Inferring Fine Scale Behaviors Within Hierarchical Hidden Markov Models

Evan Sidrow

May 18, 2020

Abstract

In the field of animal movement, recent advances in high-frequency tagging technology have made available a vast amount of data which can exhibit simultaneous behavioral processes occurring at different time scales. One way to model this data is to use a hierarchical hidden Markov model (HHMM), where the system is modeled as a nested structure of hidden Markov models (HMMs). At very short time scales, however, observations can exhibit complicated dependence structures which cannot be easily captured by traditional HMMs. This work investigates how to incorporate fine-scale processes into the larger structure of HHMMs while maintaining computational efficiency. We apply our method to dive data collected from a northern resident killer whale off the coast of British Columbia, Canada.

The field of animal movement is in the midst of a "data renaissance" where advancements in tagging technology have given rise to an explosion of data available for statistical modeling. In particular, tagging technologies are capable of recording observations at rates of tens of hertz, resulting in time series containing millions of observations over the course of several hours. This results in a vast amount of data which often exhibits many different simultaneous behavioral processes occurring at different time scales.

One solution to this issue is to use a hierarchical hidden Markov model (HHMM). HHMMs model the entire system as a nested structure of hidden Markov models (HMM) where each HMM corresponds to one behavioral process. One nice property of HHMMs is that its likelihood is relatively easy to compute, facilitating fast maximum likelihood estimates for its associated parameters.

At the shortest time scales, however, observations often exhibit complicated dependence structures which cannot be easily captured by a traditional HMMs. To address this issue, it is possible to model small-scale animal behavior as the solution to some stochastic differential equation, but these methods tend to be computationally intractable and require approximate inference techniques such as Markov-chain Monte Carlo (MCMC).

This work investigates how to incorporate fine-scale processes into the larger structure of hierarchical hidden Markov models while maintaining computational efficiency. We bridge the gap between discrete hidden Markov models and continuous-time stochastic process models by showing that the two are equivalent under certain conditions. In addition, we extract features from highly structured sub-dive behaviors that otherwise could not be modeled with a simple HMM. Finally, we apply our method to dive data collected from a Northern resident killer whale off the coast of British Columbia, Canada.

1 Background

Hidden Markov models are useful when inferring a single unobserved process, but biological processes often involve multiple simultaneous hidden processes which can occur and at different time scales. For example, a preliminary observation of the killer whale dive data shown in figure (**??**) shows that the behavior of this killer whale changes between approximately hour-long periods of predominately short, shallow dives and long, deep dives. Leos-Barajas et al. encounter a similar issue when modeling the movement of a harbor porpoise in the North Sea, and use it as a motivating example when they introduce hierarchical hidden Markov models.

1.1 Hidden Markov Models

Hidden Markov models (or HMMs) are comprised of an unobserved Markov chain $X = (X_1, \dots, X_T)$ and a sequence of observations $Y = (Y_1, \dots, Y_T)$. Each random variable in the unobserved chain X_t can take one of N possible values, and X has corresponeding probability transition matrix $\Gamma \in \mathbb{R}^{N \times N}$ and initial distribution $\delta \in \mathbb{R}^N$:

$$\delta_i = P(X_1 = i)$$

$$\Gamma_{ij} = P(X_{t+1} = j | X_t = i) \qquad \forall t \in \{1, \dots, T-1\}$$

Further, each random varaible X_t emits an observation Y_t that depends only on the value of X_t: $p(y_t|x_t, x_{t-1}, \dots, x_1, y_{t-1}, \dots, y_1) = p(y_t|x_t)$. A visualization of this structure can be seen if figure (**??**). In the field of animal movement, the unobserved chain X usually represents the latent behaviour of an animal (e.g. foraging, resting, migrating, etc.), while the observations Y are often a series of step lengths and turning angles for land animals or depth data for marine animals.

The probability transition matrix Γ and the parameters of the emission distributions, θ, can be estimated by maximizing the likelihood of the observed data y, $\mathcal{L}_{\text{HMM}}(y)$, with respect to the Γ and θ. In addition, $\mathcal{L}_{\text{HMM}}(y)$ can be calculated using the *forward algorithm*:

$$\mathcal{L}_{\text{HMM}}(y) = \delta P(y_1) \prod_{t=2}^{T} \Gamma P(y_t) \mathbf{1}$$

where:

$$P(y_t) = \text{diag}(p(y_t|x_t = x_1), ..., p(y_t|x_t = x_N))$$

and $\mathbf{1}$ is an N-dimensional column vector of ones.

In order to enusre identifiability and right-stochasticity after optimizing $\mathcal{L}_x(y)$, Γ is parameterized using $\eta \in \mathbb{R}^{N \times N}$ and the following link function:

$$\Gamma_{ij} = \frac{\exp(\eta_{ij})}{\sum_{k=1}^{N} \exp(\eta_{ik})}, \qquad \eta_{ii} = 0 \quad \forall i \in \{1, \dots, N\}$$

This allows for unconstrained optimization over η and removes the constraint that Γ be right-stochastic. $\mathcal{L}_x(y)$ can be maximized using any numerical optimizer.

1.2 Hierarchical Hidden Markov Models

A hierarchical hidden Markov model (or HHMM) is a variation of a hidden Markov model in which each hidden state of the original HMM X_t emits both an observation Y_t *and* another, fine-scale hidden markov model of length S_t^*. This fine-scale HMM is comprised of a Markov chain $X_t^* = (X_{1,t}^*, \dots, X_{S_t,t}^*)$ and observations $Y_t^* = (Y_{1,t}^*, \dots, Y_{S_t,t}^*)$. As before, each fine-scale observation $Y_{s,t}^*$ depends only on the value of its corresponding hidden state, $X_{s,t}^*$. $X_{s,t}^*$ can take one of $N^{*(X_t)}$ values and is characterized by an initial distribtuion $\delta^{*(X_t)} \in \mathbb{R}^{N^{*(X_t)}}$ and probability transition matrix $\Gamma^{*(X_t)} \in \mathbb{R}^{N^{*(X_t)} \times N^{*(X_t)}}$:

$$\delta_i^{*(x_t)} = P(X_{1,t}^* = i | X_t = x_t)$$

$$\Gamma_{ij}^{*(x_t)} = P(X_{s+1,t}^* = j | X_{s,t}^* = i, X_t = x_t) \qquad \forall s \in \{1, \dots, S_t^* - 1\}$$

Finally, the emission probabilities $p(y_{s,t}^*|x_{s,t}^*)$ are parameterized by $\theta^{*(X_t)}$. Note the parameters of the fine-scale hidden markov model, $N^{*(X_t)}$, $\Gamma^{*(X_t)}$, $\delta^{*(X_t)}$, and $\theta^{*(X_t)}$ all depend upon the hidden state of the *crude-scale* hidden markov model X_t. A visualization of this structure can be seen in figure (**??**).

Due to the nested structure of a hierarchical hidden Markov model, the likelihood of an HHMM is still easy to calculate using the forward algorithm:

$$\mathcal{L}_{\text{HHMM}}(y, y^*) = \delta P^*(y_1, y_1^*) \prod_{t=2}^{T} \Gamma P^*(y_t, y_t^*) \mathbf{1}$$

where:

$$P^*(y_t, y_t^*) = \text{diag}(p(y_t|x_t = x_1)\mathcal{L}_{\text{HMM}}(y_t^*|X_t = x_1), ..., p(y_t|x_t = x_N)\mathcal{L}_{\text{HMM}}(y_t^*|X_t = x_N))$$

.

Note that this formualtion assume that the crude-scale observations at a given time Y_t and the fine-scale observation time series Y_t^* are independent conditioned on X_t.

For more information on specific considerations for HHMMs such as incorporating covariates into the probability transition matrix, model selection and model checking, see Adam et al. [1]

1.3 State decoding

Once a HMM or HHMM model is fit using the process described above, it is common to find the most likely sequence of hidden states $\hat{X}$ conditioned on the learned parameters by using the a dynamic programing algorithm called the Viterbi algorithm [5]. In the case of HHMMs, this can be followed by running the Viterbi algorithm again on each sub-dive state to find the mostly likely sequence of fine-scale hidden states $\hat{X}_t^*$ conditioned on the learned parameters and the estimated crude-state value $\hat{X}_t$. Note that while $\hat{X}$ is a maximum likelihood estimate of X, $\hat{X}_t^*$ is *not* necessarily a maximum likelihood estimate of $\hat{X}_t^*$ because it is conditioned on the value of $\hat{X}_t$.

While the Viterbi algorithm is the de-facto standard in the current ecology literature, we suggest to instead find the *probability* of each crude-level state (conditioned on the learned parameters) using the *forward-backward algorithm*. Although not commonly used in the ecology literature, the forward-backward algorithm has the same time complexity as the forward algorithm and is also used to find the *psuedoresiduals* of a given model, which is an important tool for model validation. In addition, For HHMMs in particular, the forward-backward algorithm can be used recursively to find the probability of the fine-level states $X_{s,t}^*$ exactly by marginalizing out X_t:

$$P(X_{s,t}^* = x_{s,t}^*) = \sum_{n=1}^{N} P(X_t = x_n) P(X_{s,t}^* = x_{s,t}^* | X_t = x_n)$$

Where $P(X_t = x_n)$ can be found using the forward-backward algorithm on the crude-level markov chain and $P(X_{s,t}^* = x_{s,t}^* | X_t = x_n)$ can be found by running the forward-backward algorithm on the fine-level HMM for every possible value of X_t.

2 Autocorrelation within Fine Scale Behaviors

One of the key assumption of both HMMs and HHMMs is *conditional indpendence* between observations. In particular, given the state X_t or $X_{s,t}^*$, Y_t or $Y_{s,t}^*$ (respectively) is assumed to be independent from all other observations. Therefore,traditional HMMs can fail is when the observations Y exhibit significant auto-correlation. Unfortunately this is often the case for fine-scale animal behaviors. Examples include fluking in marine mammals in Vancouver, BC (see the results section) and swiming behavior in horn sharks in southern California [1].

One way to deal with autocorrelation in fine-scale behavioral processes is to use a state-switching continuous model such as the one introduced by Michelot et al [4], which models the movement of an animal as an Ornstein-Uhlenbeck process with parameters that depend upon the underlying state of the animal. Continuous time models are advantagous because of their flexibiliy: they can be built up from aribitrarily complex stochastic differential equations and they allow for uneven step lengths in the observations sequence $\mathbf{Y}$. However, most continuous time models require MCMC algorithms to perform inference and as a result are not easily incorporated into the HHMM structure.

Another option is to use the CarHMM, or *conditionally auto-regressive hidden Markov model*, introduced by Lawler et al [3], in which autocorrelation is explicity modeled into the emission distributions of the HMM while maintaining the strucutre needed to run the forward algorithm for fast direct likelihood maximization. In particular, the CarHMM introduces autocorrleation into the HMM by assuming that an observation Y_t has mean $(1-\phi_{x_t}) \cdot \mu_{x_t} + \phi_{x_t} \cdot y_{t-1}$, instead of μ_{x_t} in the case of a classic HMM. Note that the autocorrelation term ϕ_{x_t} depends upon the behavioral state of the animal. This model easily fits into the HHMM structure, but initially it seems to lack the flexibility and natural interpretation of continuous-time models.

2.1 Equivalency of CarHMM and one-dimensional Ornstein-Uhlenbeck state-switching process

In fact, under certain conditions, the CarHMM and the state-switching continuous model are equivalent. In particular, the conditions are that (1) the underlying behavioral state of the continuous-time model must follow a Markov chain rather than a Markov process, and (2) the emission distributions of the CarHMM must be gaussian.

This allows the theoretically grounded continuous-time state-switching model to be used in the computational convieneint HMM (and therefore HHMM) framework. In addition, it gives new intpretation to the learned parameters of the CarHMM in the context of an Ornstein-Uhlenbeck process.

A one-dimensional state-switching Ornstein-Uhlenbeck process y^* is the solution to the following stochastic differential equation:

$$dy_t^* = \beta_{x_t^*}(\gamma_{x_t^*} - y_t^*)dt + \omega_{x_t^*} dW_t$$

where x_t^* is the fine-scale behavior of the animal at time t, $\beta_{x_t^*}$ relates to rate at which the process returns to its mean value, $\gamma_{x_t^*}$ is the long-term mean value of the process, $\omega_{x_t^*}$ is related to short-term variance, and W is a Wiener process.

As before, x_t^* is described by an unobserved Markov process. The solution to this equation is known to be the following [4]:

$$y^*_{t+\delta} \sim \mathcal{N}\left((1-e^{-\beta_{x_t^*}\delta})\gamma_{x_t^*} + e^{-\beta_{x_t^*}\delta}y_t^*, \quad \frac{\omega^2_{x_t^*}}{2\beta_{x_t^*}}(1-e^{-2\beta_{x_t^*}\delta})\right)$$

Now, suppose that δ is constant for all observations, as is the case for hidden Markov models. In addition, introduce the following transformations:

$$\mu_{x_t^*} = \gamma_{x_t^*}, \qquad \phi_{x_t^*} = e^{-\beta_{x_t^*}\delta}, \qquad \sigma^2_{x_t^*} = \frac{\omega^2_{x_t^*}}{2\beta_{x_t^*}}(1-e^{-2\beta_{x_t^*}\delta})$$

Then, we have the following:

$$y^*_{t+\delta} \sim \mathcal{N}\left((1-\phi_{x_t^*})\mu_{x_t^*} + \phi_{x_t^*}y_t^*, \quad \sigma^2_{x_t^*}\right)$$

If δ is fixed and x_t^* is adjusted to follow a Markov chain rather than a Markov process, then this model is equivalent to the CarHMM with normal emission probabilities. Note that all of the parameter transformations above are one-to-one, so it is easy to go from the CarHMM to the continuous model and back again. This allows for the pricipled construction of the continuous-time model to be combined with the computational convienence of the CarHMM.

3 Capturing Non-Markovian Behavior via the Fourier Transform

Although the CarHMM and continuous-time model can effectively capture autocorrelation within a process, both still assume Markovian dynamics, i.e. that the observation Y_t depends only the value of the behaviorial state X_t and the previous observation Y_{t-1}. However, there are many animal movement process which violate the markov propertey. In particular, on a fine scale, swimming behavior of marine mammals can exhibit periodic behavior as the animal repeatedly flukes to propel itself forward. Work has been done in the past to model non-markovian dynamics in the *behavioral* process [2], but addressing non-markovian dynamics within observations $Y*$ is still a relatively unstudied area. With improvements in tagging technology allowing for data collection at very high frequencies, noisy and non-markovian fine scale behavior appears to be on the rise.

To address this issue, we recommend borrowing techniques from the signal processing literature to compress the data and summarize its essential elements. In particular, we suggest a method which can capture periodic behavior by decompsing a signal into its fourier components.

Suppose a one-dimensional fine-scale process of length T^*, y^* exhibits significant periodic and structured behavior that is clearly non-markovian. It is possible to transform this fine-scale process y^* to a shorter, but higher dimensional process $\hat{y}^*$ by selecting a window of size w, dividing y^* into $\lfloor T^*/w \rfloor$ seperate intervals of length w (truncating y^* appropriately), and taking the discrete fourier transform (DFT) of each interval. The resulting time series $\hat{y}^*$ will be a complex-valued, w-dimensional time series of length $\lfloor T^*/w \rfloor$.

Picking the window length w should be done with care. w should be long enough to capture the periodic behavior of the underlying process (at least twice as long as the length of a period), but short enough so that the resolution of the process remains high and behavioral changes can be captured via the HMM.

A visualization of transforming a one-dimensional sequence y^* to a w-dimensional complex sequence $\hat{y}^*$ can be seen in figure (**??**).

4 Simulation Study

To test this method, dive and acceleration data was simulated using the following procedure:

1. 100 dive durations were simulated using a gamma distribution using an HMM generative model and the following parameters:

$$\Gamma = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}$$

$$Y_t|X_t \sim \text{Gamma}$$

$$\mathbb{E}(Y_t|X_t=1) = 15, \quad \mathbb{E}(Y_t|X_t=2) = 60$$

$$\mathbb{V}(Y_t|X_t=1) = 25, \quad \mathbb{V}(Y_t|X_t=2) = 100$$

Figure (**??**) also shows the calculated ODBA as a function of time for one a specific dive. The acceleration exhibits sinusoidal behavior at several points in time which cannot be modeled using HMMs straightforwardly. As a result, the data was split into two-second intervals (100 data points each), and each interval was summarized by the Fourier transform of the ODBA and the velocity at the *end* of the time interval. Velocity was taken at the end of the time interval because the behavior of the whale during the time interval has no impact on the velocity of the whale at the beginning of the time interval. Because velocity is highly correlated with the average acceleration within an interval, the mean was subtracted out of each ODBA time series before the Fourier transform was taken. To reduce the dimension of the ODBA Fourier transform, Fourier coefficients corresponding to frequencies less than or equal to 5 hertz were summed and frequencies above 5 hertz were ignored. One nice property of this process is that it reduces the number of data points by a factor of 100, which drastically speeds up parameter estimation. The optimal length of the time interval over which to take the Fourier transform is a tuning parameter that is difficult to find. An ideal time interval should be long enough give useful Fourier coefficients, but short enough to preserve as much information about the vertical velocity of the animal as possible.

4.1 Lag Plot

In order to find the appropriate HMM to model whale behavior, a lag plot was made for both velocity and the ODBA fourier sum. The results of doing so are shown in figure (**??**). Unfortunately, the number of behavioral states is not clear from the lag plot, but it is clear that the velocity exhibits a large degree of autocorrelation. While the ODBA Fourier sum also exhibits some autocorrelation, the relationship is less strong, so autocorrelation was not incorporated in the ODBA Fourier sums emission distribution.

References

[1] Timo Adam, Christopher Griffiths, Vianey Leos Barajas, Emily Meese, Christopher Lowe, Paul Blackwell, David Righton, and Roland Langrock. Joint modelling of multi-scale animal movement data using hierarchical hidden markov models. *Methods in Ecology and Evolution*, 10, 06 2019.

[2] Roland Langrock, Ruth King, Jason Matthiopoulos, Len Thomas, Daniel Fortin, and Juan Morales. Flexible and practical modeling of animal telemetry data: Hidden markov models and extensions. *Ecology*, 93:2336–42, 11 2012.

[3] Ethan Lawler, Kim Whoriskey, William Aeberhard, Chris Field, and Joanna Flemming. The conditionally autoregressive hidden markov model (carhmm): Inferring behavioural states from animal tracking data exhibiting conditional autocorrelation. *Journal of Agricultural, Biological and Environmental Statistics*, 05 2019.

[4] Thèo Michelot and Paul Blackwell. State-switching continuous-time correlated random walks. *Methods in Ecology and Evolution*, 01 2019.

[5] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, pages 260–269, April 1967.

Appendix

A Equivalency of CarHMM and one-dimensional state-switching Ornstein-Uhlenbeck process

A one-dimensional state-switching Ornstein-Uhlenbeck process $\mathbf{x}$ is the solution to the following stochastic differential equation:

$$dx_t = \beta_{b_t}(\gamma_{b_t} - x_t)dt + \omega_{b_t} dW_t$$

where b_t is the behavior of the animal at time t, β_{b_t} relates to rate at which the process returns to its mean value, γ_{b_t} is the long-term mean value of the process, ω_{b_t} is related to short-term variance, and W is a Brownian motion process. b_t is described by an unobserved Markov process. The solution to this equation is known to be the following [4]:

$$x_{t+\delta} \sim \mathcal{N}\left((1 - e^{-\beta_{b_t}\delta})\gamma_{b_t} + e^{-\beta_{b_t}\delta}x_t, \quad \frac{\omega_{b_t}^2}{2\beta_{b_t}}(1 - e^{-2\beta_{b_t}\delta})\right)$$

suppose that δ is constant for all observations, and introduce the following transformations:

$$d_t = x_t, \qquad \mu_{RL,b_t} = \gamma_{b_t}, \qquad \phi_{b_t} = e^{-\beta_{b_t}\delta}, \qquad \sigma_{b_t}^2 = \frac{\omega_{b_t}^2}{2\beta_{b_t}}(1 - e^{-2\beta_{b_t}\delta})$$

Then, we have the following:

$$d_{t+\delta} \sim \mathcal{N}\left((1 - \phi_{b_t})\gamma_{b_t} + \phi_{b_t} d_t, \quad \sigma_{b_t}^2\right)$$

If δ is fixed and b_t is adjusted to follow a Markov chain rather than a Markov process, then this model is equivalent to the CarHMM with normal emission probabilities for the step length sequence. Note that all of the parameter transformations above are one-to-one, so it is easy to go from the CarHMM to the continuous model and back again.