richer get richer" property of the DP. On the other hand we can see the nonparametric nature of DP: there is always a positive probability to introduce a new cluster, whose magnitude is controlled by α. A larger value of α will lead to an a-priori larger number of clusters. The expressions in (4.3) and (4.4) are often explained in the metaphors of the "polya urn scheme" or the "Chinese restaurant" process (CRP), and we refer to Blackwell et al. [1973] and Pitman [2006] for more details and their implications.

Since the DP can be regarded as a prior distribution over distributions, we are able to talk about the posterior distribution associated with the DP, which has nice structure that makes it computationally convenient. Based on the conjugacy between the Dirichlet prior and the multinomial likelihood and the definition of DP given in (4.1), we can easily see that if $G \sim DP(\alpha, H)$ and $\{\theta_i\}_{i=1}^n | G \overset{iid}{\sim} G$, then

$$G|\{\theta_i\}_{i=1}^n \sim DP(\alpha + n, (\alpha H + \sum_{i=1}^{n} \delta_{\theta_i})/(\alpha + n)). \tag{4.5}$$

That is, the DP prior for G is conjugate with respect to i.i.d. sampling from G. It is important to note that the posterior base measure is exactly the predictive distribution given in (4.3), taking the form of a weighted average between the base measure H and the empirical distribution of $\{\theta_i\}_{i=1}^n$. As the sample size n gets larger (i.e. $n >> \alpha$), the posterior DP will have a large concentration parameter and a base measure that is dominated by the empirical distribution, which exhibits a kind of posterior consistency property. Moreover, the posterior of G in (4.5) can

be explicitly constructed as

$$\begin{aligned} G' &\sim DP(\alpha, H), \\ \beta_0, \ldots, \beta_{K_n} &\sim Dir(\alpha, n_1, \ldots, n_{K_n}), \\ G|\{\theta_i\}_{i=1}^n &= \beta_0 G' + \sum_{i=1}^{K_n} \beta_i \delta_{\theta_i^*}, \end{aligned} \tag{4.6}$$

where the θ_i^* is defined as before representing the unique values of the θ_i. We refer to Pitman [1996] for more details. This posterior representation proved to be very useful in the development of inference algorithms in DP and HDP, see, e.g. Teh et al. [2006].

4.2.2 Dirichlet process mixture models

The discrete and clustering nature of the DP makes it useful for defining a nonparametric prior for the components in a mixture model, leading to a so-called Dirichlet process mixture model (DPMM). Using the notation of section 4.2.1, the generative process of the DPMM can be specified as:

$$\begin{aligned} G|\alpha, H &\sim DP(\alpha, H), \\ \theta_i|G &\sim G, \quad i = 1, \ldots, n, \\ y_i|\theta_i &\sim f(y_i|\theta_i), \quad i = 1, \ldots, n, \end{aligned} \tag{4.7}$$

where $f(\cdot|\theta)$ denotes a generic probability distribution parameterized by θ. The resulting distribution of y_i induced by the DP prior is thus

$$F(y_i) = \int f(y_i|\theta)\, dG(\theta) = \sum_{j=1}^{\infty} \beta_j f(y_i|\theta_j),$$

where $G = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \sim DP(\alpha, H)$. Therefore the DPMM can be seen as an extension of the finite mixture model with an infinite number of components, where great flexibility is allowed. On the other hand, it can be derived by considering a $Dir(\alpha/K, \ldots, \alpha/K)$ prior for the mixture weights in a finite mixture model, $(\beta_1, \ldots, \beta_K)$, and then taking the limit as $K \to \infty$ (see e.g. Rasmussen et al. [1999]). A useful representation alternative to (4.7) can be obtained by making use

of the stick-breaking representation of the DP and the cluster assignment variables:

$$\begin{aligned} \beta|\alpha \sim GEM(\alpha), \quad \theta_i|H \sim H, \quad i = 1, \ldots, \infty, \\ S_i|\beta \sim \beta, \quad i = 1, \ldots, n, \\ y_i|\{\theta_i\}_{i=1}^{\infty}, S_i \sim f(y_i|\theta_{S_i}), \quad i = 1, \ldots, n. \end{aligned} \tag{4.8}$$

Note that clustering of data is implicitly achieved through sharing of the same cluster assignment variable across data points, a key property of the DP that we introduced earlier in section 4.2.1.

4.2.3 Hierarchical Dirichlet process

The HDP is useful for problems concerning data from multiple related groups, each of which could be modelled via a DP and we want to tie these individual DPs together in an appropriate way for modelling purposes and to enhance statistical strength. A (2-layer) HDP is built from two levels of DPs as follows (generalization to a higher number of layers follow the same rationale):

$$\begin{aligned} G_0|\gamma, H \sim DP(\gamma, H), \\ G_j|\alpha, G_0 \sim DP(\alpha, G_0), \quad j = 1, \ldots, J, \end{aligned} \tag{4.9}$$

where γ is the concentration parameter for the top-level DP and J denotes the number of groups that we want to consider jointly. We see that the random probability measures G_j from the bottom layer are connected as they share the same base measure G_0, which is itself distributed according to a DP given by the top layer. An important consequence of this construction is that $\{G_j\}_{j=1}^{J}$ will share the same set of atoms as G_0 (viewed as an atomic measure) due to the discrete nature of G_0, and this would not be the case if G_0 is continuous. The concentration parameters γ and α govern the variability of G_0 around H and G_j around G_0, respectively.

Analogous to the DP, draws from a HDP can be obtained by its stick-breaking construction. Since $G_0 \sim DP(\gamma, H)$, we know that it can be expressed as $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}$, where $\beta = \{\beta_i\}_{i=1}^{\infty} \sim GEM(\gamma)$ and $\{\theta_i\}_{i=1}^{\infty} \overset{iid}{\sim} H$. Similarly, each G_j can be constructed as $G_j = \sum_{k=1}^{\infty} \tilde{\pi}_{jk} \delta_{\theta^*_{jk}}$, where $\tilde{\pi}_j = \{\tilde{\pi}_{jk}\}_{k=1}^{\infty} \sim GEM(\alpha)$ and $\{\theta^*_{jk}\}_{k=1}^{\infty} \overset{iid}{\sim} G_0$. Note that the θ^*_{jk} can take the same value for different values of k since G_0 is discrete. Using the definition of the DP and properties of the Dirichlet distribution, Teh et al. [2006] showed that G_i can be constructed using unique atoms

as

$$\pi_j = \{\pi_{jk}\}_{k=1}^{\infty}|\alpha,\beta \sim DP(\alpha,\beta), \quad \{\theta_k\}_{k=1}^{\infty}|H \overset{iid}{\sim} H,$$
$$G_j = \sum_{k=1}^{\infty}\pi_{jk}\delta_{\theta_k}, \quad j = 1,\ldots,J, \tag{4.10}$$

where π_j can be constructed via stick-breaking as

$$\pi_{jk}^{'} \sim Be(\alpha\beta_k, \alpha(1-\sum_{i=1}^{k}\beta_i)), \quad k = 1,\ldots,\infty,$$
$$\pi_{j1} = \pi_{j1}^{'}, \quad \pi_{jk} = \pi_{jk}^{'}\prod_{l=1}^{k-1}(1-\pi_{jl}^{'}), \quad k = 2,\ldots,\infty. \tag{4.11}$$

It follows that $E[\pi_{jk}] = \beta_k$ and $Var(\pi_{jk}) = Var(\beta_k) + E[\beta_k(1-\beta_k)/(1+\alpha)]$ for $j = 1,\ldots,J$. An alternative representation of the HDP that integrates out the random measures G_0 and G_j can be obtained in terms of the Chinese restaurant franchise (CRF), which is a direct generalisation of the polya urn scheme or the CRP for the DP. Since our inference algorithms introduced in later sections do not require an explicit instantiation of the CRF, we shall not discuss this further and we refer to Teh et al. [2006] for more details.

The posterior distribution of HDP is more complicated than that of the DP due to the hierarchical structure and the discreteness of the globally shared base measure G_0. For each $j = 1,\ldots,J$, consider $\{t_{ji}\}_{i=1}^{N_j} \overset{iid}{\sim} \tilde{\pi}_j$, where N_j is a positive integer representing the number of observations in group j, and let t_j^* denote the number of unique values among $\{t_{ji}\}_{i=1}^{N_j}$ and let $\{k_{jt}\}_{t=1}^{t_j^*} \overset{iid}{\sim} \beta$. Let K_J denotes the number of unique values of $\{k_{jt}\} = \{k_{jt_{ji}}\}_{j=1,\ldots,J;i=1,\ldots,N_j}$ and $\{\theta^{**}\} = \{\theta_1^{**},\ldots,\theta_{K_J}^{**}\}$ be i.i.d. samples from H. Then it follows from the stick breaking construction of the HDP that $\theta_{jt}^* = \theta_{k_{jt}}^{**} \sim G_0$ and $\theta_{ji} = \theta_{jt_{ji}}^* \sim G_j$. Here the observations $\{\theta_{jt}^*\}$ and $\{\theta_{ji}\}$ are implicitly described via $\{\theta^{**}\}$, $\{t_{ji}\}$ and $\{k_{jt}\}$ so that the effect of β and $\tilde{\pi}_j$ on the resulting clustering behaviour of the θ_{ji} can be decoupled. Let us further define $m_{jk} = |\{\{\theta_{jt}^*\}_{t=1}^{t_j^*} : \theta_{jt}^* = \theta_k^{**}\}|$ and $n_{jk} = |\{\{\theta_{ji}\}_{i=1}^{N_j} : \theta_{ji} = \theta_k^{**}\}|$ where $|\cdot|$ denotes the cardinality of the argument. Observing that G_0 depends on $\{\theta^{**}\}$, $\{t_{ji}\}$ and $\{k_{jt}\}$ only via θ_{jt}^* which are themselves i.i.d. draws from G_0, the posterior of G_0 takes the usual form of the posterior of a DP:

$$G_0|\gamma,H,\{\theta_{jt}^*\} = DP(\gamma + m_{..}, (\gamma H + \sum_{k=1}^{K_J} m_{.k}\delta_{\theta_k^{**}})/(\gamma + m_{..})) \tag{4.12}$$

where $m_{.k} = \sum_{j=1}^{J} m_{jk}$ and $m_{..} = \sum_{k=1}^{K_J} m_{.k}$. Using the standard result of DP (see

(4.6)), G_0 can be explicitly constructed as

$$\begin{aligned} G_0^{'} &\sim DP(\gamma, H), \\ \beta_0, \ldots, \beta_{K_J} &\sim Dir(\gamma, m_{.1}, \ldots, m_{.K_J}), \\ G_0|\gamma, H, \{\theta_{jt}^*\} &= \beta_0 G_0^{'} + \sum_{i=1}^{K_J} \beta_i \delta_{\theta_i^{**}}. \end{aligned} \tag{4.13}$$

Conditional on α and G_0, the posterior for the G_j are independent and depend on $\{\theta^{**}\}$, $\{t_{ji}\}$ and $\{k_{jt}\}$ only via θ_{ji}, which are themselves i.i.d. draws from G_j. Therefore we have

$$G_j|\alpha, G_0, \{\theta_{ji}\} = DP(\alpha + n_{j.}, (\alpha G_0 + \sum_{k=1}^{K_J} n_{jk}\delta_{\theta_k^{**}})/(\alpha + n_{j.})), \tag{4.14}$$

where $n_{j.} = \sum_{k=1}^{K_J} n_{jk}$ and G_0 is specified as in (4.13). An explicit representation of a draw from (4.14) can thus be obtained as:

$$\begin{aligned} G_j^{'} &\sim DP(\alpha\beta_0, G_0^{'}), \\ \pi_{j0}, \ldots, \pi_{jK_J} &\sim Dir(\alpha\beta_0, \alpha\beta_1 + n_{j1}, \ldots, \alpha\beta_{K_J} + n_{jK_J}), \\ G_j|\alpha, G_0, \{\theta_{ji}\} &= \pi_{j0} G_j^{'} + \sum_{k=1}^{K_J} \pi_{jk} \delta_{\theta_k^{**}}. \end{aligned} \tag{4.15}$$

The posterior structure of HDP can also be explained in the metaphor of the CRF and we refer to Hjort et al. [2010] for a detailed discussion.

4.2.4 Hierarchical Dirichlet process mixture models

The HDP mixture model (HDPM) extends the DPMM to jointly model multiple DPMMs, one for each group, where the clusters underlying the data in one group can be reused for data from another group. To help better illustrate the essence of the HDPM, we introduce three different but equivalent specifications of the HDPM, each of which uses a different representation of the HDP prior and can be useful in certain inference settings. Using the definition in (4.9) the generative process of the

HDPM can be specified as

$$\begin{aligned} G_0|\gamma, H &\sim DP(\gamma, H), \\ G_j|\alpha, G_0 &\sim DP(\alpha, G_0), \quad j = 1, \ldots, J \\ \theta_{ji}|G_j &\sim G_j, \quad j = 1, \ldots, J, \quad i = 1, \ldots, N_j, \\ y_{ji}|\theta_{ji} &\sim f(y_{ji}|\theta_{ji}), \quad j = 1, \ldots, J, \quad i = 1, \ldots, N_j. \end{aligned} \tag{4.16}$$

Note that this is a direct extension of the construction of the DPMM in (4.7), where here we have J instead of one group of observations and the hierarchical prior (4.9) is used to define the base measure in (4.7) (first row) for each group. Alternatively, we can represent the HDPM using the stick-breaking construction given in (4.10) and the cluster assignment variables:

$$\begin{aligned} \beta|\gamma &\sim GEM(\gamma), \\ \pi_j|\alpha, \beta &\sim DP(\alpha, \beta), \quad j = 1, \ldots, J, \\ S_{ji}|\pi_j &\sim \pi_j, \quad j = 1, \ldots, J, \quad i = 1, \ldots, N_j, \quad \{\theta_k\}_{k=1}^{\infty}|H \overset{iid}{\sim} H, \\ y_{ji}|\{\theta_k\}_{k=1}^{\infty}, S_{ji} &\sim f(y_{ji}|\theta_{S_{ji}}), \quad j = 1, \ldots, J, \quad i = 1, \ldots, N_j. \end{aligned} \tag{4.17}$$

This specification is analogous to (4.8) for the DPMM. A third way to define a HDPM is by using the auxiliary variables t_{ji} and k_{jt} introduced in the previous section

$$\begin{gathered} \beta|\gamma \sim GEM(\gamma), \quad \tilde{\pi}_j|\alpha \sim GRM(\alpha), \\ t_{ji}|\tilde{\pi}_j \sim \tilde{\pi}_j, \quad \{k_{jt}\}_{t=1}^{t_j^*}|\beta \sim \beta, \quad j = 1, \ldots, J, \quad i = 1, \ldots, N_j, \quad \{\theta_k\}_{k=1}^{\infty}|H \sim H, \\ y_{ji}|\{\theta_k\}_{k=1}^{\infty}, \{t_{ji}\}, \{k_{jt}\} \sim f(y_{ji}|\theta_{k_{jt_{ji}}}), \quad j = 1, \ldots, J, \quad i = 1, \ldots, N_j. \end{gathered} \tag{4.18}$$

This representation is particularly useful for CRF-based inference algorithms (see e.g. Teh et al. [2006]).

4.2.5 Posterior inference via Markov chain Monte Carlo

Posterior inference for DP (DPMM) and HDP (HDPM) commonly relies on two classes of methods: sampling-based, i.e. MCMC, or optimization based, i.e. variational inference. Here we focus on the former approach and refer to Xuan et al. [2019] and references therein for more details regarding the latter.

Most existing MCMC methods for DP/HDP adopted a Gibbs sampling framework, where we may distinguish three different types of sampling strategies. The first strategy, which we refer to as the collapsed or marginal method, makes use

of the CRP/CRF representation of the DP/HDP and/or conjugacy of the observation distribution $f(\cdot)$ and the base distribution H to infer the clustering structure of the data, with the infinite dimensional random measure G (or G_0 and $\{G_j\}$) integrated out. Representative works in this context are Neal [2000] and Teh et al. [2006], who provide fundamental implementations for inference in DPMM and HDPM, respectively. While these algorithms are relatively easy to implement for conjugate models, generalization to the non-conjugate scenario can be very difficult. Moreover, they can suffer from poor convergence of the MCMC in complex models [Chang and Fisher III, 2013; Chang and III, 2014].

The second strategy, which we refer to as the uncollapsed or conditional method, works with the stick-breaking representations of the DP/HDP and retains the random measure G (or G_0 and $\{G_j\}$) as the state of the Markov chain. The key methodological challenge arises from the infinite dimensional nature of these random measures and the fact that we can only keep a finite number of parameters in the sampling process. Several MCMC methods that admit the correct posterior under the DP/HDP prior have been developed in this context. For DPMMs, a slice sampling scheme is proposed in the seminal work of Walker [2007] and in parallel, Papaspiliopoulos and Roberts [2008] proposed an alternative MCMC method known as the retrospective sampling method. Papaspiliopoulos [2008] combines the ideas of the slice and retrospective sampling to develop a more efficient block Gibbs sampler which is implemented in Yau et al. [2011]. More recently, building on earlier works of Walker [2007] and Kalli et al. [2011], Ge et al. [2015] derived an improved slice sampling method by exploiting the posterior structure of the DP which also admits parallel inference and scales well with large data set. For HDPMs, an efficient distributed slice sampling scheme is developed in Ge et al. [2015] which extends that for the DPMM. We also note that exact parallelized MCMC methods that incorporate an uncollapsed Gibbs sampler and split-merge moves have been developed for both DPMMs [Chang and Fisher III, 2013] and HDPMs [Chang and III, 2014]. While the split-merge moves significantly increase the sampling complexity, no apparent advantages in terms of sampling efficiency could be observed in comparison to the alternative slice sampler developed in Ge et al. [2015].

The uncollapsed framework can also be approached by using some kind of finite truncation of the DP prior such that it generates probability measures of the form $G_K = \sum_{k=1}^{K} \beta_k \delta_{\theta_k}$ where K is a predetermined truncation point (see e.g. Ishwaran and Zarepour [2000]; Ishwaran and James [2001, 2002]; Ishwaran and Zarepour [2002]). With this approximation standard MCMC methods can be used for posterior simulation and in certain scenarios theoretical guarantees are available

regarding the accuracy of the approximation. However, an appropriate truncation point can still be difficult to choose in practice and the error implied by the approximation is hard to quantify in general (see Hjort et al. [2010] for more comments).

The third strategy, which we refer to as partially collapsed methods, emerges in more recent works. It utilizes both the CRP/CRF and the stick-breaking representations in an attempt to improve sampling efficiency in high dimensional settings while enabling parallel and distributed inference. Implementation of such samplers for DPMMs and HDPMs can be found in Yerebakan and Dundar [2017] and Dubey et al. [2020], respectively. However, the computational efficiency of these algorithms comes at a cost: they require approximations when simulating the cluster assignment variables and moreover, they rely on conjugate priors which can be too restrictive.

4.3 Nonparametric modelling in multivariate hidden Markov models using HDP mixtures

In this section we will describe the proposed nonparametric HMM based on the HDP mixtures that enables flexible modelling of the emission distributions in a general multivariate setting. In the following, we first describe the generative process of the model, and then introduce the associated sampling and inference methods, and finally present a simulation study to illustrate the effectiveness of the approach.

4.3.1 Model formulation

To set up the HMM, let $\mathbf{y} = (y_1, \ldots, y_T)$ denote the observed process with $y_t \in \mathbf{R}^p$, let $\mathbf{x} = (x_1, \ldots, x_T)$ denote the corresponding state process distributed as an N-state time-homogeneous Markov chain with transition matrix $\Pi = [\pi_{i,j}]_{i,j=1,\ldots,N}$ and initial distribution δ. Then the proposed model is specified hierarchically as follows

$$\begin{aligned}
\pi_i = \{\pi_{i,j}\}_{j=1}^N | \rho \sim Dir(\rho, \ldots, \rho), \quad i = 1, \ldots, N, \\
x_1 \sim \delta, \quad x_t | x_{t-1}, \{\pi_k\}_{k=1}^N \sim \pi_{x_{t-1}}, \quad t = 2, \ldots, T, \\
\beta | \gamma \sim GEM(\gamma), \\
\phi_k = \{\phi_{k,i}\}_{i=1}^\infty | \alpha, \beta \sim DP(\alpha, \beta), \quad k = 1, \ldots, N, \\
s_t | \{\phi_k\}_{k=1}^N, x_t \sim \phi_{x_t}, \quad t = 1, \ldots, T, \quad \{\theta_k\}_{k=1}^\infty | H_\lambda \overset{iid}{\sim} H_\lambda, \\
y_t | \{\theta_k\}_{k=1}^\infty, s_t \sim f(y_t | \theta_{s_t}), \quad t = 1, \ldots, T,
\end{aligned} \tag{4.19}$$

where ρ, γ and $\alpha > 0$, $\phi_k = (\phi_{k,1}, \phi_{k,2}, \ldots)$, $\Phi = (\phi_1, \ldots, \phi_N)$, $\mathbf{s} = (s_1, \ldots, s_T)$, $\Theta = (\theta_1, \theta_2, \ldots)$, H_λ is a distribution indexed by some hyperparameters λ, $f(\cdot|\theta)$ is

a generic density parameterized by θ. Following earlier works (e.g. Zhou et al. [2021]) we additionally assign gamma hyperpriors to the concentration parameters γ and α, i.e. $\gamma \sim Ga(a_\gamma, b_\gamma)$ and $\alpha \sim Ga(a_\alpha, b_\alpha)$, and prefix ρ and λ in a noninformative and data driven way, respectively. The initial distribution δ may also be fixed in advance. The first two lines in the hierarchy of equations stated in (4.19) defines the hidden state Markov chain process which is the same as that in a parametric Bayesian HMM. The nonparametric part lies in lines 3-6 of (4.19) where we make use of the stick-breaking representation of the HDP to specify the emission distribution as a HDP mixture models. By integrating out the cluster assignment variable s_t we can see that the emission distributions take the form

$$f(y_t|x_t) = \sum_{k=1}^{\infty} \phi_{x_t,k} f(y_t|\theta_k),$$

where the mixture component parameters are shared across states and the weights are state-specific and coupled via the HDP. It is worth noting the similarities and differences of the construction of the HDPM in (4.17) and that in (4.19). The number of groups J in a HDPM now plays the role of the number of states N, which is pre-fixed. However, the "group membership" of y_t in (4.19) is now random and is determined by the unobserved latent variable x_t, whereas in (4.17) this is known a-priori. The representation of the HMM given in (4.19) facilitates a flexible yet parsimonious specification of the model, and moreover, permits an efficient and exact posterior simulation scheme which will be introduced in the next subsection.

4.3.2 Markov chain Monte Carlo methodology

Here we introduce the MCMC algorithm for simulating from the joint posterior density of $(\gamma, \beta, \alpha, \Phi, \Pi, \mathbf{x}, \mathbf{s}, \Theta)$, which can be written as

$$\begin{aligned} f(\gamma, \beta, \alpha, \Phi, \Pi, \mathbf{x}, \mathbf{s}, \Theta|\mathbf{y}) &\propto f(\gamma|a_\gamma, b_\gamma) f(\alpha|a_\alpha, b_\alpha) f(\beta|\gamma) \\ \times \prod_{k=1}^{N} f(\phi_k|\alpha, \beta) \prod_{k=1}^{N} f(\pi_k|\rho) \prod_{k=1}^{\infty} H_\lambda(\theta_k) \delta_{x_1} \prod_{t=2}^{T} \pi_{x_{t-1},x_t} \prod_{t=1}^{T} \phi_{x_t,s_t} f(y_t|\theta_{s_t}). \end{aligned} \tag{4.20}$$

There are two key facts from (4.19) and (4.20) which underlie the design of our algorithm: (i) we can jointly update the state sequence $\mathbf{x}$ using an efficient dynamic programming algorithm available for standard HMMs, namely the FFBS, conditioned on the rest of the parameters; (ii) conditioned on the simulated state sequence the model is reduced to a HDPM where a state-of-the-art slice sampling scheme for the HDP can be modified and used [Ge et al., 2015]. Note that despite

the infinite dimensional nature of the parameter space, for given data only finitely many mixture components in the emission distributions will ever be "activated" (have at least one allocated observation) and required for simulation. To enable the use of the slice sampling technique, we augment the parameter space by introducing auxiliary variables (slice variables) $\mathbf{u} = (u_1, \ldots, u_T)$ such that the joint posterior becomes

$$f(\gamma, \beta, \alpha, \Phi, \Pi, \mathbf{x}, \mathbf{s}, \Theta, \mathbf{u}|\mathbf{y}) \propto f(\gamma|a_\gamma, b_\gamma) f(\alpha|a_\alpha, b_\alpha) f(\beta|\gamma)$$
$$\times \prod_{k=1}^{N} f(\phi_k|\alpha, \beta) \prod_{k=1}^{N} f(\pi_k|\rho) \prod_{k=1}^{\infty} H_\lambda(\theta_k) \delta_{x_1} \prod_{t=2}^{T} \pi_{x_{t-1}, x_t} \prod_{t=1}^{T} \mathbf{I}(u_t < \phi_{x_t, s_t}) f(y_t|\theta_{s_t}), \tag{4.21}$$

It is easy to check that by integrating out $\mathbf{u}$ we recover the original target distribution in (4.20). Importantly, these variables will play the role of dynamically truncating the number of HDP mixture components required for sampling. Let K^* denotes the number of currently activated mixture components, which is given by the number of distinct values in the current HDP cluster assignment variables $\mathbf{s}$. The structure of one sweep of the proposed algorithm (Algorithm 4.1) for simulating from (4.21) is outlined below. Where possible, we shall remove variables in the conditioning set of a conditional distribution based on conditional independence or explicit integration.

- sample $\mathbf{x}|\mathbf{s}, \Pi, \Phi, \delta$,
- sample $\Pi|\mathbf{x}$,
- sample $\Phi, u|\alpha, \beta, \mathbf{s}, \mathbf{x}$,
- sample $\mathbf{s}|\Theta, \mathbf{y}, u, \mathbf{x}, \Phi$,
- sample $\{m_{j,k}\}_{j=1,\ldots,N;k=1,\ldots,K^*}|\alpha, \beta, \mathbf{s}, \mathbf{x}$
- sample $\beta|\{m_{j,k}\}_{j=1,\ldots,N;k=1,\ldots,K^*}, \gamma$,
- sample $\{\theta_k\}_{k=1}^{K^*}|\mathbf{y}, \mathbf{s}$,
- sample $\alpha|\{m_{j,k}\}_{j=1,\ldots,N;k=1,\ldots,K^*}, \mathbf{s}, \mathbf{x}$,
- sample $\gamma|\{m_{j,k}\}_{j=1,\ldots,N;k=1,\ldots,K^*}$,

where the $\{m_{j,k}\}$ are another set of auxiliary variables introduced to facilitate sampling parameters associated with the HDP. More details regarding each of the sampling steps are provided as follows.

Step 1: sampling x. This is achieved by using the standard FFBS procedure (see Chapter 1), which is shown to be more efficient than a element-wise update [Scott, 2002]. Let

$$\alpha_1(k) := f(x_1 = k, s_1|\Phi, \delta) = \delta_k \phi_{k,s_1}, \quad k = 1, \ldots, N,$$

$$\begin{aligned}\alpha_t(k) &:= f(x_t = k, \{s_i\}_{i=1}^t|\Phi, \Pi) \\ &= \phi_{k,s_t} \sum_{x_{t-1}} \alpha_{t-1}(x_{t-1})\pi_{x_{t-1},k}, \quad t = 2, \ldots, T, \quad k = 1, \ldots, N.\end{aligned}$$

Then we can simulate $\mathbf{x}$ from its full conditional distribution by first sampling x_T from

$$f(x_T = k|\{s_i\}_{i=1}^T, \Phi, \Pi) \propto \alpha_T(k), \quad k = 1, \ldots, N,$$

and then iteratively sampling x_t, $t = T-1, \ldots, 1$, from

$$f(x_t = k|x_{t+1}, \{s_i\}_{i=1}^T, \Phi, \Pi) \propto \alpha_t(k)\pi_{k,x_{t+1}}, \quad k = 1, \ldots, N.$$

Step 2: sampling $\mathbf{\Pi}$. Conditional on the state sequence, the rows of the transition matrix $\{\pi_k\}$ are conditionally independent and

$$(\pi_{i,1}, \ldots, \pi_{i,N})|\mathbf{x} \sim Dir(\rho + n_{i,1}, \ldots, \rho + n_{i,N}), \quad i = 1, \ldots, N,$$

where $n_{i,j}$ denotes the number of transitions between state i and j in $\mathbf{x}$. This result follows from the conjugacy between the Dirichlet prior and a multinomial likelihood.

Step 3: jointly sampling $\mathbf{\Phi}, \mathbf{u}$. This is motivated by the fact that by integrating out the slice variables $\mathbf{u}$, we have, for $j = 1, \ldots, N$,

$$(\phi_{j,1}, \ldots, \phi_{j,K^*}, \phi_j^*)|\alpha, \beta, \mathbf{s}, \mathbf{x} \sim Dir(\alpha\beta_1 + n_{j,1}^{'}, \ldots, \alpha\beta_{K^*} + n_{j,K^*}^{'}, \alpha(1 - \sum_{i=1}^{K^*} \beta_i)),$$

where $\phi_{j,1}, \ldots, \phi_{j,K^*}$ are the mixture weights associated with the activated mixture components in the emission distribution for state i, $\phi_j^* = 1 - \sum_{i=1}^{K^*} \phi_{j,i}$ which collapses the weights associated with the inactive mixture components and $n_{i,j}^{'} = \sum_{t=1}^{T} \mathbf{I}(x_t = i, s_t = j)$. We can then sample the slice variables $\mathbf{u}$ conditioned on Φ

$$u_t|\Phi, \mathbf{x}, \mathbf{s} \sim \mathbf{U}(0, \phi_{x_t,s_t}), \quad t = 1, \ldots, T,$$

where $\mathbf{U}(a, b)$ denotes a uniform distribution with support (a, b).

Step 4: sampling s. Let $\beta^* = 1 - \sum_{i=1}^{K^*} \beta_i$. Using the stick-breaking representation in (4.2) and (4.11), we first create new mixture components by recursively

splitting the residual atoms until $\phi_j^* < \min_{t:x_t=j} u_t$ for $\forall j = 1, \ldots, N$, by proceeding as follows

$$\begin{gathered}
K^* := K^* + 1, \\
\beta'_{K^*} \sim Be(1, \gamma), \\
\beta_{K^*} = \beta^* \beta'_{K^*}, \quad \theta_{K^*} \sim H_\lambda, \quad \beta^* := \beta^*(1 - \beta'_{K^*}), \\
\phi'_{j,K^*} \sim Be(\alpha\beta_{K^*}, \alpha(1 - \sum_{i=1}^{K^*} \beta_i)), \quad j = 1, \ldots, N, \\
\phi_{j,K^*} = \phi_j^* \phi'_{j,K^*}, \quad \phi_j^* := \phi_j^*(1 - \phi'_{j,K^*}), \quad j = 1, \ldots, N.
\end{gathered}$$

We then sample the HDP cluster assignment variables using

$$f(s_t = i|\Theta, y_t, u_t, x_t, \Phi) \propto f(y_t|\theta_i)\mathbf{I}(u_t < \phi_{x_t,i}).$$

Note that by ensuring $\phi_j^* < \min_{t:x_t=j} u_t$ for $\forall j = 1, \ldots, N$, it is guaranteed that $\phi_{x_t,k} < u_t$ for $\forall k > K^*$. Therefore the resulting K^* effectively provides an upper limit on the number of mixture components that need updating at each sweep and it can also be shown that it would be finite almost surely [Walker, 2007; Ge et al., 2015]. After sampling $\mathbf{s}$, we further update the K^* (as the number of distinct values in $\mathbf{s}$) and accordingly relabel $\mathbf{s}$, Φ and $\{\theta_k\}_{k=1}^{K^*}$, and then reconstruct ϕ_j^* by collapsing the non-active mixture weights.

Step 5: sampling β. This is achieved using the theory of HDP presented in Teh et al. [2006]. We first sample a set of auxiliary random variables $\{m_{j,k}\}_{j=1,\ldots,N;k=1,\ldots,K^*}$ described in section 4.2.3 from the following conditional distributions

$$f(m_{j,k} = m|\mathbf{s}, \alpha, \beta) \propto S(n'_{j,k}, m)(\alpha\beta_k)^m,$$

where $S(\cdot, \cdot)$ denotes the unsigned Stirling numbers of the first kind. Conditional on $m_{j,k}$, we use the result from (4.13) to sample

$$(\beta_1, \ldots, \beta_{K^*}, \beta^*) \sim Dir(m_{\cdot 1}, \ldots, m_{\cdot K^*}, \gamma)$$

where $m_{\cdot k} = \sum_{j=1}^{N} m_{j,k}$. In this way we can bypass the need to introduce the variables $\{t_{ji}\}$ and $\{k_{jt}\}$ described in section 4.2.3 into our state space.

Step 6: sampling $\{\theta_\mathbf{k}\}_{\mathbf{k=1}}^{\mathbf{K^*}}$. Note that we only need to update the parameters that are associated with the active mixture components as the full conditional distributions for the remaining mixture component parameters are given by their

priors. The full conditional distribution of θ_k is given by

$$f(\theta_k|\mathbf{y},\mathbf{s}) \propto H_\lambda(\theta_k) \prod_{t:s_t=k} f(y_t|\theta_k), \quad k = 1,\ldots,K^*.$$

For illustration consider the case where f is a multivariate normal density parameterized by the mean vector μ and covariance matrix Σ (i.e. $\theta_k = (\mu_k, \Sigma_k)$) and H_λ is a $N_p(\mu_0, \Sigma_0) \times IW(\Delta, V)$ product measure with hyperparameters $\lambda = (\mu_0, \Sigma_0, \Delta, V)$, where IW stands for the inverse Wishart distribution with density given by $IW(\Sigma|\Delta, V) \propto |\Sigma|^{-\frac{(V+p+1)}{2}} \exp(-\frac{1}{2}tr(\Delta\Sigma^{-1}))$. Of course more general non-conjugate models can be applied here. Let $Y_k = \{y_t : s_t = k\}$, then using standard results in multivariate statistics we have that for each k

$$\mu_k|\Sigma_k, Y_K \sim N_p(\tilde{\mu}_k, \tilde{\Sigma}_k),$$

where $\tilde{\Sigma}_k = (\Sigma_0^{-1} + |Y_k|\Sigma_k^{-1})^{-1}$, $\tilde{\mu}_k = \tilde{\Sigma}_k(\Sigma_0^{-1}\mu_0 + \Sigma_k^{-1}\sum_{y_t \in Y_k} y_t)$, and

$$\Sigma_k|\mu_k, Y_k \sim IW(\tilde{\Delta}_k, \tilde{V}_k),$$

where $\tilde{V}_k = V + |Y_k|$ and $\tilde{\Delta}_k = \Delta + \sum_{y_t \in Y_k}(y_t - \mu_k)(y_t - \mu_k)^T$.

Step 7: sampling α. This is achieved using the auxiliary variable sampling scheme proposed in Teh et al. [2006]. First note that the conditional distribution of α given the rest of parameters can be derived as

$$f(\alpha|\{m_{i,j}\}, \{n'_{i,j}\}) \propto \alpha^{a_\alpha + m_{..} - 1} \exp(-\alpha b_\alpha) \prod_{j=1}^{N} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n'_{j,.})},$$

where $\{n'_{i,j}\}$ and $\{m_{i,j}\}$ are computed from steps 3 and 5, respectively, Γ is the gamma function and $n'_{j,.} = \sum_{i=1}^{K^*} n'_{j,i}$. We introduce additional auxiliary variables $\mathbf{w} = (w_1, \ldots, w_N)$ and $\mathbf{e} = (e_1, \ldots, e_N)$ such that the augmented joint density becomes

$$f(\alpha, \mathbf{w}, \mathbf{e}|\{m_{i,j}\}, \{n'_{i,j}\}) \propto \alpha^{a_\alpha + m_{..} - 1} \exp(-\alpha b_\alpha) \prod_{j=1}^{N} w_j^{\alpha} (1 - w_j)^{n'_{j,.} - 1} (\frac{n'_{j,.}}{\alpha})^{e_j}.$$

Then samples of α can be obtained as follows

$$\alpha|\mathbf{w}, \mathbf{e} \sim Ga(a_\alpha + m_{..} - \sum_{j=1}^{N} e_j, b_\alpha - \sum_{j=1}^{N} \log(w_j)),$$

$$w_j|\alpha \sim Be(\alpha + 1, n'_{j,.}), \quad e_j|\alpha \sim Ber(\frac{n'_{j,.}}{\alpha + n'_{j,.}}), \quad j = 1, \ldots, N,$$

where $Ber(p)$ denotes a Bernoulli distribution with parameter p.

Step 8: sampling γ. This is achieved using the result of Escobar and West [1995]. Let $f(\gamma)$ denotes the prior for γ (i.e. $Ga(a_\gamma, b_\gamma)$). We know from Teh et al. [2006] that γ is independent of the rest of the parameters given K^* and $m_{..}$, and we have

$$f(\gamma|K^*, m_{..}) \propto f(K^*|\gamma, m_{..})f(\gamma) = \frac{\gamma^{K^*}\Gamma(\gamma)f(\gamma)}{\Gamma(\gamma + m_{..})},$$

which is the marginal of $f(\gamma, \eta|K^*, m_{..}) \propto f(\gamma)\gamma^{K^*-1}(\gamma + m_{..})\eta^\gamma(1-\eta)^{m_{..}-1}$. We can thus sample γ via

$$\eta|\gamma, K^*, m_{..} \sim Be(\gamma + 1, m_{..}),$$

$$\gamma|\eta, K^*, m_{..} \sim \pi_\eta Ga(a_\gamma + K^*, b_\gamma - \log(\eta)) + (1 - \pi_\eta)Ga(a_\gamma + K^* - 1, b_\gamma - \log(\eta)),$$

where $\pi_\eta = (a_\gamma + K^* - 1)/(m_{..}(b_\gamma - \log(\eta)) + a_\gamma + K^* - 1)$. We refer to Escobar and West [1995] and Teh et al. [2006] for more details.

4.3.3 Model identification

Here we discuss two potential identifiability issues associated with the proposed model. The first one is related to the emission distributions which are specified via HDPM. Clearly the mixture component parameters, weights and also the cluster assignment variables are not identifiable as the "labelling" of mixture components can change from iteration to iteration. However, this indeterminacy does not pose a special problem to our inference as these mixture component themselves don't have substantive interpretations and merely serve as a tool to provide sufficient flexibility in modelling the emission density.

The second identification issue, known as the label switching problem, occurred in general Bayesian HMMs and it is important to address it as it can directly impact the reliability of the posterior inference results. It arises due to the fact that we can arbitrarily permute the state labels of a Bayesian HMM resulting in the same joint posterior density of model parameters. In the literature of Bayesian

nonparametric HMMs, however, this issue is sometimes overlooked or left undiscussed (see, e.g. Sgouralis and Pressé [2017]). It is not a concern only if the objects of interest are label invariant [Geweke, 2007]. Otherwise, some postprocessing of the posterior samples is needed prior to making inference. Here, we propose to use the Kullback–Leibler (KL) relabelling algorithm developed in Stephens [2000] which has been successfully applied for spline-based HMMs in chapter 2, and is also well-suited to the model proposed here. More specifically, for each of the B collected MCMC samples, we construct a $T \times N$ dimensional classification probability matrix whose (i,k)-th entry in our context is given by $P_{i,k}^{(t)} = (\tilde{\pi}_k^{(t)} \phi_{k,s_i}^{(t)})/(\sum_{j=1}^{N} \tilde{\pi}_j^{(t)} \phi_{j,s_i}^{(t)})$, $t = 1, \ldots, B$, where $(\tilde{\pi}_1^{(t)}, \ldots, \tilde{\pi}_N^{(t)})$ is the stationary distribution associated with the transition matrix $\Pi^{(t)}$. The algorithm then involves iteratively searching a specific permutation of state labels to minimize the KL divergence between classification probabilities averaged over the B MCMC iterations, $q_{i,k} = (\sum_{t=1}^{B} P_{i,k}^{(t)})/B$, and the classification probabilities obtained in each MCMC iteration. The "optimized" permutation identified for each MCMC sample can then be used to relabel the samples to achieve a consistent ordering of the labels. See chapter 2 and references therein for more details of the algorithm. In our implementation we make use of the R package *Label.switching* of Papastamoulis [2016a] to perform this minimization procedure.

4.3.4 Simulation study

We demonstrate the performance of the proposed inference algorithm using a simulation study where we prefix the number of states N at its true value and evaluate the estimation performance with regard to the following three aspects:

1. Retrieval of true model parameters: The parameters of the HDP-based emission models are not identifiable so we focus on the parameters associated with the hidden state process, the transition probabilities $\pi_{i,j}$. We report point and interval estimates for the $\pi_{i,j}$ which are based on the posterior mean and 95% credible interval, respectively.

2. Decoding accuracy: We employ the commonly used normalized Hamming distance between the inferred and true state sequence as the metric for quantifying the decoding accuracy (see Fox et al. [2011]; Zhou et al. [2021]). To infer the state sequence, we first estimate the posterior state probability from MCMC draws as

$$\widehat{\mathbf{Pr}}(x_t = k|\mathbf{y}) \approx \frac{1}{B} \sum_{i=1}^{B} \mathbf{I}(x_t^{(i)} = k), \quad t = 1, \ldots, T, \quad k = 1, \ldots, N,$$

where $x_t^{(i)}$ denotes the simulated value of x_t from the i-th MCMC draw. We then estimate the state at time t via local decoding as

$$x_t^* = \underset{k=1,\ldots,N}{\operatorname{argmax}} \widehat{\mathbf{Pr}}(x_t = k|\mathbf{y}),$$

which is the most probable state at time t given all observations. In our experiments we found that this local decoding approach gives superior decoding results compared to directly selecting a simulated state sequence from MCMC samples based on other criteria (e.g. maximize the complete data likelihood).

3. Recovery of the true emission distributions: We estimate each emission density via its posterior predictive density, $f_i(\cdot|\mathbf{y})$, $i = 1, \ldots, N$, which is a natural density estimate in the Bayesian setting and can be easily evaluated by simulation. More specifically, for each $\{\phi_k^{(j)}\}$ and $\{\theta_k^{(j)}\}$ drawn by the MCMC algorithm, we use the posterior decomposition property of the HDP introduced in section 4.2.3 to first generate the mixture component parameter using

$$\begin{cases} \theta_{new} = \theta_k^{(j)} & \text{w.p. } \phi_{i,k}^{(j)}, \quad k = 1, \ldots, K^{*(j)} \\ \theta_{new} \sim H_\lambda & \text{w.p. } \phi_i^{*(j)} \end{cases}$$

where $K^{*(j)}$ denotes the number of activated mixture components in the j-th iteration, and then sample $y_{new} \sim f(y_{new}|\theta_{new})$.

In our example we place a noninformative $Ga(1, 1)$ prior on the HDP concentration parameters γ and α. The hyperparameters μ_0 and Σ_0 in the Gaussian prior are set to the empirical mean and covariance of the data, respectively. For the inverse Wishart prior we set the degree of freedom $V = 5$ and let the expected covariance to be equal to the empirical covariance of the data (i.e. $\Delta = (V - p - 1)$ times the sample covariance). Our results are based on 30k iterations of the algorithm, with the first 10k samples discarded as burn in. Convergence of the sampler is examined by monitoring the traces of the number of the activated mixture components, the concentration parameters, the complete data likelihood and the normalized Hamming distance between the simulated and the true state sequence.

We simulate a data set of length $T = 2000$ from a 3-state bivariate HMM such that the emission densities exhibit multimodality and non-linear within state dependence structures. We specify the emission distributions using mixtures of two

bivariate normal distributions:

$$y_t|x_t = 1 \sim 0.6\mathbf{N}_2(\begin{pmatrix}-15\\10\end{pmatrix}, \begin{pmatrix}100 & 60\\60 & 140\end{pmatrix}) + 0.4\mathbf{N}_2(\begin{pmatrix}10\\40\end{pmatrix}, \begin{pmatrix}100 & 30\\30 & 110\end{pmatrix}),$$
$$y_t|x_t = 2 \sim 0.35\mathbf{N}_2(\begin{pmatrix}-5\\35\end{pmatrix}, \begin{pmatrix}70 & -50\\-50 & 100\end{pmatrix}) + 0.65\mathbf{N}_2(\begin{pmatrix}20\\10\end{pmatrix}, \begin{pmatrix}100 & -55\\-55 & 130\end{pmatrix}),$$
$$y_t|x_t = 3 \sim 0.5\mathbf{N}_2(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}60 & -5\\-5 & 40\end{pmatrix}) + 0.5\mathbf{N}_2(\begin{pmatrix}30\\30\end{pmatrix}, \begin{pmatrix}60 & 35\\35 & 80\end{pmatrix}),$$

and transition matrix

$$\Pi = \begin{pmatrix}0.85 & 0.1 & 0.05\\0.075 & 0.85 & 0.075\\0.05 & 0.1 & 0.85\end{pmatrix}.$$

We run the proposed MCMC algorithm as described earlier. Figure 4.1 presents some trace plots that are helpful in assessing the convergence of the sampler. We can see that the chain reaches stationarity within the first 5k iterations and no apparent convergence issue is detected. Posterior summaries for entries of the transition probability matrix is

$$\begin{pmatrix}0.85_{(0.816,0.881)} & 0.117_{(0.088,0.148)} & 0.033_{(0.017,0.053)}\\0.08_{(0.058,0.105)} & 0.833_{(0.8,0.862)} & 0.087_{(0.064,0.113)}\\0.053_{(0.032,0.077)} & 0.088_{(0.062,0.117)} & 0.859_{(0.826,0.89)}\end{pmatrix}.$$

where the point estimates are the posterior means of the parameters and the associated credible intervals (shown in brackets) are obtained from the 2.5% and 97.5% empirical percentiles of the corresponding posterior samples. Clearly the intervals for all estimates contain their respective true values. Using our proposed local decoding approach, we also achieved a reasonably good match between the inferred and the true state sequences, with a normalized Hamming distance of 0.071, which is equivalent to a decoding accuracy of 92.9%. Our density estimates for the emission distributions along with their empirical counterparts are displayed in Figure 4.2, where the posterior predictive distributions are able to accurately reproduce the key features of the emission distributions in each state.

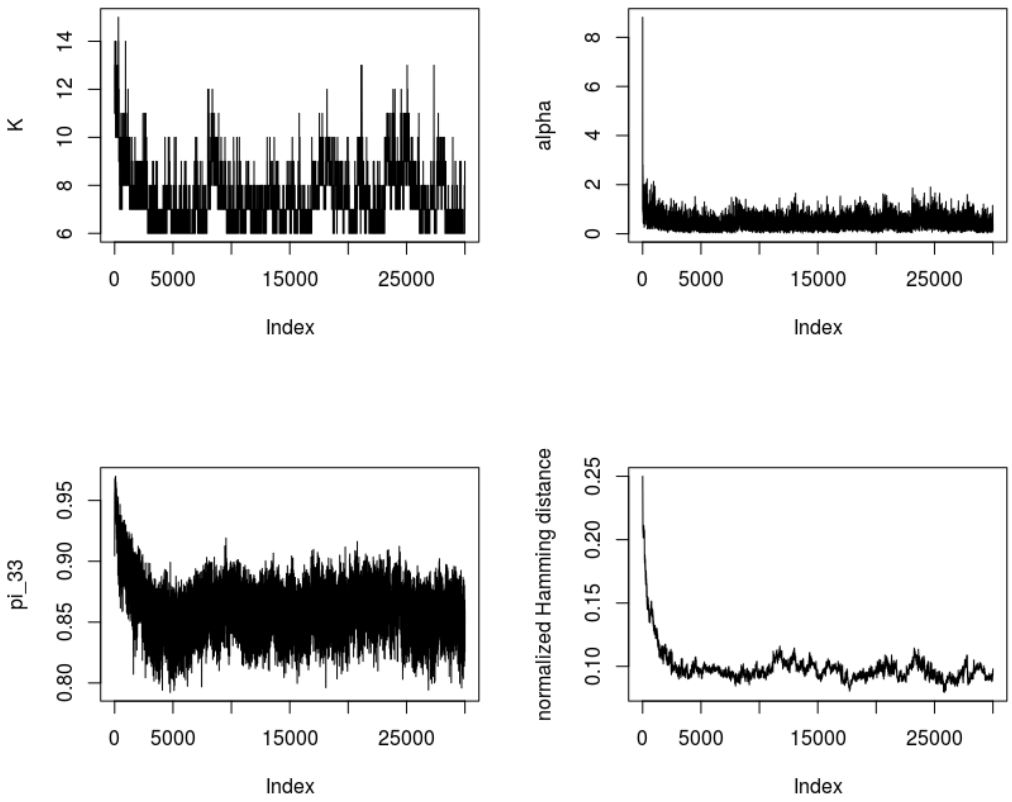

Figure 4.1: Convergence diagnostics for simulation model 1. Top left, top right, bottom left and bottom right panels show the trace plots for the the number of the activated mixture components, the concentration parameter α, the transition probability $\pi_{3,3}$ and the normalized Hamming distance between the simulated and the true state sequence, respectively.

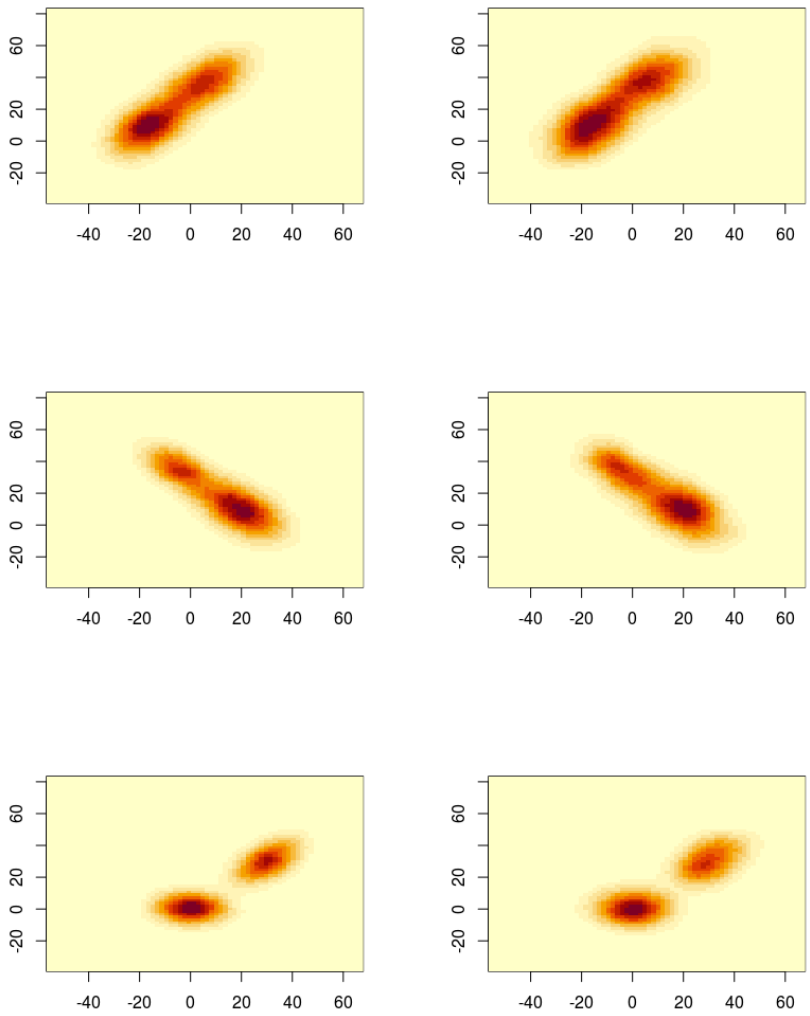

Figure 4.2: Estimate of the posterior predictive density and the simulated data. Left panel: contour plots obtained based on 20k samples of y_{new} for each state; right panel: contour plots obtained based on the simulated data allocated to each state (ground truth). All contour plots were obtained using the R function *kde2d* with default bandwidths, where darker shades representing higher density regions.

4.4 Toward fully nonparametric hidden Markov models with HDPs

In this section we extend the HDP based HMM developed in section 4.3 to specify the transition matrix nonparametrically via the HDP prior, permitting the number of states to be simultaneously inferred from the data along with other model parameters. We shall first review earlier attempts in this direction and then present our model and inference algorithms, whose performance is shown via a simulation study.

4.4.1 HDP-HMM and its extensions

The HDP-HMM proposed in Teh et al. [2006] provides a basic and useful framework for constructing Bayesian nonparametric HMMs with an a-priori unbounded number of states. It makes use of the HDP to define a prior over the rows of the transition matrix, the state-specific transition distributions, so that the state space becomes infinite-dimensional and the transition distributions are coupled by sharing the same set of atoms. The resulting hidden state process can be defined hierarchically as

$$\begin{aligned} \pi_0 = \{\pi_{0,i}\}_{i=1}^{\infty} | \sigma \sim GEM(\sigma), \\ \pi_j = \{\pi_{j,i}\}_{i=1}^{\infty} | \pi_0, c \sim DP(c, \pi_0), \quad j = 1, \ldots, \infty, \\ x_t | x_{t-1}, \{\pi_j\} \sim \pi_{x_{t-1}}, \quad t = 2, \ldots, T, \end{aligned} \tag{4.22}$$

where c, $\sigma > 0$ and the initial distribution of the Markov chain is usually assumed to be prefixed or set to π_0. Here, π_0 can be understood as a global transition distribution that ties the state-specific transition distributions π_j such that $\mathbf{E}[\pi_{j,k}|\pi_0] = \pi_{0,k}$, while the concentration parameter c controls the variability of the π_j around π_0. From (4.22) we may identify a potential limitation of this model: the probabilities of self-transitions are not distinguished from those of out of state transitions, which can lead to state processes that have unrealistically rapid switching between states and generation of redundant states (see Fox et al. [2011] for more detailed investigations). Fox et al. [2011] proposed a remedy for this issue by modifying the DP prior for the transition distributions as

$$\pi_j | \pi_0, c, \kappa \sim DP(c + \kappa, \frac{c\pi_0 + \kappa\delta_j}{c + \kappa}), \quad j = 1, \ldots, \infty, \tag{4.23}$$

where κ is a positive "sticky" parameter for augmenting the prior probability of self-transitions. In particular, (4.23) implies that the expected self-transition prob-

abilities are now given by

$$\mathbf{E}[\pi_{j,k}|\pi_0, \kappa, c] = \frac{c}{c+\kappa}\pi_{0,k} + \frac{\kappa}{c+\kappa}\mathbf{I}(k=j),$$

which is strictly greater than $\pi_{0,k}$ when $k = j$ for any positive values of κ. When $\kappa = 0$ this model reduced to the original HDP-HMM. The sticky model is, however, still inflexible to some extent due the sharing of the sticky parameter in (4.23) across states, which induces an undesirable coupling effect between the prior on state transition probabilities and that on self-persistence probability (see Zhou et al. [2021] for more detailed discussions and illustrations).

More recently, Zhou et al. [2021] propose a new prior for the transition matrix, termed as the "disentangled" sticky HDP (d-sHDP) prior, to simultaneously address the aforementioned limitations with the HDP and sticky HDP prior. The key modification lies in the replacement of the prior in (4.23) by

$$\begin{gathered}\kappa_j|\rho_1, \rho_2 \sim Be(\rho_1, \rho_2), \quad \overline{\pi}_j|\pi_0, c \sim DP(c, \pi_0), \\ \pi_j = \kappa_j\delta_j + (1-\kappa_j)\overline{\pi}_j, \quad j = 1, \ldots, \infty,\end{gathered} \tag{4.24}$$

where ρ_1, $\rho_2 > 0$ are hyperparameters. It is important to note that the sticky HDP prior in (4.23) is a special case of the d-sHDP prior in (4.24) with $(\rho_1, \rho_2) = (\kappa, c)$. This can be seen by reexpressing the DP in (4.23) using the DP decomposition property as in (4.6). Under this specification we can see that for sticky HDP prior parameter c appears both in the priors for $\overline{\pi}_j$ and in those for κ_j whereas for the d-sHDP prior the sticky parameter is modelled with two free parameters, implying the extra expressive power offered by the d-sHDP prior. By introducing additional latent indicator variables $\mathbf{w} = (w_1, \ldots, w_T)$, the resulting hidden state process defined under (4.24) can be equivalently specified as

$$\begin{gathered}w_t|\{\kappa_j\}_{j=1}^{\infty}, x_{t-1} \sim Ber(\kappa_{x_{t-1}}), \quad t = 1, \ldots, T, \\ x_t|w_t, \{\overline{\pi}_j\}_{j=1}^{\infty}, x_{t-1} \sim w_t\delta_{x_{t-1}} + (1-w_t)\overline{\pi}_{x_{t-1}}, \quad t = 2, \ldots, T,\end{gathered} \tag{4.25}$$

where the κ_j and $\overline{\pi}_j$ are defined as in (4.24). Clearly by integrating out the w_t we recover the d-sHDP prior in (4.24). This later formulation would facilitate the design of the associated posterior simulation algorithms.

Posterior inference for the HDP-HMM and its variants generally relies on sampling based methods, where we may distinguish three different MCMC sampling strategies. The first relies on the CRF representation of the HDP and conjugate priors for the emission model where the infinite dimensional transition matrix and

the emission parameters are integrated out, focusing explicitly on inferring the state sequence. The direct assignment samplers developed in e.g. Teh et al. [2006], Fox et al. [2011] and Zhou et al. [2021] all belong to this category. However, this method is relatively restrictive in its setting (i.e. require conjugacy) and also in the inference outputs (e.g. samples for the transition probabilities are not available). In addition, it can suffer from slow mixing as the states are sampled one-at-a-time. A more popular alternative is the degree-K weak limit sampler which relies on a finite approximation of the DP/HDP prior [Fox et al., 2011; Bauwens et al., 2017; Zhou et al., 2021]. With relatively large choices of K (in our context it refers to an upper bound of N) the truncated model is able to offer a reasonably good approximation and it converges to the DP/HDP prior as $K \to \infty$ [Ishwaran and Zarepour, 2002]. This approximation permits joint sampling of the state sequence via the FFBS and the resulting sampler usually has a much better mixing rate than the first approach. On the other hand, the need to pre-specify the degree K may pose a limitation as it directly controls the approximation error which is difficult to quantify in practice [Hjort et al., 2010]. The third method is the so-called beam sampler originally developed in Van Gael et al. [2008] for inference in HDP-HMMs, and was later extended in Song [2014], Dufays [2016] and Hou [2017] to the case of sticky HDP-HMM/infinite Markov-switching models but is not yet available for the d-sHDP HMM variant. It uses a slice sampling technique to sample from the exact posterior distribution, where the number of activated regimes (have at least one allocated observation) is stochastically and dynamically truncated to be finite at each MCMC iteration thanks to the introduction of the slice variables and the FFBS routine is available to jointly update the hidden states. Although it may have a relative slower mixing rate compared with the weak limit sampler, it is usually computationally much more efficient as the number of activated regimes during the course of MCMC simulation is usually much less than in the case where a relatively large truncation level K is used. Simple adaptation of this beam sampler for enhancing the mixing rate also exists [Dufays, 2016]. We refer to Mouchet et al. [2019], Song and Woźniak [2020] and references therein for related discussions on these different sampling methods.

4.4.2 Model formulation

We can now build a fully nonparametric HMM by an effective combination of the HDPM based emission model introduced in section 4.3 and the d-sHDP prior described in section 4.4.1, generalizing the state-of-the-art sticky HDP-HMM to offer greater modelling flexibility. The generative process for the resulting full model can

be described as

$$
\begin{gathered}
\pi_0|\sigma \sim GEM(\sigma), \\
\kappa_j|\rho_1,\rho_2 \sim Be(\rho_1,\rho_2), \quad \overline{\pi}_j = \{\overline{\pi}_{j,k}\}_{k=1}^{\infty}|\pi_0, c \sim DP(c,\pi_0), \quad j = 1,\ldots,\infty, \\
w_t|\{\kappa_j\}_{j=1}^{\infty}, x_{t-1} \sim Ber(\kappa_{x_{t-1}}), \quad t = 2,\ldots,T, \\
x_t|w_t, \{\overline{\pi}_j\}_{j=1}^{\infty}, x_{t-1} \sim w_t\delta_{x_{t-1}} + (1-w_t)\overline{\pi}_{x_{t-1}}, \quad t = 1,\ldots,T, \\
\beta|\gamma \sim GEM(\gamma), \\
\phi_k = \{\phi_{k,i}\}_{i=1}^{\infty}|\alpha,\beta \sim DP(\alpha,\beta), \quad k = 1,\ldots,\infty, \\
s_t|\{\phi_k\}_{k=1}^{\infty}, x_t \sim \phi_{x_t}, \quad t = 1,\ldots,T, \quad \{\theta_k\}_{k=1}^{\infty}|H_\lambda \sim H_\lambda, \\
y_t|\{\theta_k\}_{k=1}^{\infty}, s_t \sim f(y_t|\theta_{s_t}), \quad t = 1,\ldots,T.
\end{gathered}
\tag{4.26}
$$

For notational and computational convenience we assume that the hidden Markov chain starts at a dummy state $x_0 = 1$ and we prefix $w_1 = 0$. The first four lines in (4.26) specify the hidden state process via the d-sHDP as defined in (4.24) and (4.25) while the last four lines in the hierarchy define the HDPM-based emissions as in (4.19) (lines 3-6), except that here the number of states N is a-priori unbounded. As for γ and α, we further assign Gamma hyperpriors to the concentration parameters σ and c associated with the state process, i.e. $\sigma \sim Ga(a_\sigma, b_\sigma)$ and $c \sim Ga(a_c, b_c)$, and we assume that ρ_1 and ρ_2 are a-priori independent, each of which is assigned a vague Gamma prior $Ga(1,1)$.

4.4.3 Posterior inference

We now describe an asymptotically exact MCMC method that extends existing beam samplers for simulating from the joint posterior density of $(\rho_1, \rho_2, \sigma, c, \gamma, \beta, \alpha, \pi_0, \{\overline{\pi}_j\}, \{\kappa_j\}, \{\phi_j\}, \mathbf{x}, \mathbf{s}, \{\theta_j\})$, avoiding any finite truncated approximations to the DP/HDPs underlying the model. Our modelling structure facilitates a block Gibbs sampler that alternates between updating the parameters for the emission model $(\gamma, \beta, \alpha, \{\phi_j\}, \mathbf{s}, \{\theta_j\})$ and those for the state process $(\rho_1, \rho_2, \sigma, c, \pi_0, \{\overline{\pi}_j\}, \{\kappa_j\}, \mathbf{w}, \mathbf{x})$. The key computational challenge in this fully nonparametric scenario lies in the efficient simulation of the state sequence $\mathbf{x}$, given the potentially unbounded state space and the nonparametric nature of the emission distributions. Once a sample of $\mathbf{x}$ is obtained, the HDPM-based emission parameters can be updated in essentially the same way as the MCMC algorithms described in section 4.3, with N now being random but for each MCMC iteration N is determined by counting the number of activated regimes in the sample $\mathbf{x}$.

Here, we extend the idea of beam sampling to simulate parameters associated

with the state process conditional on the emission parameters. We augment the model in (4.26) by introducing auxiliary variables $\tilde{\mathbf{u}} = (\tilde{u}_1, \ldots, \tilde{u}_T)$ such that the conditional density of $\tilde{u}_t$ given the rest of parameters is

$$f(\tilde{u}_t|\{\overline{\pi}_j\}, x_t, x_{t-1}, w_t) = \begin{cases} \mathbf{I}(0 < \tilde{u}_t < 1) & w_t = 1 \\ \frac{\mathbf{I}(0<\tilde{u}_t<\overline{\pi}_{x_{t-1},x_t})}{\overline{\pi}_{x_{t-1},x_t}} & w_t = 0 \end{cases} \quad t = 1, \ldots, T. \tag{4.27}$$

Clearly the inclusion of $\tilde{\mathbf{u}}$ does not alter the marginal distribution over the other model parameters as by integrating out $\tilde{\mathbf{u}}$ we return to the original model. Importantly, conditional on $\tilde{\mathbf{u}}$ (and other model parameters) the number of state trajectories with positive probability is finite since

$$\begin{aligned} &f(x_t|\tilde{u}_t, x_{t-1}, \{\overline{\pi}_j\}, w_t) \propto f(x_t, \tilde{u}_t|x_{t-1}, \{\overline{\pi}_j\}, w_t) \\ &= (w_t\delta_{x_{t-1}}(x_t) + (1-w_t)\overline{\pi}_{x_{t-1},x_t})(w_t\mathbf{I}(0 < \tilde{u}_t < 1) + (1-w_t)\frac{\mathbf{I}(0 < \tilde{u}_t < \overline{\pi}_{x_{t-1},x_t})}{\overline{\pi}_{x_{t-1},x_t}}). \end{aligned} \tag{4.28}$$

When $w_t = 1$, the right-hand side of (4.28)$= \delta_{x_{t-1}}(x_t)\mathbf{I}(0 < \tilde{u}_t < 1)$, enforcing $x_t = x_{t-1}$. When $w_t = 0$, the right-hand side of (4.28)$= \mathbf{I}(0 < \tilde{u}_t < \overline{\pi}_{x_{t-1},x_t})$ where there are only finitely many x_t satisfying $\overline{\pi}_{x_{t-1},x_t} > \tilde{u}_t$ due to the constraint that $\sum_{k=1}^{\infty} \overline{\pi}_{x_{t-1},k} = 1$. Therefore the FFBS is applicable here for jointly updating the state sequence and moreover, only parameters that are associated with the currently active states (have at least one allocated observation) need to be updated at each MCMC iteration. Let N^* denote the number of currently active states, $\pi_0^{N^*} = (\pi_{0,1}, \ldots, \pi_{0,N^*}, \pi_0^*)$ and $\overline{\pi}_j^{N^*} = (\overline{\pi}_{j,1}, \ldots, \overline{\pi}_{j,N^*}, \overline{\pi}_j^*)$, where $\pi_0^* = 1 - \sum_{k=1}^{N^*} \pi_{0,k}$ and $\overline{\pi}_j^* = 1 - \sum_{k=1}^{N^*} \overline{\pi}_{j,k}$, $j = 1, \ldots, N^*$. Our proposed sampling steps for $(\rho_1, \rho_2, \sigma, c, \pi_0, \{\overline{\pi}_j\}, \{\kappa_j\}, \mathbf{x})$ (conditional on the state of the emission model) has the following structure:

- sample $\tilde{\mathbf{u}}|\{\overline{\pi}_j^{N^*}\}_{j=1}^{N^*}, \mathbf{x}$,
- sample $\mathbf{x}|\mathbf{s}, \tilde{\mathbf{u}}, \mathbf{w}, \{\overline{\pi}_j^{N^*}\}_{j=1}^{N^*}, \{\kappa_j\}_{j=1}^{N^*}, \{\phi_j\}_{j=1}^{N^*}$,
- sample $\mathbf{w}|\mathbf{x}, \{\overline{\pi}_j^{N^*}\}_{j=1}^{N^*}, \{\kappa_j\}_{j=1}^{N^*}$,
- sample $\{\kappa_j\}_{j=1}^{N^*}|\mathbf{x}, \mathbf{w}$,
- sample $\{m_{j,k}\}_{j,k=1}^{N^*}|\mathbf{x}, \mathbf{w}, c, \pi_0^{N^*}$,
- sample $\pi_0^{N^*}, \{\overline{\pi}_j^{N^*}\}_{j=1}^{N^*}|c, \sigma, \mathbf{w}, \{m_{j,k}\}_{j,k=1}^{N^*}$,
- sample $c|\{m_{j,k}\}_{j,k=1}^{N^*}, \mathbf{x}, \mathbf{w}$

- sample $\sigma|\{m_{j,k}\}_{j,k=1}^{N^*}$
- sample $\rho_i|\{\kappa_j\}_{j=1}^{N^*}, \quad i=1,2,$

where the $\{m_{j,k}\}$ are another set of auxiliary variables introduced to facilitate the sampling process that allow us to bypass the need to additionally invoke the CRF representation. More details regarding each of the sampling steps are provided as follows.

Step 1: sampling $\tilde{\mathbf{u}}$. For each $t=1,\ldots,T$ we sample $\tilde{u}_t$ according to (4.27).

Step 2: sampling $\mathbf{x}$. We first expand $\pi_0^{N^*}$, $\overline{\pi}_j^{N^*}$, $\{\phi_j\}_{j=1}^{N^*}$ and $\{\kappa_j\}_{j=1}^{N^*}$, by sampling the "unoccupied" parameters from their respective prior in a similar fashion to step 4 of Algorithm 4.1 until $\overline{\pi}_j^* < \min_{t:w_t=0}\tilde{u}_t$ for $\forall j=1,\ldots,N^*$:

$$
\begin{gathered}
N^* := N^*+1, \quad \epsilon \sim Be(1,\sigma), \\
\pi_{0,N^*} = \epsilon\pi_0^*, \quad \pi_0^* := (1-\epsilon)\pi_0^*, \quad \kappa_{N^*} \sim Be(\rho_1,\rho_2), \\
(\phi_{N^*,1},\ldots,\phi_{N^*,K^*},\phi_{N^*}^*) \sim Dir(\alpha\beta_1,\ldots,\alpha\beta_{K^*},\alpha(1-\sum_{i=1}^{K^*}\beta_i)) \\
\epsilon_j \sim Be(c\pi_{0,N^*},c\pi_0^*), \quad j=1,\ldots,N^*-1, \\
\overline{\pi}_{j,N^*} = \epsilon_j\overline{\pi}_j^*, \quad \overline{\pi}_j^* := (1-\epsilon_j)\overline{\pi}_j^*, \quad j=1,\ldots,N^*-1, \\
(\overline{\pi}_{N^*,1},\ldots,\overline{\pi}_{N^*,N^*}) \sim Dir(c\pi_{0,1},\ldots,c\pi_{0,N^*},c\pi_0^*).
\end{gathered}
$$

We proceed to jointly update $\mathbf{x}$ via a FFBS procedure. Omitting model parameters $\{\overline{\pi}_j\},\{\kappa_{\mathbf{j}}\}$ and $\{\phi_{\mathbf{j}}\}$ in the conditioning set for notational simplicity, we define

$$
\begin{aligned}
&\alpha_1(k) := f(x_1=k,s_1|\tilde{u}_1,w_1,x_0) \propto f(x_1=k,\tilde{u}_1|w_1,x_0)f(s_1|x_1=k) \\
&= w_1(\delta_{x_0}(k)\phi_{k,s_1}) + (1-w_1)\mathbf{I}(0<\tilde{u}_t<\overline{\pi}_{x_0,k})\phi_{k,s_1}), \quad k=1,\ldots,N^*,
\end{aligned}
$$

and for $t=2,\ldots,T$; $k=1,\ldots,N^*$,

$$
\begin{aligned}
&\alpha_t(k) := f(x_t=k,\{s_i\}_{i=1}^t|\{\tilde{u}_i\}_{i=1}^t,\{w_i\}_{i=1}^t) \\
&\propto \sum_{x_{t-1}} f(x_{t-1},x_t=k,\{s_i\}_{i=1}^t,\tilde{u}_t,w_t|\{\tilde{u}_i\}_{i=1}^{t-1},\{w_i\}_{i=1}^{t-1}) \\
&= \begin{cases} \phi_{k,s_t}\sum_{x_{t-1}}\alpha_{t-1}(x_{t-1})\kappa_{x_{t-1}}\delta_{x_{t-1}}(k) & \text{if } w_t=1 \\ \phi_{k,s_t}\sum_{x_{t-1}:\tilde{u}_t<\overline{\pi}_{x_{t-1},k}}\alpha_{t-1}(x_{t-1})(1-\kappa_{x_{t-1}}) & \text{if } w_t=0 \end{cases}
\end{aligned}
$$

Then we simulate $\mathbf{x}$ by first sampling x_T from

$$f(x_T = k|\{s_i\}_{i=1}^T, \{\tilde{u}_i\}_{i=1}^T, \{w_i\}_{i=1}^T) \propto \alpha_T(k), \quad k = 1, \ldots, N^*,$$

and then iteratively sample x_t, $t = T-1, \ldots, 1$, from

$$\begin{aligned} & f(x_t = k|x_{t+1}, \{s_i\}_{i=1}^T, \{\tilde{u}_i\}_{i=1}^T, \{w_i\}_{i=1}^T) \\ \propto\ & f(x_t = k, \{s_i\}_{i=1}^t|\{\tilde{u}_i\}_{i=1}^t, \{w_i\}_{i=1}^t) f(w_{t+1}|x_t = k) f(x_{t+1}, \tilde{u}_{t+1}|x_t = k, w_{t+1}) \\ =\ & w_{t+1}\alpha_t(k)\kappa_k\delta_k(x_{t+1}) + (1 - w_{t+1})\alpha_t(k)(1 - \kappa_k)\mathbf{I}(0 < \tilde{u}_{t+1} < \overline{\pi}_{k,x_{t+1}}) \end{aligned}$$

Note that here N^* provides an upper bound on the number of unique states that can be generated via the FFBS since it is guaranteed that $\overline{\pi}_{i,j} < \min_{t:w_t=0} \tilde{u}_t$ for $\forall i = 1, \ldots, N^*$ and $j > N^*$ (by construction $\max_{i=1,\ldots,N^*} \overline{\pi}_i^* < \min_{t:w_t=0} \tilde{u}_t$). After sampling $\mathbf{x}$, we update N^* as the number of active states in $\mathbf{x}$, and relabel and reconstruct $\mathbf{x}$, $\pi_0^{N^*}$, $\{\overline{\pi}_j^{N^*}\}_{j=1}^{N^*}$, $\{\phi_j\}_{j=1}^{N^*}$ and $\{\kappa_j\}_{j=1}^{N^*}$ accordingly where parameters associated with the non-active states are either discarded (in the case of $\{\phi_j\}_{j=1}^{N^*}$ and $\{\kappa_j\}_{j=1}^{N^*}$) or collapsed (in the case of $\pi_0^{N^*}$ and $\{\overline{\pi}_j^{N^*}\}_{j=1}^{N^*}$).

Step 3: sampling w. Since w_t is binary, we can directly sample w_t, $t = 2, \ldots, T$, from its full conditional distribution which is given by

$$\mathbf{Pr}(w_t|x_t, x_{t-1}, \tilde{u}_t) \propto f(w_t|x_{t-1}) f(x_t, \tilde{u}_t|w_t, x_{t-1}),$$

with

$$\begin{aligned} \mathbf{Pr}(w_t = 1|x_t, x_{t-1}, \tilde{u}_t) &\propto \kappa_{x_{t-1}}\delta_{x_{t-1}}(x_t), \\ \mathbf{Pr}(w_t = 0|x_t, x_{t-1}, \tilde{u}_t) &\propto (1 - \kappa_{x_{t-1}})\mathbf{I}(0 < \tilde{u}_t < \overline{\pi}_{x_{t-1},x_t}). \end{aligned}$$

Step 4: sampling $\{\kappa_j\}_{j=1}^{N^*}$. Using the Beta-Binomial conjugacy property, the full conditional distribution for κ_j is given by

$$\kappa_j|\mathbf{x}, \mathbf{w} \sim Be(\rho_1 + \sum_{i:x_{i-1}=j} w_i, \rho_2 + \sum_{i:x_{i-1}=j} (1 - w_i)), \quad j = 1, \ldots, N^*.$$

Step 5: sampling $\{m_{j,k}\}_{j,k=1,\ldots,N^*}$. We follow Fox et al. [2011] to sample these auxiliary variables (in the CRF context $m_{j,k}$ corresponds to the number of tables in restaurant j that serve dish k). For each pair $(j, k) \in \{1, \ldots, N^*\}^2$, initialize $m_{j,k} = 0$ and let $n_{j,k} = \sum_{t=1}^T \mathbf{I}(x_{t-1} = j, x_t = k, w_t = 0)$. Then for $i = 1, \ldots, n_{j,k}$: sample $m \sim Ber(c\pi_{0,k}/(i - 1 + c\pi_{0,k}))$, and update $m_{j,k} := m_{j,k} + 1$ if $m = 1$.

Step 6: sampling $\pi_0^{N^*}$ and $\{\overline{\pi}_j^{N^*}\}_{j=1}^{N^*}$. Conditional on the $m_{j,k}$, we can

sample $\pi_0^{N^*}$ and $\{\overline{\pi}_j^{N^*}\}_{j=1}^{N^*}$ following the theory of HDP [Teh et al., 2006; Fox et al., 2011] as

$$(\pi_{0,1}, \ldots, \pi_{0,N^*}, \pi_0^*) \sim Dir(m_{\cdot 1}, \ldots, m_{\cdot N^*}, \sigma),$$
$$(\overline{\pi}_{j,1}, \ldots, \overline{\pi}_{j,N^*}, \overline{\pi}_j^*) \sim Dir(c\pi_{0,1} + n_{j,1}, \ldots, c\pi_{0,N^*} + n_{j,N^*}, c\pi_0^*), \quad j = 1, \ldots, N^*.$$

Steps 7 and 8: sampling $\sigma, \mathbf{c}$. Sampling for the DP concentration parameters σ and c can be achieved by the same auxiliary sampling scheme used in steps 7 and 8 of Algorithm 4.1, where K^*, $\{m_{j,k}\}$ and $\{n_{i,j}^{'}\}$ are replaced by N^*, $\{m_{j,k}\}$ and $\{n_{i,j}\}$ defined above in step 5, respectively (with slight abuse of notation). We therefore omit the details here.

Step 9: sampling ρ_1, ρ_2. We employ a random walk Metropolis-Hasting algorithm to update each of the hyperparameters as their associated full conditional distributions do not have standard distributional forms. We first update ρ_1 via a log-normal random walk

$$\log \rho_1^{'} = \log \rho_1 + \epsilon_{\rho_1}, \quad \epsilon_{\rho_1} \sim \mathbf{N}(0, \lambda_{\rho_1}),$$

where λ_{ρ_1} is a tuning parameter adjusted to achieve a satisfactory sampling efficiency with an acceptance rate of around 0.5 [Gelman et al., 1997]. The candidate $\rho_1^{'}$ is accepted with probability

$$\min\Big(1, \frac{\prod_{i=1}^{N^*} f(\kappa_i|\rho_1^{'}, \rho_2) f(\rho_1^{'}) f(\rho_1|\rho_1^{'})}{\prod_{i=1}^{N^*} f(\kappa_i|\rho_1, \rho_2) f(\rho_1) f(\rho_1^{'}|\rho_1)}\Big),$$

where the proposal ratio $f(\rho_1|\rho_1^{'})/f(\rho_1^{'}|\rho_1) = \rho_1^{'}/\rho_1$. We update ρ_2 analogously, conditional on the updated value of ρ_1.

The whole MCMC algorithm completes by sampling parameters for the emission model conditional on $N = N^*$ and $\mathbf{x}$, using steps 3-7 of Algorithm 4.1. After the sampling process, posterior samples need to be post-processed to address the label switching issue, which can be achieved by using the relabelling algorithm introduced in section 4.3.3 conditioned on a fixed value of N.

4.4.4 Simulation study

We simulate a data set of length $T = 2000$ from a 3-state trivariate HMM. We purposely choose the parameters such that the emission densities exhibit multimodality and relatively complex within state dependence structures, where the overlaps between these densities are moderate-to-high. The emission distribution for the first

two states are specified as mixtures of two normal distributions

$$y_t|x_t = i \sim w_{i1}\mathbf{N}_3(\mu_{i,1}, \Sigma_{i,1}) + (1 - w_{i1})\mathbf{N}_3(\mu_{i,2}, \Sigma_{i,2}), \quad i = 1, 2,$$

with $w_{11} = 0.55$, $w_{21} = 0.35$, $\mu_{1,1} = \begin{pmatrix} -15 \\ 10 \\ -10 \end{pmatrix}$, $\mu_{1,2} = \begin{pmatrix} 10 \\ 40 \\ 15 \end{pmatrix}$, $\mu_{2,1} = \begin{pmatrix} -5 \\ 35 \\ 10 \end{pmatrix}$, $\mu_{2,2} = \begin{pmatrix} 20 \\ 20 \\ -15 \end{pmatrix}$, $\Sigma_{11} = \begin{pmatrix} 100 & 60 & -20 \\ 60 & 140 & 30 \\ -20 & 30 & 100 \end{pmatrix}$, $\Sigma_{12} = \begin{pmatrix} 100 & 30 & -30 \\ 30 & 110 & -30 \\ -30 & -30 & 70 \end{pmatrix}$, $\Sigma_{21} = \begin{pmatrix} 70 & -50 & 30 \\ -50 & 100 & -10 \\ 30 & -10 & 140 \end{pmatrix}$ and $\Sigma_{22} = \begin{pmatrix} 100 & -55 & 45 \\ -55 & 130 & 30 \\ 45 & 30 & 110 \end{pmatrix}$, and that for state 3 is a mixture of three normal distributions

$$y_t|x_t = 3 \sim \sum_{j=1}^{3} w_{3j}\mathbf{N}_3(\mu_{3,j}, \Sigma_{3,j}),$$

with $w_{31} = w_{32} = w_{33} = 1/3$, $\mu_{3,1} = \begin{pmatrix} -10 \\ -5 \\ 20 \end{pmatrix}$, $\mu_{3,2} = \begin{pmatrix} 30 \\ 30 \\ 15 \end{pmatrix}$, $\mu_{3,3} = \begin{pmatrix} 5 \\ -10 \\ 0 \end{pmatrix}$, $\Sigma_{3,1} = \begin{pmatrix} 60 & -5 & 20 \\ -5 & 40 & -10 \\ 20 & -10 & 130 \end{pmatrix}$, $\Sigma_{3,2} = \begin{pmatrix} 60 & 35 & -20 \\ 35 & 80 & -40 \\ -20 & -40 & 150 \end{pmatrix}$ and $\Sigma_{3,3} = \begin{pmatrix} 100 & -35 & 30 \\ -35 & 80 & -40 \\ 30 & -40 & 150 \end{pmatrix}$. The transition matrix of this HMM is specified as

$$\Pi = \begin{pmatrix} 0.75 & 0.15 & 0.1 \\ 0.075 & 0.85 & 0.075 \\ 0.025 & 0.25 & 0.95 \end{pmatrix},$$

where states exhibit different levels of self-persistence and transition patterns.

To implement the proposed MCMC algorithm, we place a slightly informative $Ga(0.5, 1)$ prior on γ and σ to discourage the generation of overly complex models (i.e. too many states and large number of mixture components for the emission distributions) and we use a $Ga(10, 2)$ prior for α and c which encourages some sort of "similarity" across states. The Gaussian and inverse Wishart prior for the emission model are chosen empirically as described in section 4.3.4. Our results are based on 30k iterations of the algorithm, with the first 10k samples discarded as burn in. Figure 4.3 displays some diagnostic plots that are helpful for assessing the convergence of the sampler. The chain seems to reach stationarity within the first 5k iterations and no apparent convergence issue is detected. The true model with $N = 3$ is identified as the posterior mode with a posterior probability of 0.767, followed by $N = 4$ with a posterior probability of 0.206. Conditional on the number of states estimated by the posterior mode, the transition probabilities are reconstructed as

$$\pi_j^3 = (\pi_{j,1}, \pi_{j,2}, \pi_{j,3}, \pi_j^*) = \kappa_j\delta_j + (1 - \kappa_j)\overline{\pi}_j^3, \quad j = 1, 2, 3,$$

and the posterior means are estimated as $\hat{\pi}_1^3 = (0.778, 0.131, 0.09, 0.001)$, $\hat{\pi}_2^3 = (0.055, 0.885, 0.06, 0)$ and $\hat{\pi}_3^3 = (0.022, 0.016, 0.962, 0)$. Note that the last entry of each transition probability vector, the residual probability π_j^*, stands for the probability of transiting to a new unseen state from state j. Using the proposed local decoding algorithm, the estimated state sequence matches well with the ground truth, with a normalized Hamming distance of 0.026, which is equivalent to a very satisfying decoding accuracy of 97.4%. Figure 4.4 illustrates predictive density estimates (in terms of 2-dimensional marginals) for a specific state (state 3) along with their empirical counterparts, where we observe a reasonably good agreement, indicating a good fit of the nonparametric model.

4.5 Sleep analysis using acceleration and heart rate data from the Apple Watch

Activity and heart rate are among the most commonly used physiological signals for sleep monitoring and evaluation in a free-living condition due to the popularity of multi-sensory wearable devices and their informativeness regarding certain sleep stages (see Imtiaz [2021] and references therein). However, existing algorithms for these data that achieve state-of-the-art classification accuracy for sleep detection and staging, such as neural networks and gradient boosting decision trees (see e.g. Walch et al. [2019] and Roberts et al. [2020]), requires extensive tuning and supervised training with the the polysomnography (PSG) which is very costly to collect and label. Therefore the applicability of these supervised methods can be very limited. In addition, the generalizability of these supervised algorithms can be of concern as different cohorts can exhibit very different sleep and physiological patterns [Liu et al., 2020]. Here we find that even within the same cohort, the inter-subject variability can be very large. Self supervised methods that do not require labeled PSG data have also been proposed recently for sleep recognition with promising performance on real data [Zhao et al., 2020], however, they require a rather careful set-up and tuning of an upstream pre-training model, whose performance can have a big impact on the downstream classification algorithms. Here we investigate the use of Bayesian nonparametric HMMs for inferring sleep structure in an individualized and unsupervised manner without pre-training. HMMs are naturally capable of capturing the temporal dependency as well as dynamic patterns in these physiological signals, and are superior to alternative clustering-based unsupervised methods which ignore temporal structure in the data (see e.g. Lüdtke et al. [2021]). To our knowledge, (parametric) HMMs have been applied for analyzing biological signals

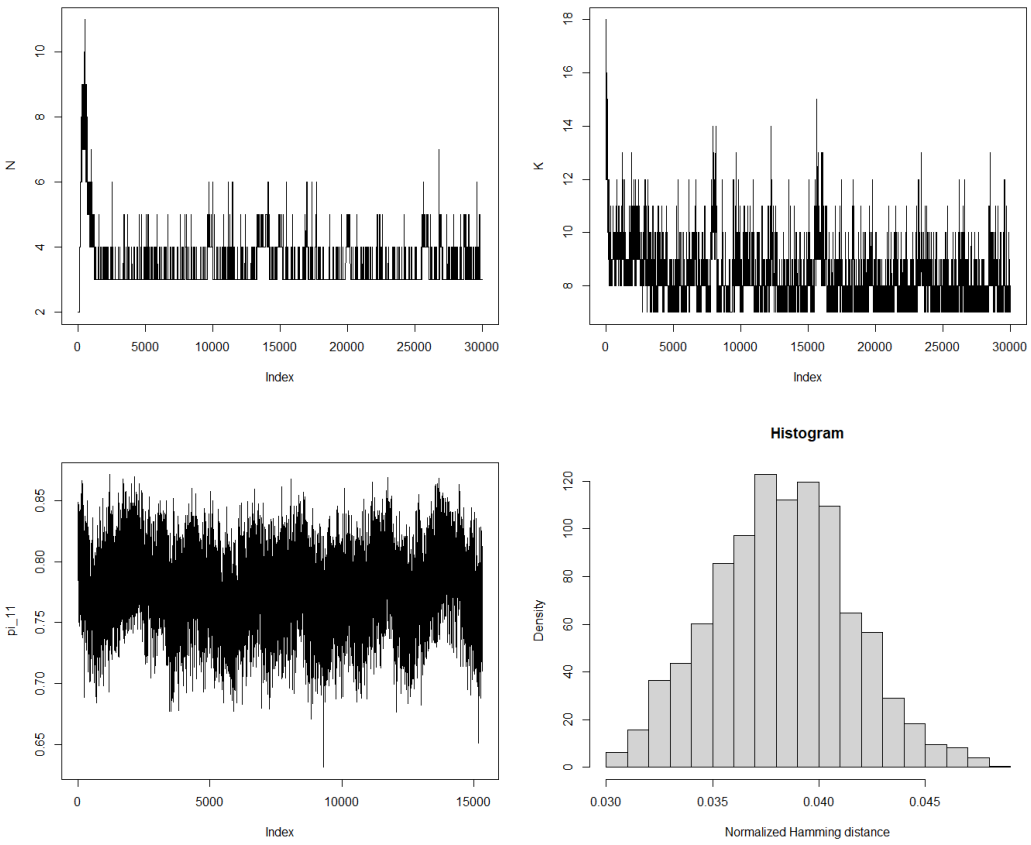

Figure 4.3: Convergence diagnostics for simulation model in section 4.4.4. Top panel shows the trace plots for the the number of the active states (left) and mixture components for the emission densities (right); bottom left and bottom right panels show the trace plot for the transition probability $\pi_{1,1}$ and the histogram of the normalized Hamming distance between the simulated and the true state sequence, respectively, both conditional on the modal number of states $N = 3$.

collected from the PSG [Pan et al., 2012; Langrock et al., 2013] and for actigraphy based sleep wake classification under a supervised learning framework [Lüdtke et al., 2021], but have not been investigated in the context of unsupervised sleep staging.

4.5.1 Data description

We considered the Apple Watch data set from Walch et al. [2019], which is openly available in Walch [2019] from the PhysioNet platform [Goldberger et al., 2000]. It contains raw acceleration (in units of g, i.e. $9.8m/s^2$, measured by a triaxial

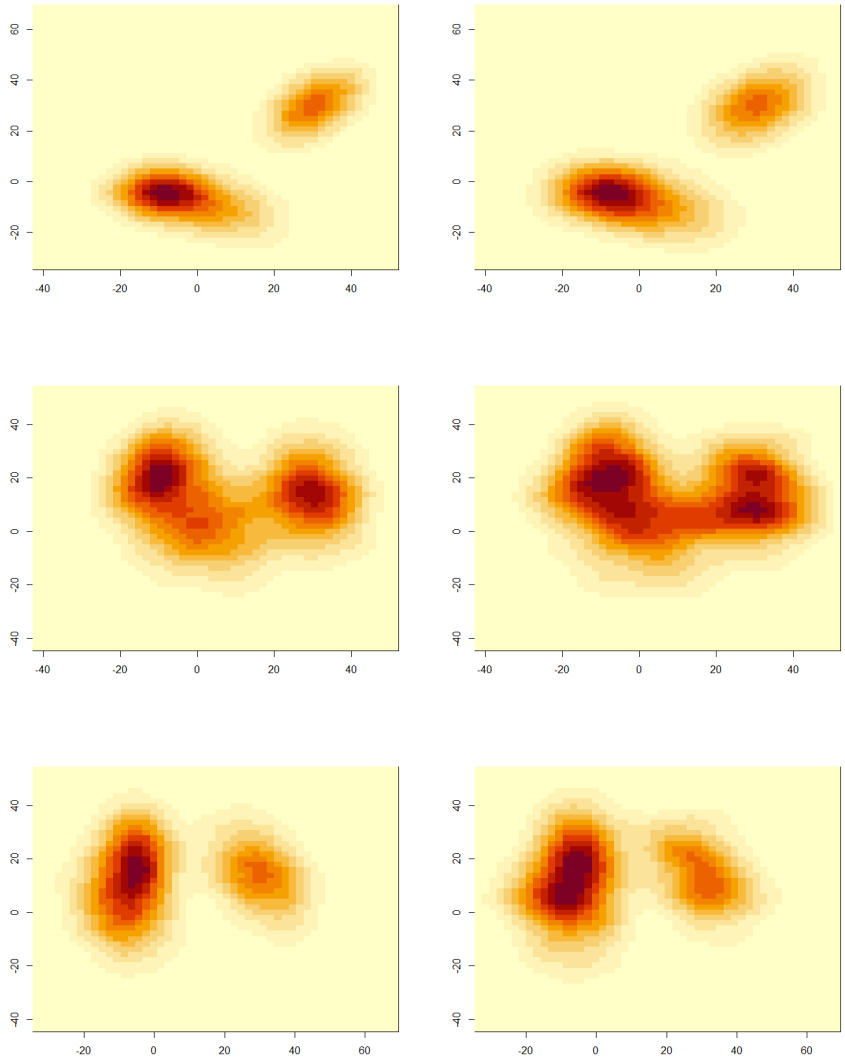

Figure 4.4: Estimate of the posterior predictive density and the simulated data. Left panel: marginal contour plots obtained based on 20k samples of y_{new} for state 3; right panel: corresponding marginal contour plots obtained based on the simulated data allocated to state 3 (ground truth). All contour plots were obtained using the R function *kde2d* with default bandwidths, where darker shades representing higher density regions.

accelerometer at about 50 HZ) and heart rate (beats per minute, measured by photoplethysmography at every several seconds) recorded from the Apple Watch, as well as labeled sleep stages scored from the co-recorded PSG for one night for a total of 31 healthy subjects free of sleep disorders. We refer to Walch et al. [2019]

for more details regarding the data set. In our analysis we exclude 11 subjects from the original cohort due to incomplete data. For the remaining 20 subjects, we preprocess the acceleration and heart rate data as follows. We convert the raw 3-dimensional acceleration signal at each time point to its Euclidean norm to obtain a summary activity metric [Roberts et al., 2020], and the resulting 1-dimensional signal is further averaged over 30s non-overlapping intervals. The raw heart rate data is also averaged over 30s windows. We choose this specific resolution since it agrees with that of sleep staging from the PSG data, and moreover, the data quality from the Apple Watch at this resolution is established in Roberts et al. [2020] in that both signals are strongly correlated with data from the reference devices. Figure 4.5 shows the transformed acceleration (AC) and heart rate (HR) data for three example subjects, with PSG-derived sleep stages indicated in colors.

4.5.2 Sleep modelling with fully nonparametric HMMs

We first analyze the bivariate AC and HR data for each subject using our fully nonparametric HMM, where we examine the ability of the HMM in retrieving the true underlying sleep stages determined from the PSG. We notice that the shifts in the mean levels and trends in the AC and HR signals do not appear to have clear associations with the sleep stages, and they can cause the algorithm to generate an excessive number of states that were not interpretable. We hence propose to "stationarize" both AC and HR signals first before fitting the model by applying a log transformation, followed by first differencing. The transformed data thus approximate the percentage changes or growth rates in the original data at each time point (as $\log(x/y) \approx x/y - 1$ for x/y close to 1). Unless specified otherwise, we shall report results for the transformation of the AC and HR. In this analysis we place a $Gamma(0.3, 1)$ prior for γ and σ and a $Gamma(10, 2)$ prior for α and c. Hyperparameters of the base measures for the emission model are chosen empirically as before. Our results are based on 30k iterations of the MCMC sampler, 15k of which are discarded as burn-in. Postprocessing of the posterior samples for tackling the label switching issue is achieved as described in section 4.3.3 using the R package *Label.switching*.

Figure 4.5 displays the inferred hidden state sequences (piecewise horizontal line) for our example subjects in the cohort, obtained via local decoding conditional on the posterior modal number of states. We can see from the upper and middle panel that the rapid eye movement (REM) sleep stage is linked to 1 or 2 states of the fitted models which are characterized by relatively high volatility in the signals, especially HR, with a lower level of persistency, whereas the Non-REM sleep stages,

which include sleep stages N1, N2 and N3, can be connected to one state with a higher state persistency. Indeed, the REM sleep stage is usually associated with fluctuating cardiovascular activity, often resulting in an increase in heart rate with a high variability [Boe et al., 2019]. The wake epochs are generally characterized by a large variation in both signals and can be linked to the states estimated with the most dispersed emission distributions. However, individual sleep stages N1, N2 and N3 were not identified by the current model base on these two signals. This is not surprising as by visual inspection we cannot identify systematic patterns in AC and HR during these sleep stages and the empirical distributions conditional on these sleep stages are highly overlapped. Even in a supervised learning framework, they have been noted to be generally very challenging or impossible to identify, depending on the sensor modality, the cohort and also the classification algorithms used [Imtiaz, 2021].

It should be pointed out that high inter-subject variability and complex patterns in terms of physiological changes (in AC and HR) in different sleep stages are observed within the cohort. In particular, the REM and Non-REM sleep may not be distinguishable on the basis of AC and HR for some subjects. In addition, there are also shortage and imbalance of sleep samples in certain sleep stages (i.e. a subject may only spent a little proportion of their sleep in some sleep stages such as N1 or N3), which can further complicate the inference especially when an HMM is learned for every individual with only one night's of data, as is the case here. Sleep stage N2 is usually much more prevalent than any of the other sleep stages, and certain sleep stages such as N3 or REM may not even appear during a sleep bout. In these scenarios the inferred HMM states can be difficult to interpret and the recognition power regarding the sleep stages can be very low (see bottom panel of Figure 4.5 for an illustration). In fact, these complications pose challenges to both existing supervised and unsupervised methods in that a model can work "well" for sleep staging some subjects, yet perform poorly for others and it will be important to understand if the degree of predictability may be linked to covariates such as age, gender, etc. Therefore there remains scope for further improvement on both data and modelling sides. Regarding the latter, it may be fruitful to consider a longitudinal extension to the current modelling approach which would allow sharing of information across subjects. Nonparametric HMMs combined with semi or self-supervised learning techniques that require little or no labelled sensor data may also offer potential in improving the accuracy of sleep staging.

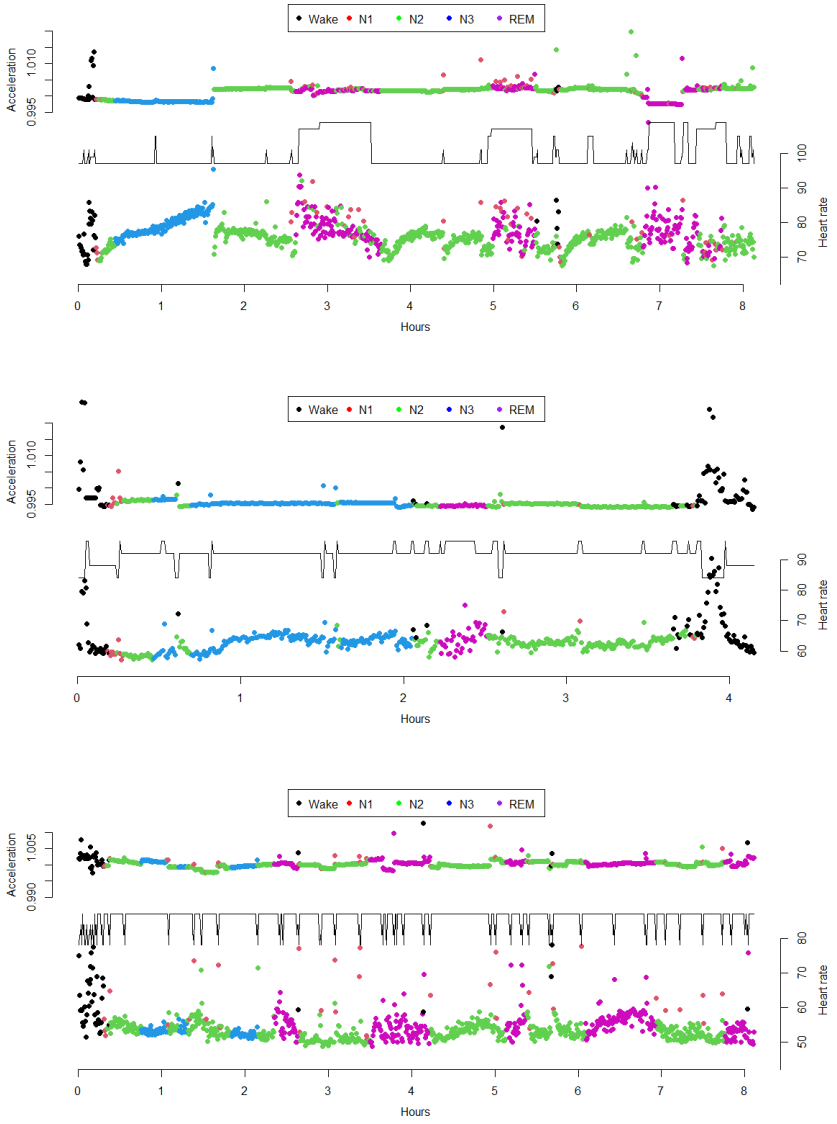

Figure 4.5: Results for example subjects ID 1818417 (top panel), 4018081 (middle panel) and 3997827 (bottom panel). For each subject the dots represent the 30s averaged AC and HR data over the monitoring period of 1 night, where color indicates the PSG-derived sleep stage at each 30s epoch. The piecewise horizontal line denotes the state sequence estimated from the fully nonparametric model, conditional on posterior modal number of states (5,7 and 4, respectively).

4.5.3 Classification of circadian sleep-wake cycle

Results from our fully nonparametric analysis (some are not shown here) indicate that under the unsupervised framework, wake is generally much more distinguish-

able than any of the other sleep stages, motivating us to investigate further the performance of our proposed HMM for the particular task of 2-state sleep/wake identification (within the sleep bout). Although this is the lowest resolution of sleep staging, it is of real interest as the classification result is required for the computation of some key metrics on sleep quality such as sleep efficiency and sleep latency [Nakazaki et al., 2014]. Here, we consider and compare three different models with a fixed number of states, namely bivariate HMMs using both AC and HR, with either 2 or 3 states and a 2-state univariate HMM based on AC only. We also investigated other possible HMM configurations but they tended to be inferior to the models considered in terms of the classification power and model interpretability. In this analysis we place a $Gamma(1,1)$ prior for γ and a $Gamma(1,1)$ prior for α and hyperparameters are chosen empirically as before. The inferred state sequence is obtained based on $20k$ iterations of the MCMC sampler described in section 4.3, with the first $5k$ discarded as burn-in. For all models the state with the most dispersed emission distribution is assigned to represent the wake state while the other state(s) is interpreted as sleep. We summarize the results for the cohort of 20 subjects in Figure 4.6, where we compare the classification performance of the three candidate models in terms of three commonly used performance metrics: overall accuracy, sensitivity for sleep (proportion of true sleep epochs identified correctly) and specificity for wake (proportion of true wake epochs identified correctly). It is interesting to note that none of the models was uniformly best regarding the chosen performance metrics. In general, the 3-state bivariate model achieves the highest overall accuracy and sensitivity but suffers from very low specificity, while the 2-state bivariate model has the lowest overall accuracy and sensitivity but achieves the highest specificity. The performance of the univariate AC-based model with 2 states lies in-between, indicating that the sleep-wake cycle may well be identified from AC alone. While in principle heart rate does provide additional useful information on sleep/wake, its contribution towards sleep/wake classification appeared to be subject-specific was found to be very subtle under the present modelling framework, which is in agreement with the findings from Walch et al. [2019]. We also note that our novel solution of using nonparametric HMMs for sleep/wake identification achieves the state-of-the-art on this specific data set in comparison to alternative unsupervised classification methods (see e.g. Ramnath and Katkoori [2020]). Moreover, even comparing to the best performed supervised, i.e. training with sleep stages from PSG, neural net classifier in Walch et al. [2019] (accuracy= 90%, sensitivity= 93%, specificity= 59.6%), our proposed models only perform slightly worse. Our relatively low specificity and its relatively high variability across subjects is understandable

as there are wake epochs with little/no movement and therefore less differentiable from sleep (e.g. occurred when the subject is trying to fall into sleep). Also, the occurrence of wake during sleep is generally low (on average only accounts for about 10% of the data). We expect that our flexible nonparametric HMM may better fulfill its potential in situations where the cohort has sleep disorders (more wake epochs during sleep), and it would be interesting to study if a clearer advantage of including HR to the model can be obtained in this more challenging scenario.

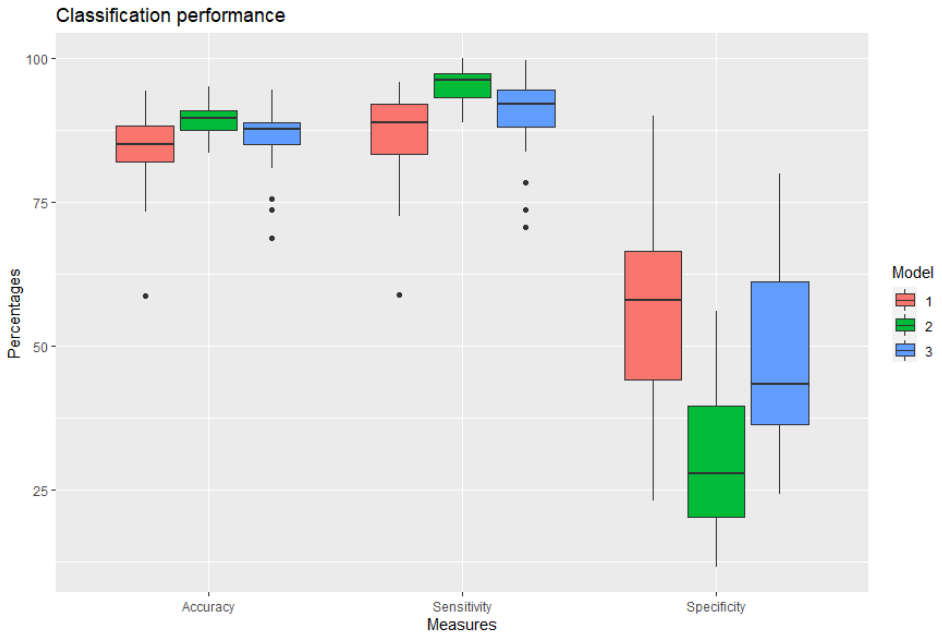

Figure 4.6: Classification performance of Models 1 (bivariate 2-state HMM), 2 (bivariate 3-state HMM) and 3 (2-state HMM based on AC only) in terms of overall accuracy, sensitivity for sleep and specificity for wake.

4.6 Discussion

In this chapter, we explore the use of Bayesian nonparametric techniques, in particular the hierarchical Dirichlet processes, as building blocks for constructing nonparametric Bayesian HMMs in a multivariate setting which generalize existing Bayesian nonparametric HMMs to offer extra flexibility. We first investigate the use of HDP mixture models for nonparametrically modelling the emission distributions in a multivariate HMM with fixed number of states, and we make use of the slice sampling technique to develop an efficient MCMC methodology for asymptotically exact pos-

terior inference. We then extend the HDPM-based HMM to allow for automatic learning of the number of states by specifying the hidden state process via the disentangled sticky HDP, and we develop an exact and computationally accessible MCMC method for inference in the resulting model via an extension of the beam sampling technique. The performance of the proposed algorithms are illustrated via two different simulation studies and we apply our proposed models to motion and heart rate data collected from Apple watch for learning human sleep dynamics in an unsupervised context.

It should be pointed out that despite the success and popularity of using HDP or its variants for specifying the HMM transition matrix, there is still a lack of theoretical guarantees (e.g. posterior consistency) with regard to the posterior on number of states implied by such models. When the data is generated from a finite mixture model, the mixture of finite mixture (MFM) model (i.e. a finite mixture models with a prior on the number of components), which can be regarded as a variable-dimension counterpart of the DPMM, is known to be consistent for number of component and the mixing distribution [Nobile, 1994]. Importantly, as revealed by Miller [2014]; Miller and Harrison [2018], there are interesting connections between the properties of the DPMM and MFM that allows efficient sampling-based inference techniques developed for the former to be adapted for the latter, for which posterior inference is conventionally achieved by RJMCMC method [Richardson and Green, 1997]. Therefore an interesting direction of further investigation is to consider extending the the use of the variable-dimension counterpart of the HDP for modelling the HMM transition matrix, where efficient sampling algorithms and theoretical insights may be obtained by extending existing results.

Chapter 5

Summary and outlook

In this thesis, we explored the use of splines and the DP/HDP, both nonparametric modelling techniques that enjoy tremendous success in applied statistical modelling and machine learning, as building blocks to construct nonparametric HMMs in both univariate and multivariate settings under a Bayesian modelling framework. The resulting models generalize existing nonparametric methods for HMMs, permitting greater modelling flexibility for complex real data while largely retaining the interpretability of the conventional HMMs. We developed novel and computationally feasible MCMC-based methods for learning and inference in such models, and we illustrate their great potential to modelling physiological data from animal movement and digital health.

In chapter 2, we introduced the first main contribution of the thesis: developing and investigating the first Bayesian methodology for inference in spline-based HMMs, where the emission distributions are modelled via Bayesian free-knot splines. We advocate the use of B-splines due to their nice mathematical properties, the fact that they can be effectively incorporated into the HMM framework as they retain an attractively simple model formulation and the computational efficiency of the standard HMM machinery. We introduced use of a trans-dimensional Markov chain inference algorithm to jointly infer the HMM parameters including the knot configuration of the B-splines, conditional on the number of states, N. Model selection regarding the cardinality N can be performed based on the marginal likelihood, which can be estimated using a truncated harmonic mean estimator under a parallel sampling framework. Using an extensive simulation study, we demonstrated the significant advantages of our proposed approach in terms of estimation accuracy, parsimoniousness and stability in comparison to alternative spline-based methods, namely the frequentist P-spline-based approach of Langrock et al. [2015]

and a Bayesian P-spline approach which is investigated for the first time in this thesis. Importantly, a more stable and parsimonious estimation of the nonparametric model allows the applier to perform model selection to compare model performance across a different and increasing number of states. The latter has been difficult to address in the past due to convergence problems and in some relevant previous work has been essentially avoided by pre-selecting the number of states. Our method also compares favourably over the Bayesian nonparametric model developed in Yau et al. [2011] as we are able to model more general emission distributions and address the model selection problem. We showed how our methodology may be used in an explorative way in searching for suitable parametric models in modelling animal movement data

In chapter 3, we built on and extended the spline-based modelling framework proposed in chapter 2 to develop a hierarchical conditional hidden Markov modelling approach which allows us to analyse the dynamics within a specific state(s) of a main-HMM at a finer level with another hidden Markov process(es), referred to as the sub-HMM. In this way we were able to achieve inferences that are otherwise not possible with a single HMM. We developed a fully Bayesian framework for jointly learning the main and sub-HMMs. Regarding the former the MCMC method developed in chapter 2 can be directly used. For inference in the sub-HMM, we modify the algorithm in chapter 2 by introducing the key notion of conditional likelihood, through which the specific state of the main-HMM is conditioned and, moreover, it can be efficiently computed thanks to the availability of a forward algorithm. We demonstrated the potential usefulness of the proposed method by analyzing activity data for a cohort of 44 subjects from the MESA data set. Our flexible hierarchical framework enables us to retrospectively analyze the time-varying features of a person's sleep–wake cycle and quantify the sleep periods in a coherent and systematic way. The sub-HMM further allows us to systematically characterize an individual's stochastic dynamic behaviour of transitions between, and sojourn times within, substates that may be associated with deeper and lighter or interrupted sleep stages. To our knowledge this is the first probabilistic modelling framework which may be applied to jointly identify and characterise sleep periods on an individual basis.

In chapter 4, we exploited the strengths of the HDP and a suitable integration with HMMs to develop new Bayesian nonparametic HMMs that generalize existing models to offer greater modelling flexibility. We first investigate the use of HDP-based mixture models for flexible yet parsimonious modelling of the emission distributions in a multivariate HMM with finite state space. The infinite dimensionality of the resulting parameter space was tackled by developing a novel MCMC

method which combines the slice sampling technique for efficient and exact sampling from the HDP mixture model, and a dynamic programming algorithm for HMMs for joint simulation of the hidden states. We then relaxed the assumption of fixed cardinality N to allow it to be automatically adapted to the sophistication of the data as there are scenarios when this value is unavailable a-priori. To this end, we make use of a disentangled sticky HDP prior to specify the transition matrix nonparametrically, leading to a fully nonparametric HMM that generalize existing HDP-based HMMs in the literature. An asymptotically exact MCMC method was developed for the resulting model via an extension of the beam sampling technique and its feasibility was assessed via a simulation study. Finally, we illustrated the use of the proposed method for joint analyzing motion and heart rate data collected from Apple watch for unsupervised learning of sleep macrostructure.

It should be noted that the spline-based HMM developed in chapter 2 and the HDP-based HMMs developed in this chapter have their own merits and limitations. The former enjoys a relatively simpler modelling framework (e.g. a finite parameter space) and is naturally capable of modelling complex univariate emissions with varying degree of smoothness over the domain. Standard HMM inference algorithms such as the forward algorithm are directly applicable in this context which makes it convenient to perform various inference tasks and permit an extension as shown in chapter 3. However, generalization of the spline-based HMM for multivariate observations is a challenging task without significant simplifying assumptions, and the resulting algorithm can be computationally prohibitive. In the bivariate scenario we expect the Bayesian P-spline-based HMM (the univariate case was explored in chapter 2) that uses tensor product of univariate B-splines (with pre-fixed knot configuration) and spatial smoothness priors to be a potentially feasible solution. On the other hand, the latter nonparametric HMMs are suitable for modelling multivariate data and no adaptation of the algorithm is required when the dimensionality of the data changes. For the fully nonparametric version of the model the ability to generate new states that accommodate for previously unseen patterns in the data may also be an advantage in certain applied problems. However, the relatively complex model structure may result in a loss of interpretability of the model parameters as compared to the conventional HMMs, and furthermore, convergence diagnostics can be more difficult to perform.

The works developed here suggest a number of future research directions that would be interesting to explore. For instance, the Bayesian methodology for spline-based HMM developed in chapter 2 can be extended in a relatively straightforward manner to Markov switching generalized additive models as studied in Lan-

grock et al. [2017, 2018] using frequentist approaches, where the splines would be used for modelling the functional effects of the covariates instead of the emissions. Note that in this context we no longer need to work with standardized spline basis functions, which would simplify the design of the RJMCMC algorithm and the efficiency/mixing of the resulting algorithm may be further improved (the knot insertion rule of De Boor (2001) would become exact instead of approximate, see chapter 2). Such an extension would contribute to the literature by providing the first Bayesian treatment of such spline-based Markov switching models. We believe that the advantages of using a Bayesian approach over a frequentist penalized approach as observed in chapter 2 would be extended to the case here. A longitudinal extension of the HDP-based Bayesian nonparametric HMMs developed in chapter 4 for jointly analysing multiple multivariate physiological data sets is another interesting future work. While parametric HMMs with continuous/discrete random effects are commonly used [Altman, 2007; de Chaumaray et al., 2020], their representation power is limited and significant computational burden arises in the likelihood evaluation or model selection. It may be useful to consider adopting a DP mixture of HMM framework, extending earlier attempts (e.g. Qi et al. [2007]) to characterise the heterogeneous behaviour of transition patterns across subjects, in combination with a HDP prior for flexibly modelling the emission distributions which are shared globally. The use of Bayesian free knot spline or the Bayesian penalised spline technique may also be investigated for incorporating subject-level covariates into the model. Another future work which I would like to consider is to develop a flexible hidden semi-Markov modelling framework (HSMM) that generalizes existing parametric HSMMs [Economou et al., 2014; Hadj-Amar et al., 2020b] to allow for nonparametric emissions and automatic learning of the state complexity, while being more flexible and computationally efficient than the existing nonparametric version of the HSMM (HDP-HSMM) [Johnson and Willsky, 2013]. To this end it may be fruitful to consider a reformulation of the HSMM as HMMs with extended state space following earlier work of Langrock and Zucchini [2011], and to explore an appropriate use of HDP priors (or their variants) to model the state transition and emission distributions nonparametrically.

Bibliography

Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, and Tanzeem Choudhury. Circadian computing: sensing, modeling, and maintaining biological rhythms. In *Mobile health*, pages 35–58. Springer, 2017.

Luigi Acerbi, Kalpana Dokka, Dora E Angelaki, and Wei Ji Ma. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLoS computational biology*, 14(7):e1006110, 2018.

Timo Adam, Christopher A Griffiths, Vianey Leos-Barajas, Emily N Meese, Christopher G Lowe, Paul G Blackwell, David Righton, and Roland Langrock. Joint modelling of multi-scale animal movement data using hierarchical hidden Markov models. *Methods in Ecology and Evolution*, 10(9):1536–1550, 2019a.

Timo Adam, Roland Langrock, and Christian H Weiß. Penalized estimation of flexible hidden Markov models for time series of counts. *Metron*, 77(2):87–104, 2019b.

Jung Ae Lee and Jeff Gill. Missing value imputation for physical activity data measured by accelerometer. *Statistical methods in medical research*, 27(2):490–506, 2018.

Grigory Alexandrovich, Hajo Holzmann, and Anna Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434, 2016.

Rachel MacKay Altman. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210, 2007.

Rachel MacKay Altman and A John Petkau. Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine*, 24(15):2335–2344, 2005.

Sonia Ancoli-Israel, Roger Cole, Cathy Alessi, Mark Chambers, William Moorcroft, and Charles P Pollak. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 26(3):342–392, 2003.

Sonia Ancoli-Israel, Jennifer L Martin, Terri Blackwell, Luis Buenaver, Lianqi Liu, Lisa J Meltzer, Avi Sadeh, Adam P Spira, and Daniel J Taylor. The sbsm guide to actigraphy monitoring: clinical and research applications. *Behavioral sleep medicine*, 13(sup1):S4–S38, 2015.

David Ardia, Nalan Baştürk, Lennart Hoogerheide, and Herman K Van Dijk. A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics & Data Analysis*, 56(11):3398–3414, 2012.

Yves Atchade, Gersende Fort, Eric Moulines, and Pierre Priouret. Adaptive Markov chain Monte Carlo: theory and methods. *Bayesian time series models*, 1, 2011.

Yves F Atchadé. An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodology and Computing in applied Probability*, 8(2): 235–254, 2006.

David Barber, A Taylan Cemgil, and Silvia Chiappa. *Bayesian time series models*. Cambridge University Press, 2011.

Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni. *Latent Markov models for longitudinal data*. Chapman and Hall/CRC, 2019.

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

Luc Bauwens, Jean-François Carpantier, and Arnaud Dufays. Autoregressive moving average infinite hidden Markov-switching models. *Journal of Business & Economic Statistics*, 35(2):162–182, 2017.

Matthew J Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. The infinite hidden Markov model. *Advances in neural information processing systems*, 1: 577–584, 2002.

Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom), 2003.

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V Vaughn, et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176:2012, 2012.

Peter J Bickel, Ya'acov Ritov, and Tobias Ryden. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

Clemens Biller. Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics*, 9(1):122–140, 2000.

David Blackwell, James B MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

Alexander J Boe, Lori L McGee Koch, Megan K O'Brien, Nicholas Shawen, John A Rogers, Richard L Lieber, Kathryn J Reid, Phyllis C Zee, and Arun Jayaraman. Automating sleep stage classification using wireless, wearable sensors. *NPJ digital medicine*, 2(1):1–9, 2019.

Richard J Boys and Daniel A Henderson. A comparison of reversible jump MCMC algorithms for dna sequence segmentation using hidden Markov models. *Comp. Sci. and Statist*, 33:35–49, 2001.

Vincent Bremhorst and Philippe Lambert. Flexible estimation in cure survival models using Bayesian p-splines. *Computational Statistics & Data Analysis*, 93: 270–284, 2016.

M Bres. The behaviour of sharks. *Reviews in Fish Biology and Fisheries*, 3(2): 133–159, 1993.

Andreas Brezger and Stefan Lang. Generalized structured additive regression based on Bayesian p-splines. *Computational Statistics & Data Analysis*, 50(4):967–991, 2006.

Stephen P Brooks, Paolo Giudici, and Gareth O Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–39, 2003.

Jodie Buckby, Ting Wang, Jiancang Zhuang, and Kazushige Obara. Model checking for hidden Markov models. *Journal of Computational and Graphical Statistics*, 29(4):859–874, 2020.

Jan Bulla and Andreas Berzel. Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics*, 23(1):1–18, 2008.

Antonio Canale, Pierpaolo De Blasi, et al. Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404, 2017.

Olivier Cappé, Christian P Robert, and Tobias Rydén. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3): 679–700, 2003.

Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. New York: Springer, 2005.

Bradley P Carlin and Siddhartha Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484, 1995.

Gilles Celeux and Jean-Baptiste Durand. Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564, 2008.

Gilles Celeux, Florence Forbes, Christian P Robert, D Mike Titterington, et al. Deviance information criteria for missing data models. *Bayesian analysis*, 1(4): 651–673, 2006.

Joshua CC Chan and Angelia L Grant. Fast computation of the deviance information criterion for latent variable models. *Computational Statistics & Data Analysis*, 100:847–859, 2016.

Jason Chang and John W Fisher III. Parallel sampling of dp mixture models using sub-clusters splits. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pages 620–628, 2013.

Jason Chang and John W Fisher III. Parallel sampling of hdps using sub-cluster splits. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pages 235–243, 2014.

Cathy WS Chen, Richard H Gerlach, and Ann MH Lin. Multi-regime nonlinear capital asset pricing models. *Quantitative Finance*, 11(9):1421–1438, 2011.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6):877–888, 2015.

Yang Chen, Kuang Shen, Shu-Ou Shan, and SC Kou. Analyzing single-molecule protein transportation experiments via hierarchical hidden Markov models. *Journal of the American Statistical Association*, 111(515):951–966, 2016.

Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

Siddhartha Chib. Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75(1):79–97, 1996.

Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

Hyoyoung Choo-Wosoba, Paul S Albert, and Bin Zhu. A hidden Markov modeling approach for identifying tumor subclones in next-generation sequencing studies. *Biostatistics*, 2020.

Ole F Christensen, Gareth O Roberts, and Jeffrey S Rosenthal. Scaling limits for the transient phase of local metropolis–hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268, 2005.

Peter Congdon. Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational statistics & data analysis*, 50(2):346–357, 2006.

Adrian Corduneanu and Christopher M Bishop. Variational Bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

Michele Costa and Luca De Angelis. Model selection in hidden Markov models: a simulation study. 2010.

Carl De Boor. A practical guide to splines. 2001. *Appl. Math. Sci*, 2001.

Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978.

Yohann De Castro, Elisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Transactions on Information Theory*, 63(8):4758–4777, 2017.

Marie Du Roy de Chaumaray, Matthieu Marbac, and Fabien Navarro. Mixture of hidden Markov models for accelerometer data. *The Annals of Applied Statistics*, 14(4):1834–1855, 2020.

MCM De Gunst and O Shcherbakova. Asymptotic behavior of bayes estimators for hidden Markov models with application to ion channels. *Mathematical Methods of Statistics*, 17(4):342–356, 2008.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

David GT Denison, Christopher C Holmes, Bani K Mallick, and Adrian FM Smith. *Bayesian methods for nonlinear classification and regression*, volume 386. John Wiley & Sons, 2002.

DGT Denison, BK Mallick, and AFM Smith. Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2): 333–350, 1998.

Stacy L DeRuiter, Roland Langrock, Tomas Skirbutas, Jeremy A Goldbogen, John Calambokidis, Ari S Friedlaender, Brandon L Southall, et al. A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *The Annals of Applied Statistics*, 11(1):362–392, 2017.

Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265. ACM, 1986.

Thomas J DiCiccio, Robert E Kass, Adrian Raftery, and Larry Wasserman. Computing bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439):903–915, 1997.

Ilaria DiMatteo, Christopher R Genovese, and Robert E Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.

Finale Doshi-Velez, David Pfau, Frank Wood, and Nicholas Roy. Bayesian nonparametric methods for partially-observable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):394–407, 2013.

Randal Douc, Eric Moulines, Tobias Rydén, et al. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of statistics*, 32(5):2254–2304, 2004.

Randal Douc, Eric Moulines, Jimmy Olsson, Ramon Van Handel, et al. Consistency of the maximum likelihood estimator for general hidden Markov models. *the Annals of Statistics*, 39(1):474–513, 2011.

Randal Douc, Jimmy Olsson, and François Roueff. Posterior consistency for partially observed Markov models. *Stochastic Processes and their Applications*, 130(2):733–759, 2020.

Kumar Avinava Dubey, Michael Zhang, Eric Xing, and Sinead Williamson. Distributed, partially collapsed MCMC for Bayesian nonparametrics. In *International Conference on Artificial Intelligence and Statistics*, pages 3685–3695. PMLR, 2020.

Arnaud Dufays. Infinite-state Markov-switching for dynamic volatility. *Jnl of Financial Econometrics*, 14(2):418–460, 2016.

Jay C Dunlap, Jennifer J Loros, and Patricia J DeCoursey. *Chronobiology: biological timekeeping.* Sinauer Associates, 2004.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Przemyslaw Dymarski. *Hidden Markov Models: Theory and Applications.* BoD–Books on Demand, 2011.

Theodoros Economou, Trevor C Bailey, and Zoran Kapelan. Mcmc implementation for Bayesian hidden semi-Markov models with illustrative applications. *Statistics and Computing*, 24(5):739–752, 2014.

Matthew C Edwards, Renate Meyer, and Nelson Christensen. Bayesian nonparametric spectral density estimation using b-spline priors. *Statistics and Computing*, 29(1):67–78, 2019.

Elise Epaillard and Nizar Bouguila. Proportional data modeling with hidden Markov models based on generalized Dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. *Pattern Recognition*, 55:125–136, 2016.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

Jennifer A Evans and Alec J Davidson. Health consequences of circadian disruption in humans and animal models. *Progress in molecular biology and translational science*, 119:283–323, 2013.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

Russell G Foster. Sleep, circadian rhythms and health. *Interface Focus*, 10(3): 20190098, 2020.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056, 2011.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Nial Friel and Jason Wyse. Estimating the evidence–a review. *Statistica Neerlandica*, 66(3):288–308, 2012.

Sylvia Frühwirth-Schnatter. Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7(1):143–167, 2004.

Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*, volume 796. Springer, 2006.

Elisabeth Gassiat and Stéphane Boucheron. Optimal error exponents in hidden Markov models order estimation. *IEEE Transactions on Information Theory*, 49 (4):964–980, 2003.

Elisabeth Gassiat and Christine Keribin. The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM: Probability and Statistics*, 4:25–52, 2000.

Elisabeth Gassiat, Judith Rousseau, et al. About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20(4):2039–2075, 2014.

Élisabeth Gassiat, Alice Cleynen, and Stéphane Robin. Inference in finite state space non parametric hidden Markov models and applications. *Statistics and Computing*, 26(1-2):61–71, 2016a.

Elisabeth Gassiat, Judith Rousseau, et al. Nonparametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193–212, 2016b.

Hong Ge, Yutian Chen, Moquan Wan, and Zoubin Ghahramani. Distributed inference for Dirichlet process mixture models. In *International Conference on Machine Learning*, pages 2276–2284. PMLR, 2015.

Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955. URL `https://books.google.co.uk/books?id=ZXL6AQAAQBAJ`.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Andrew Gelman, Walter R Gilks, and Gareth O Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

John Geweke. Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics & Data Analysis*, 51(7):3529–3550, 2007.

Subhashis Ghosal and Aad W Van Der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, pages 1233–1263, 2001.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220, 2000.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Peter J Green, Krzysztof Łatuszyński, Marcelo Pereyra, and Christian P Robert. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862, 2015.

Oswaldo Gressani and Philippe Lambert. Fast Bayesian inference using laplace approximations in a flexible promotion time cure model based on p-splines. *Computational Statistics & Data Analysis*, 124:151–167, 2018.

Oswaldo Gressani and Philippe Lambert. Laplace approximations for fast Bayesian inference in generalized additive models based on p-splines. *Computational Statistics & Data Analysis*, 154:107088, 2021.

Beniamino Hadj-Amar, Bärbel Finkenstädt, Mark Fiecas, and Robert Huckstepp. Identifying the recurrence of sleep apnea using a spectral hidden Markov model. *arXiv e-prints*, pages arXiv–2001, 2020a.

Beniamino Hadj-Amar, Jack Jewson, and Mark Fiecas. Bayesian approximations to hidden semi-Markov models. *arXiv preprint arXiv:2006.09061*, 2020b.

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384, 1989.

James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020.

Wolfgang Karl Härdle, Ostap Okhrin, and Weining Wang. Hidden Markov structures for dynamic copulae. *Econometric Theory*, pages 981–1015, 2015.

David I Hastie and Peter J Green. Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 66(3):309–338, 2012.

David I Hastie, Silvia Liverani, and Sylvia Richardson. Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and computing*, 25(5):1023–1037, 2015.

Keegan E Hines, John R Bankston, and Richard W Aldrich. Analyzing single-molecule time series via nonparametric Bayesian inference. *Biophysical journal*, 108(3):540–556, 2015.

Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

Wytske A Hofstra and Al W de Weerd. How to assess circadian rhythm in humans: a review of literature. *Epilepsy & Behavior*, 13(3):438–444, 2008.

Tracy Holsclaw, Arthur M Greene, Andrew W Robertson, and Padhraic Smyth. Bayesian nonhomogeneous Markov models via pólya-gamma data augmentation with applications to rainfall modeling. *The Annals of Applied Statistics*, 11(1): 393–426, 2017.

Jim A Horne and Olov Östberg. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International journal of chronobiology*, 1976.

Chenghan Hou. Infinite hidden Markov switching vars with application to macroeconomic forecast. *International Journal of Forecasting*, 33(4):1025–1043, 2017.

Weiming Hu, Guodong Tian, Yongxin Kang, Chunfeng Yuan, and Stephen Maybank. Dual sticky hierarchical Dirichlet process hidden Markov model and its application to natural language description of motions. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2355–2373, 2017.

Qi Huang, Dwayne Cohen, Sandra Komarzynski, Xiao-Mei Li, Pasquale Innominato, Francis Lévi, and Bärbel Finkenstädt. Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data. *Journal of The Royal Society Interface*, 15(139):20170885, 2018.

Ying Hung, Yijie Wang, Veronika Zarnitsyna, Cheng Zhu, and CF Jeff Wu. Hidden markov models with applications in cell adhesion experiments. *Journal of the American Statistical Association*, 108(504):1469–1479, 2013.

Syed Anas Imtiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5):1562, 2021.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Hemant Ishwaran and Lancelot F James. Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical statistics*, 11(3):508–532, 2002.

Hemant Ishwaran and Mahmoud Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

Girardin Jean-Louis, Hans von Gizycki, Ferdinand Zizi, Jeffrey Fookson, Arthur Spielman, Joao Nunes, Robert Fullilove, and Harvey Taub. Determination of sleep and wakefulness with the actigraph data analysis software (adas). *Sleep*, 19 (9):739–743, 1996.

William H Jefferys and James O Berger. Ockham's razor and Bayesian analysis. *American Scientist*, 80(1):64–72, 1992.

Seonghyun Jeong, Taeyoung Park, and David A van Dyk. Bayesian model selection in additive partial linear models via locally adaptive splines. *arXiv preprint arXiv:2008.06213*, 2020.

Matthew James Johnson and Alan S Willsky. Bayesian nonparametric hidden semi-Markov models. 2013.

Astrid Jullion and Philippe Lambert. Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian p-splines models. *Computational statistics & data analysis*, 51(5):2542–2558, 2007.

Maria Kalli, Jim E Griffin, and Stephen G Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.

Kai Kang, Jingheng Cai, Xinyuan Song, and Hongtu Zhu. Bayesian hidden Markov models for delineating the pathology of alzheimer's disease. *Statistical methods in medical research*, 28(7):2112–2124, 2019.

Chang-Jin Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1-2):1–22, 1994.

Ja-Yong Koo. Bivariate b-splines for tensor logspline density estimation. *Computational statistics & data analysis*, 21(1):31–42, 1996.

Philippe Lambert and Vincent Bremhorst. Inclusion of time-varying covariates in cure survival models with an application in fertility studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1):333–354, 2020.

Stefan Lang and Andreas Brezger. Bayesian P-splines. *Journal of computational and graphical statistics*, 13(1):183–212, 2004.

Roland Langrock and W Zucchini. Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, 55(1):715–724, 2011.

Roland Langrock, Ruth King, Jason Matthiopoulos, Len Thomas, Daniel Fortin, and Juan M Morales. Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology*, 93(11):2336–2342, 2012a.

Roland Langrock, Iain L MacDonald, and Walter Zucchini. Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models. *Journal of Empirical Finance*, 19(1):147–161, 2012b.

Roland Langrock, Bruce J Swihart, Brian S Caffo, Naresh M Punjabi, and Ciprian M Crainiceanu. Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in medicine*, 32(19):3342–3356, 2013.

Roland Langrock, Tiago A Marques, Robin W Baird, and Len Thomas. Modeling the diving behavior of whales: a latent-variable approach with feedback and semi-Markovian components. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(1):82–100, 2014.

Roland Langrock, Thomas Kneib, Alexander Sohn, and Stacy L DeRuiter. Nonparametric inference in hidden Markov models using p-splines. *Biometrics*, 71(2): 520–528, 2015.

Roland Langrock, Thomas Kneib, Richard Glennie, and Théo Michelot. Markov-switching generalized additive models. *Statistics and Computing*, 27(1):259–270, 2017.

Roland Langrock, Timo Adam, Vianey Leos-Barajas, Sina Mews, David L Miller, and Yannis P Papastamatiou. Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, 72(3):179–200, 2018.

Luc Lehéricy. State-by-state minimax adaptive estimation for nonparametric hidden Markov models. *The Journal of Machine Learning Research*, 19(1):1432–1477, 2018.

Luc Lehéricy et al. Consistent order estimation for nonparametric hidden Markov models. *Bernoulli*, 25(1):464–498, 2019.

Vianey Leos-Barajas, Eric J Gangloff, Timo Adam, Roland Langrock, Floris M Van Beest, Jacob Nabe-Nielsen, and Juan M Morales. Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):232–248, 2017.

Brian G Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1):127–143, 1992.

Francis Lévi and Alper Okyar. Circadian clocks and drug delivery systems: impact and opportunities in chronotherapeutics. *Expert opinion on drug delivery*, 8(12): 1535–1541, 2011.

Francis Lévi, Alper Okyar, Sandrine Dulong, Pasquale F Innominato, and Jean Clairambault. Circadian timing in cancer treatments. *Annual review of pharmacology and toxicology*, 50:377–421, 2010.

Michael Li and Benjamin M Bolker. Incorporating periodic variability in hidden Markov models for animal movement. *Movement ecology*, 5(1):1–12, 2017.

Xinyue Li, Yunting Zhang, Fan Jiang, and Hongyu Zhao. A novel machine learning unsupervised algorithm for sleep/wake identification using actigraphy. *Chronobiology international*, 37(7):1002–1015, 2020a.

Yong Li, Jun Yu, and Tao Zeng. Deviance information criterion for latent variable models and misspecified models. *Journal of Econometrics*, 216(2):450–493, 2020b.

Yuanxi Li, Stephen Swift, and Allan Tucker. Modelling and analysing the dynamics of disease progression from cross-sectional studies. *Journal of biomedical informatics*, 46(2):266–274, 2013.

Jiaxing Liu, Yang Zhao, Boya Lai, Hailiang Wang, and Kwok Leung Tsui. Wearable device heart rate and activity data in an unsupervised approach to personalized sleep monitoring: Algorithm validation. *JMIR mHealth and uHealth*, 8(8):e18370, 2020.

Fernando Llorente, Luca Martino, David Delgado, and Javier Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *arXiv preprint arXiv:2005.08334*, 2020.

Stefan Lüdtke, Wiebke Hermann, Thomas Kirste, Heike Beneš, and Stefan Teipel. An algorithm for actigraphy-based sleep/wake scoring: Comparison with polysomnography. *Clinical Neurophysiology*, 132(1):137–145, 2021.

Laura M Lyall, Cathy A Wyse, Nicholas Graham, Amy Ferguson, Donald M Lyall, Breda Cullen, Carlos A Celis Morales, Stephany M Biello, Daniel Mackay, Joey Ward, et al. Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91 105 participants from the uk biobank. *The Lancet Psychiatry*, 5(6):507–514, 2018.

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

David JC MacKay. Ensemble learning for hidden Markov models. Technical report, Citeseer, 1997.

Arnab Kumar Maity, Sanjib Basu, and Santu Ghosh. Bayesian criterion-based variable selection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2021.

Jean-Michel Marin and Christian P Robert. Importance sampling methods for Bayesian discrimination between embedded models. *arXiv preprint arXiv:0910.2325*, 2009.

Jean-Michel Marin, Kerrie Mengersen, and Christian P Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507, 2005.

Tristan Marshall and Gareth Roberts. An adaptive approach to langevin MCMC. *Statistics and Computing*, 22(5):1041–1057, 2012.

Antonello Maruotti and Antonio Punzo. Initialization of hidden markov and semi-markov models: A critical evaluation of several strategies. *International Statistical Review*, 2021.

Antonello Maruotti, Jan Bulla, Francesco Lagona, Marco Picone, and Francesca Martella. Dynamic mixtures of factor analyzers to characterize multivariate air pollutant exposures. *The Annals of Applied Statistics*, 11(3):1617–1648, 2017.

Patricio Maturana-Russel and Renate Meyer. Bayesian spectral density estimation using p-splines with quantile-based knot placement. *Computational Statistics*, pages 1–23, 2021.

Kenneth Jordan Mccallum and Ji-Ping Wang. Quantifying copy number variations using a hidden Markov model with inhomogeneous emission distributions. *Biostatistics*, 14(3):600–611, 2013.

Brett T McClintock, Roland Langrock, Olivier Gimenez, Emmanuelle Cam, David L Borchers, Richard Glennie, and Toby A Patterson. Uncovering ecological state dynamics with hidden Markov models. *Ecology letters*, 23(12):1878–1903, 2020.

Clare A McGrory and DM Titterington. Variational Bayesian analysis for hidden Markov models. *Australian & New Zealand Journal of Statistics*, 51(2):227–244, 2009.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

Laurent Mevel and Lorenzo Finesso. Asymptotical statistics of misspecified hidden Markov models. *IEEE Transactions on Automatic Control*, 49(7):1123–1132, 2004.

Théo Michelot, Roland Langrock, Thomas Kneib, and Ruth King. Maximum penalized likelihood estimation in semiparametric mark-recapture-recovery models. *Biometrical Journal*, 58(1):222–239, 2016.

Jeffrey W Miller. *Nonparametric and variable-dimension Bayesian mixture models: Analysis, comparison, and new methods*. PhD thesis, Division of Applied Mathematics at Brown University, 2014.

Jeffrey W Miller and Matthew T Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113 (521):340–356, 2018.

David Minors, Torbjorn Akerstedt, Greg Atkinson, Merryl Dahlitz, Simon Folkard, Francis Levi, Christine Mormont, David Parkes, and James Waterhouse. The difference between activity when in bed and out of bed. i. healthy subjects and selected patients. *Chronobiology international*, 13(1):27–34, 1996.

Bhavya Mor, Sunita Garhwal, and Ajay Kumar. A systematic review of hidden Markov models and their applications. *Archives of computational methods in engineering*, 28(3), 2021.

Jeffrey S Morris, Cassandra Arroyo, Brent A Coull, Louise M Ryan, Richard Herrick, and Steven L Gortmaker. Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *Journal of the American Statistical Association*, 101(476):1352–1364, 2006.

Maxime Mouchet, Sandrine Vaton, and Thierry Chonavel. Statistical characterization of round-trip times with nonparametric hidden Markov models. In *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pages 43–48. IEEE, 2019.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Kyoko Nakazaki, Shingo Kitamura, Yuki Motomura, Akiko Hida, Yuichi Kamei, Naoki Miura, and Kazuo Mishima. Validity of an algorithm for determining sleep/wake states using a new actigraph. *Journal of physiological anthropology*, 33(1):1–8, 2014.

Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2): 125–139, 2001.

Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

Nam Thanh Nguyen, Dinh Q Phung, Svetha Venkatesh, and Hung Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 955–960. IEEE, 2005.

Agostino Nobile. *Bayesian analysis of finite mixture distributions*. Carnegie Mellon University, 1994.

E Ortiz-Tudela, A Mteyrek, A Ballesta, PF Innominato, and F Levi. Cancer chronotherapeutics: experimental, theoretical, and clinical aspects. *Circadian clocks*, pages 261–288, 2013.

Elisabet Ortiz-Tudela, Antonio Martinez-Nicolas, Manuel Campos, María Ángeles Rol, and Juan Antonio Madrid. A new integrated variable based on thermometry, actimetry and body position (tap) to evaluate circadian system status in humans. *PLoS computational biology*, 6(11):e1000996, 2010.

Elisabet Ortiz-Tudela, Ida Iurisci, Jacques Beau, Abdoulaye Karaboue, Thierry Moreau, Maria Angeles Rol, Juan Antonio Madrid, Francis Lévi, and Pasquale F Innominato. The circadian rest-activity rhythm, a potential safety pharmacology endpoint of cancer chemotherapy. *International journal of cancer*, 134(11):2717–2725, 2014.

Elisabet Ortiz-Tudela, Pasquale F Innominato, Maria Angeles Rol, Francis Lévi, and Juan Antonio Madrid. Relevance of internal time and circadian robustness for cancer patients. *BMC cancer*, 16(1):1–12, 2016.

Shing-Tai Pan, Chih-En Kuo, Jian-Hong Zeng, and Sheng-Fu Liang. A transition-constrained discrete hidden Markov model for automatic sleep staging. *Biomedical engineering online*, 11(1):1–19, 2012.

Omiros Papaspiliopoulos. A note on posterior sampling from Dirichlet mixture models. *manuscript, Department of Economics, Universitat Pompeu Fabra*, 2008.

Omiros Papaspiliopoulos and Gareth O Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

Panagiotis Papastamoulis. label.switching: An r package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets*, 69(1):1–24, 2016a. ISSN 1548-7660. doi: 10.18637/jss.v069.c01. URL `https://www.jstatsoft.org/v069/c01`.

Panagiotis Papastamoulis. label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets*, 69(1):1–24, 2016b. doi: 10.18637/jss.v069.c01.

Carrie L Partch, Carla B Green, and Joseph S Takahashi. Molecular architecture of the mammalian circadian clock. *Trends in cell biology*, 24(2):90–99, 2014.

Toby A Patterson, Marinelle Basson, Mark V Bravington, and John S Gunn. Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, 78(6):1113–1123, 2009.

Joe Scutt Phillips, Toby A Patterson, Bruno Leroy, Graham M Pilling, and Simon J Nicol. Objective classification of latent behavioral states in bio-logging data using multivariate-normal hidden Markov models. *Ecological Applications*, 25(5):1244–1258, 2015.

Massimo Piccardi and Óscar Pérez. Hidden Markov models with kernel density estimation of emission probabilities and their use in activity recognition. In *CVPR*. Citeseer, 2007.

Jim Pitman. Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996.

Jim Pitman. Poisson–Dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11(5):501–514, 2002.

Jim Pitman. *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.

Jennifer Pohle, Roland Langrock, Floris M van Beest, and Niels Martin Schmidt. Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):270–293, 2017.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Hartmut Prautzsch, Wolfgang Boehm, and Marco Paluszny. *Bézier and B-spline techniques*, volume 6. Springer, 2002.

Antonio Punzo and Antonello Maruotti. Clustering multivariate longitudinal observations: The contaminated gaussian hidden Markov model. *Journal of Computational and Graphical Statistics*, 25(4):1097–1098, 2016.

Yuting Qi, John William Paisley, and Lawrence Carin. Music analysis using hidden Markov mixture models. *IEEE Transactions on Signal Processing*, 55(11):5209–5224, 2007.

Mirja Quante, Emily R Kaplan, Michael Cailler, Michael Rueschman, Rui Wang, Jia Weng, Elsie M Taveras, and Susan Redline. Actigraphy-based sleep estimation in adolescents and adults: a comparison with polysomnography using two scoring algorithms. *Nature and science of sleep*, 10:13, 2018.

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

Vishalini Laguduva Ramnath and Srinivas Katkoori. A smart iot system for continuous sleep state monitoring. In *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 241–244. IEEE, 2020.

Carl Edward Rasmussen et al. The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

Sylvia Richardson and Peter J Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Christian P Robert and DM Titterington. Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, 8(2):145–158, 1998.

Christian P Robert and Darren Wraith. Computational methods for Bayesian model choice. In *Aip conference proceedings*, volume 1193, pages 251–262. American Institute of Physics, 2009.

Christian P Robert, Gilles Celeux, and Jean Diebolt. Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics & Probability Letters*, 16 (1):77–83, 1993.

Christian P Robert, Tobias Ryden, and David M Titterington. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):57–75, 2000.

Christian P Robert, Jean-Michel Marin, et al. On some difficulties with a posterior probability approximation technique. *Bayesian Analysis*, 3(2):427–441, 2008.

Daniel M Roberts, Margeaux M Schade, Gina M Mathew, Daniel Gartenberg, and Orfeu M Buxton. Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. *Sleep*, 43(7):zsaa045, 2020.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

Gareth O Roberts and Jeffrey S Rosenthal. Harris recurrence of metropolis-within-gibbs and trans-dimensional Markov chains. *The Annals of Applied Probability*, pages 2123–2139, 2006.

Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive MCMC. *Journal of computational and graphical statistics*, 18(2):349–367, 2009.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Carlos E Rodríguez and Stephen G Walker. Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, 2014.

Till Roenneberg and Martha Merrow. The circadian clock and human health. *Current biology*, 26(10):R432–R443, 2016.

Jeffrey S Rosenthal. Amcmc: An r interface for adaptive MCMC. *Computational Statistics & Data Analysis*, 51(12):5467–5470, 2007.

Marc D Ruben, David F Smith, Garret A FitzGerald, and John B Hogenesch. Dosing time matters. *Science*, 365(6453):547–549, 2019.

Tobias Rydén. Estimating the order of hidden Markov models. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(4):345–354, 1995.

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

Tobias Rydén, Timo Teräsvirta, and Stefan Åsbrink. Stylized facts of daily return series and the hidden Markov model. *Journal of applied econometrics*, 13(3): 217–244, 1998.

Giada Sacchi and Ben Swallow. Toward efficient Bayesian approaches to inference in hierarchical hidden Markov models for inferring animal behavior. *Frontiers in Ecology and Evolution*, 9:249, 2021.

Richard S Savage, Zoubin Ghahramani, Jim E Griffin, Bernard J De la Cruz, and David L Wild. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics*, 26(12):i158–i167, 2010.

Larry Schumaker. *Spline functions: basic theory*. Cambridge University Press, 2007.

Steven L Scott. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American statistical Association*, 97(457): 337–351, 2002.

Steven L Scott, Gareth M James, and Catherine A Sugar. Hidden Markov models for longitudinal comparisons. *Journal of the American Statistical Association*, 100 (470):359–369, 2005.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Ioannis Sgouralis and Steve Pressé. An introduction to infinite HMMs for single-molecule data analysis. *Biophysical journal*, 112(10):2021–2029, 2017.

Emmanuel Sharef, Robert L Strawderman, David Ruppert, Mark Cowen, Lakshmi Halasyamani, et al. Bayesian adaptive b-spline estimation in proportional hazards frailty models. *Electronic journal of statistics*, 4:606–642, 2010.

Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

Michael Smith and Robert Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

Michael Smith and Robert Kohn. A Bayesian approach to nonparametric bivariate regression. *Journal of the American Statistical Association*, 92(440):1522–1535, 1997.

Kyung-Ah Sohn, Eric P Xing, et al. A hierarchical Dirichlet process mixture model for haplotype reconstruction from multi-population data. *Annals of Applied Statistics*, 3(2):791–821, 2009.

Xinyuan Song, Kai Kang, Ming Ouyang, Xuejun Jiang, and Jingheng Cai. Bayesian analysis of semiparametric hidden Markov models with latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(1):1–20, 2018.

Yong Song. Modelling regime switching and structural breaks with an infinite hidden Markov model. *Journal of Applied Econometrics*, 29(5):825–842, 2014.

Yong Song and Tomasz Woźniak. Markov switching. *arXiv preprint arXiv:2002.03598*, 2020.

Luigi Spezia. Reversible jump and the label switching problem in hidden Markov models. *Journal of Statistical Planning and Inference*, 139(7):2305–2315, 2009.

Luigi Spezia. Bayesian analysis of multivariate gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis*, 31(1): 1–11, 2010.

Luigi Spezia, Martyn N Futter, and Mark J Brewer. Periodic multivariate normal hidden Markov models for the analysis of water quality time series. *Environmetrics*, 22(3):304–317, 2011.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van der Linde. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 485–493, 2014.

Statisticat and LLC. *LaplacesDemon: Complete Environment for Bayesian Inference*, 2021. R package version 16.1.6.

Phyllis K Stein and Yachuan Pu. Heart rate variability, sleep and sleep disorders. *Sleep medicine reviews*, 16(1):47–66, 2012.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

Surya T Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pages 90–110, 2006.

Amir H Harati Nejad Torbati and Joseph Picone. A doubly hierarchical Dirichlet process hidden Markov model with a non-ergodic structure. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1):174–184, 2015.

Augustin Touron, Thi Thu Huong Hoang, and Sylvie Parey. Bivariate modelling of precipitation and temperature using a non-homogeneous hidden Markov model. *arXiv preprint arXiv:1810.09682*, 2018.

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th international conference on Machine learning*, pages 1088–1095, 2008.

Elodie Vernet et al. Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics*, 9(1):717–752, 2015.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2): 260–269, 1967.

Stevenn Volant, Caroline Bérard, Marie-Laure Martin-Magniette, and Stéphane Robin. Hidden Markov models with mixtures as emission distributions. *Statistics and Computing*, 24(4):493–504, 2014.

Olivia Walch. Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography. *PhysioNet*, 2019. URL `https://doi.org/10.13026/hmhs-py35`.

Olivia Walch, Yitong Huang, Daniel Forger, and Cathy Goldstein. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, 42(12):zsz180, 2019.

Stephen G Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54, 2007.

Eva Charlotte Winnebeck, Dorothee Fischer, Tanya Leise, and Till Roenneberg. Dynamics and ultradian structure of human sleep in real life. *Current Biology*, 28 (1):49–59, 2018.

Vitali Witowski, Ronja Foraita, Yannis Pitsiladis, Iris Pigeot, and Norman Wirsik. Using hidden Markov models to improve quantifying physical activity in accelerometer data–a simulation study. *PloS one*, 9(12):e114089, 2014.

Luo Xiao, Lei Huang, Jennifer A Schrack, Luigi Ferrucci, Vadim Zipunnikov, and Ciprian M Crainiceanu. Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics*, 16(2):352–367, 2015.

Junyu Xuan, Jie Lu, and Guangquan Zhang. A survey on Bayesian nonparametric learning. *ACM Computing Surveys (CSUR)*, 52(1):1–36, 2019.

Christopher Yau, Omiros Papaspiliopoulos, Gareth O Roberts, and Christopher Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57, 2011.

Halid Ziya Yerebakan and Murat Dundar. Partially collapsed parallel gibbs sampler for Dirichlet process mixture models. *Pattern Recognition Letters*, 90:22–27, 2017.

Shun-Zheng Yu. *Hidden Semi-Markov models: theory, algorithms and applications*. Morgan Kaufmann, 2015.

Yu Ryan Yue, Paul L Speckman, and Dongchu Sun. Priors for Bayesian adaptive spline smoothing. *Annals of the Institute of Statistical Mathematics*, 64(3):577–613, 2012.

Elena Zanini, Emma Eastoe, MJ Jones, David Randell, and Philip Jonathan. Flexible covariate representations for extremes. *Environmetrics*, 31(5):e2624, 2020.

Bing Zhai, Ignacio Perez-Pozuelo, Emma AD Clifton, Joao Palotti, and Yu Guan. Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–33, 2020.

Eric E Zhang and Steve A Kay. Clocks not winding down: unravelling circadian networks. *Nature reviews Molecular cell biology*, 11(11):764–776, 2010.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Aite Zhao, Junyu Dong, and Huiyu Zhou. Self-supervised learning from multi-sensor data for sleep recognition. *IEEE Access*, 8:93907–93921, 2020.

Ding Zhou, Yuanjun Gao, and Liam Paninski. Disentangled sticky hierarchical Dirichlet process hidden Markov model. In Frank Hutter, Kristian Kersting, Jefrey Lijffijt, and Isabel Valera, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 612–627, Cham, 2021. Springer International Publishing. ISBN 978-3-030-67658-2.

W Zucchini and IL MacDonald. Illustrations of the use of pseudo-residuals in assessing the fit of a model. In *Proceedings of the 14th International Workshop on Statistical Modelling*, volume 3, pages 409–416. Graz, Austria, Statistical Modeling Society, 1999.

Walter Zucchini and Iain L MacDonald. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC, 2009.

Walter Zucchini, Iain L MacDonald, and Roland Langrock. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC, 2016.