Bayesian inference for hidden Markov models under genuine multimodality with application to ecological time series

Marco A. Gallegos-Herrada, Vianey Leos-Barajas, and Jeffrey S. Rosenthal

Department of Statistical Sciences

University of Toronto

April 24, 2026

Abstract

Bayesian inference in hidden Markov models (HMMs) can be challenging due to the presence of multimodality in the likelihood function, and consequently in the joint posterior distribution, even after correcting for label switching. The parallel tempering (PT) algorithm, a state-space agumentation method, is a widely used approach for dealing with multimodal distributions. Nevertheless, standard implementation of the PT algorithm may not always be sufficient to effectively explore the high-dimensional, complex multimodal posterior distributions that arise in HMMs. In this work, we demonstrate common pitfalls when implementing the PT algorithm for HMMs, approaches to remedy them, and introduce new non-informative prior distributions that facilitate effective posterior distribution exploration. We analyse time series of blue whale dive data with two 3-state HMMs in a Bayesian framework, one of which includes a categorical covariate in the transition probability matrix to account for the effect of sound stimuli on the whale's behavior. We demonstrate how effective implementation of the modified PT algorithm for Bayesian inference leads to effective exploration of the resultant multimodal posterior distribution and how that affects inference for the underlying movement patterns of the blue whales.

1 Introduction

Discrete-time, finite-state hidden Markov models (HMMs) are a class of state-space models that assume that an observed time series is driven by a latent state process. They have been applied broadly across many domains, such as in ecology to connect observed movements to latent behaviors and capture-recapture data to estimate population dynamics [McClintock et al., 2020], in medicine to describe the dynamics of disease progression [Williams et al., 2020], in finance to classify market regimes such as bear and bullish periods [Maheu and McCurdy, 2000] and in speech recognition [Juang and Rabiner, 1991]. The natural interpretation of the hidden states as proxies for underlying

behaviors allows practitioners to gain insights into unobserved processes based on observed data over time.

Model fitting can be performed, for example, via direct maximization of the likelihood function, using the HMM-specific expectation-maximization algorithm, known as the Baum-Welch algorithm [Baum et al., 1970], or in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods, among other inferential approaches. When working with the (unnormalized) joint posterior distribution, one can implement general-purpose MCMC approaches such as Hamiltonian Monte Carlo or variants of the Metropolis–Hastings algorithm. Alternatively, Gibbs-type MCMC algorithms can be employed when the hidden states are estimated jointly with the parameters [Cappé et al., 2003]. Regardless of the model-fitting approach, a common challenge in HMMs is the permutation invariance of hidden state labels, known as label switching, where different state labelings yield the same likelihood value. As a result, HMMs likelihood exhibit multimodality. While label switching can be mitigated by applying ordering constraints to certain parameters, multimodality beyond label switching can still persist.

In the presence of multimodality in the likelihood function, [Zucchini et al., 2017] suggest using multiple starting values when implementing numerical maximization algorithms to assess whether the same maximum is consistently identified, with the goal of finding the global maximum. However, [Chen, 2023] describes how the presence of multiple modes affects construction of confidence intervals and subsequent coverage in a frequentist framework where different approaches, such as Wald or likelihood-based, can lead to different outcomes. In a Bayesian framework, uncertainty in parameter estimates is represented through the joint posterior distribution. If the likelihood function is multimodal and induces a multimodal posterior, inference can be conducted by constructing intervals for each local maximum, thereby accounting for uncertainty across all high-density regions. In particular, it is important that all regions of the parameter space with non-negligible posterior mass, after correcting for label switching, are thoroughly explored.

Failing to adequately account for the uncertainty associated with multiple local maxima in the joint posterior distribution can lead to biased, and potentially substantially different, parameter estimates. As such, multimodality should be treated as an inferential challenge, rather than only a computational one. A class of state-space augmentation methods has been developed to facilitate the exploration of multimodality in complex, high-dimensional settings. In particular, methods such as Parallel Tempering (PT) [Geyer, 1991] are specifically designed for this purpose. However, direct application of the general PT algorithm to HMMs may not work as intended due to computational challenges, particularly in the construction of the tempered distributions used to extend the state space, typically defined as π^β for $\beta \in (0, 1)$, where π is the target distribution, as well as in the specification of priors for model parameters. In both cases, the standard construction of the tempered replicas or an inadequate prior specification can lead in shifts in probability mass for the hotter replicas, that is, those with $\beta \to 0$, toward regions where parameter configurations correspond to near non-identifiability, resulting in poor mixing and a failure to adequately explore the original state space.

In this paper, we focus on implementation of the PT algorithm for HMMs within a Bayesian framework and how it can be effectively applied to this class of models. Specifically, we provide implementation guidelines for defining the tempered replicas and key components of the PT algorithm, together with prior specifications that mitigate shifts in probability mass towards regions associated with near non-identifiability in the hotter replicas, and thereby ensuring robust exploration of the joint posterior distribution. We also introduce a new non-informative prior specification for

the transition probability matrix entries when incorporating categorical covariates.

This paper is organized as follows. In Section 2, we introduce the mathematical formulation of HMMs, the PT algorithm and its key components, as well as the notion of genuine multimodality and the estimation of mode weights. In Section 3, we provide implementation guidelines for applying the PT algorithm to HMMs.In Section 4, we present the results of applying the PT algorithm to HMMs, following the proposed implementation guidelines, to ecological time series data, specifically blue whale dive data. In Section 5, we summarize our work, discuss its limitations, and outline potential directions for future research.

2 Background

In this section, we introduce the probabilistic definition of hidden Markov models, the Parallel Tempering algorithm and its main components, as well as notions of implementation effectiveness and a formal definition of genuine multimodality. We also describe how to estimate the weights of each high-density region, or mode.

2.1 Hidden Markov models

A discrete-time, finite-state hidden Markov model (HMM) is a bivariate stochastic process composed of a state process $\{S_t\}_{t=1}^T$ and an observation process $\{Y\}_{t=1}^T$ [Zucchini et al., 2017]. The observations $\{Y_t\}_{t=1}^T$ are taken to be conditionally independent given the states $\{S_t\}_{t=1}^T$ and assumed to be generated by a set of state-dependent distributions, $\{f(Y_t|S_t = n)\}_{n=1}^N$ for $N \in \mathbb{Z}^+$. The state process is taken to be a first-order Markov chain that evolves over time by an $N \times N$ transition probability matrix $\mathbf{\Gamma}$ with entries $\gamma_{ij} = \Pr(S_t = j|S_{t-1} = i)$ for $i, j \in \{1, \ldots, N\}$. Finally, the initial state distribution $\boldsymbol{\delta}$ has entries $\delta_n = \Pr(S_1 = n)$ for $n \in \{1, \ldots, N\}$. See figure 1 for a graphical representation of the structure of a basic HMM.

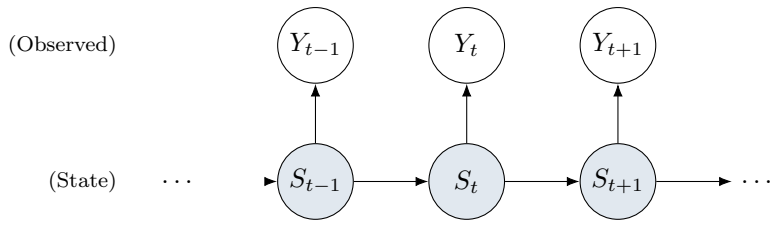


Figure 1: Graphical representation of an HMM.

The likelihood of an HMM can be written as a matrix product of the initial state distribution vector $\boldsymbol{\delta}$, transition matrix $\mathbf{\Gamma}$ and state-dependent distributions,

$$L_T = \boldsymbol{\delta}^\top \boldsymbol{D}(y_1)\mathbf{\Gamma}\boldsymbol{D}(y_2)\cdots\mathbf{\Gamma}\boldsymbol{D}(y_T)\mathbf{1}, \tag{1}$$

where $D(y_t)$ is a diagonal matrix with entries $f(y_t \mid s_t)$, for $t = 1, \ldots, T$, and $\mathbf{1}$ is an $N-$dimensional vector with 1 entries. Given the recursive nature of the likelihood, it can be effectivley evaluated with the forward algorithm [Zucchini et al., 2017].

HMMs can accommodate multivariate observation processes and incorporation of covariates in the transition probability matrix, among other possible extensions. For a k-dimensional multivariate observation process, $\{\boldsymbol{Y}_t\}_{t=1}^T$, we assume conditional contemporaneous independence such

that $f(\boldsymbol{Y}_t|S_t) = \prod_{k=1}^{K} f(Y_{tk}|S_t)$. HMMs can also be extended to incorporate covariates in the transition probability matrix via a multinomial logit link function,

$$\gamma_{ij}(z_t) = \frac{\exp\left(\eta_{ij}(z_t)\right)}{\sum_{l=1}^{n}\exp(\eta_{il}(z_t))} \qquad \eta_{ij}(z_t) = \begin{cases} 0 & \text{if } i = j \\ \alpha_0^{(ij)} + \alpha_1^{(ij)} z_t & \text{if } i \neq j \end{cases}. \tag{2}$$

2.2 Parallel Tempering

A popular algorithm used to explore multimodal probability distribution functions is the parallel tempering (PT) algorithm, also known as replica exchange or Metropolis-coupled Markov chain Monte Carlo, independently introduced in statistics [Geyer, 1991] and physics [Hukushima and Nemoto, 1996], and has been succesfully applied in the fields of biology [Müller and Bouckaert, 2019], chemistry [Lin et al., 2003], physics [Diaz et al., 2020] and machine learning [Chandra et al., 2018], [Desjardins et al., 2014]. See also [Swendsen and Wang, 1986] for an earlier related proposal. Parallel tempering is a state-space augmentation method that extends the parameter space of the target distribution π using auxiliary distributions or tempered replicas π_β, where in a Bayesian framework π is the joint posterior distribution. The tempered replicas are commonly defined as power-tempered versions of target distribution,

$$\pi_\beta(\boldsymbol{x}) \propto [\pi(\boldsymbol{x})]^\beta. \tag{3}$$

These power transformations make the modes of π less separated as the power values decrease, resulting in high-density regions that are closer together. The tempered replicas are indexed by a sequence of inverse temperatures $0 \leq \beta_M < \beta_{M-1} < \ldots < \beta_1 < \beta_0 = 1$, referred to as the inverse temperature schedule or inverse temperature ladder. The coldest replica π_{β_0} is defined to be the original distribution of interest π, whereas the hottest replica π_{β_n} is a transformation of the distribution π that can be fully explored using standard MCMC approaches. Nonetheless, depending on the structure of the distribution, tempering can be performed in different ways. For example, in a Bayesian problems, it is common practice to temper only the likelihood component, so the sequence of tempered replicas interpolates prior and posterior. Many other schemes are possible; see [Geyer and Thompson, 1995], [Paquet et al., 2009], [Cameron and Pettitt, 2014],[Surjanovic et al., 2023], for alternative constructions of tempered replicas.

The PT algorithm runs a Markov chain on the augmented state space $\mathcal{X}^{M+1}$ with the invariant distribution

$$\pi_M(\boldsymbol{x_0}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_M}) \propto \pi_{\beta_0}(\boldsymbol{x_0})\pi_{\beta_1}(\boldsymbol{x_1}) \cdots \pi_{\beta_M}(\boldsymbol{x_M}).$$

The Markov chain targeting π_M alternates between two Markovian moves: within-temperature moves and swap moves. During the within-temperature moves, each $\boldsymbol{x_i}$ is updated U times, where $\boldsymbol{x_i}$ is the state from the i-th chain targeting π_{β_i}. Performing multiple within-temperature moves between each temperature swap proposal, that is, $U > 1$, may improve efficiency; [Roberths and Rosenthal, 2026] argues that $O(\sqrt{d})$ such moves are optimal in some contexts, where d is the dimension of the state space $\mathcal{X}$. For the swap moves, there are two main schemes: stochastic even-odd (SEO) swaps and deterministic even-odd (DEO) swaps. In the SEO scheme, referred to as reversible PT, a pair of inverse temperatures β_k and β_{k+1} is chosen uniformly from all the possible pairs of adjacent inverse temperatures, and the swap proposal is constructed by swapping

the corresponding chain states $\boldsymbol{x}_i$, $\boldsymbol{x}_{i+1}$,

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k+1}, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_M).$$

and the swap move is accepted with probability

$$A_k = 1 \wedge \frac{\pi_{\beta_{k+1}}(\boldsymbol{x}_k)\pi_{\beta_k}(\boldsymbol{x}_{k+1})}{\pi_{\beta_k}(\boldsymbol{x}_k)\pi_{\beta_{k+1}}(\boldsymbol{x}_{k+1})}.$$

The DEO scheme, introduced by [Okabe et al., 2001] and referred to as non-reversible PT, differs from uniform selection of adjacent inverse-temperature pairs. Instead, it deterministically alternates between even and odd swap moves after within-temperature steps. See [Syed et al., 2022] for further details.

Tracking the exchange of chain states during swap moves serves as a metric for assessing the performance of the PT algorithm. The completion of round trips, that is, information from the coldest replica reaching the hottest replica and returning back to the coldest, indicates how effectively the information is being passed on. See figure 2 for an illustration of the occurrence of a round trip, when using $M = 3$. For reversible PT, minimizing the expected round-trip time, this is, the number of iterations it takes on average to complete a round trip, has been shown to be equivalent to optimizing swap acceptance rates under some regularity conditions [Nadler and Hansmann, 2007].

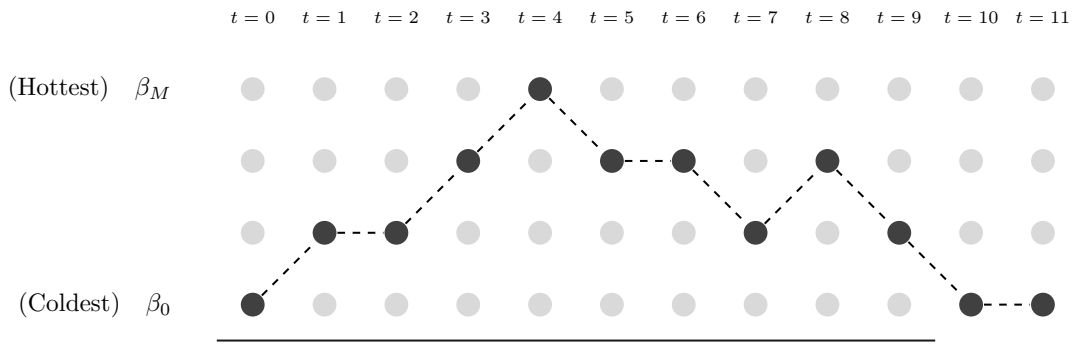


Figure 2: Illustration of a round trip occurrence at time $t = 10$

Methods for setting the inverse temperature schedule

The number of round trips strongly depends on the choice of inverse temperature levels $\beta_0, \ldots, \beta_M$. Several methods have been proposed in the literature for constructing the temperature schedule. A common approach is a geometric schedule [Kofke, 2002], consisting of a geometric progression between the coldest (β_0) and hottest (β_M) inverse temperatures. In this case, the inverse temperatures are defined as $\beta_m = \beta_0 R^m$, where $R = (\beta_M/\beta_0)^{1/M}$.

Other approaches aim to achieve a uniform swap acceptance rate across tempered replicas. [Predescu et al., 2004] showed that a geometric schedule aligns with a uniform swap acceptance rate when the heat capacity is constant across replicas. [Nadler and Hansmann, 2007] demonstrated that constant swap acceptance rates are part of the optimal solution when maximizing flow current under certain regularity conditions. Similarly, [Atchadé et al., 2011], [Roberts and Rosenthal, 2014], showed that, under strong regularity conditions on the target distribution, the optimal constant

swap acceptance rate between adjacent inverse temperatures is $A^* = 0.234$. Consistently, [Kone and Kofke, 2005] and [Lingenheil et al., 2009] obtained the same optimal rate when optimizing the diffusion of a tempered replica along the inverse temperature schedule.

Other methods focus on constructing an inverse temperature schedule that concentrates around simulation bottlenecks. [Katzgraber et al., 2006] proposed optimizing the current flow by identifying bottlenecks through measurements of local diffusivity of tempered replicas and placing temperature values more densely in those regions. Inverse-linear schedules have also been used in practice. In this approach, inverse temperatures are defined as $\beta_m = \beta_0 + (\beta_M - \beta_0)(m/M)$. [Rozada et al., 2019] showed that this strategy performs well for sparse spin-glass problems.

Genuine multimodality

The invariance to relabeling the states from the likelihood in HMMs induces a system of equivalence classes $\mathcal{P}$ on the parameter space Θ. Two elements $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$ of Θ result in belonging to the same equivalence class if there exists a permutation of the hidden state labels such that, after reordering the elements according to that permutation, they are equal. Formally,

$$\boldsymbol{\theta_1} \sim \boldsymbol{\theta_2} \iff \exists \nu \in \text{Perm}(S) \text{ such that } \boldsymbol{\theta_1} = \nu(\boldsymbol{\theta_2}),$$

where $\text{Perm}(S)$ denotes the set of all possible permutations of S items. Assigning exchangeable priors to the initial state distribution, the rows of the transition probability matrix, and the state-dependent parameters extends the relabeling invariance of the likelihood to the posterior distribution [Frühwirth-Schnatter, 2001]. Consequently, the posterior distribution $p(\cdot \mid y)$ of an N-state HMM will, at least theoretically, exhibit $N!$ symmetric modes, where a mode is defined as a local maximum in $p(\cdot \mid y)$. Grün and Leisch [2009] defines $p(\cdot \mid y)$ to be genuinely multimodal if

$$\exists\, \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathcal{M} \text{ such that } \boldsymbol{\theta_1} \neq \nu(\boldsymbol{\theta_2}) \quad \forall \text{ Perm}(S),$$

with $\mathcal{M}$ the set of modes of $p(\cdot \mid y)$. Let $\hat{\Theta} = \text{ident}(\Theta) \subseteq \Theta$ be the subset of parametrizations containing only one permutation from each equivalence class, and let $\hat{\mathcal{M}} \subseteq \mathcal{M}$ denote the subset of modes located in $\hat{\Theta}$. To conduct robust inference, it is crucial to account for the case $|\hat{\mathcal{M}}| > 1$, indicating there is more than one genuine mode. When multiple genuine modes are present in the posterior parameter space, their exploration via the PT algorithm, or any other MCMC method, must respect the corresponding mode weights, i.e. the probability mass associated with each mode. If mode $h \in \hat{\mathcal{M}}$ is well separated from the others such that it is located in a subset $\hat{\Theta}_h \subseteq \hat{\Theta}$ containing only mode h and no other genuine mode, then the weight w_h can be approximated by estimating $\mathbb{P}_{p(\cdot|y)}(\boldsymbol{\theta} \in \hat{\Theta}_h)$. Specifically, given K samples from the coldest chain, with the first B samples discarded as burn-in, the running weight estimate $\hat{w}_{h,K}$ for mode h at iteration K is computed using the samples $\boldsymbol{\theta^{(B)}}, \boldsymbol{\theta^{(B+1)}}, \ldots, \boldsymbol{\theta^{(K)}}$,

$$\hat{w}_{h,K} = \frac{1}{N - B + 1} \sum_{i=B}^{N} \mathbf{1}_{\{\boldsymbol{\theta}^{(i)} \in \hat{\Theta}_h\}}(\boldsymbol{\theta^{(i)}}).$$

In the case that the subset $\hat{\Theta}_h$ can be defined by thresholds for only one of its dimensions, this is, $\hat{\Theta}_h = \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_V) : L < \theta_s < U\}$, then $\hat{w}_{h,K}$ can be simply computed by

$$\hat{w}_{h,K} = \frac{1}{N-B+1} \sum_{i=B}^{N} \mathbf{1}_{\{L<\theta_s^{(i)}<U\}}(\theta_s^{(i)}).$$

3 Methodology

In this section, we introduce novel prior specifications for the unconstrained parameters arising from the transformation of the transition probabilities in Equation 2, which induce equal weighting on the simplex of the transition probability rows. We also propose a prior formulation for the coefficients associated with incorporating a categorical covariate into the transition probabilities that similarly ensures equal weighting on the simplex. We then detail the implementation of the PT algorithm for HMMs, including the tempering of the replicas, the construction of the inverse temperature schedule $\{\beta_0, \beta_1, \ldots, \beta_M\}$, and practical considerations for implementing the PT algorithm.

3.1 Bayesian framework for HMMs

A full specification for HMMs in a Bayesian framework requires assigning prior distributions for all parameters in the likelihood function given in Equation 1. Setting prior distributions for the parameters of the state-dependent distributions depends on the forms of $f(\cdot)$. A common practice when specifying weakly informative priors is to use exchangeable priors on the state-dependent distribution parameters across all hidden states; see [Frühwirth-Schnatter, 2001] for a detailed discussion. For the initial state distribution vector $\boldsymbol{\delta}$, a common choice of non-informative prior is a Dirichlet distribution with parameter vector $\mathbf{1}_N$. Similarly, for the entries of a time-homogeneous $\boldsymbol{\Gamma}$, each row is given prior $\boldsymbol{\gamma_{i\cdot}} \sim \mathit{Dirichlet}(\mathbf{1}_N)$, for $i = \{1, \ldots, N\}$. However, this prior cannot be used directly when the transition probabilities are reparameterized into unconstrained components, such as when incorporating covariates in Equation 2, and it is not straightforward how priors for $\boldsymbol{\alpha_0^{(ij)}}, \boldsymbol{\alpha_1^{(ij)}}$ affect the behavior of $\boldsymbol{\gamma_{i\cdot}}$.

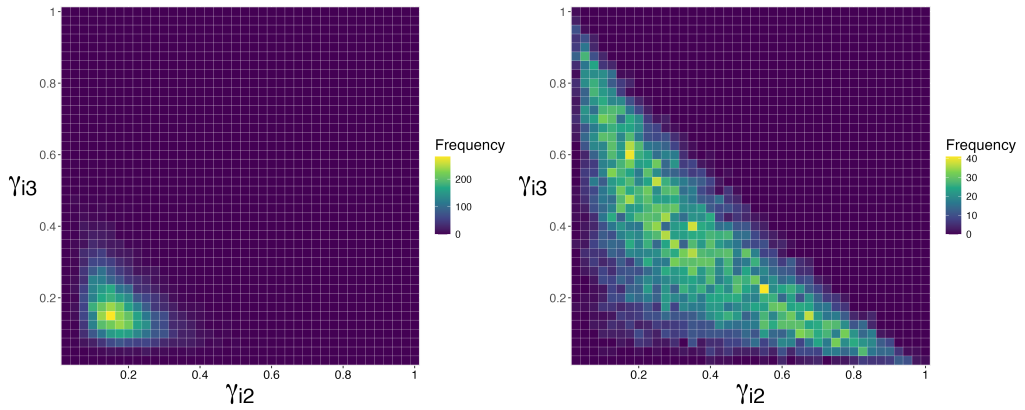
For the individual components of the vector $\boldsymbol{\alpha_0^{(ij)}}$ a normal distribution can be used, $\mathcal{N}\left(\mu_{\alpha_0^{(ij)}}, \sigma_{\alpha_0^{(ij)}}\right)$. However, specifying $\mu_{\alpha_0^{(ij)}}$ and $\sigma_{\alpha_0^{(ij)}}$ to reflect non-informative or weakly informative priors on the scale of the unconstrained parameters can induce highly skewed, informative distributions for $\boldsymbol{\gamma_{i\cdot}}$ a priori. See figure 3 for a visualization of the joint distribution corresponding to two elements of the i-th row $\boldsymbol{\gamma_{i\cdot}} = (\gamma_{i1}, \ldots, \gamma_{iN})$ induced by normal priors when $N = 3$.

We introduce a prior distribution for the unconstrained baseline components $\alpha_0^{(ij)}$ that induces a Dirichlet($\mathbf{1}_N$) distribution on $\boldsymbol{\gamma_{i\cdot}}$. Assume the following:

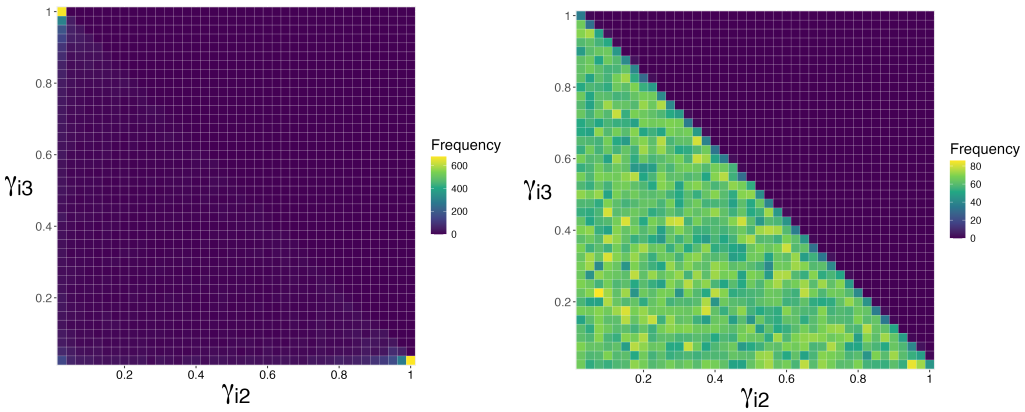
$$\begin{aligned} -\alpha_0^{(ij)} \mid \zeta_{ii} &\sim \text{Gumbel}(\zeta_{ii}, 1), \\ -\zeta_{ii} &\sim \text{Gumbel}(0, 1), \end{aligned}$$

where ζ_{ii} is a latent variable for the i-th row $\boldsymbol{\gamma}_{i\cdot}$. This leads to $(\gamma_{i1}, \ldots, \gamma_{iN}) \sim \text{Dirichlet}(\mathbf{1}_N)$, for $i = 1, \ldots, N$. For the complete derivation of the proposed prior distributions and how they induce a Dirichlet distribution, see Appendix A.1.

A similar approach can be used to induce a non-informative distribution on the transition probability rows for the coefficients resulting from incorporating a categorical covariate in the transition probability matrix, as shown in Equation 2. In particular, when the covariate is categorical, the formulation in Equation 2, originally introduced for a continuous covariate, is modified with the



(a) $\alpha_0^{(ij)} \sim N(0, \sigma = 0.5)$ (b) $\alpha_0^{(ij)} \sim N(0, \sigma = 1)$



(c) $\alpha_0^{(ij)} \sim N(0, \sigma = 3)$ (d) $(\gamma_{i1}, \gamma_{i2}, \gamma_{i3}) \sim \text{Dir}(\mathbf{1}_N)$

Figure 3: Joint distribution of $(\gamma_{i2}, \gamma_{i3})$ induced by different choices of priors for the working parameters $\alpha_0^{(ij)}$.

inclusion of auxiliary variables to avoid model non-identifiability. In this case, being $c_0, \ldots, c_L$ the possible categories of z_t, the linear formulation becomes

$$\eta_{ij}(z_t) = \alpha_0^{(ij)} + \sum_{l=1}^{L} \alpha_{1l}^{(ij)}, \mathbf{1}_{\{z_t = c_l\}}, \quad \text{for } i \neq j, \tag{4}$$

To construct a prior for the coefficient related to the covariate that induces uniform weights to the transition probability rows, we leverage the prior setting defined for the baseline components. Specifically, given the construction of the baseline components $\alpha_0^{(ij)}$ with the auxiliary variables ζ_{ii}, the following prior assumption induces equal weights for the transition probability rows for any outocome of z_t:

$$-\alpha_{1l}^{(ij)} \mid \alpha_0^{(ij)}, \zeta_{ii} \sim Gumbel(\alpha_0^{(ij)} + \zeta_{ii}, 1).$$

Additionally, we assume the $\alpha_{1l}^{(ij)}$ coefficients to be independent conditional on $\alpha_0^{(ij)}, \zeta_{ii}$. The dependence on the auxiliary variables ζ_{ii} guarantees identifiability, and from the prior formulation it follows that $(\gamma_{i1}(z_t), \ldots, \gamma_{iN}(z_t)) \sim \text{Dirichlet}(\mathbf{1}_N)$ for any outcome of z_t, which can be interpreted as no additional information is introduced about whether the treatment or latent variable affects the entries of the transition probability matrix compared to the baseline. For the detailed derivation of this prior setting, see Appendix A.1.

3.2 Parallel tempering for hidden Markov models

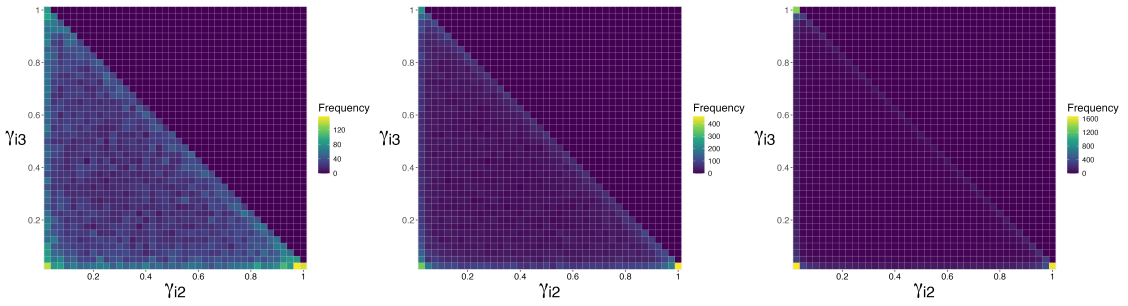
Direct application of the general PT algorithm for Bayesian inference of HMMs can lead to computational difficulties when tempering the prior distributions for the working parameters $\alpha_0^{(ij)}, \alpha_1^{(ij)}$ associated with the transition probabilities. For the normal priors and Gumbel priors introduced in Section 3.1 for $\alpha_0^{(ij)}$, tempering shifts probability mass in the induced distribution of the transition probabilities γ_{ij} toward transition probability matrices with diagonal entries close to one. In other words, it favors configurations near the identity matrix, which can produce near non-ergodic transition structures. This may lead to nearly non-identifiability, which makes the model very difficult to identify. Figure 4 illustrates how the probability mass of the induced distribution in the transition probabilities concentrate in certain regions when tempering the priors for the working parameters $\alpha_0^{(ij)}$. Moreover, when these regions have non-negligible probability mass in hotter tempered replicas, the high-density regions of the tempered replicas may not overlap, inducing broken ergodicity [Stein and Newman, 1995] and rendering the swap acceptance rate unreliable as a metric for the PT algorithm performance. To address this issue, power posteriors can be used as tempered replicas instead of power-tempered versions of π. Namely, if the distribution of interest $\pi(\boldsymbol{x})$ has the form

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}), \tag{5}$$

where $p(\boldsymbol{y} \mid \boldsymbol{x})$ is the likelihood, $\boldsymbol{y}$ are the observations, and and $p(\boldsymbol{x})$ is the joint prior distribution, the power-posteriors are constructed by tempering the likelihood component only [Brooks et al., 2011]:

$$\pi_\beta(\boldsymbol{x}|\boldsymbol{y}) \propto [p(\boldsymbol{y} \mid \boldsymbol{x})]^\beta p(\boldsymbol{x}). \tag{6}$$

See Algorithm 1 for the implementation of the reversible PT algorithm when using power-posterior as tempered replicas. For the computation of the power-posteriors for HMMs, see Appendix A.4.



(a) $\beta = 0.667$ (b) $\beta = 0.5$ (c) $\beta = 0.25$

Figure 4: Induced joint distribution of $(\gamma_{i2}, \gamma_{i3})$ when the prior distribution of working parameters α_{ij} are tempered at different inverse temperature levels when number of hidden states is $N = 3$.

For the within-temperature steps, we use the component-wise Metropolis–Hastings (CWMH) algorithm. The CWMH algorithm is a variant of Metropolis–Hastings that updates one sub-block of parameters at a time. The dimension of the parameter space in HMMs grows quickly as additional model features are incorporated or as the number of hidden states increases, making the CWHMH algorithm ideal for the within-temperature move steps. See Appendix A.3 for the implementation of the CWMH algorithm for HMMs.

For the inverse temperature schedule, the inverse temperatures are selected to achieve a swap acceptance rate of 0.234 between adjacent inverse temperatures, as discussed in Section 2.2. The hottest replica π_{β_M} for the application of the PT algorithm to HMMs is proposed to be a power posterior with $\beta_M > 0$. Specifically, since we require β_M to be small enough for the corresponding hottest tempered replica to be unimodal, or at least easily sampled using standard MCMC routines, the choice of β_M was based on candidate hottest replicas that satisfy one of these conditions. To identify an appropriate hottest inverse temperature value, we ran standard MCMC routines to explore the joint posterior distribution of the corresponding power posterior for each candidate. We selected β_M such that the resulting high-density regions of the marginal densities were sufficiently close, sometimes becoming effectively unimodal, allowing for consistent and frequent transitions between them.

Practical considerations when setting up PT

A proper scaling of the proposal distribution $q_{\beta_m}(\cdot \mid \boldsymbol{x_m})$ for the chains targeting the tempered replicas π_{β_m} plays a crutial role. There is a vast literature on how to choose an appropriate scaling for different scenarios and proposal schemes; see, for example, [Brooks et al., 2011], Chapter 4. Although the swap acceptance rates are, in principle, not directly affected by poor mixing of the Markov chains targeting the tempered replicas, and thus the inverse temperature schedule can be constructed independently of the tuning of $q_{\beta_m}(\cdot \mid \boldsymbol{x}_m)$, inadequate within-temperature mixing can substantially increase the number of iterations required to explore each tempered replica effectively. This has a direct impact on the running weight estimates, leading to higher computational costs for accurately estimating the weight of each mode.

The choice of the hottest inverse temperature β_M is another important factor. The smaller its value, the more likely π_{β_M} can be explored succesfully via standard MCMC routines, particularly when considering power posteriors of π. If the joint prior distribution accounts for the potential

Algorithm 1 PT(observations $\left(\boldsymbol{y}_{\boldsymbol{1:T_w}}^{(\boldsymbol{w})}\right)_w$, latent process $\left(z_{1:T_w}^{(w)}\right)_w$, covariate indicator c, inverse temperature schedule $\beta_{0:M}$, initialising chain values $\boldsymbol{X^0} = \left(\boldsymbol{x}_{\boldsymbol{m}}^{(\boldsymbol{0})}\right)$, number of iterations H)

Require: A within-temperature proposal mechanism for each tempered replica $q_m(\boldsymbol{x_m}, \cdot)$.

for $h = 1, \ldots, H$ **do**
  for $m = 0, \ldots M$ **do**
    Update: $\boldsymbol{x}_{\boldsymbol{m}}^{(\boldsymbol{h})} \sim q_m(\boldsymbol{x}_{\boldsymbol{m}}^{(\boldsymbol{h-1})}, \cdot)$. ▷ Within-temperature moves
  end for
  $\boldsymbol{X^h} \leftarrow \{\boldsymbol{x}_{\boldsymbol{0}}^{(\boldsymbol{h})}, \ldots, \boldsymbol{x}_{\boldsymbol{M}}^{(\boldsymbol{h})}\}$
  $k \sim U\{0, \ldots, M-1\}$ ▷ Choose uniformly a pairs of adjacent inverse temperatures

  Compute tempered replicas ▷ For definition of LogPowerPosterior$(\cdot)$, see Appendix A.4

  $\ell_k \leftarrow \text{LogPowerPosterior}\left(\left(\boldsymbol{y}_{\boldsymbol{1:T_w}}^{(\boldsymbol{w})}\right)_w, \left(z_{1:T_w}^{(w)}\right)_w, \boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{h})}, c, \beta_k\right)$
  $\ell_{k+1} \leftarrow \text{LogPowerPosterior}\left(\left(\boldsymbol{y}_{\boldsymbol{1:T_w}}^{(\boldsymbol{w})}\right)_w, \left(z_{1:T_w}^{(w)}\right)_w, \boldsymbol{x}_{\boldsymbol{k+1}}^{(\boldsymbol{h})}, c, \beta_{k+1}\right)$
  $\hat{\ell}_k \leftarrow \text{LogPowerPosterior}\left(\left(\boldsymbol{y}_{\boldsymbol{1:T_w}}^{(\boldsymbol{w})}\right)_w, \left(z_{1:T_w}^{(w)}\right)_w, \boldsymbol{x}_{\boldsymbol{k+1}}^{(\boldsymbol{h})}, c, \beta_k\right)$
  $\hat{\ell}_{k+1} \leftarrow \text{LogPowerPosterior}\left(\left(\boldsymbol{y}_{\boldsymbol{1:T_w}}^{(\boldsymbol{w})}\right)_w, \left(z_{1:T_w}^{(w)}\right)_w, \boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{h})}, c, \beta_{k+1}\right)$
  $a_k \leftarrow \min\Big(0,\ (\tilde{\ell}_i + \tilde{\ell}_j) - (\ell_i + \ell_j)\Big)$ ▷ Metropolis ratio to accept/reject swap proposal
  $U \sim \text{Unif}(0, 1)$
  if $\log(U) < a_k$ **then**
    $\boldsymbol{X^h} = \{\boldsymbol{x}_{\boldsymbol{0}}^{(\boldsymbol{h})}, \ldots, \boldsymbol{x}_{\boldsymbol{k+1}}^{(\boldsymbol{h})}, \boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{h})}, \ldots, \boldsymbol{x}_{\boldsymbol{M}}^{(\boldsymbol{h})}\}$ ▷ Swap chain states if swap accepted
  end if
end for
return $\{\boldsymbol{X^0}, \boldsymbol{X^1}, \ldots, \boldsymbol{X^H}\}$.

within-correlation present in the observed process, a natural candidate for the hottest inverse temperature is $\beta_N = 0$. This choice can simplify the selection of the hottest inverse temperature, as it corresponds to using the joint prior distribution as the hottest replica. However, when this is not the case in the context of HMMs, setting $\beta_M = 0$ will likely lead to negligible swap acceptance rates between β_M and any reasonable choice of β_{M-1}.

Additionally, the smaller β_M, the more intermediate inverse temperatures between β_0 and β_M may be required to preserve a uniform swap acceptance rate. [Nadler and Hansmann, 2007] showed that, under regularity conditions and for a large number of inverse temperatures, the exchange of information between replicas through swap proposals exhibits behaviour analogous to a simple random walk. In that scenario, the expected number of accepted swaps required to complete a single round trip is of order $O(M^2)$ [Diaconis et al., 2000] as M goes to infinity. Furthermore, [Syed et al., 2022] showed that, under the DEO scheme, this reduces to $O(M)$. Nonetheless, aiming for large M is not optimal. [Roberts and Rosenthal, 2025] showed that, under strong assumptions and in the limit as the dimension $d \to \infty$ of the state space $\mathcal{X}$, the maximum efficiency of the non-reversible PT algorithm is approximately 42% higher than that of the reversible PT algorithm.

Selecting a hottest replica π_{β_M} for which there remains a substantial gap between high-density regions can also significantly increase the computational cost of estimating the weights of each mode using running weight estimates. We suggest the use of analytical results together with exploratory Markov chain routines targeting candidate π_{β_M} to examine the geometry of the distribution before finalizing the choice of the hottest replica.

4 Application to ecological time series

We use the implementation guidelines developed in Section 3.1 to implement the reversible PT algorithm and analyze blue whale dive data using HMMs. The dataset was previously analyzed within an HMM framework in [DeRuiter et al., 2017] to quantify the impact of sound stimuli on blue whale behaviour. Inference was conducted using maximum likelihood estimation, and multiple local maxima were reported during the model fitting procedure, making this a suitable setting to test and compare our approach. We propose two Bayesian HMMs to analyse the blue whale dive data: a baseline 5-dimensional 3-state HMM with a time-homogeneous transition probability matrix, and an extension that incorporates sound stimuli during dives as a covariate in the transition probability matrix.

4.1 Bayesian HMMs for blue whale movements

Animal-borne tags were deployed to track the movements of 37 blue whales, and the collected information was processed into multiple data streams. See Supplement 2 from DeRuiter et al. [2017] for a full description of the data collection and data processing steps. While the blue whales were tracked, they were also exposed to sound stimuli, including mid-frequency military sonars and pseudo-random noise in the same frequency range in order to understand how the stimuli affects their movements, and thus behaviors.

Five data streams from the full dataset were considered for analysis: the number of lunges, dive duration, post-dive surface duration, maximum depth, and step length (in the horizontal dimension). For construction of the state-dependent distributions, we assume contemporaneous conditional independence; see [Zucchini et al., 2017] for explicit assumption formulation. Consistent with the model applied in DeRuiter et al. [2017], continuous, non-negative data streams were modeled with a Gamma distribution, parameterized in terms of the mean and standard deviation, while data streams with positive integer outcomes were modeled with a Poisson distribution:

$$\begin{aligned}
\text{Number of lunges: } & y_{t1} \mid s_t \sim Poisson(\lambda_{s_t}), \\
\text{Dive duration: } & y_{t2} \mid s_t \sim Gamma(\mu_{2s_t}, \sigma_{2s_t}), \\
\text{Surface duration: } & y_{t3} \mid s_t \sim Gamma(\mu_{3s_t}, \sigma_{3s_t}), \quad s_t = 1, 2, 3. \\
\text{Maximum depth: } & y_{t4} \mid s_t \sim Gamma(\mu_{4s_t}, \sigma_{4s_t}), \\
\text{Step length: } & y_{t5} \mid s_t \sim Gamma(\mu_{5s_t}, \sigma_{5s_t}),
\end{aligned} \tag{7}$$

For the baseline 3-state HMM, a time-homogeneous process for $\mathbf{\Gamma}$ was assumed.The reparametrization from Equation 2 was used for the transition probabilities, and the working parameters α_{ij} were estimated instead of the transition probabilities on their original scale. Weakly informative priors were chosen for the state-dependent distribution parameters; see Appendix A.5 for the explicit formulations. For the initial state distribution vector $\boldsymbol{\delta}$ it was assumed to follow a Dirichlet($\mathbf{1}_N$) as prior. For the working parameters $\alpha_0^{(ij)}$, the priors were specified as introduced in Section 3.1. In total, the baseline 3-state HMM required estimating 36 parameters, plus 3 latent variables related to the prior of $\alpha_0^{(ij)}$.

For the extended baseline model, the exogenous variable indicating the presence of sound stimuli to which whales were exposed during tracking was incorporated as a covariate in the transition probabilities via a multinomial logit transformation, as formulated in equation 2. This extension

breaks the time-homogeneity assumption of the hidden process, resulting in a non-homogeneous HMM [Zucchini et al., 2017]. The specifications for the state-dependent distributions remained the same as in the baseline model, and the same priors formulated in the Appendix A.5 were used. Since the variable indicating the presence of sound stimuli is binary, we retain the notation from Equation 2 in Section 2.1 throughout this section, without introducing auxiliary variables as in Equation 4 in Section 3.1. The priors for the covariate coefficients $\alpha_1^{(ij)}$ introduced in Section 3.1 were adopted for these parameters. In total, the extended 3-state HMM required estimation of 42 parameters, plus 3 latent variables related to the prior on the working parameters.

4.2 Parallel tempering algorithm

The inverse-temperature schedules for implementing the PT algorithm for the two proposed Bayesian HMMs were tuned to achieve a uniform swap acceptance rate across tempered replicas, as introduced in Section 2.2. Specifically, the schedules were tuned to achieve a swap acceptance rate between adjacent temperatures of 0.23. Initially, the Robbins–Monro algorithm [Robbins and Monro, 1951] was used to estimate the next hottest inverse temperature, targeting a swap acceptance rate of 0.234. However, this approach proved unstable due to the introduction of sensitive tuning parameters. As a result, the schedule was constructed by adding one inverse temperature at a time. Once the current set achieved the desired swap acceptance rate between adjacent inverse temperatures, the process continued until reaching an adequate hottest temperature. For each proposed new hottest inverse temperature, the PT algorithm was run for 50,000 iterations using the extended schedule, and the swap acceptance rate between the proposed and current hottest temperatures was monitored to determine whether the candidate should be closer to or farther from the current value. The new hottest inverse temperature was selected once this rate was sufficiently close to 0.234, specifically between 0.22 and 0.24.

For the within-temperature move steps, a single update was performed between swap proposals, that is $U = 1$. Additionally, the component-wise Metropolis-Hastings algorithm was used to marginally explore the parameter space of the tempered replicas. The proposal distributions for each implementation were tuned to achieve a step size that induced a proposal acceptance rate between 0.2 and 0.4 for each sub-block of the full parameter vector.

The parallel tempering algorithm was implemented in C++ via the Rcpp R package, which provides R functions as well as C++ classes that enable seamless integration of R and C++ [Eddelbuettel and Francois, 2011]. Additionally, parallel computing was incorporated via the RcppParallel R package, reducing computational cost per algorithm implementation by a factor of 10. The PT algorithm was implemented on the Digital Research Alliance of Canada (DRAC) servers by submitting jobs in batches of 400,000 iterations for each PT algorithm implementation. For both models, 10 PT algorithm implementations were carried out, each aiming for 2,000,000 iterations, and all PT algorithm implementations per model were initialized using random starting values. Once all batches were completed, the states from the coldest chain were extracted from all PT algorithm implementations and merged in a post-sampling process.

For both models, the first 1,000,000 iterations from the coldest chain of each PT algorithm implementation were discarded as burn-in, and the remaining 1,000,000 iterations were retained as samples. After removing burn-in, a visual inspection of the marginal posterior distributions revealed label switching in all 10 coldest chains for both models. A further visual inspection of the marginal posterior distributions based on samples from the coldest chain was conducted to identify

parameters with well-separated high-density regions that could be used to correct label switching via artificial ordering constraints. We found that the estimated mean maximum depth μ_{4n} for each hidden state n exhibited clearly separated high-density regions. Based on this, the label switching present in the posterior samples was corrected by reordering the hidden state labels according to an ordering constraint on the state-dependent distribution parameters associated with maximum dive depth for all coldest chains in both models, specifically on the estimated mean maximum depth for each hidden state:

$$\mu_{41} < \mu_{42} < \mu_{43}. \tag{8}$$

For both the baseline 3-state HMM and its extension, multiple high-density regions were identified after correcting for label switching, indicating genuine multimodality. For each model, summary statistics for the marginal posterior distributions, such as the posterior median and 95% credible intervals, were computed mode-wise. That is, for each mode, the summary statistics were calculated using only samples from that mode, without incorporating parameter estimates from other high-density regions. We relied on the auxiliary variables $\hat{w}_{h,K}$, introduced in Section 2.2, which are used to estimate the weight of a given mode h, to filter for samples corresponding to each specific mode.

Baseline 3-state HMM

For the baseline model, a 12-temperature inverse schedule was constructed with the hottest temperature set to $\beta_M = 0.019$. See Table 1 for the complete 12-temperature schedule. The job wall-clock time of each 400,000-iterations batch per PT algorithm implementation ranged between 5–15 hours, and with the job submission queue wait times, the full number of iterations concluded in four days.

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
1	0.667	0.444	0.298	0.2	0.139	0.096	0.068	0.05	0.039	0.027	0.019

Table 1: 12-inverse temperature schedule selected for the baseline 3-state HMM

Consistent swap acceptance rates across all runs, along with traceplot of the information from the coldest replica moving accross the tempered replcas were indicative of a succesful implementation of the algorithm. Additionally, a standard deviation of 4.77 in the total number of round trips across the different PT algorithm implementations suggests that the PT algorithm performed consistently across all runs. See table 2 for details. To see how the information from the coldest tempered replica moved across all the different tempered replicas for all PT algorithm implementations, see plot, included in Appendix A.2. Additionally, $\hat{R}$ and effective sample sizes were computed for the estimated parameters using samples from the coldest chain, after correcting for label switching in the post-sampling process via the ordering constraint defined in Equation 8; for these metrics, see Appendix A.8.

PT implementation ID	1	2	3	4	5	6	7	8	9	10
Total round trips	90	90	78	80	87	90	89	89	92	91

Table 2: Total number of round trips per PT algorithm implementation corresponding to the baseline 3-state HMM

After correcting for label switching across all chains, two high-density regions, denoted as

mode A and B, were identified in the coldest chain states. Running weight estimates were used for the estimation of the mode weight associated to the high-density region B. The running weight estimates for mode B, $\hat{w}_{B,K}$, were computed using the estimated values of μ_{43}, which corresponds to the mean paramater for the state-depedent distribution associated with the maximum depth per dive for hidden state $n = 3$. Table 3 shows the running weight estimates for mode B across all PT algorithm implementations and Figure 5 illustrates their convergence. The average of these estimates is 0.181 with a standard deviation of 0.01976.

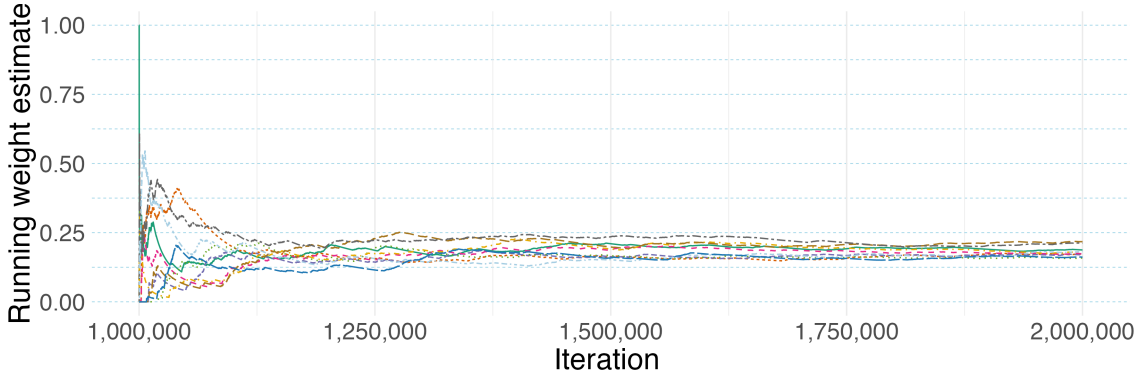


Figure 5: Running weight estimates $\hat{w}_{B,K}$ of mode B accross all 10 PT algorithm implementations

PT implementation ID	1	2	3	4	5	6	7	8	9	10
$\hat{w}_{B,K}$	0.187	0.172	0.170	0.174	0.158	0.180	0.218	0.212	0.182	0.162

Table 3: Running weight estimates $\hat{w}_{B,K}$ of mode B

The chain states from the coldest chain corresponding to the PT algorithm implementation with ID 1 were extracted for the computation of the marginal posterior distributions, as well as the 66% and 95% credible intervals. Figure 6 illustrates the estimated marginal posterior distribution of the state-dependent distribution parameters associated with the number of lunges and the entries of the transition probability matrix.

3-state HMM with sound stimuli covariate

The temperature schedule used to implement the parallel tempering algorithm for this extension resulted in a 13-temperature inverse schedule, with the hottest temperature set to $\beta_M = 0.015$. See Table 4 for the complete 13-temperature schedule. The job wall-clock time of each 400,000-iterations batch per PT algorithm implementation took about 8–20 hours, and with the job submission queue wait times, the full number of iterations concluded in six days.

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
1	0.690	0.476	0.328	0.227	0.156	0.110	0.078	0.057	0.043	0.031	0.022	0.015

Table 4: 13−inverse temperature schedule selected for the extended baseline 3-state HMM

Consistent swap acceptance rates across runs, along with traceplots showing the movement of information from the coldest replica accross the tempered replicas, indicated a successful implementation of the PT algorithm. Additionally, a standard deviation of 6.51 in the total number of

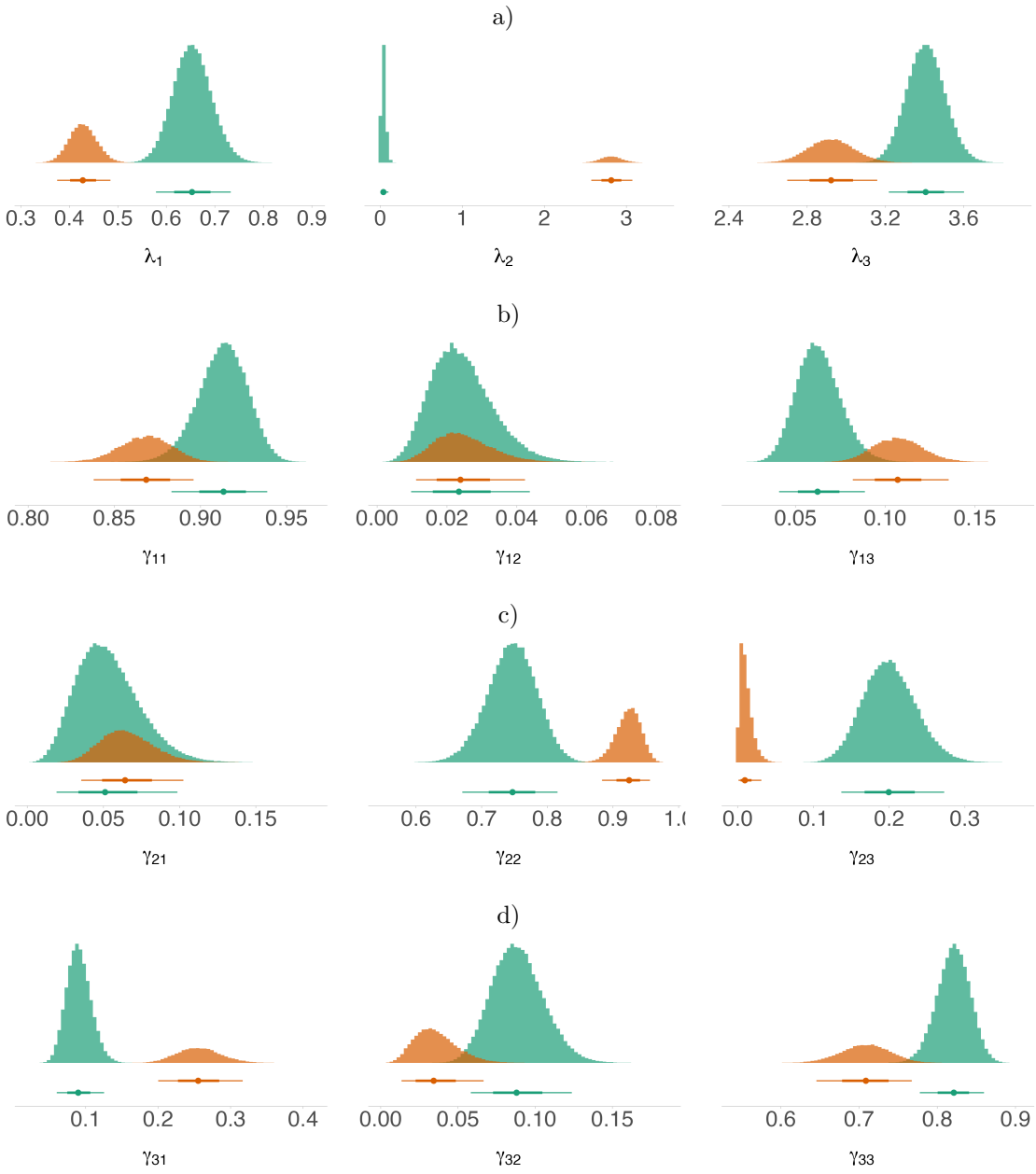


Figure 6: Histograms of the marginal posterior distributions estimated from the coldest chain. Row a) corresponds to the rate parameter λ_n associated with the number of lunges data stream for each hidden state n, while rows b)—d) correspond to the transition probabilities γ_{ij}. The color indicates the mode to which the high-density region correspond. The lines below the histograms are the mode-wise 95% credible interval. The dot in the line indicates the posterior median, whereas the thicker line inside the 95% credible interval indicates the mode-wise 66% credible interval.

round trips across the different PT algorithm implementations suggests that the PT algorithm performed consistently across all runs. Details are provided in Table 5. Plots showing how information from the coldest replica moved across all tempered replicas for all PT algorithm implementations are included in Appendix A.2. Additionally, $\hat{R}$ and effective sample sizes were computed for the estimated parameters using samples from the coldest chain, after correcting for label switching in the post-sampling process via the ordering constraint defined in Equation 8; for these metrics, see Appendix A.8.

PT implementation ID	1	2	3	4	5	6	7	8	9	10
Total round trips	75	53	60	70	57	60	60	66	63	67

Table 5: Total number of round trips per PT algorithm implementation corresponding to the extended 3-state HMM with covariates in the transition probabilities

PT implementation ID	1	2	3	4	5	6	7	8	9	10
$\hat{w}_{\tilde{B},K}$	0.0037	0.0041	0.0041	0.0027	0.0048	0.0043	0.0035	0.0044	0.0034	0.0035

Table 6: Running weight estimates $\hat{w}_{\tilde{B},K}$ for mode $\tilde{B}$

After correcting for label switching using the ordering constraints defined in equation 8, two high-density regions, denoted as mode $\tilde{A}$ and $\tilde{B}$, were identified in the coldest chain states. Running weight estimates were used for the estimation of the mode weight associated to the high-density region $\tilde{B}$. The running weight estimates for mode $\tilde{B}$, $\hat{w}_{\tilde{B},K}$, were computed using the estimated values of μ_{43}, which corresponds to the mean paramater for the state-depedent distribution associated with the maximum depth per dive for hidden state $n = 3$. Table 6 shows the running weight estimates for mode $\tilde{B}$ across all PT algorithm implementations and Figure 7 illustrates their convergence. The average of these estimates is 0.00385 with a standard deviation of 0.00061.

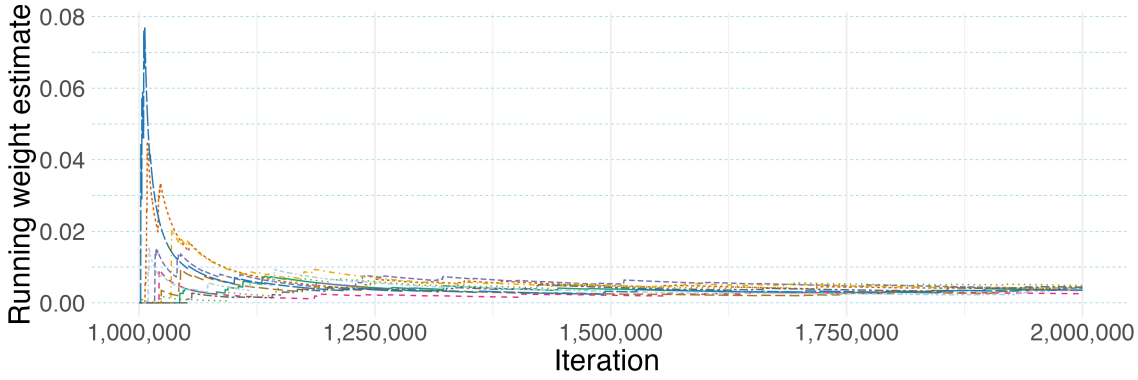


Figure 7: Running weight estimates of mode $\tilde{B}$ accross all 10 PT algorithm implementations

Following the same procedure as for the baseline 3-state HMM, the chain states from the coldest chain corresponding to the PT algorithm implementation with ID 1 were extracted for the computation of the marginal posterior distributions. Figure 8 illustrates the same marginal posterior distributions shown in Figure 6, but for the extended baseline 3-state HMM. Additionally, Figure 9 shows the marginal posterior distribution of the transition probabilities when sound stimuli is present.

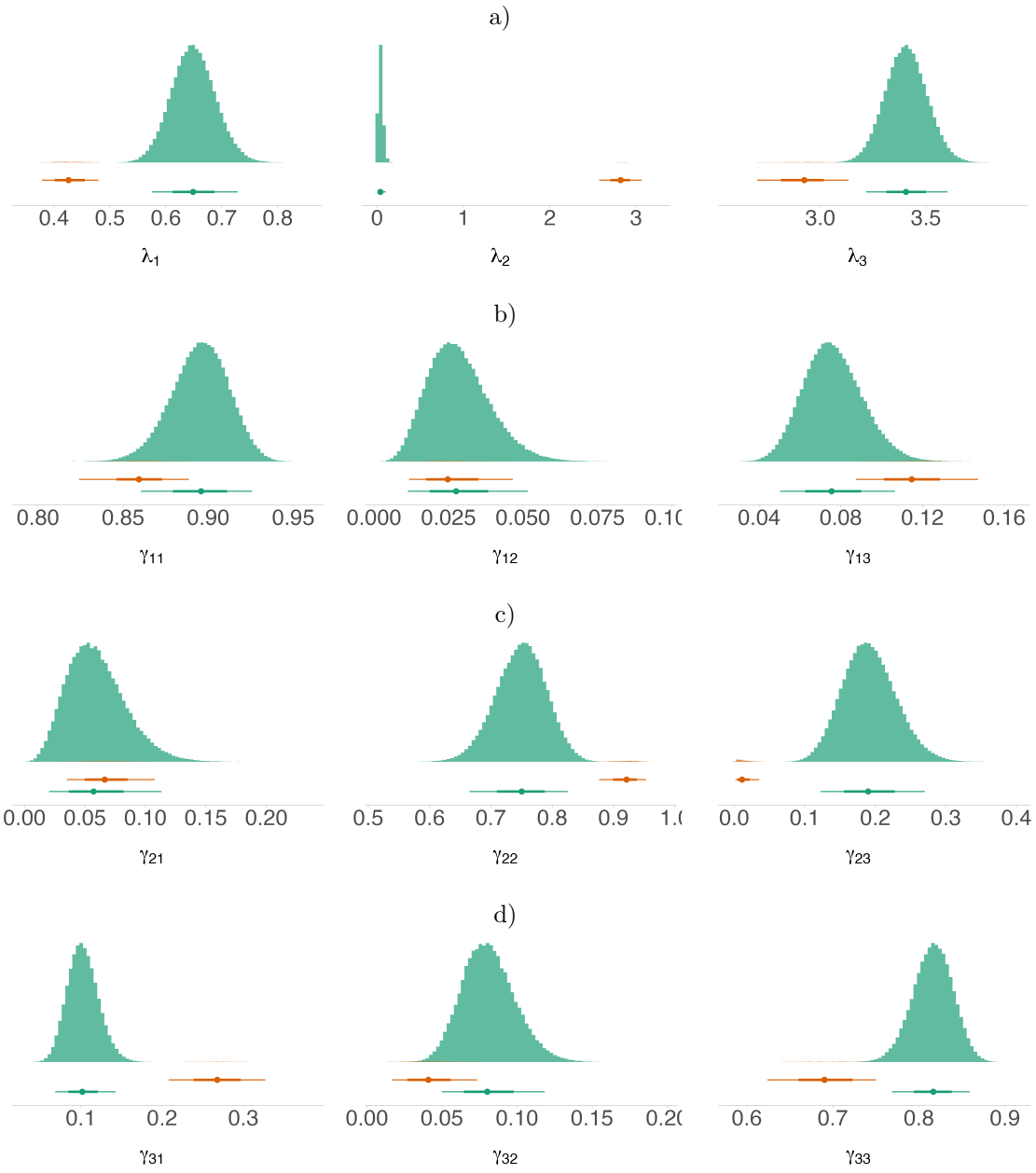


Figure 8: Histograms of the marginal posterior distributions estimated from the coldest chain. Row a) corresponds to the rate parameter λ_n associated with the number of lunges data stream for each hidden state n, while rows b)—d) correspond to the transition probabilities γ_{ij}. The color indicates the mode to which the high-density region corresponds. The lines below the histograms are the mode-wise 95% credible interval. The dot in the line indicates the posterior median, whereas the thicker line inside the 95% credible interval indicates the mode-wise 66% credible interval.

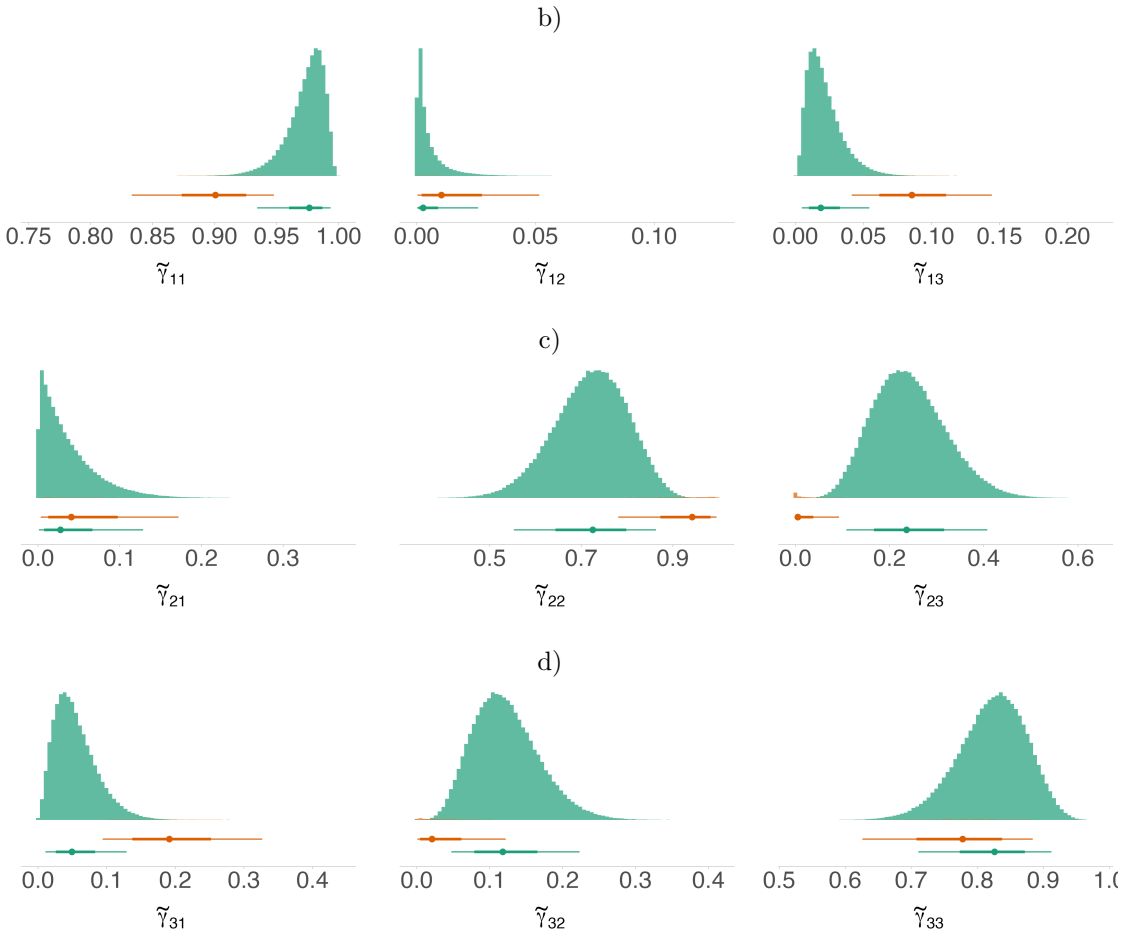


Figure 9: Histograms of the marginal posterior distributions of the transition probabilities in the presence of sound stimuli, estimated from the coldest chain related to the PT algorithm implementation with ID 1. The color indicates the mode to which the high-density region correspond. The lines below the histograms are the mode-wise 95% credible interval. The dot in the line indicates the posterior median, whereas the thicker line inside the 95% credible interval indicates the mode-wise 66% credible interval.

4.3 Results

For the baseline 3-state HMM, the 95% credible intervals of approximately two thirds of the estimated marginal state-dependent distribution parameters do not overlap. The parameters for which the 95% credible intervals overlap are $\mu_{22}, \sigma_{23}, \mu_{31}, \sigma_{31}, \mu_{32}, \sigma_{32}, \sigma_{33}, \mu_{41}, \sigma_{41}, \sigma_{52}$. For the transition probabilities, the 95% credible intervals for γ_{22} and γ_{23} do not overlap, while those for γ_{11} overlap slightly. For the initial state probability estimates, all 95% credible intervals overlap. See Table 7 for the posterior median parameter estimates and mode-wise 95% credible intervals for these parameters. For the extended 3-state model with covariates, again approximately two thirds of the 95% credible intervals for the state-dependent distribution parameters overlap across the two high-density regions, and these correspond to the same parameters as in the baseline model without covariates. For the entries of the transition probability matrix, when sound stimuli is absent, the 95% credible intervals for γ_{22} and γ_{23} do not overlap, and the same is true for γ_{31} and γ_{33}. When sound stimuli is present, the credible intervals for γ_{33} across the two high-density regions do not overlap. See Appendix A.6 for the posterior median parameter estimates and mode-wise 95% credible intervals for these parameters.

Table 7: Posterior median parameter estimates and 95% credible intervals (in parentheses) computed mode-wise from samples of the PT algorithm implementation ID 1 for the baseline 3-state HMM. Results are organized by state-dependent parameters, transition probabilities, initial state distribution, and baseline working parameters corresponding to the reparameterization introduced in Equation 2

State-dependent parameters

	State $n=1$		State $n=2$		State $n=3$	
Variable	Mode A	Mode B	Mode A	Mode B	Mode A	Mode B
λ_n	0.653 (0.579,0.733)	0.428 (0.375,0.485)	0.037 (0.004,0.098)	2.817 (2.573,3.072)	3.405 (3.217,3.601)	2.92 (2.7,3.158)
μ_{2n}	150 (142,159)	182 (171,193)	357 (324,395)	404 (385,421)	523 (510,536)	569 (553,585)
σ_{2n}	89 (81,97)	131 (120,143)	224 (196,259)	120 (107,134)	124 (115,135)	124 (113,137)
μ_{3n}	68 (63,75)	72 (66,78)	96 (90,102)	106 (101,111)	154 (147,162)	181 (172,191)
σ_{3n}	66 (60,74)	67 (61,74)	37 (32,43)	38 (35,43)	70 (64,77)	69 (62,79)
μ_{4n}	33 (30,35)	30 (28,33)	80 (69,93)	112 (108,116)	173 (167,179)	216 (209,222)
σ_{4n}	24 (22,27)	22 (20,25)	73 (61,88)	26 (23,29)	60 (55,65)	41 (37,46)
μ_{5n}	207 (190,224)	314 (291,339)	712 (666,763)	354 (315,396)	414 (383,448)	524 (475,580)
σ_{5n}	151 (134,169)	268 (244,295)	288 (255,329)	235 (198,281)	293 (264,326)	380 (333,437)

Transition probabilities and initial state distribution

	State $n=1$		State $n=2$		State $n=3$	
Variable	Mode A	Mode B	Mode A	Mode B	Mode A	Mode B
γ_{1n}	0.913 (0.883,0.939)	0.869 (0.838,0.896)	0.023 (0.01,0.044)	0.024 (0.011,0.042)	0.062 (0.041,0.089)	0.107 (0.082,0.135)
γ_{2n}	0.051 (0.019,0.099)	0.064 (0.035,0.102)	0.747 (0.671,0.816)	0.925 (0.883,0.956)	0.2 (0.137,0.273)	0.01 (0.001,0.031)
γ_{3n}	0.09 (0.06,0.126)	0.255 (0.2,0.317)	0.088 (0.058,0.124)	0.034 (0.013,0.067)	0.821 (0.778,0.86)	0.709 (0.646,0.767)
δ_n	0.179 (0.08,0.323)	0.271 (0.15,0.421)	0.17 (0.065,0.318)	0.224 (0.109,0.377)	0.641 (0.477,0.788)	0.495 (0.341,0.651)

Baseline working parameters

Variable	Mode A	Mode B
α_{12}	-3.664 (-4.555,-3.02)	-3.593 (-4.368,-3.003)
α_{13}	-2.685 (-3.128,-2.304)	-2.094 (-2.387,-1.827)
α_{21}	-2.683 (-3.694,-1.97)	-2.667 (-3.293,-2.158)
α_{23}	-1.321 (-1.771,-0.912)	-4.578 (-6.755,-3.366)
α_{31}	-2.213 (-2.64,-1.84)	-1.021 (-1.339,-0.718)
α_{32}	-2.234 (-2.674,-1.854)	-3.029 (-3.984,-2.327)

4.4 Comparing uncertainty quantification results in the presence of multiple modes

As mentioned in Section 4.1, the blue whale dive data were previously analyzed in DeRuiter et al. [2017] with inference conducted using maximum likelihood estimation. To quantify uncertainty around the point estimates, they constructed Wald-based intervals. In that study, seven data streams were used for the observable process, whereas in our analysis we used five of the seven data streams. Their baseline model was a 3-state HMM with the same assumptions as the baseline model used in this paper. During model fitting, they reported the presence of multiple local maxima.

The estimates of the state-dependent distribution parameters reported in [DeRuiter et al., 2017] were generally captured within one of the two high-density regions identified for the baseline 3-state HMM fitted here, specifically for the mode with the highest estimated weight, mode A, with the exception of five parameters: μ_{21}, μ_{32}, σ_{32}, μ_{42}, and σ_{51}. For the entries of the transition probability matrix, the credible intervals of the same mode as the state-dependent distribution parameters, mode A, captured all of the estimates presented in [DeRuiter et al., 2017]. For reference, the results from [DeRuiter et al., 2017] were reproduced following their procedure and can be found in Appendix A.7. Because the number of data streams differ between our baseline model and that of [DeRuiter et al., 2017], differences in parameter estimates were expected to occur. However, the uncertainty associated with the estimates from one of the two modes we found is not reported in DeRuiter et al. [2017], leading to different conclusions between their work and ours, namely, the mode with the lowest estimated weight, mode B.

Comparing the results from DeRuiter et al. [2017] and our full Bayesian analysis highlights the difficulties with quantifying uncertainty for our parameters in the presence of a multimodal likelihood, and subsequently posterior distribution. In a frequentist framework, Chen [2023] highlights how different confidence intervals can be constructed depending on the choice of interval estimation. However, a full Bayesian analysis is quite straightforward as we are only interested in computing or sampling from the joint posterior distribution, irrespective of how many modes it may have. To construct credible intervals from the joint posterior distribution, we can use the estimated weights for each mode, together with the uncertainty within each mode.

For the blue whale movement data, we consider the average of the estimated weights across all PT algorithm implementations, rather than relying on a single run. Specifically, we take the estimated weight of mode A to be 0.819 and of mode B as 0.181. From here, we ensure to construct credible intervals from the marginal distributions as a mixture of mode-specific credible intervals that are weighted by either 0.819 if belonging to mode A or 0.181 if belonging to mode B. Interpretation of these intervals in a Bayesian setting is also straightforward as making probabilistic statements over disjoint intervals need not be different than if only a single mode existed. The intuitive construction of credible intervals per mode to account for uncertainty, together with the estimation of mode weights to assign probabilities to each interval, highlights an advantage of the Bayesian framework over frequentist approaches. In particular, it provides a natural way to construct credible intervals within each high-density region, thereby accounting for the full uncertainty in the parameter space, which can be directly interpreted since uncertainty is expressed as a distribution.

These differences have important implications for inference. If posterior summaries are based on a single mode, conclusions may differ substantially. Using posterior means from mode A, an

observation generated from hidden state 1 would be characterized by a mean number of lunges rate of 0.653, a mean dive duration of 150 seconds, a mean post-dive surface duration of 68 seconds, a mean maximum depth of 33 meters, and a mean step length of 207 meters. In contrast, using mode B, the same hidden state would correspond to a mean number of lunges rate of 0.428, which is an average decrease of 0.225 units compared to mode A, a mean dive duration of 182 seconds, an average increase of 32 seconds, a mean post-dive surface duration of 72 seconds, an average increase of 4 seconds, a mean maximum depth of 30 meters, an average decrease of 3 meters, and a step length of 314 meters, which is an average increase of 107 meters compared to mode A. Similar discrepancies arise for results in hidden state 2 and 3. In contrast, our Bayesian credible intervals provide that state 1 is characterized by a bimodal distribution in which 0.819 percent of our lunges are generated from a distribution with estimated mean 0.653 and 0.181 percent from a distribution with estimated 0.428, with uncertainty intervals provided easily from the joint posterior distribution.

5 Discussion

In this work, we provide implementation guidelines for applying the parallel tempering (PT) algorithm to hidden Markov models (HMMs) within a Bayesian framework, with the goal of conducting inference in the presence of genuine multimodality in the posterior distribution. For the PT algorithm implementation, a temperature schedule targeting a 0.23 swap acceptance rate was proposed; the tempered replicas were defined as power-posterior versions of the target distributions, and within-temperature moves for marginal exploration are proposed to be performed using a component-wise Metropolis–Hastings algorithm. A new prior formulation was also introduced for the working parameters $\alpha_0^{(ij)}$ and the coefficients $\alpha_{1l}^{(ij)}$ associated when incorporating a categorical covariate into the transition probability matrix using Equation 2 in Section 2.1, inducing a uniform distribution on the simplex corresponding to the rows of the transition probability matrix. The developed framework was applied to ecological time-series data consisting of blue whale dives from a group of blue whales that were exposed to sound stimuli. For each dive, five variables summarizing blue whale movement were recorded: number of lunges, dive duration, surface duration, maximum depth, and step length. These data streams were modeled using a baseline 3-state HMM and an extended version that incorporated the occurrence of sound stimuli in the entries of the transition probability matrix. For both models, the PT algorithm was implemented following the guidelines developed in Section 3, with 10 independent PT algorithm runs conducted for each model. Consistent swap acceptance rates across runs, along with information from the coldest replica moving back and forth to the hottest replica, provided evidence of a successful implementation. Label switching correction was conducted in a post-sampling process for the coldest chains for all PT algorithm implementations, revealing genuine multimodality in the joint posterior distribution. Specifically, two high-density regions were identified for both models. Consistent running weight estimates across runs further indicated that the PT algorithm successfully explored both modes.

The prior distribution introduced in Section 3.1 can be extended to multiple exogenous categorical covariates; however the number of parameters to estimate does not increase linearly. Instead, it grows with the total number of combinations of outcomes across the multiple categorical covariates. Specifically, if we have C categorical covariates $\{Z_t^{(1)}\}, \ldots, \{Z_t^{(C)}\}$ with corresponding number of outcomes $O_1, \ldots, O_C$, then maintaining the provided Gumbel prior requires constructing a new latent categorical covariate whose values correspond to all possible combinations of the outcomes.

This results in $\prod_{i=1}^{C} O_i$ outcomes for the new covariate, rather than $\sum_{i=1}^{C} O_i$, which would be the case when using popular prior assumption approaches for covariates.

The inverse temperature schedules were constructed to target a swap acceptance rate of 0.234 across all adjacent temperatures. While the optimal inverse temperature schedule conveys constant swap acceptance rates, it has not been verified theoretically that 0.234 is the correspond to the value for the optimal solution in the case of hidden Markov models. Consequently, the optimal swap acceptance rate for hidden Markov models remains an open research question.

Artificial identifiability constraints, this is, ordering constraints, were used to correct for label switching in a post-sampling process. Since the high-density regions were well separated, this approach was suitable for our models. However, this situation is rare. [Jasra et al., 2005] highlights that finding appropriate identifiability constraints in multivariate problems is often nearly impossible. There are cases in which the posterior distribution is genuinely multimodal, and no identifiability constraints can successfully isolate both major and minor modes (see, e.g., Grün and Leisch [2009]). There is an extensive literature on relabelling methods for addressing label switching. Classical approaches include label-invariant loss functions Celeux et al. [2000], the pivotal reordering algorithm Marin et al. [2005], Kullback–Leibler divergence–based algorithms Stephens [2000], and the equivalence classes representatives (ECR) algorithm Papastamoulis and Iliopoulos [2010]. More recent approaches are based on optimal transport Monteiller et al. [2019]. An extensive review of artificial identifiability constraints, relabelling algorithms, and label-invariant loss functions is provided in Jasra et al. [2005]. Additionally, Papastamoulis [2016] compiles several of these methods into the R package label.switching, which can be used to assess whether there are differences compared to the artificial ordering constraints applied in this problem.

For the model with covariates, the weight of one of the two modes is very small, ranging between 0.00268 and .00482 across all PT algorithm implementations. Nevertheless, we believe this highlights a strength of the parallel tempering algorithm, which is its ability to capture uncertainty across all high-density regions, even those with relatively small weights. The results for both the baseline and extended model demonstrate that accounting for uncertainty associated with either local or global modes can lead to substantially different uncertainty quantification, with implications for inference.

Our results show that accounting for the uncertainty associated with multiple high-density regions in the posterior parameter space, when present, can lead to substantially different parameter estimates when summarizing the posterior within each mode independently. Furthermore, since uncertainty is represented as a distribution in a Bayesian framework, this allows us to account for uncertainty across all high-density regions by assigning each mode a weight using the running weight estimates introduced in Section 2.2. This highlights an advantage of the Bayesian approach over frequentist methods, as we can construct credible intervals within each high-density region, thereby accounting for the full uncertainty in the parameter space. This reinforces our belief that genuine multimodality in the joint posterior distribution should be treated as an inferential challenge in addition to a computational one. Nonetheless, it is important to note that the uncertainty captured by local maxima does not necessarily correspond to meaningful information for inference. Model validation is essential to determine whether such multimodality arises as a computational artefact or reflects substantively relevant uncertainty. The development of diagnostic and validation methods for HMMs under genuine multimodality is a crucial next step and a potential direction for future research.

Acknowledgements

During this project, Marco A. Gallegos-Herrada received funding from the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI). He also received support from the Canadian Statistical Sciences Institute (CANSSI). Vianey Leos-Barajas and Jeffrey S. Rosenthal were financially supported by the Natural Sciences and Engineering Research Council of Canada. Computational resources were provided by the Digital Research Alliance of Canada (DRAC). We give thanks to Stacy De Ruiter for many conversations that proved very useful to development of the paper.

References

Yves F. Atchadé, Gareth O. Roberts, and Jeffrey S. Rosenthal. Towards optimal scaling of metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21(4):555–568, October 2011. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-010-9192-1. URL http://link.springer.com/10.1007/s11222-010-9192-1.

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. ISSN 0003-4851. URL https://www.jstor.org/stable/2239727. Publisher: Institute of Mathematical Statistics.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, New York, May 2011. ISBN 978-0-429-13850-8. doi: 10.1201/b10905.

Ewan Cameron and Anthony Pettitt. Recursive Pathways to Marginal Likelihood Estimation with Prior-Sensitivity Analysis. *Statistical Science*, 29(3), August 2014. ISSN 0883-4237. doi: 10.1214/13-STS465. URL http://arxiv.org/abs/1301.6450. arXiv:1301.6450 [stat].

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, August 2003. ISSN 1369-7412. doi: 10.1111/1467-9868.00409. URL https://doi.org/10.1111/1467-9868.00409.

G Celeux, MA Hurn, and CP Robert. Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association*, 95:957–970, 2000. ISSN 1537-274X.

Rohitash Chandra, Konark Jain, Ratneel V. Deo, and Sally Cripps. Langevin-gradient parallel tempering for Bayesian neural learning, November 2018. URL http://arxiv.org/abs/1811.04343. arXiv:1811.04343 [cs].

Yen-Chi Chen. Statistical Inference with Local Optima. *Journal of the American Statistical Association*, 118(543):1940–1952, July 2023. ISSN 0162-1459. doi: 10.1080/01621459.2021.2023550. URL https://doi.org/10.1080/01621459.2021.2023550. _eprint: https://doi.org/10.1080/01621459.2021.2023550.

Stacy L. DeRuiter, Roland Langrock, Tomas Skirbutas, Jeremy A. Goldbogen, John Calambokidis, Ari S. Friedlaender, and Brandon L. Southall. A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *The Annals of Applied Statistics*, 11(1):362–392, March 2017. ISSN 1932-6157, 1941-7330. doi: 10.1214/16-AOAS1008. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-11/issue-1/A-multivariate-mixed-hidden-Markov-model-for-blue-whale-behaviour/10.1214/16-AOAS1008.full.

Guillaume Desjardins, Heng Luo, Aaron Courville, and Yoshua Bengio. Deep Tempering, October 2014. URL http://arxiv.org/abs/1410.0123. arXiv:1410.0123 [cs].

Persi Diaconis, Susan Holmes, and Radford M. Neal. Analysis of a nonreversible Markov chain sampler. *The Annals of Applied Probability*, 10(3), August 2000. ISSN 1050-5164. doi: 10.1214/aoap/1019487508. URL https://projecteuclid.org/journals/annals-of-applied-probability/volume-10/issue-3/Analysis-of-a-nonreversible-Markov-chain-sampler/10.1214/aoap/1019487508.full.

A. Diaz, C. A. Argüelles, G. H. Collin, J. M. Conrad, and M. H. Shaevitz. Where Are We With Light Sterile Neutrinos? *Physics Reports*, 884:1–59, November 2020. ISSN 03701573. doi: 10.1016/j.physrep.2020.08.005. URL http://arxiv.org/abs/1906.00045. arXiv:1906.00045 [hep-ex].

Dirk Eddelbuettel and Romain Francois. Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, 40:1–18, April 2011. ISSN 1548-7660. doi: 10.18637/jss.v040.i08. URL https://doi.org/10.18637/jss.v040.i08.

Sylvia Frühwirth-Schnatter. Markov chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association*, 96(453):194–209, March 2001. ISSN 0162-1459. doi: 10.1198/016214501750333063. URL https://doi.org/10.1198/016214501750333063. _eprint: https://doi.org/10.1198/016214501750333063.

Charles Geyer and Elizabeth Thompson. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *Journal of The American Statistical Association - J AMER STATIST ASSN*, 90:909–920, September 1995. doi: 10.1080/01621459.1995.10476590.

Charles J. Geyer. Markov Chain Monte Carlo Maximum Likelihood. Interface Foundation of North America, 1991. URL http://conservancy.umn.edu/handle/11299/58440. Accepted: 2010-02-24T20:38:06Z.

Bettina Grün and Friedrich Leisch. Dealing with label switching in mixture models under genuine multimodality. *Journal of Multivariate Analysis*, 100(5):851–861, May 2009. ISSN 0047-259X. doi: 10.1016/j.jmva.2008.09.006. URL https://www.sciencedirect.com/science/article/pii/S0047259X08001929.

Koji Hukushima and Koji Nemoto. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, June 1996. ISSN 0031-9015, 1347-4073. doi: 10.1143/JPSJ.65.1604. URL http://journals.jps.jp/doi/10.1143/JPSJ.65.1604.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1), February 2005. ISSN 0883-4237. doi: 10.1214/088342305000000016. URL https://projecteuclid.org/journals/statistical-science/volume-20/issue-1/Markov-Chain-Monte-Carlo-Methods-and-the-Label-Switching-Problem/10.1214/088342305000000016.full.

B. H. Juang and L. R. Rabiner. Hidden Markov Models for Speech Recognition. *Technometrics*, 33 (3):251–272, August 1991. ISSN 0040-1706. doi: 10.1080/00401706.1991.10484833. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1991.10484833. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1991.10484833.

Helmut G. Katzgraber, Simon Trebst, David A. Huse, and Matthias Troyer. Feedback-optimized parallel tempering Monte Carlo. *Journal of Statistical Mechanics: Theory and Experiment*, 2006 (03):P03018–P03018, March 2006. ISSN 1742-5468. doi: 10.1088/1742-5468/2006/03/P03018. URL http://arxiv.org/abs/cond-mat/0602085. arXiv:cond-mat/0602085.

David A. Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *The Journal of Chemical Physics*, 117(15):6911–6914, October 2002. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.1507776. URL https://pubs.aip.org/jcp/article/117/15/6911/446885/On-the-acceptance-probability-of-replica-exchange.

Aminata Kone and David A. Kofke. Selection of temperature intervals for parallel-tempering simulations. *Journal of Chemical Physics*, 122:206101–206101, May 2005. ISSN 0021-9606. doi: 10.1063/1.1917749. URL https://ui.adsabs.harvard.edu/abs/2005JChPh.122t6101K. ADS Bibcode: 2005JChPh.122t6101K.

Chai-Yu Lin, Chin-Kun Hu, and Ulrich H. E. Hansmann. Parallel tempering simulations of HP-36. *Proteins*, 52(3):436–445, August 2003. ISSN 1097-0134. doi: 10.1002/prot.10351.

Martin Lingenheil, Robert Denschlag, Gerald Mathias, and Paul Tavan. Efficiency of exchange schemes in replica exchange. *Chemical Physics Letters*, 478(1):80–84, August 2009. ISSN 0009-2614. doi: 10.1016/j.cplett.2009.07.039. URL https://www.sciencedirect.com/science/article/pii/S0009261409008604.

John M. Maheu and Thomas H. McCurdy. Identifying Bull and Bear Markets in Stock Returns. *Journal of Business & Economic Statistics*, 18(1):100–112, 2000. ISSN 0735-0015. doi: 10.2307/1392140. URL https://www.jstor.org/stable/1392140. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Jean-Michel Marin, Kerrie Mengersen, and Christian Robert. Bayesian Modelling and Inference on Mixtures of Distributions. *Handbook of Statistics*, 25, December 2005. ISSN 9780444515391. doi: 10.1016/S0169-7161(05)25016-2.

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. ISSN 1461-0248. doi: 10.1111/ele.13610. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.13610. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.13610.

Pierre Monteiller, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, Justin M Solomon, and Mikhail Yurochkin. Alleviating Label Switching with Optimal Transport. October 2019.

Nicola F. Müller and Remco R. Bouckaert. Adaptive parallel tempering for BEAST 2, April 2019. URL http://biorxiv.org/lookup/doi/10.1101/603514.

Walter Nadler and Ulrich H. E. Hansmann. Generalized ensemble and tempering simulations: A unified view. *Physical Review E*, 75(2):026109, February 2007. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.75.026109. URL https://link.aps.org/doi/10.1103/PhysRevE.75.026109.

Tsuneyasu Okabe, Masaaki Kawata, Yuko Okamoto, and Masuhiro Mikami. Replica-exchange Monte Carlo method for the isobaric–isothermal ensemble. *Chemical Physics Letters*, 335(5): 435–439, March 2001. ISSN 0009-2614. doi: 10.1016/S0009-2614(01)00055-0. URL https://www.sciencedirect.com/science/article/pii/S0009261401000550.

Panagiotis Papastamoulis. label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs. *Journal of Statistical Software*, 69(Code Snippet 1), 2016. ISSN 1548-7660. doi: 10.18637/jss.v069.c01. URL http://arxiv.org/abs/1503.02271. arXiv:1503.02271 [stat].

Panagiotis Papastamoulis and George Iliopoulos. An Artificial Allocations Based Solution to the Label Switching Problem in Bayesian Analysis of Mixtures of Distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331, January 2010. ISSN 1061-8600. doi: 10.1198/jcgs.2010.09008. URL https://doi.org/10.1198/jcgs.2010.09008. _eprint: https://doi.org/10.1198/jcgs.2010.09008.

Ulrich Paquet, Ole Winther, and Manfred Opper. Perturbation Corrections in Approximate Inference: Mixture Modelling Applications. *Journal of Machine Learning Research*, 10(43):1263–1304, 2009. ISSN 1533-7928. URL http://jmlr.org/papers/v10/paquet09a.html.

Cristian Predescu, Mihaela Predescu, and Cristian V. Ciobanu. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *The Journal of Chemical Physics*, 120(9):4119–4128, March 2004. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.1644093. URL http://arxiv.org/abs/physics/0310101. arXiv:physics/0310101.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177729586. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-3/A-Stochastic-Approximation-Method/10.1214/aoms/1177729586.full.

Gareth O. Roberths and Jeffrey S. Rosenthal. Optimal balancing of different MCMC updates, with an application to tempering algorithms. In preparation. 2026.

Gareth O. Roberts and Jeffrey S. Rosenthal. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *The Annals of Applied Probability*, 24(1), February 2014. ISSN 1050-5164. doi: 10.1214/12-AAP918. URL http://arxiv.org/abs/1401.3559. arXiv:1401.3559 [math].

Gareth O. Roberts and Jeffrey S. Rosenthal. Quantifying the Speed-Up from Non-Reversibility in MCMC Tempering Algorithms, January 2025. URL http://arxiv.org/abs/2501.16506. arXiv:2501.16506 [math].

Ignacio Rozada, Maliheh Aramon, Jonathan Machta, and Helmut G. Katzgraber. Effects of setting temperatures in the parallel tempering Monte Carlo algorithm. *Physical Review E*, 100(4): 043311, October 2019. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.100.043311. URL https://link.aps.org/doi/10.1103/PhysRevE.100.043311.

D. L. Stein and C. M. Newman. Broken ergodicity and the geometry of rugged landscapes. *Physical Review E*, 51(6):5228–5238, June 1995. doi: 10.1103/PhysRevE.51.5228. URL https://link.aps.org/doi/10.1103/PhysRevE.51.5228.

Matthew Stephens. Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):795–809, 2000. URL http://www.jstor.org/stable/2680622.

Nikola Surjanovic, Saifuddin Syed, Alexandre Bouchard-Côté, and Trevor Campbell. Parallel Tempering With a Variational Reference, January 2023. URL http://arxiv.org/abs/2206.00080. arXiv:2206.00080 [stat].

Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters*, 57(21):2607–2609, November 1986. doi: 10.1103/PhysRevLett.57.2607. URL https://link.aps.org/doi/10.1103/PhysRevLett.57.2607. Publisher: American Physical Society.

Saifuddin Syed, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Non-Reversible Parallel Tempering: a Scalable Highly Parallel MCMC Scheme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):321–350, April 2022. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12464. URL http://arxiv.org/abs/1905.02939. arXiv:1905.02939 [stat].

Jonathan P. Williams, Curtis B. Storlie, Terry M. Therneau, Clifford R. Jack Jr, and Jan Hannig. A Bayesian Approach to Multistate Hidden Markov Models: Application to Dementia Progression. *Journal of the American Statistical Association*, 115(529):16–31, January 2020. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2019.1594831. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2019.1594831.

Walter Zucchini, Iain L. MacDonald, and Roland Langrock. *Hidden Markov Models for Time Series: An Introduction Using R, Second Edition*. Chapman and Hall/CRC, New York, 2 edition, December 2017. ISBN 978-1-315-37248-8. doi: 10.1201/b20790.

A Appendix

A.1 Derivation of priors for $\alpha_0^{(ij)}$ and $\alpha_1^{(ij)}$

The rationale for constructing the priors for the components $\alpha_0^{(ij)}$ presented in Section 3.1 is as follows. If we consider

$$\gamma_{ij} = \frac{-\log(U_{ij})}{\sum_{k=1}^{N} -\log(U_{ik})}, \quad j = 1, \ldots, N,$$

with $U_{ij} \sim U(0,1)$, we have that $(\gamma_{i1}, \ldots, \gamma_{iN})$ follows a Dirichlet($\mathbf{1}_N$) distribution.Let $\zeta_{ij} = \log(-\log(U_{ij}))$. Then, we have

$$\log\left(\frac{\gamma_{ij}}{\gamma_{ii}}\right) = \zeta_{ij} - \zeta_{ii}, \quad i \neq j. \tag{9}$$

From the reparametrization of the transition probabilities shown in Equation 2, we have that $\alpha_0^{(ij)} = \zeta_{ij} - \zeta_{ii}$ and $-\zeta_{ij} \sim \text{Gumbel}(0,1)$, $i \neq j$. Assuming $\zeta_{ii} \perp \zeta_{ij}$, $i \neq j$, we have

$$\begin{aligned} -\alpha_{ij} \mid \zeta_{ii} &\sim \text{Gumbel}(\zeta_{ii}, 1), \\ -\zeta_{ii} &\sim \text{Gumbel}(0,1), \end{aligned}$$

which leads to $(\gamma_{i1}, \ldots, \gamma_{iN}) \sim \text{Dirichlet}(\mathbf{1}_N)$, for $i = 1, \ldots, N$. A consequence this prior seting is that the marginal distribution of $\alpha_0^{(ij)}$ is Logistic$(0,1)$, with marginal distributions of $\alpha_0^{(ij)}$ correlated. The dependence on the latent component ζ_{ii} guarantees identifiability, as this induces a one-to-one transformation between the baseline components and the transition probabilities in the original scale.

To construct a prior that induces equal weights over the simplex corresponding to the transition probability rows when a categorical covariate z_t is incorporated, we adapt the prior formulation from Equation 2 to the categorical case described in Equation 4. Specifically, we leverage the prior specification for $\alpha_0^{(ij)}$. Given the construction of priors for the parameters $\alpha_0^{(ij)}$, we define

$$\alpha_{1l}^{(ij)} = \zeta_{ij}^{(l)} - \zeta_{ij},$$

with $-\zeta_{ij}^{(l)} \sim \text{Gumbel}(0,1)$. This implies

$$\begin{aligned} \alpha_0^{(ij)} + \alpha_1^{(ij)} = \zeta_{ij}^{(l)} - \zeta_{ii} &\sim \textit{Logistic}(0,1), \\ -(\alpha_0^{(ij)} + \alpha_1^{(ij)}) \mid \zeta_{ii} &\sim \text{Gumbel}(\zeta_{ii}, 1), \\ -\alpha_1^{(ij)} \mid \alpha_0^{(ij)}, \zeta_{ii} &\sim \text{Gumbel}(\alpha_0^{(ij)} + \zeta_{ii}, 1). \end{aligned}$$

A.2 PT diagnostics

Key metrics for assessing the effectiveness of the PT algorithm include the swap acceptance rate between adjacent temperatures and the occurrence of round trips, where information from the coldest replica reaches the hottest replica and returns to the coldest. The latter indicates how effectively information is propagated across replicas. To evaluate round trips, trace plots can be used to visually assess the movement of information across the tempered replicas and identify any issues. Below, we present trace plots of the coldest replica moving across all tempered replicas for both the baseline 3-state HMM and its extension with covariates in the transition probabilities for all the PT algorithm implementations.

Baseline 3-state HMM

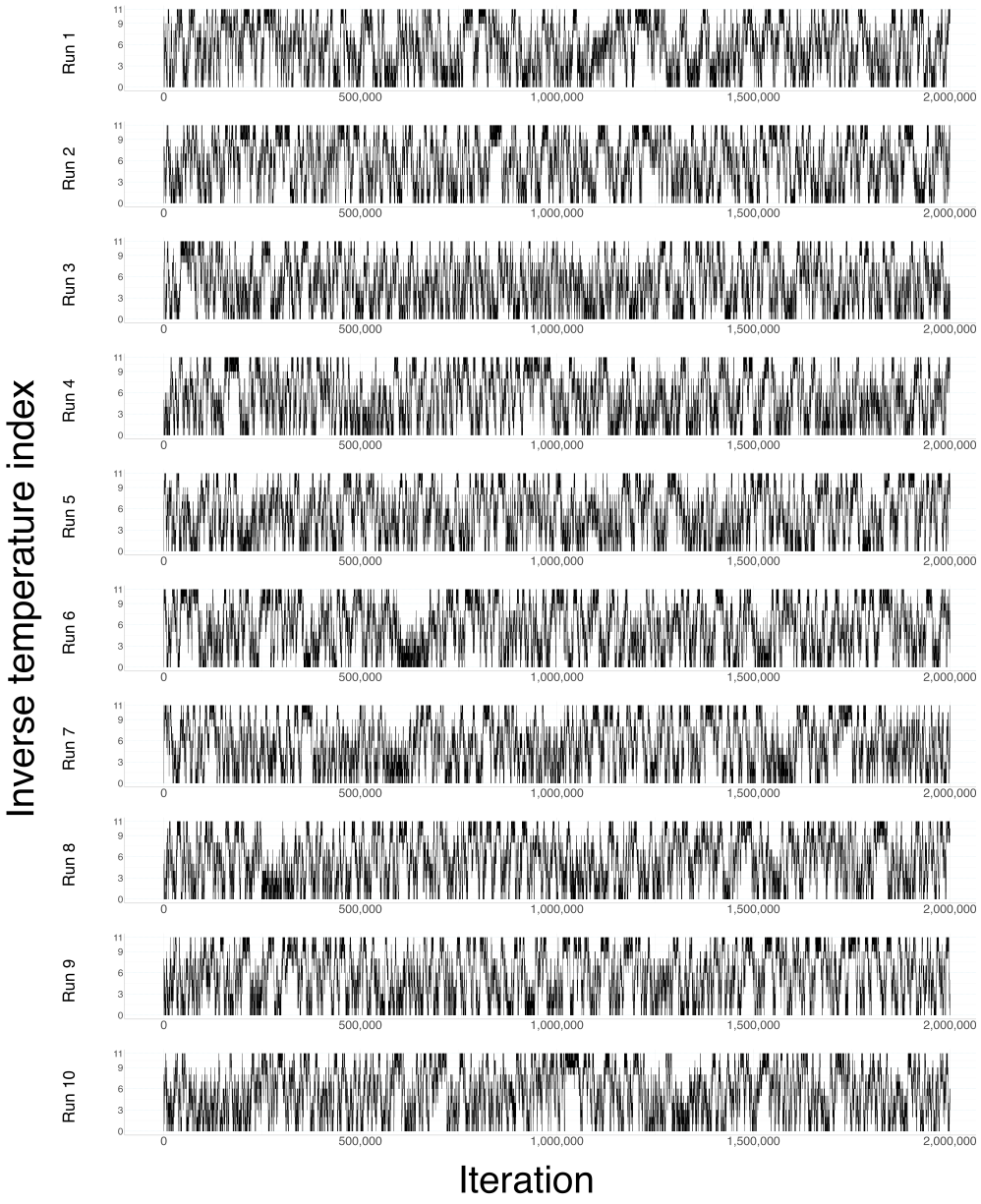


Figure 10: Traceplot of information from the coldest replica moving across all tempered replicas for the 10 PT algorithm implementations of the baseline 3-state HMM

3-state HMM with sound stimuli covariate

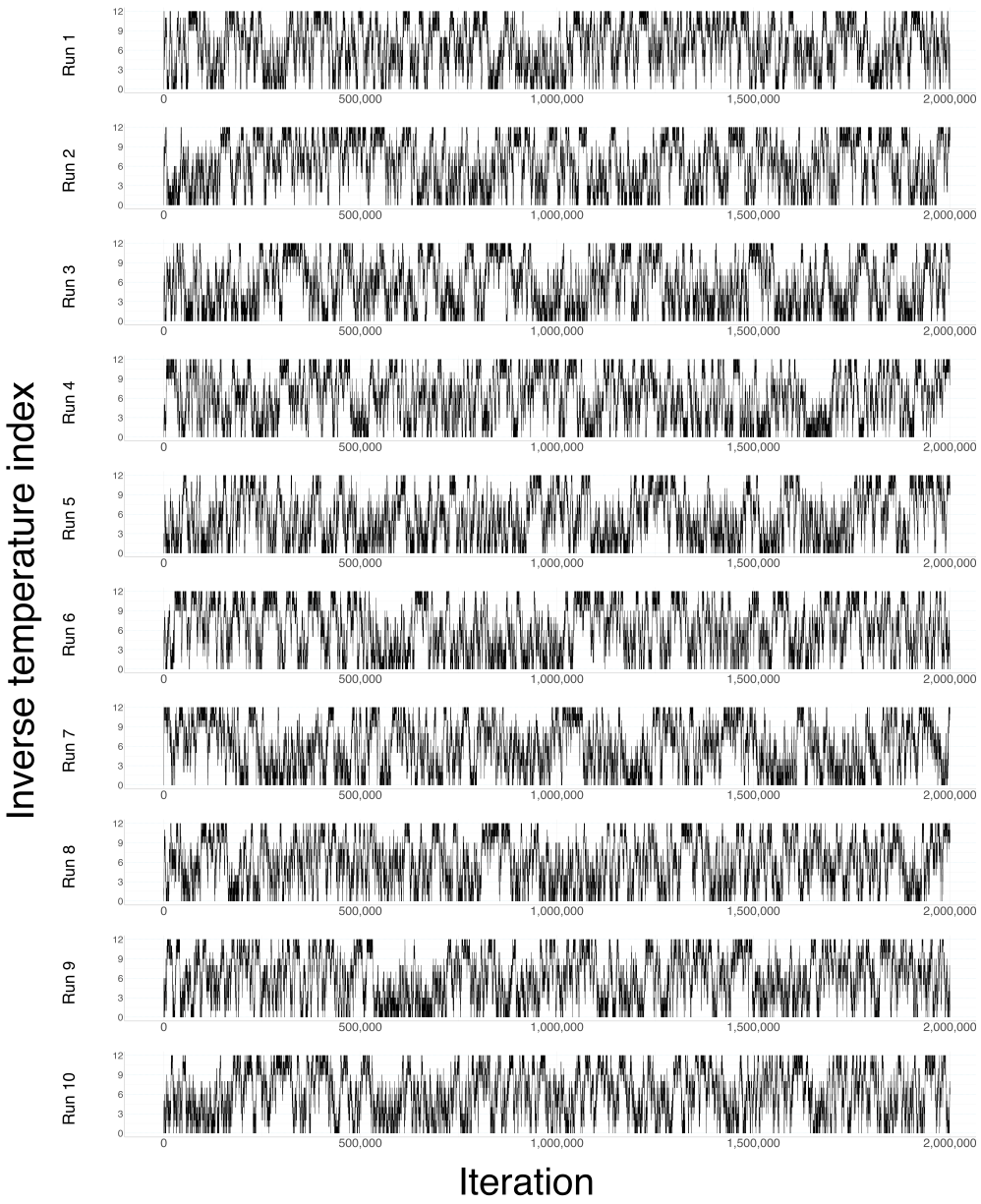


Figure 11: Traceplot of information from the coldest replica moving across all tempered replicas for the 10 PT algorithm implementations of the extended 3-state HMM with covariates

A.3 Pseudocode for implementing CWMH for HMMs

The following pseudocode outlines the implementation of the Component-Wise Metropolis–Hastings (CWMH) algorithm for sampling from an N-state HMM with a single P-dimensional observation sequence, with the option to incorporate a binary covariate. Although the algorithm is presented for a single observation sequence, it can be readily generalized by following the Algorithm **??** in Appendix A.4. Similarly, the algorithm can be extended to accommodate a covariate with L categories by iterating over the $L-1$ levels, since one level can be absorbed into the baseline working parameter.

Algorithm 2 CWMH(P-dimensional observation sequence $\boldsymbol{y_{1:T}}$, working parameters $(\alpha_{0:1}^{(ij)})_{ij}$, initial state distribution $\boldsymbol{\delta}$, state-dependent parameters $\boldsymbol{\phi} = (\boldsymbol{\phi_1}, \ldots, \boldsymbol{\phi_P})$, covariate indicator c)

Require: A proposal mechanism $q(\cdot, \cdot)$ for each sub-block of parameters.

includeCovariate $\leftarrow$ FALSE
if $c == 1$ **then**
    includeCovariate $\leftarrow$ TRUE
end if

1. Update baseline working parameters

for $i = 1, \ldots, N$ **do**
    Propose $\left(\alpha_0^{(ij)}\right)_j^* \sim q\left(\left(\alpha_0^{(ij)}\right)_j, \cdot\right)$
    $\boldsymbol{\theta} \leftarrow \left(\left(\alpha_0^{(ij)}\right)_1^*, \ldots, \left(\alpha_0^{(ij)}\right)_i, \ldots \left(\alpha_0^{(ij)}\right)_N, \boldsymbol{\delta}, \boldsymbol{\alpha_1^{(ij)}}, \boldsymbol{\phi}\right)$
    $\boldsymbol{\theta}^* \leftarrow \left(\left(\alpha_0^{(ij)}\right)_1^*, \ldots, \left(\alpha_0^{(ij)}\right)_i^*, \ldots \left(\alpha_0^{(ij)}\right)_N, \boldsymbol{\delta}, \boldsymbol{\alpha_1^{(ij)}}, \boldsymbol{\phi}\right)$
    Compute:

$$A = \log p(\theta^* \mid \boldsymbol{y_{1:T}}) - \log p(\theta \mid \boldsymbol{y_{1:T}}) + \log q\left(\left(\alpha_0^{(ij)}\right)_j^*, \left(\alpha_0^{(ij)}\right)_j\right) - \log q\left(\left(\alpha_0^{(ij)}\right)_j, \left(\alpha_0^{(ij)}\right)_j^*\right)$$

    $U \sim \text{Unif}(0, 1)$
    if $\log(U) > A$ **then**
        $\left(\alpha_0^{(ij)}\right)_j^* \leftarrow \left(\alpha_0^{(ij)}\right)_j$
    end if
end for
Let $\left(\boldsymbol{\alpha_0^{(ij)}}\right)^* \leftarrow \left(\left(\alpha_0^{(ij)}\right)^*\right)_{ij}$

$U \sim \text{Unif}(0, 1)$
if $\log(U) > A$ **then**
    $\boldsymbol{\delta}^* \leftarrow \boldsymbol{\delta}$
end if

2. Update initial distribution

Propose $\boldsymbol{\delta}^* \sim q(\boldsymbol{\delta}, \cdot)$
$\boldsymbol{\theta}^* \leftarrow \left(\left(\boldsymbol{\alpha_0^{(ij)}}\right)^*, \boldsymbol{\delta}, \boldsymbol{\alpha_1^{(ij)}}, \boldsymbol{\phi}\right)$
$\boldsymbol{\theta}^* \leftarrow \left(\left(\boldsymbol{\alpha_0^{(ij)}}\right)^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha_1^{(ij)}}, \boldsymbol{\phi}\right)$

Compute:

$$A = \log p(\theta^* \mid \boldsymbol{y_{1:T}}) - \log p(\theta \mid \boldsymbol{y_{1:T}}) + \log \frac{q(\boldsymbol{\delta}^*, \boldsymbol{\delta})}{q(\boldsymbol{\delta}, \boldsymbol{\delta}^*)}$$

3. Update covariate parameters (if applicable)

if includeCovariate **then**
  for $i = 1, \ldots, N$ **do**
    Propose $\left(\alpha_1^{(ij)}\right)_j^* \sim q\left(\left(\alpha_1^{(ij)}\right)_j, \cdot\right)$
    $\boldsymbol{\theta} \leftarrow \left(\left(\boldsymbol{\alpha_0^{(ij)}}\right)^*, \boldsymbol{\delta}^*, \left(\alpha_1^{(ij)}\right)_1^*, \ldots, \left(\alpha_1^{(ij)}\right)_i, \ldots \left(\alpha_1^{(ij)}\right)_N, \boldsymbol{\phi}\right)$
    $\boldsymbol{\theta}^* \leftarrow \left(\left(\boldsymbol{\alpha_0^{(ij)}}\right)^*, \boldsymbol{\delta}^*, \left(\alpha_1^{(ij)}\right)_1^*, \ldots, \left(\alpha_1^{(ij)}\right)_i^*, \ldots \left(\alpha_1^{(ij)}\right)_N, \boldsymbol{\phi}\right)$
    Compute:

$$A = \log p(\boldsymbol{\theta}^* \mid \boldsymbol{y_{1:T}}) - \log p(\boldsymbol{\theta} \mid \boldsymbol{y_{1:T}}) + \log \frac{q\left(\left(\alpha_1^{(ij)}\right)_j^*, \left(\alpha_0^{(ij)}\right)_j\right)}{q\left(\left(\alpha_1^{(ij)}\right)_j, \left(\alpha_1^{(ij)}\right)_j^*\right)}$$

    $U \sim \text{Unif}(0, 1)$
    if $\log(U) > A$ **then**
      $\left(\alpha_1^{(ij)}\right)_j^* \leftarrow \left(\alpha_1^{(ij)}\right)_j$
    end if
  end for
  Let $\left(\boldsymbol{\alpha_1^{(ij)}}\right)^* \leftarrow \left(\left(\alpha_1^{(ij)}\right)^*\right)_{ij}$
end if

4. Update state-dependent parameters

for $p = 1, \ldots, P$ **do**
  Propose $\phi_p^* \sim q(\phi_p, \cdot)$
  $\boldsymbol{\theta} \leftarrow \left(\left(\boldsymbol{\alpha_0^{(ij)}}\right)^*, \boldsymbol{\delta}^*, \left(\boldsymbol{\alpha_1^{(ij)}}\right)^*, \phi_1^*, \ldots, \boldsymbol{\phi}_p, \ldots, \boldsymbol{\phi}_P\right)$
  $\boldsymbol{\theta}^* \leftarrow \left(\left(\boldsymbol{\alpha_0^{(ij)}}\right)^*, \boldsymbol{\delta}^*, \left(\boldsymbol{\alpha_1^{(ij)}}\right)^*, \boldsymbol{\phi}_1^*, \ldots, \boldsymbol{\phi}_p^*, \ldots, \boldsymbol{\phi}_P\right)$
  Compute:

$$A = \log p(\boldsymbol{\theta}^* \mid \boldsymbol{y_{1:T}}) - \log p(\boldsymbol{\theta} \mid \boldsymbol{y_{1:T}}) + \log \frac{q(\phi_k^*, \phi_k)}{q(\phi_k, \phi_k^*)}$$

  $U \sim \text{Unif}(0, 1)$
  if $\log(U) > A$ **then**
    $\phi_p^* \leftarrow \phi_p$
  end if
end for
Let $\boldsymbol{\phi}^* \leftarrow (\boldsymbol{\phi_1^*}, \ldots, \boldsymbol{\phi_P^*})$

return $\left(\boldsymbol{\alpha_0^{(ij)}}\right)^*, \boldsymbol{\delta}^*, \left(\boldsymbol{\alpha_1^{(ij)}}\right)^*, \boldsymbol{\phi}^*$

A.4 Log-likelihood computation for HMMs via the forward algorithm and log-power posteriors

The algorithm 3 provides pseudocode for computing the log-likelihood of an N-state HMM using the forward algorithm [Zucchini et al., 2017], with the option to incorporate a binary covariate. The algorithm is presented for a single observation sequence; however, it can be applied independently to multiple sequences, in which case the log-likelihood is obtained by summing the contributions from each sequence. Additionally, it can be extended to accommodate a categorical covariate with L possible outcomes.

Algorithm 4 shows how to compute the log-power posterior under the same structure as Algorithm 3, with the main difference being that it explicitly accounts for multiple independent observation sequences. In this case, the evaluation of the log-likelihood for multiple sequences refers to computing the log-likelihood for each sequence independently and summing the results.

Algorithm 3 LogLikelihood(observations $\boldsymbol{y_{1:T}}$, latent process $z_{1:T}$, initial state distribution $\boldsymbol{\delta}$, working parameters $(\alpha_{0:1}^{(ij)})_{ij}$, state-dependent parameters $\boldsymbol{\phi}$, covariate indicator c)

includeCovariate $\leftarrow$ FALSE
if $c == 1$ **then**
    includeCovariate $\leftarrow$ TRUE
end if

for $i = 1, \ldots, N$ **do** ▷ Compute the forward variable at time $t = 1$ in the log scale
    $\psi_i(1) \leftarrow \log(\delta_i) + \log[p(\boldsymbol{y_1} \mid S_1 = i, \boldsymbol{\phi})]$
end for

if !includeCovariate **then**
    for $i = 1, \ldots, N$ **do** ▷ Compute i-th row of t.p.m. without covariates using equation 2
        $\Gamma_{i\cdot} \leftarrow$ multinomial-logit$((\alpha_{0:1}^{(ij)})_j, 0)$
    end for
end if

for $t = 2, \ldots, T$ **do** ▷ Compute the forward probabilities at time $t > 2$ in the log scale
    if includeCovariate **then**
        for $i = 1, \ldots, N$ **do** ▷ Compute i-th row of t.p.m. with covariates using equation 2
            $\Gamma_{i\cdot} \leftarrow$ multinomial-logit$((\alpha_{0:1}^{(ij)})_j, z_t)$
        end for
    end if
    for $j = 1, \ldots, N$ **do**
        $b_j(t) \leftarrow p(\boldsymbol{y_t} \mid S_t = j, \boldsymbol{\phi})$
        $\psi_j(t) \leftarrow$ log-sum-exp $[\psi_1(t-1) + \log(\Gamma_{1j}), \ldots, \psi_N(t-1) + \log(\Gamma_{Nj})] + \log(b_j(t))$
    end for
end for
$\ell \leftarrow$ log-sum-exp $[\psi_1(T), \ldots, \psi_N(T)]$ ▷ Compute log-likelihood

return ℓ

Algorithm 4 LogPowerPosterior(observations $\left(\boldsymbol{y}_{\boldsymbol{1:T_w}}^{\boldsymbol{(w)}}\right)_w$, latent process $\left(z_{1:T_w}^{(w)}\right)_w$, initial state distribution $\boldsymbol{\delta}$, working parameters $(\alpha_{0:1}^{(ij)})_{ij}$, state-dependent parameters $\boldsymbol{\phi}$, covariate indicator c, inverse temperature β_m)

Require: Prior densities $p(\boldsymbol{\delta})$, $p\left((\alpha_{0:1}^{(ij)})\right)$, $p(\boldsymbol{\phi})$

$\ell \leftarrow 0$ ▷ Initialize log-likelihood

for $w = 1, \ldots, W$ **do**

  $\ell \leftarrow \ell + \text{LogLikelihood}\left(\boldsymbol{y}_{\boldsymbol{1:T_w}}^{\boldsymbol{(w)}}, z_{1:T_w}^{(w)}, \boldsymbol{\delta}, (\alpha_{0:1}^{(ij)})_{ij}, \boldsymbol{\phi}, c\right)$ ▷ Log-likelihood for sequence w

end for

$\ell_{\text{temp}} \leftarrow \ell / \beta_m$ ▷ Apply likelihood tempering

$\ell_{\text{temp}} \leftarrow \ell_{\text{temp}} + \log\left[p(\boldsymbol{\delta})\right] + \log\left[p\left((\alpha_{0:1}^{(ij)})\right)\right] + \log\left[p(\boldsymbol{\phi})\right]$ ▷ Add log-priors (not tempered)

return ℓ_{temp}

A.5 Prior distributions

The prior distributions for the state-dependent distribution parameters are as follows:

Number of lunges:

$$\lambda_n \sim Gamma(1.5, 0.5) \quad n = 1, 2, 3$$

Dive duration:

$$\mu_{2n} \sim Gamma(3, .01) \quad n = 1, 2, 3$$
$$\sigma_{2n} \sim Gamma(3, .01) \quad n = 1, 2, 3$$

Surface duration:

$$\mu_{3n} \sim Gamma(3, .01) \quad n = 1, 2, 3$$
$$\sigma_{3n} \sim Gamma(3, .01) \quad n = 1, 2, 3$$

Maximum depth:

$$\mu_{4n} \sim Gamma(3, .01) \quad n = 1, 2, 3$$
$$\sigma_{4n} \sim Gamma(3, .01) \quad n = 1, 2, 3$$

Step length:

$$\mu_{5n} \sim Gamma(3, .01) \quad n = 1, 2, 3$$
$$\sigma_{5n} \sim Gamma(3, .01) \quad n = 1, 2, 3$$

For the initial state distribution vector, it was assumed $\boldsymbol{\delta} \sim Dirichlet(\mathbf{1}_N)$. For the reparametrized transition probabilities, we have that the conditional priors for the baseline coefficients are

$$-\alpha_0^{(ij)} \mid \zeta_i \sim Gumbel(\zeta_i, 1) \quad i, j = 1, 2, 3,\ i \neq j,$$

with $\zeta_i \sim Gumbel(0, 1)$, $i = 1, 2, 3$. For the extended model incorporating sound stimuli as a covariate, the conditional priors for the covariate coefficients $\alpha_1^{(ij)}$ are

$$-\alpha_1^{(ij)} \mid \alpha_0^{(ij)}, \zeta_i \sim Gumbel(\alpha_0^{(ij)} + \zeta_i, 1).$$

A.6 Posterior estimates for the 3-state HMM with sound stimuli as a covariate

For each of the two modes identified in the baseline 3-state HMM extended with covariates in the transition probabilities, posterior median estimates and 95% credible intervals were computed from one million iterations after correcting for label switching, as described in Section 4.2. The tables below present the values for PT algorithm implementation ID 1, with 95% credible intervals shown in parentheses. Results are organized by state-dependent parameters, transition probabilities, initial state distribution, and baseline and covariate working parameters corresponding to the reparameterization introduced in Equation 2, including transition probabilities under the presence of sound stimuli ($z_t = 1$).

State-dependent parameters

	State $n=1$		State $n=2$		State $n=3$	
Variable	Mode $\tilde{A}$	Mode $\tilde{B}$	Mode $\tilde{A}$	Mode $\tilde{B}$	Mode $\tilde{A}$	Mode $\tilde{B}$
λ_n	0.649 (0.575,0.728)	0.426 (0.379,0.479)	0.037 (0.004,0.1)	2.82 (2.571,3.063)	3.405 (3.216,3.601)	2.925 (2.701,3.133)
μ_{2n}	150 (142,159)	182 (172,194)	360 (326,398)	401 (384,417)	523 (510,536)	568 (554,584)
σ_{2n}	89 (82,97)	131 (122,143)	226 (198,262)	118 (107,133)	124 (115,135)	124 (113,136)
μ_{3n}	68 (63,75)	72 (66,77)	96 (90,102)	105 (100,111)	154 (147,162)	181 (171,191)
σ_{3n}	66 (60,74)	67 (61,73)	36 (31,42)	38 (34,43)	70 (63,77)	68 (61,77)
μ_{4n}	33 (31,36)	30 (28,33)	80 (69,94)	112 (107,116)	173 (167,179)	216 (207,222)
σ_{4n}	24 (22,27)	22 (20,25)	74 (62,89)	26 (23,29)	60 (55,65)	42 (37,47)
μ_{5n}	210 (194,227)	314 (291,336)	717 (670,767)	345 (307,399)	414 (383,448)	524 (473,578)
σ_{5n}	154 (138,172)	267 (242,293)	288 (255,329)	227 (193,284)	293 (264,327)	380 (333,428)

Transition probabilities and initial state distribution

	State $n=1$		State $n=2$		State $n=3$	
Variable	Mode $\tilde{A}$	Mode $\tilde{B}$	Mode $\tilde{A}$	Mode $\tilde{B}$	Mode $\tilde{A}$	Mode $\tilde{B}$
γ_{1n}	0.896 (0.86,0.926)	0.859 (0.824,0.889)	0.028 (0.011,0.052)	0.025 (0.012,0.047)	0.076 (0.05,0.107)	0.115 (0.087,0.147)
γ_{2n}	0.057 (0.02,0.113)	0.066 (0.035,0.108)	0.75 (0.665,0.825)	0.921 (0.877,0.953)	0.19 (0.123,0.27)	0.011 (0.002,0.035)
γ_{3n}	0.102 (0.068,0.143)	0.268 (0.208,0.327)	0.08 (0.05,0.118)	0.041 (0.017,0.073)	0.817 (0.769,0.859)	0.69 (0.624,0.75)
δ_n	0.182 (0.08,0.325)	0.258 (0.151,0.43)	0.169 (0.064,0.317)	0.22 (0.106,0.354)	0.641 (0.476,0.786)	0.509 (0.331,0.656)

Baseline and covariate working parameters

Variable	Mode $\tilde{A}$	Mode $\tilde{B}$
α_{12}	-3.482 (-4.384,-2.831)	-3.545 (-4.315,-2.92)
α_{13}	-2.474 (-2.906,-2.096)	-2.015 (-2.315,-1.717)
α_{21}	-2.576 (-3.649,-1.825)	-2.629 (-3.298,-2.104)
α_{23}	-1.374 (-1.892,-0.917)	-4.425 (-6.245,-3.239)
α_{31}	-2.082 (-2.512,-1.7)	-0.945 (-1.269,-0.649)
α_{32}	-2.321 (-2.823,-1.892)	-2.831 (-3.749,-2.177)
β_{12}	-2.423 (-4.621,-0.149)	-1.02 (-4.098,0.825)
β_{13}	-1.486 (-2.847,-0.333)	-0.339 (-1.178,0.31)
β_{21}	-0.697 (-3.954,1.276)	-0.555 (-2.836,1.324)
β_{23}	0.253 (-0.765,1.186)	-1.034 (-5.939,2.924)
β_{31}	-0.734 (-2.343,0.412)	-0.467 (-1.314,0.312)
β_{32}	0.376 (-0.692,1.288)	-0.762 (-3.957,1.479)

Transition probabilities during the presence of sound stimuli ($z_t = 1$)

	State $n=1$		State $n=2$		State $n=3$	
Variable	Mode $\tilde{A}$	Mode $\tilde{B}$	Mode $\tilde{A}$	Mode $\tilde{B}$	Mode $\tilde{A}$	Mode $\tilde{B}$
$\tilde{\gamma}_{1n}$	0.976 (0.934,0.993)	0.901 (0.833,0.948)	0.003 (0,0.026)	0.01 (0,0.052)	0.019 (0.005,0.054)	0.086 (0.041,0.144)
$\tilde{\gamma}_{2n}$	0.027 (0.001,0.129)	0.041 (0.003,0.172)	0.725 (0.553,0.863)	0.942 (0.781,0.995)	0.236 (0.108,0.407)	0.005 (0,0.093)
$\tilde{\gamma}_{3n}$	0.049 (0.011,0.129)	0.191 (0.094,0.327)	0.118 (0.047,0.223)	0.021 (0.001,0.122)	0.826 (0.711,0.912)	0.778 (0.626,0.884)

A.7 7-dimensional 3-state HMM parameter estimates using maximum likelihood estimation

Parameter estimates from the baseline model in [DeRuiter et al., 2017] were replicated by following the same procedure, namely numerically maximizing the model's likelihood using the `nlm` function in R. The replicated results are shown below.

State-dependent parameters

Variable	State $n = 1$	State $n = 1$	State $n = 3$
μ_{2n}	140 (132,148)	334 (303,365)	516 (503,529)
σ_{2n}	80 (73,88)	212 (186,241)	130 (120,140)
μ_{3n}	70 (64,77)	86 (78,95)	151 (144,158)
σ_{3n}	68 (61,76)	55 (47,65)	69 (63,75)
μ_{4n}	32 (30,35)	68 (58,78)	170 (164,176)
σ_{4n}	24 (21,26)	65 (55,77)	60 (56,65)
μ_{5n}	187 (173,202)	675 (627,723)	406 (375,437)
σ_{5n}	132 (119,147)	305 (268,347)	287 (258,318)
κ	1.005 (0.845,1.195)	3.023 (2.505,3.65)	0.816 (0.656,1.015)
a	0.977 (0.844,1.13)	0.501 (0.425,0.591)	1.673 (1.459,1.918)
b	2.106 (1.815,2.444)	5.432 (4.166,7.083)	1.564 (1.37,1.786)
λ_n	0.67 (0.594,0.757)	0.02 (0.004,0.091)	3.358 (3.175,3.552)

Initial state distribution

Variable	State $n = 1$	State $n = 1$	State $n = 3$
γ_{1n}	0.93 (0.876,0.96)	0.014 (0.005,0.038)	0.055 (0.035,0.085)
γ_{2n}	0.019 (0.006,0.059)	0.785 (0.686,0.851)	0.196 (0.143,0.255)
γ_{3n}	0.072 (0.049,0.103)	0.099 (0.072,0.134)	0.829 (0.764,0.879)
δ_n	0.168 (0.183,0.435)	0.171 (0.077,0.094)	0.66 (0.489,0.723)

A.8 $\hat{R}$ and effective sample size (ESS) of estimated parameters

The $\hat{R}$ and ESS were computed using one million samples for each parameter in both the baseline 3-state HMM and its extension with covariates in the transition probabilities. These values were obtained after correcting for label switching in a post-sampling process. The $\hat{R}$ values were rounded to three decimal places, while the ESS values were rounded to two decimal places. Both quantities were calculated using the functions `rhat` and `ess_basic` from the R package `posterior` (version 1.6.0).

Baseline 3-state HMM

State-dependent parameters

	State $n=1$		State $n=2$		State $n=3$	
Variable	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS
λ_n	1.001	3713.77	1.001	3127.21	1.001	4038.41
μ_{2n}	1.001	3583	1.001	6001.93	1.001	3603.27
σ_{2n}	1.001	3342.97	1.001	3530.03	1.000	748704.95
μ_{3n}	1.000	21901.48	1.001	5168.68	1.001	3574.69
σ_{3n}	1.000	259757.41	1.000	102791.06	1.000	533488.24
μ_{4n}	1.001	8707.18	1.001	3763.77	1.001	3230.83
σ_{4n}	1.001	8772.34	1.001	3477.5	1.001	3502.17
μ_{5n}	1.001	3297.53	1.001	3201.78	1.001	3747.97
σ_{5n}	1.001	3279.76	1.001	5933.18	1.001	4053.81

Initial state distribution

	State $n=1$		State $n=2$		State $n=3$	
Variable	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS
δ	1.000	14639.17	1.000	42526.81	1.001	9994.35

Baseline working parameters

Variable	$\hat{R}$	ESS
α_{12}	1.000	452751.1
α_{13}	1.001	5499.35
α_{21}	1.000	617852.98
α_{23}	1.001	3466.2
α_{31}	1.001	3710.96
α_{32}	1.001	5335.35

3-state HMM with sound stimuli covariate

State-dependent parameters

	State $n = 1$		State $n = 2$		State $n = 3$	
Variable	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS
λ_n	1.000	317520.9	1.001	44298.21	1.000	387285.09
μ_{2n}	1.000	183615.78	1.000	205753.75	1.000	247091.23
σ_{2n}	1.000	126922.04	1.000	161924.27	1.000	796138.77
μ_{3n}	1.000	248422.18	1.000	399463.62	1.000	219407.28
σ_{3n}	1.000	262077.67	1.000	350528.63	1.000	592682.15
μ_{4n}	1.000	251749.29	1.000	107433.96	1.000	101629.6
σ_{4n}	1.000	252209.65	1.000	109726.76	1.000	222800.43
μ_{5n}	1.000	102849.21	1.000	91681.94	1.000	194589.05
σ_{5n}	1.000	97392.6	1.000	397324.33	1.000	239858.9

Initial state distribution

	State $n = 1$		State $n = 2$		State $n = 3$	
Variable	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS
δ	1.000	649477.98	1.000	620897.06	1.000	560350.02

Baseline working parameters

Variable	$\hat{R}$	ESS
α_{12}	1.000	511634.37
α_{13}	1.000	996941.52
α_{21}	1.000	462488.08
α_{23}	1.000	100783.9
α_{31}	1.000	343611.57
α_{32}	1.000	425643.67

Baseline working parameters

Variable	$\hat{R}$	ESS
β_{12}	1.000	56804.45
β_{13}	1.000	294474.78
β_{21}	1.000	243709.09
β_{23}	1.000	363338.05
β_{31}	1.000	752708.09
β_{32}	1.000	481134.05