Hidden Markov models with an unknown number of states and a repulsive prior on the state parameters

Ioannis Rotous
School of Mathematics, Statistics and Actuarial Science, University of Kent
email: ir237@kent.ac.uk

and

Alex Diana

Department of Mathematics, Statistics and Actuarial Science, University of Essex
email: ad23269@essex.ac.uk

and

Alessio Farcomeni
Department of Economics and Finance, Tor Vergata University of Rome
email: alessio.farcomeni@uniroma2.it

and

Eleni Matechou
School of Mathematics, Statistics and Actuarial Science, University of Kent
email: e.matechou@kent.ac.uk

and

Andréa Thiebault
Institut des Neurosciences Paris-Saclay (NeuroPSI), CNRS UMR 9197, Université Paris-Saclay
email: andrea.thiebault@cnrs.fr

Summary: Hidden Markov models (HMMs) offer a robust and efficient framework for analyzing time series data, modelling both the underlying latent state progression over time and the observation process conditional on the latent state. However, a critical challenge lies in determining the appropriate number of underlying states, often unknown in practice. In this paper, we employ a Bayesian framework, treating the number of states as a random variable and employing reversible jump Markov chain Monte Carlo to sample from the posterior distributions of all parameters, including the number of states. Additionally, we introduce repulsive priors for the state parameters in HMMs, and hence avoid overfitting issues and promote parsimonious models with dissimilar state components. We demonstrate our proposed framework on two ecological case studies: GPS tracking data on muskox in Antarctica and acoustic data on Cape gannets in South Africa, Our results demonstrate how our framework effectively explores the model space, defined by models with different latent state dimensions, while leading to latent states that are distinguished better and hence are more interpretable, enabling understanding and interpretation of complex dynamic systems.

Key words: GPS tracking, acoustic data, interactive point process, reversible jump MCMC

1. Introduction

Hidden Markov models (HMMs) are a powerful and well-established framework for analyzing time series data in cases where the studied system transitions between a set of hidden states over time. HMMs jointly model two processes: the underlying latent process of the hidden states, and the observation process, conditional on the states, as shown in Figure 1 (Cappé et al., 2009; Zucchini and MacDonald, 2009). HMMs enable efficient modelling of the evolution of the latent states across time and, conditional on those latent states, explicit modelling of the data observation process, even in complex systems and processes with multiple latent states and complicated observation processes (Popov et al., 2017).

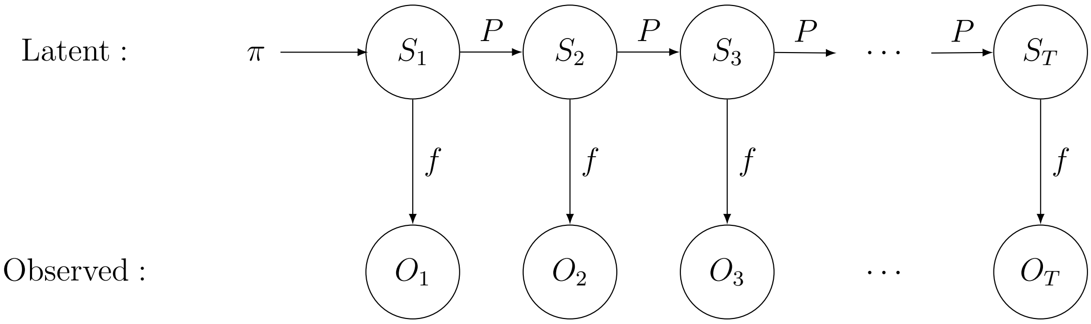

Figure 1: Illustration of a hidden Markov model evolution across $t = 1, 2, ..., T$, time points, with latent states S_i and corresponding observations O_i, characterized by an initial latent distribution π, transition probabilities P, and emission distribution f.

HMMs commonly employ a first-order Markov chain, where the evolution of the latent states depends only on the previous time point. Additionally, conditional on the latent state, they emit observables at the current time point, independent of the rest of the observables. Further details can be found in Section 2.1, which describes the joint distribution of observables with latent states (Equation (1)). The efficacy of HMMs relies on the separation of the latent and observation processes and the use of algorithms that efficiently marginalize over the latent states, such as the forward/backward algorithm for computing the likelihood function and the Viterbi algorithm for finding the most likely sequence of hidden states (Zucchini and MacDonald, 2009; Bartolucci et al., 2013). Hence, HMMs have proven to be powerful and easy-to-use tools, and are widely utilized in various fields, such as finance (Rydén et al., 1998), biology (Leroux and Puterman, 1992), social science (Rabiner, 1989; Zucchini and MacDonald, 2009), medicine (Farcomeni, 2017), and ecology (Schmidt et al., 2016; Patterson et al., 2017), among others.

One of the key challenges in applying HMMs is the decision on the appropriate number of underlying states in the system. In practice, the true number of states is often unknown. It is standard practice to fix the number of latent states or fit models that consider different numbers of latent states (Robert et al., 1993; Chib, 1996; Robert and Titterington, 1998) in either a classical (Huang et al., 2017), or Bayesian framework (Berkhof et al., 2003), and compare them with appropriate criteria to select the number of states. However, there are several issues with this practice, for example the model needs to be fitted multiple times, and in the end, a single model is used for interpretation, but without accounting for the uncertainty in the model selection process itself (McLachlan et al., 2019). Alternatively, in a Bayesian framework, the number of latent sates can be treated as an additional random variable, and hence reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995) methods can be employed to sample from a posterior distribution in this case where the model dimension is not fixed; indeed RJMCMC has been used extensively within an HMM context (Robert et al., 2000; Cappé et al., 2003, 2009; Russo et al., 2022). We note that HMMs can equivalently be viewed from the perspective of dynamic mixture models (Spezia, 2020), both in discrete and continuous space (Bartolucci and Pandolfi, 2011; Reynolds et al., 2009), with the number of mixture components corresponding to the number of states. In this paper, we refer to HMMs and corresponding states, but the concepts equally apply to dynamic mixture models and corresponding components.

Within a Bayesian framework, state allocation (S_t in Figure 1) can be sampled within the MCMC algorithm (Stephens and Phil, 1997), so that the complete data likelihood is used (King, 2014). This approach however can lead to a substantial number of sampled latent variables. Instead, state allocation can be marginalised out of the model, as is standard practice within the HMM machinery (Russo et al., 2022), so that the observed data likelihood is used for inference, and this is the approach we employ in this paper. However, in either case, HMMs are prone to overfitting, and hence such algorithms can lead to an unnecessarily large number of similar states (Duan and Dunson, 2018). Recent advancements in the field of mixture modelling have introduced the use of repulsive priors, which promote parsimony in the model (Petralia et al., 2012; Quinlan et al., 2021; Natarajan et al., 2023). These repulsive prior distributions serve to impose constraints on the proximity of state parameters, which discourages similar states from being created. Unavoidably, this particular form of penalty applied to the parameter space also affects the selection of the number of states (Natarajan et al., 2023). An additional advantage associated with incorporating a repulsive prior into HMMs, whether with fixed or variable dimensions, is the mitigation of overfitting. In certain instances, conventional mixture models may excessively fine-tune their components to capture noise in the data, resulting in poor generalization. Extensive research, as documented in the literature (Petralia et al., 2012; Quinlan et al., 2021; Beraha et al., 2022), has demonstrated that addressing these issues can be effectively achieved by introducing repulsion constraints among distribution parameters within the model, thereby promoting dissimilarity among its components.

To introduce a repulsion constraint within an HMM framework we use interaction point processes, referred to as a repulsive prior in this paper, which is a class of distributions on a set of points that actively promotes repulsion between points that are close together. Specifically, we consider distributions belonging to the family of pairwise interaction point processes called the Strauss point process prior, first described in Strauss (1975), as priors on the state parameters of the emission distribution f as illustrated in Figure 1. This approach enables us to achieve more effective and interpretable modelling without pre-specifying the number of states, thereby improving various aspects of analysis, including inference, model selection, and our overall understanding of dynamics.

HMMs have been extensively used in the ecological literature (Gimenez et al., 2009; Schmidt et al., 2016; Patterson et al., 2017), since they are well-suited to capturing the underlying latent state structure in ecological systems, enabling researchers to seamlessly integrate the observed data with the unobserved latent states (Glennie et al., 2023). In ecological systems, these latent states can correspond to life stages (McClintock et al., 2020) or behavioural states (Schmidt et al., 2016; Nicol et al., 2023). We demonstrate our approach using two ecological case studies: GPS data on muskox, *Ovibos moschatus*, in Antarctica, also analysed in Pohle et al. (2017), who used model selection criteria to select the number of states in their HMM, and acoustic data on Cape gannets, *Morus capensis*, in South Africa, analysed in Thiebault et al. (2021) where behavioural state classification was performed manually to train a subsequent model. Our results demonstrate how our framework effectively explores the model space, defined by models with different latent state dimensions, while leading to latent states that are distinguished better and hence are more interpretable.

The article is structured as follows: Section 2 introduces the general concepts of HMMs and repulsive priors, and gives a broad overview of the model-fitting approach developed in this paper, with technical details provided in the Supplementary Material. Section 3 presents the results of a case study using GPS data on muskox, as illustrated in Pohle et al. (2017) and Section 4 the results of a case study using acoustic data on Cape gannets. Finally, the paper concludes with a discussion in Section 5.

2. Models

2.1 *Hidden Markov Models*

A first-order Hidden Markov Model is a stochastic process consisting of a set of hidden/latent states S and observations O. The state process is assumed to be an N-state Markov chain $P(S_t|S_1, S_2, ..., S_{t-1}) = P(S_t|S_{t-1})$ with $S_t \in \{1, 2, ..., N\}$. The evolution of the hidden states across time is described by the transition matrix P, where P_{ij} is the probability of transitioning from state i to state j for all t, i.e.,

$$P(S_t = j|S_{t-1} = i) = P_{ij}.$$

The probability of being in a particular state at the first time point can be modeled using an initial state distribution π i.e. $P(S_1 = i) = \pi_i$. At each time step, we observe O_t, whose distribution only depends on the current value S_t,

$$f(O_t|O_1, ..., O_{t-1}, S_1, ..., S_t) = f(O_t|S_t).$$

Therefore, the model for a particular sequence of observations given the hidden states, $f(O_1, O_2, ..., O_T|S_1, S_2, ..., S_T)$, can be factorised as $\prod_{t=1}^{T} f(O_t|S_t)$, and the joint model of a particular sequence of hidden states and observations is equal to

$$P(O_1, O_2, ..., O_T, S_1, S_2, ..., S_T) = \pi_{S_1} f(O_1|S_1) \prod_{t=2} P_{S_{t-1},S_t} f(O_t|S_t) \tag{1}$$

The emission distribution, f, which describes how the observations are generated conditional on the states, is a function of corresponding state-specific parameters, θ_i, where θ_i can be a scalar or a vector of parameters, and it is on these parameters that we place the repulsive prior distributions proposed in this paper.

2.2 *Repulsive prior*

We make use of a point process belonging to the family of pairwise interaction point processes called the Strauss process, as described by Strauss (1975). Pairwise interaction point processes can be constructed on the parameters θ_i, as described in Section 2.1 by defining a point process with density of the form

$$h(\theta_1, \theta_2, ..., \theta_N|\xi_1, \xi_2) = \frac{1}{Z_\xi} \prod_{i=1}^{N} \phi_1(\theta_i|\xi_1) \prod_{1 \leqslant i<j \leqslant N} \phi_2(\theta_i, \theta_j|\xi_2) \tag{2}$$

with respect to a homogeneous Poisson process.

It is common to take $\phi_1(\theta_i|\xi_1 = \xi) = \xi\mathbb{I}[\theta_i \in R]$, where ξ is the intensity of the points and R is the region where θ is defined. In the case of the Strauss process, $\phi_2(\theta_i, \theta_j|\xi_2 = \{a, d\}) = a^{\mathbb{I}[\|\theta_i - \theta_j < d\|]}$. This term denotes the interaction term between the locations θ_i, θ_j for parameters α, d and norm distance $\|\cdot\|$. Parameter α ranges from 0 to 1 and controls the penalty magnitude between the points θ_i and θ_j; the smaller the α the stronger the penalty, if $\alpha = 1$ there is no penalty, and we retrieve a Poisson point process. Parameter d is the threshold such that if the distance (typically the Euclidean distance) between two components is less than d, the penalty applies. Lastly, the normalizing constant Z_ξ of Equation (2) is intractable, which makes the update of the parameter ξ challenging.

Finally, bringing together the concepts described in Sections 2.1 and 2.2, the hierarchical representation of an HMM model with a random number of states is shown in Equation (3).

$$
\begin{aligned}
&N \sim g(\cdot) \\
&O_t \sim f(O_t|\theta_{S_t}),\ t = 1, 2, ..., T \\
&\theta = (\theta_1, \theta_2, ..., \theta_N) \sim \text{StraussProcess}(\xi, \alpha, d) \\
&P(S_1 = i) = \pi_i,\ i = 1, 2, ..., N \\
&P(S_t = j|S_{t-1} = i) = P_{ij},\ i, j = 1, 2, ..., N \ \&\ t = 2, 3, ..., T \\
&P_{i.} = (P_{i1}, P_{i2}, ..., P_{iN}) \sim \text{Dirichlet}(a_1^P, a_2^P, ..., a_N^P),\ i = 1, 2, ..., N \\
&\pi = (\pi_1, \pi_2, ..., \pi_N) \sim \text{Dirichlet}(a_1^\pi, a_2^\pi, ..., a_N^\pi)
\end{aligned}
\tag{3}
$$

2.3 *Inference*

Inference is made on the parameters θ, π, P, ξ and N. Since the dimension of θ, P, π changes according to N, we employ a RJMCMC sampling algorithm that allows us to move between models with different parameter dimensions. On each iteration of the algorithm, we implement a fixed and a variable dimension move. The fixed move updates the model parameters (θ, P, π) conditional on the number of states, and the variable move updates the dimensions of the model. Finally, we update ξ with the use of the exchange algorithm (Murray et al., 2012), described in Section 2.4.

- ***Fixed dimension Moves***

We update the model parameters π, P, θ, for a fixed value N, by sampling from the corresponding posterior distributions using a Metropolis Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), since the HMM likelihood with state allocation marginalised out is not conjugate to the prior distribution(s).

- ***Variable dimension moves***

With probability 0.5, we choose between the moves Split/Combine and Birth/Death.

Split/Combine moves

Suppose that the chain at the current state has N components. In this step, we choose whether to split or combine components with probability 0.5.

In the split case, if we have a single component, then with probability one, we split that component. Otherwise, we choose uniformly one of the N components, denoted as j_* which we propose to split into j_1 and j_2, therefore proposing to split $\pi_{j_*}, P_{j_*,.}, P_{.j_*}, \theta_{j_*}$ to new model parameters $(\pi_{j_1}, P_{j_1.}, P_{.j_1}, \theta_{j_1})$ and $(\pi_{j_2}, P_{j_2.}, P_{.j_2}, \theta_{j_2})$.

The split move is accepted with probability $min\,\{1, A\}$, where

$$A = \frac{f(\{O_t\}_{t=1}^{T} \,|\, \{\pi\}_{j=1}^{N+1}, \{P_{j.}\}_{j=1}^{N+1}, \{\theta_j\}_{j=1}^{N+1})}{f(\{O_t\}_{t=1}^{T} \,|\, \{\pi\}_{j=1}^{N}, \{P_{j.}\}_{j=1}^{N}, \{\theta_j\}_{j=1}^{N})} \frac{p(\{\pi\}_{j=1}^{N+1}, \{P_{j.}\}_{j=1}^{N+1}, \{\theta_j\}_{j=1}^{N+1})p(N+1)}{p(\{\pi\}_{j=1}^{N}, \{P_{j.}\}_{j=1}^{N}, \{\theta_j\}_{j=1}^{N})p(N)} \frac{q(N+1 \to N)}{q(N \to N+1)}$$

where $q(N+1 \to N)$ and $q(N \to N+1)$ are the proposal probabilities for the transdimensional moves.

In the combine case, we choose the two components j_1 and j_2 whose distance is the smallest, and we combine them to j_*. The combine move is accepted with probability $\min\{1, A^{-1}\}$.

Birth/Death moves

The Birth/Death move is performed similarly to the Split/Combine. Likewise, if we have N components, we choose with probability 0.5 to give birth to a new component or kill an existing one.

In the birth move, we generate a new component by sampling its parameters from the prior distribution. On the other hand, for the death move, we uniformly choose a component and propose to kill it. In this case, the acceptance probability of the birth move is again $\min\{1, A\}$ whereas for the death move it is $\{1, A^{-1}\}$ with

$$A = \frac{f(\{O_t\}_{t=1}^{T} \,|\, \{\pi\}_{j=1}^{N+1}, \{P_{j.}\}_{j=1}^{N+1}, \{\theta_j\}_{j=1}^{N+1})}{f(\{O_t\}_{t=1}^{T} \,|\, \{\pi\}_{j=1}^{N}, \{P_{j.}\}_{j=1}^{N}, \{\theta_j\}_{j=1}^{N})} \frac{p(\{\pi\}_{j=1}^{N+1}, \{P_{j.}\}_{j=1}^{N+1}, \{\theta_j\}_{j=1}^{N+1})p(N+1)}{p(\{\pi\}_{j=1}^{N}, \{P_{j.}\}_{j=1}^{N}, \{\theta_j\}_{j=1}^{N})p(N)} \frac{q(N+1 \to N)}{q(N \to N+1)}$$

2.4 *Update ξ*

Parameter ξ of the Strauss process prior is updated with a Metropolis Hastings algorithm.

At each iteration, we propose ξ^* from

$$\xi_* \sim q(\xi_*|\xi) = \text{LogNormal}(\log(\xi), \tau_\xi)$$

where the current value is ξ. However, calculation of the Metropolis Hastings acceptance ratio depends on the ratio of the corresponding density (Equation (2)) at ξ and ξ_*.

$$\frac{h(\theta_1, \theta_2, ..., \theta_N|\xi, a, d)}{h(\theta_1, \theta_2, ..., \theta_N|\xi_*, a, d)} = \frac{\frac{1}{Z_{\xi_*}} \prod_{i=1}^N \xi \mathbb{I}[\theta_i \in R]}{\frac{1}{Z_\xi} \prod_{i=1}^N \xi_* \mathbb{I}[\theta_i \in R]} = \frac{Z_\xi}{Z_{\xi_*}} \frac{\prod_{i=1}^N \xi \mathbb{I}[\theta_i \in R]}{\prod_{i=1}^N \xi_* \mathbb{I}[\theta_i \in R]} \tag{4}$$

and since the fraction $\frac{Z_\xi}{Z_{\xi_*}}$ is intractable in the Strauss process prior, we perform the update using the exchange algorithm (Murray et al., 2012). The main idea of the exchange algorithm is based on the following generating process

$$\theta_{aux} \sim h(\theta_{aux}|\xi_*, a, d) = \frac{g(\theta_{aux}|\xi_*, a, d)}{Z_{\xi_*}} = \frac{\prod_{i=1}^{N'} \xi_* \mathbb{I}[\theta_{i,aux} \in R] \prod_{1 \leqslant i \leqslant j \leqslant N'} \phi_2(\theta_{i,aux}, \theta_{j,aux}|a, d)}{Z_{\xi_*}} \tag{5}$$

We generate θ_{aux} from a Strauss process based on the proposed value ξ_*. To do that, we do not require the normalized density $h(\theta_{aux}, \xi_*, a, d)$ which is intractable due to Z_{ξ_*}, on the contrary we use the unnormalised density of the Strauss process prior which is tractable, $g(\theta_{aux}|\xi_*, a, d) = \prod_{i=1}^{N'} \xi_* \mathbb{I}[\theta_{i,aux} \in R] \prod_{1 \leqslant i \leqslant j \leqslant N'} \phi_2(\theta_{i,aux}, \theta_{j,aux}|a, d)$, in order to implement the Birth and Death point process algorithm (Møller and Sørensen, 1994), which is a method for generating a random point process.

Then the proposed move to ξ_* is accepted with probability $\min(1, A)$, and as can be seen in the (6), we have overcome the calculation of the intractable ratio $\frac{Z_\xi}{Z_{\xi_*}}$.

$$A = \frac{q(\xi|\xi_*)p(\xi_*)g(\theta|\xi_*, a, d)}{q(\xi_*|\xi)p(\xi)g(\theta|\xi, a, d)} \frac{g(\theta_{aux}|\xi, a, d)}{g(\theta_{aux}|\xi, a, d)} \tag{6}$$

with $p(\xi)$ the prior distribution assigned on parameter ξ.

3. Case study 1: Muskox GPS data

We consider data on muskox movement in east Greenland analysed in Pohle et al. (2017). The data consist of 25103 hourly GPS locations, covering a period of roughly three years, giving information on the step length, L_t, which represents the distance in meters between time points $t-1$ and t, and the angle A_t, indicating the turning angle between time points $t-2$ and t, as is standard practice in GPS tracking data (Zucchini and MacDonald, 2009; Langrock, 2012; Patterson et al., 2017).

We model the step-length at time t, L_t, using a 0-inflated Gamma distribution to account for the number of 0s in the data (0.58% or 145), so that step length at time t, given state at time t is modelled as

$$f(L_t|S_t) = z_{S_t}\delta_{L_t}(0) + (1 - z_{S_t})\text{Gamma}(L_t; \mu_{S_t}, \sigma_{S_t}) \tag{7}$$

where z_{S_t} represents the probability of individuals being stationary given their corresponding state at time t and μ_{S_t} and σ_{S_t} denote the mean and standard deviation of the Gamma distribution governing the step length, conditional on state.

We model the turning angle between $t-1$ and $t+1$ happening at time point t, A_t, using a vonMises distribution with location and concentration deviation parameters m_{S_t} and k_{S_t} , respectively, so that the angle at time t, given state at time t is modelled as

$$f(A_t|S_t) = \text{vonMises}(A_t; m_{S_t}, k_{S_t}) \tag{8}$$

Therefore, the observation at time t, given state at time t, is modelled as

$$f(O_t|S_t) = f(L_t|S_t)f(A_t|S_t) \tag{9}$$

We choose to set a repulsive prior on the mean step length, $\mu_1, \mu_2, ..., \mu_N$

$$\mu = (\mu_1, \mu_2, ..., \mu_N) \sim \text{StraussProcess}(\mu_1, \mu_2, ..., \mu_N; \xi, \alpha, d) \tag{10}$$

and for comparison purposes also present results considering a standard, independent, prior distribution

$$\mu = (\mu_1, \mu_2, ..., \mu_N) \sim \text{PoissonProcess}(\mu_1, \mu_2, ..., \mu_N; \xi) \tag{11}$$

We also place the following prior distributions on the remaining model parameters

$$\begin{aligned} N &\sim \text{Uniform}\left\{1, 2, ..., N_{max}\right\} \\ \pi = (\pi_1, \pi_2, ..., \pi_N) &\sim \text{Dirichlet}(a_1^\pi, a_2^\pi, ..., a_N^\pi) \\ P_{i.} = (P_{i,1}, P_{i,2}, ..., P_{i,N}) &\sim \text{Dirichlet}(a_1^P, a_2^P, ..., a_N^P), \quad i = 1, 2, ..., N \\ z_i &\sim \text{Beta}(a^z, b^z), \quad i = 1, 2, ..., N \\ k_i &\sim \text{Uniform}(a^k, b^k), \quad i = 1, 2, ..., N \\ m_i &\sim \text{Uniform}(a^m, b^m), \quad i = 1, 2, ..., N \\ \sigma_i &\sim \text{Uniform}(a^\sigma, b^\sigma), \quad i = 1, 2, ..., N \end{aligned}$$

We run a RJMCMC algorithm for 500,000 iterations with 50,000 burn-in iterations. The details of the prior distribution choices and inference are shown in Appendix B.1, B.2 and D. We fit the model with the repulsive prior of Equation (10) and the standard prior of Equation (11) and compare our results to those obtained by Pohle et al. (2017).

The repulsive prior of Equation (10) leads to a posterior mode of four states, with posterior distribution on the number of explored states $p(2) = 0.002$, $p(3) = 0.054$, $p(4) = 0.55$, $p(5) = 0.35$, $p(6) = 0.03$, $p(7) = 0.009$, $p(8) = 0.005$, with $\sum_{i=2}^{8} p(i) = 1$, whereas the standard prior of Equation (11) leads to a posterior mode of seven states. Pohle et al. (2017) considered models with up to five states and selected the model with four states according to the integrated completed likelihood (ICL Biernacki et al., 2000) criterion. However, we note that the model with seven states actually leads to a smaller ICL (see Table 1 in Appendix section B.3), agreeing with our results in the case of a standard prior.

We plot the resulting mixture distributions of step length and angle, conditional on four states, from our model with a repulsive prior and those obtained by Pohle et al. (2017) in Figure 2.

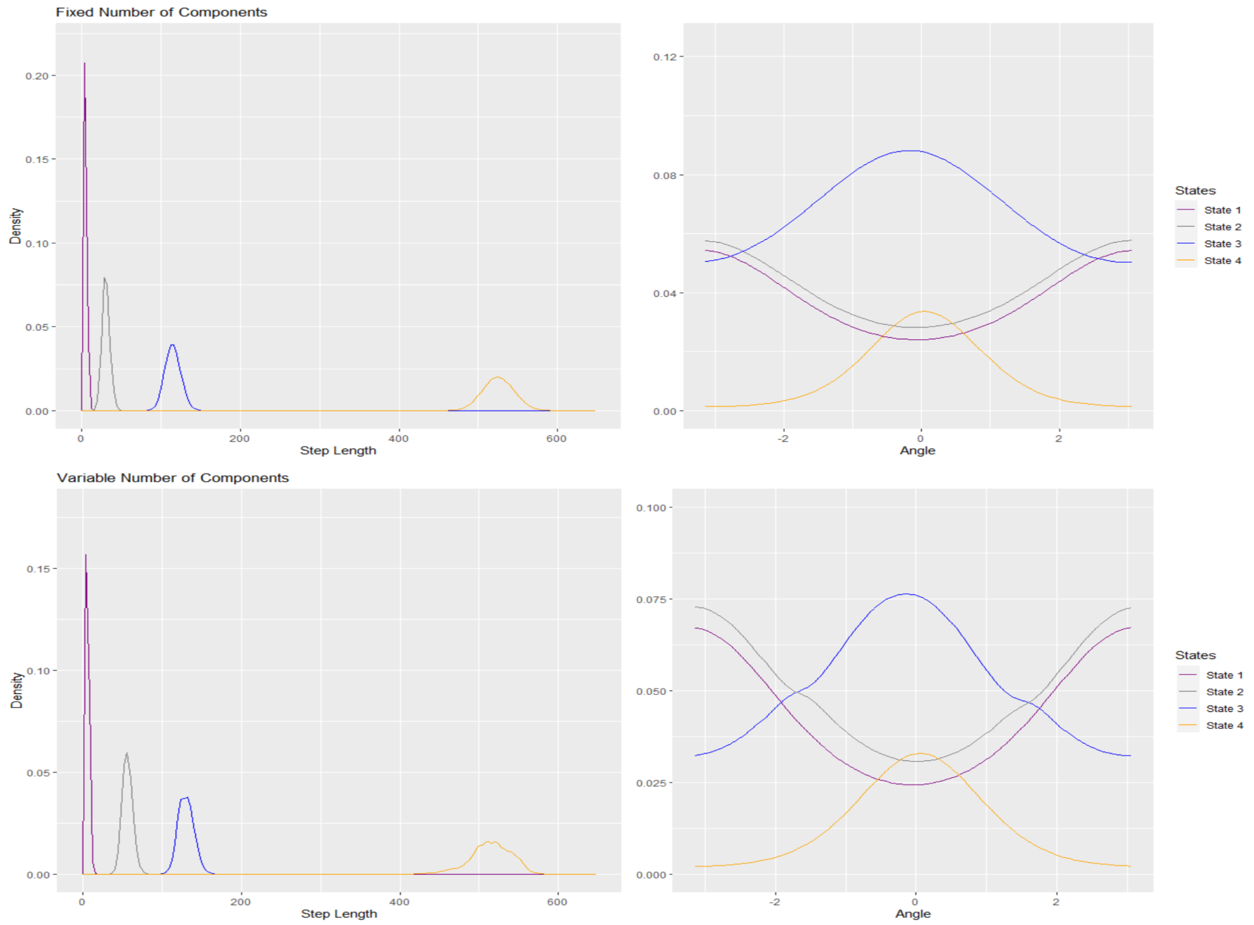

Figure 2: Averaged posterior mixture distribution of step length (left) and angle (right) for the last time point of the time series. The first row illustrates the mixture distribution of step length and angle as inferred in Pohle et al. (2017). The second row illustrates the posterior mixture distribution of the step length and angle as inferred by our model with a repulsive prior distribution.

Similarly to Pohle et al. (2017), we identify four types of step lengths, corresponding to hardly any movement (state 1), small movement (state 2), moving (state 3), and traveling (state 4). Additionally, states 3 and 4 have a much more directed movement compared to states 1 and 2 as observed from Figure 2 and discussed in Pohle et al. (2017).

Lastly, when we consider a standard prior on the mean step length parameters $\mu_1, \mu_2, ..., \mu_N$, the algorithm tends overfitting by introducing unnecessary components. We identify the following: state distributions that are almost completely overlapping and state distributions that are assigned insignificant weights (Figure 1 in Appendix Section B.3). The presence of such behaviors does not address problems frequently encountered in mixture model inference such as overfitting, where a number of state distributions are similar and do not provide additional interpretability benefits in the inference. Models exhibiting the described behaviours could be replaced by a more parsimonious model achieving similar or even better inference. Therefore, the aforementioned problems can be easily solved with the use of Strauss process prior on the model parameters as demonstrated from the results illustrated in Figure 2.

4. Case study 2 : Cape gannet acoustic data

We consider data on Cape gannets in South Africa, comprising 3078.1 seconds of acoustic time points using animal-borne devices, analyzed in Thiebault et al. (2021). The data were recorded at 22.05kHz sampling frequency and were pre-processed by downsampling the audio at 12 kHz and with a high-pass filter above 10 Hz before being segmented into 2179 intervals of 1.4 seconds (Thiebault et al., 2021). For each time segment of length 1.4 seconds we extracted 12 acoustic features based on the Mel-frequency cepstral coefficients with n measurements each (Cheng et al., 2010), which is standard practise in acoustic data analysis (Cheng et al., 2010; Ramirez et al., 2018; Noda et al., 2019; Chalmers et al., 2021). However, these 12 features are correlated with each other, and so we employ principal component analysis (PCA) to obtain a set of uncorrelated components as model inputs, instead of modelling the 12 features directly, as described in Trang et al. (2014). We consider the two first principal components (2-PC) that explain 70% of the variability of the original Mel-frequency cepstral coefficients.

We model the 2-PC using a Multivariate Normal distribution so that the observation vectors are $\underline{E_{t,i}} = \{E_{t,i,1}, E_{t,i,2}\}$, where $E_{t,i,1}$ and $E_{t,i,2}$ correspond to the measurements of the first and second PC at time t, for $t = 1, \ldots, T,\ i = 1, \ldots, n$ with

$$\underline{E_{t,i}} \sim \text{Normal}_2(\underline{E_{t,i}}; \underline{\mu_{S_t}}, \Sigma_{S_t})$$

the $\underline{\mu_{S_t}}$ corresponds to the mean vector of the 2-PC for the latent state S_t and similarly the Σ_{S_t} is the covariance matrix for the 2-PC under the latent state S_t.

We choose to place a repulsive prior on the mean parameters $\underline{\mu_1}, \ldots, \underline{\mu_N}$ so that

$$\underline{\mu} = (\underline{\mu_1}, \underline{\mu_2}, ..., \underline{\mu_N}) \sim \text{StraussProcess}(\underline{\mu_1}, \underline{\mu_2}, ..., \underline{\mu_N}; \xi, \alpha, d) \tag{12}$$

but also consider a standard prior for comparison

$$\underline{\mu} = (\underline{\mu_1}, \underline{\mu_2}, ..., \underline{\mu_N}) \sim \text{PoissonProcess}(\underline{\mu_1}, \underline{\mu_2}, ..., \underline{\mu_N}; \xi) \tag{13}$$

The prior distributions placed on the rest of parameters of the model such as the states initial probabilities, transition probabilities, covariance matrices for an unknown number of states are the following

$$
\begin{aligned}
N &\sim \text{Uniform}\,\{1, 2, ..., N_{max}\} \\
\pi = (\pi_1, \pi_2, ..., \pi_N) &\sim \text{Dirichlet}(a_1^{\pi}, a_2^{\pi}, ..., a_N^{\pi}) \\
P_i = (P_{i,1}, P_{i,2}, ..., P_{i,N}) &\sim \text{Dirichlet}(a_1^{P}, a_2^{P}, ..., a_N^{P}), \quad i = 1, 2, ..., N \\
\Sigma_i &\sim \text{Wishart}(n^{\Sigma}, \Sigma_0), \quad i = 1, 2, ..., N
\end{aligned}
$$

We run a RJMCMC algorithm for 100,000 iterations with burn-in 10,000 iterations and the details of the prior distribution choices and inference are displayed in Appendix C.1, C.2 and D. We fit the model with the repulsive prior (Equation (12)) and the standard prior (Equation (13)), with respective posterior distributions given in Appendix C.2. The results of states allocation across time within our Bayesian framework are compared to the manual allocation of Thiebault et al. (2021) in Figure 3. We focus our interpretation conditional on the two and three states, which are the models mostly visited by the RJMCMC algorith. For each model, we sample from the corresponding posterior distributions the allocation states for each time point. On the other hand, in Thiebault et al. (2021), state allocation is conducted manually with the assistance of experts by listening to the audio.

Based on the manual states allocation in Thiebault et al. (2021) three states were identified, flying, floating on water and diving, as indicated in Figure 3. There is general agreement in state allocation between our model and the manual allocation by Thiebault et al. (2021), even though in our case we have not trained our model to obtain the allocation. In the two state model, our model successfully distinguishes between the flying and floating on water states, with time points corresponding to known diving activity, according to the manual allocation, being allocated to either of the two states with no apparent pattern. In the three state model, the model also mostly identifies the diving state, and tends to allocate more

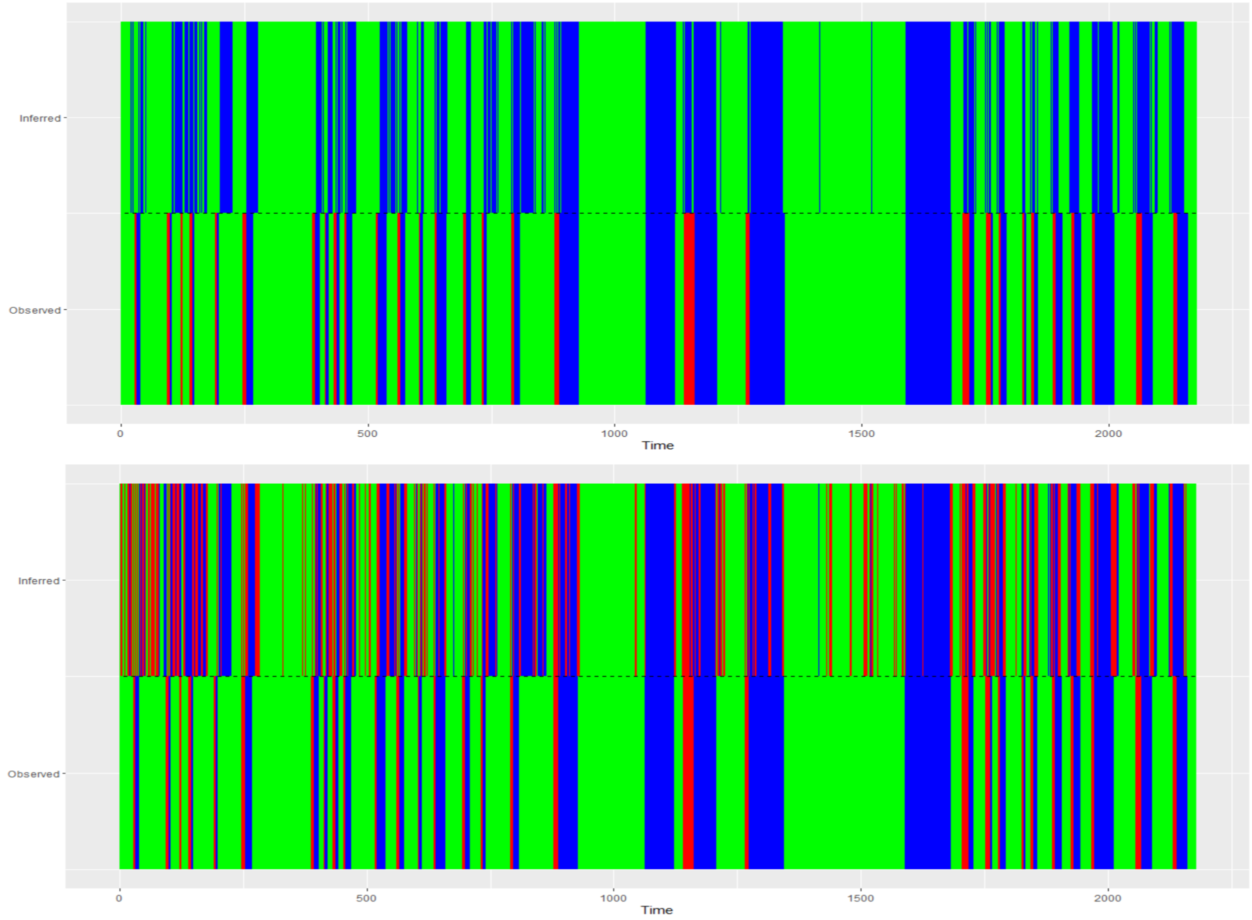

Figure 3: Comparison of the posterior classification (top row) with the observed classification (bottom row) as described in Thiebault et al. (2021).Based on the manual classification of Thiebault et al. (2021) the observed case has states as: blue color corresponds to floating on water, green on flying and red on diving. Whereas our inferred case blue color corresponds to floating on water, green on flying and red on an intermediate state which can simultaneously correspond to diving, entering to water and exiting from water

points around known diving times to that state, corresponding to the transition between flying and diving in and out of the water.

In Section C.2. of the Appendix in Figure 3 we display the uncertainty of the classification inferred by our Bayesian framework. In particular, we display the posterior allocation probabilities for each state across time. All time points are in fairly high proportion of the time ($\sim$ 25%) allocated to states other than their modal state, indicating that state allocation has a high degree of uncertainty in this case, which is expected given the noisy and multivariate nature of the data and the very short time span of 1.4 seconds between time points. Remarkably, our framework has identified the patterns associated with floating on water and flying and a third state corresponding to sounds of diving, entering to water and exiting from water, in agreement with the manual identification performed by ecologists.

In Figure 4 we display the corresponding biplot, with points coloured according to their modal state allocation on the domain defined by the 2-PC. The first PC is dominated by the 1st feature, which has a negative coefficient, whereas the second PC is a contrast between a weighted average of 5th, 8th, 10th and 11th features and 2nd feature with a loadings threshold of 0.3.

Based on Figure 4, state 1 (flying) is characterised by large scores for the first PC, whereas state 3 (floating water) by small scores, while state 2 is a strange one, because as the PC2 scores increase, the range of PC1 scores becomes wider, although always remains mid-range, indicating that when our features do not give strong evidence for the state 1 or state 3 then they get allocated to state 2.

Finally, when the standard prior is used instead (Equation (11)) the posterior distribution of the number of states is very diffuse, with 2, 3, 4, ..., 42 states almost equally supported, with the results given in the Appendix section C.2 Figure 2.

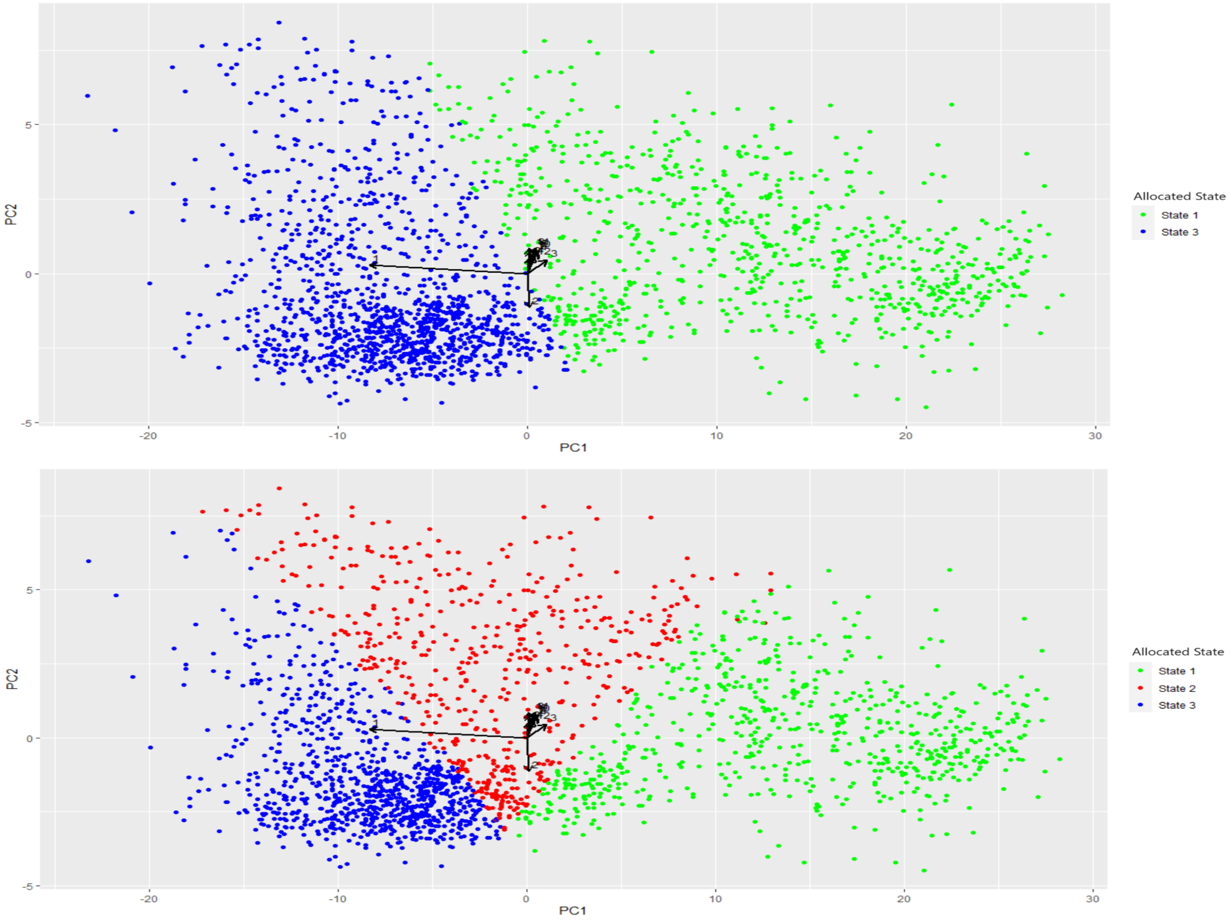

Figure 4: Biplot, with observations coloured according to their modal state allocation, in the case of two states (top row) and three states (bottom row), plotted on the domain of the first two PC.Based on the manual classification of Thiebault et al. (2021) the observed case has states as: blue color corresponds to floating on water, green on flying and red on diving. Whereas our inferred case blue color corresponds to floating on water, green on flying and red on an intermediate state which can simultaneously correspond to diving, entering to water and exiting from water

5. Conclusion

bullet points to help structure discussion

- summarizing the main findings of the study: we developed a modelling framework for inferring the number of states and corresponding distribution parameters in HMMs, building on RJMCMC and interactive point processes etc. demonstrated the model using two interesting and challenging types of ecological applications using GPS data and acoustic data. our results demonstrated the ability of our framework to yield parsimonious models with good state allocation ability in a completely unsupervised modelling framework. case studies showcase the effectiveness and practicality of our framework, with the repulsive prior penalizing the number of underlying states while effectively exploring the model sample space.
- broader implications of the work: the modelling framework readily applied to other HMM applications and to dynamic mixture models in general etc
- imitations or constraints?
- potential avenues for future research? how about cases a repulsive prior jointly to different parameters, like mean and variance, step length and angle?
- we need a strong message to close, something like the ideas presented in the paper of fitting HMMs to ecological data within a dynamic mixture modelling framework with a repulsive prior on the latent number of components provides a valuable new point of view for this widely used class of models?

References

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). Latent Markov models for longitudinal data. CRC Press.

Bartolucci, F. and Pandolfi, S. (2011). Bayesian inference for a class of latent Markov models for categorical longitudinal data. arXiv preprint arXiv:1101.0391 .

Beraha, M., Argiento, R., Møller, J., and Guglielmi, A. (2022). MCMC computations for Bayesian mixture models using repulsive point processes. Journal of Computational and Graphical Statistics **31,** 422–435.

Berkhof, J., Van Mechelen, I., and Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. Statistica Sinica pages 423–442.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. IEEE transactions on pattern analysis and machine intelligence **22,** 719–725.

Cappé, O., Moulines, E., and Rydén, T. (2009). Inference in hidden Markov models. In Proceedings of EUSFLAT conference, pages 14–16.

Cappé, O., Robert, C. P., and Rydén, T. (2003). 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,** 679–700.

Chalmers, C., Fergus, P., Wich, S., and Longmore, S. (2021). Modelling animal biodiversity using acoustic monitoring and deep learning. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–7. IEEE.

Cheng, J., Sun, Y., and Ji, L. (2010). A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. Pattern Recognition **43,** 3846–3852.

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

Duan, L. L. and Dunson, D. B. (2018). Bayesian distance clustering. arXiv preprint arXiv:1810.08537 .

Farcomeni, A. (2017). Penalized estimation in latent Markov models, with application to monitoring serum calcium levels in end-stage kidney insufficiency. Biometrical Journal

59, 1035–1046.

Gimenez, O., Bonner, S. J., King, R., Parker, R. A., Brooks, S. P., Jamieson, L. E., Grosbois, V., Morgan, B. J., and Thomas, L. (2009). Winbugs for population ecologists: Bayesian modeling using Markov chain Monte Carlo methods. Modeling demographic processes in marked populations pages 883–915.

Glennie, R., Adam, T., Leos-Barajas, V., Michelot, T., Photopoulou, T., and McClintock, B. T. (2023). Hidden Markov models: Pitfalls and opportunities in ecology. Methods in Ecology and Evolution **14,** 43–56.

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

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Huang, T., Peng, H., and Zhang, K. (2017). Model selection for gaussian mixture models. Statistica Sinica pages 147–169.

King, R. (2014). Statistical ecology. Annual Review of Statistics and Its Application **1,** 401–426.

Langrock, R. (2012). Flexible latent-state modelling of Old Faithful's eruption inter-arrival times in 2009. Australian & New Zealand Journal of Statistics **54,** 261–279.

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

McClintock, B. T., Langrock, R., Gimenez, O., Cam, E., Borchers, D. L., Glennie, R., and Patterson, T. A. (2020). Uncovering ecological state dynamics with hidden markov models. Ecology letters **23,** 1878–1903.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. Annual review of statistics and its application **6,** 355–378.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. The journal of chemical physics **21,** 1087–1092.

Møller, J. and Sørensen, M. (1994). Statistical analysis of a spatial birth-and-death process model with a view to modelling linear dune fields. Scandinavian journal of statistics pages 1–19.

Murray, I., Ghahramani, Z., and MacKay, D. (2012). MCMC for doubly-intractable distributions. arXiv preprint arXiv:1206.6848 .

Natarajan, A., De Iorio, M., Heinecke, A., Mayer, E., and Glenn, S. (2023). Cohesion and repulsion in Bayesian distance clustering. Journal of the American Statistical Association pages 1–11.

Nicol, S., Cros, M.-J., Peyrard, N., Sabbadin, R., Trépos, R., Fuller, R. A., and Woodworth, B. K. (2023). Flywaynet: A hidden semi-Markov model for inferring the structure of migratory bird networks from count data. Methods in Ecology and Evolution **14,** 265–279.

Noda, J. J., Travieso-González, C. M., Sanchez-Rodriguez, D., and Alonso-Hernández, J. B. (2019). Acoustic classification of singing insects based on MFCC/LFCC fusion. Applied Sciences **9,** 4097.

Patterson, T. A., Parton, A., Langrock, R., Blackwell, P. G., Thomas, L., and King, R. (2017). Statistical modelling of individual animal movement: an overview of key methods and a discussion of practical challenges. AStA Advances in Statistical Analysis **101,** 399–438.

Petralia, F., Rao, V., and Dunson, D. (2012). Repulsive mixtures. Advances in neural information processing systems **25,**.

Pohle, J., Langrock, R., Van Beest, F. M., and Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement.

Journal of Agricultural, Biological and Environmental Statistics **22,** 270–293.

Popov, V., Langrock, R., DeRuiter, S. L., and Visser, F. (2017). An analysis of pilot whale vocalization activity using hidden Markov models. The Journal of the Acoustical Society of America **141,** 159–171.

Quinlan, J. J., Quintana, F. A., and Page, G. L. (2021). On a class of repulsive mixture models. Test **30,** 445–461.

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

Ramirez, A. D. P., de la Rosa Vargas, J. I., Valdez, R. R., and Becerra, A. (2018). A comparative between mel frequency cepstral coefficients (MFCC) and inverse mel frequency cepstral coefficients (IMFCC) features for an automatic bird species recognition system. In 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI), pages 1–4. IEEE.

Reynolds, D. A. et al. (2009). Gaussian mixture models. Encyclopedia of biometrics **741,**.

Robert, C. P., Celeux, G., and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: a stochastic implementation. Statistics & Probability Letters **16,** 77–83.

Robert, C. P., Ryden, T., and Titterington, D. M. (2000). 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,** 57–75.

Robert, C. P. and Titterington, M. (1998). Resampling schemes for hidden Markov models and their application for maximum likelihood estimation. In Statistical Computing, volume 8, pages 145–158.

Russo, A., Farcomeni, A., Pittau, M. G., and Zelli, R. (2022). Covariate-modulated rectangular latent markov models with an unknown number of regime profiles. Statistical Modelling page 1471082X221127732.

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. Journal of applied econometrics **13,** 217–244.

Schmidt, N. M., van Beest, F. M., Mosbacher, J. B., Stelvig, M., Hansen, L. H., Nabe-Nielsen, J., and Gr ndahl, C. (2016). Ungulate movement in an extreme seasonal environment: Year-round movement patterns of high-arctic muskoxen. Wildlife Biology **22,** 253–267.

Spezia, L. (2020). Bayesian variable selection in non-homogeneous hidden Markov models through an evolutionary Monte Carlo method. Computational Statistics & Data Analysis **143,** 106840.

Stephens, M. and Phil, D. (1997). Bayesian methods for mixtures of normal distributions.

Strauss, D. J. (1975). A model for clustering. Biometrika **62,** 467–475.

Thiebault, A., Huetz, C., Pistorius, P., Aubin, T., and Charrier, I. (2021). Animal-borne acoustic data alone can provide high accuracy classification of activity budgets. Animal Biotelemetry **9,** 1–16.

Trang, H., Loc, T. H., and Nam, H. B. H. (2014). Proposed combination of PCA and MFCC feature extraction in speech recognition system. In 2014 international conference on advanced technologies for communications (ATC 2014), pages 697–702. IEEE.

Zucchini, W. and MacDonald, I. L. (2009). Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC.

Received . *Revised* . *Accepted* .