The Canadian Journal of Statistics

Vol. xx, No. yy, 2020, Pages 1–25

La revue canadienne de statistique

Modelling fine-scale correlation structures using hidden Markov models

BlindedA[1]* and BlindedB[2]

[1]*Author affiliations will go here in the accepted manuscript, but do NOT include them in your initial submission because it must be anonymous.*

[2]*Second Affiliation*

Key words and phrases: Association parameters; clustered data; mean parameters; missing data; pairwise likelihood; repeated measurements.

MSC 2010: Primary 62???; secondary 62???

Abstract:

Recent advances in high-frequency tagging technology have made a vast amount of movement data available in a variety of fields. This rich data can exhibit simultaneous behavioural processes occurring at different time scales, resulting in very complicated dependence structures. These processes can be modelled through a hierarchical hidden Markov model (HHMM), where the system is modelled as a nested structure of hidden Markov models (HMMs). At very short time scales, however, many assumptions of traditional HMMs are violated. We demonstrate how to incorporate fine-scale processes into the larger structure of HHMMs while maintaining computational efficiency. We apply our method to dive and accelerometer data collected from a northern resident killer whale off the coast of British Columbia, Canada.

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–25; 2020

Notes from Nancy

I've replaced your Background section with my Models and Parameter Estimation (NH) section, borrowing heavily from your writing.

Here are some comments and explanations.

General things (to always keep in mind - I may add to this list of rules as we go along)

1. When I recompile in overleaf, I see there are 21 warnings (in the symbol next to the recompile button). It looks like these are from sections you have written, mostly from problems with newblock. Please go through these since Latex may be making decisions that you don't like. It's always good to go through the Latex errors and warnings.
2. Don't begin a sentence with a symbol, e.g. don't write "Θ is an important parameter". It's better to write "The parameter Θ is important".
3. If you start a section with a subsection, there needs to be some text before the subsection starts. This text can be a summary of what is in the section. I see some writers don't follow my rule! But it bugs me to have a subsection just start up with no explanation of the section.
4. Journal articles do not have as many things in math display mode as a thesis or a report. This is to save space - printing is expensive. Oh, there is no printing these days, or very little! But it is the journal style that developed during the print era. Of course, if you refer to an equation, it is displayed

and numbered. A complicated equation (lots of fractions or layers with subscripts, long expressions) should be displayed since it would be hard to read in text mode, where you aren't allowed to use things like frac for fraction. Deciding whether to display or not is always a judgement call. In the end, the copy editor will make decisions, but it's good to have the submission somewhat in a journal style.

5. Bibliography things (not crucial at this point but must be done before submission): these issues are common in all stat journals, not just CJS. Biology journals tend to use numbers and list papers as they appear in the article. This is not the case in stat journals.

    5.1. Make sure you have caps where you need them - in your bib file, markov needs to be {M}arkov, for instance, to force capitalization.

    5.2. You have a bib-style that cites in the text via a number, as opposed to via author (data). CJS uses author (date). Also, CJS doesn't use full first name of authors. See the sample.pdf document that comes with the CJS template.

6. From CJS sample.pdf: "Try to avoid double subscripts, and never use triple subscripts." I think you're good – no triples, but you do have doubles. We can take a look at those at some point. This is a common expectation in stat journals.

7. From CJS sample.pdf: “Unless central or essential to the flow of the discussion, mathematical arguments should be deferred to the Appendix.” So we’ll see about that theorem you have. This is getting more common in stat journals - but it’s something you need to check.
8. From CJS sample.pdf: “When you submit the final version of your manuscript in Latex form, please include postscript files (.ps or .eps) for the figures, labelling them fig1.ps, fig2.ps, etc.” That’s a direction for the final submission, after review and acceptance. For this submission, you probably just submit one pdf file (I haven’t checked). If so, I’d put each figure and table on a separate page, not merged into text, at the end of the pdf document.
9. Vector notation - boldface or not? matrix notation - caligraphy or not? I find it easier to keep things straight if vectors are in bold, matrices in cal. It looks like CJS has no specific rules (some journals do) but seems to allow this (I looked at some published articles). I suggest using bold and caligraphy, but using newcommand in case we want to switch the notation. What do you think?

Things specific to this version.

1. I have commented out the section title lines in your subfiles (the ones you input) and placed the section title lines in main.tex. This way, Marie and I can see what sections you’re planning to include. But we can also see the

titles in the input files, so we don't get confused.

2. You have done a great job with the notation. I see that is not a simple task.
3. Some of the write-up can be a little cleaner, technically, and I think some of the explanatory material, although nice, can distract from the technical specifications of the model. It's good to put the explanatory stuff in the introduction. See my comments in the introductory section (which I've written in main.tex).

    3.1. Typically there is less detail in a journal article than in, say, a comprehensive proposal. For the journal article, you can assume that the reader either already knows about things like HMMs or can get enough of an idea from what you have said (possibly supplementing knowledge from your references). But it is useful to lay out the notation, as you have done. So ... there is a balance!

    3.2. In the end, you consider three models: CarHMM, HHMM, CarHHMM. So these need to be sharply defined, so the reader can glance back and skim.

    3.3. fine-scale, finescale, fine scale. I keep using different ones! I think if you use it as a noun, then it is fine scale. "The Markov chain is on a fine scale." I think if you use it as an adjective, it is fine-scale, as in "that is a fine-scale model". What do you think?

3.4. HMMs or HMM's, X_ts or X_t's?? I don't know, and I am inconsistent. I think either is acceptable. Which do you prefer?

1. INTRODUCTION

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, some 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. In response, researchers have introduced a variety of new statistical techniques to infer animal behavior from movement data (**?**).

One of the most prevalent techniques in recent literature is the hidden Markov model (HMM), where observations depend upon the state of an associated unobserved behavioral process following Markovian dynamics (**?**). Importantly, under the traditional HMM model, subsequent observations are assumed to be independent from one another after conditioning on the underlying behavioral process. However, this assumption is often violated in real world processes, especially when observations are taken at high frequencies. For example, the location of an animal at a given time is highly correlated with the location of that animal one second later. Several publications have dealt with this issue in the past, including the hidden movement Markov model (HMMM) (**?**) and the conditionally autoregressive hidden Markov model (CarHMM) (4). The CarHMM in particular explicitly models auto-correlation into an HMM while maintaining the structure needed to run the forward algorithm. It also only adds one additional parameter

per possible hidden state.

Another issue that arises in high-frequency data is that several simultaneous behavioral processes may occur at different time scales. In this work, we consider an example where killer whales exhibit a variety of different types of dives at a coarse scale, but also exhibit many different types of swimming behaviors within dives at a fine scale. One solution to this issue is to use a hierarchical hidden Markov model (HHMM) (5) (1). HHMMs model the entire time series in question as a nested structure of hidden Markov models (HMM) where each HMM corresponds to one behavioral process.

All HMMs presented so far assume Markovian dynamics in the underlying process (i.e. that any observation Y_t depends only on the behavioral state X_t and Y_{t-1} when conditioned on all previous time steps). At the shortest time scales, however, observations often exhibit complicated dependence structures which cannot be easily captured by traditional HMMs, CarHMMs, or HHMMs. Examples included periodic fluking behavior in killer whales off the coast of Vancouver, BC, and swimming patterns of horn sharks of the coast of Southern California (1). With improvements in tagging technology allowing for data to be collected at very high frequencies, noisy and non-Markovian fine-scale processes are likely to persist.

One solution is to model the fine-scale behavior using a continuous-time model, which involves modelling the dynamics of an animal as the solution

to a stochastic differential equation. Continuous-time models are more flexible than their discrete-time counterparts and can incorporate observations taken at irregular time intervals. Moreover, they are often computationally intractable and require approximate inference techniques such as Markov-chain Monte Carlo (MCMC) methods to perform inference.

For periodic behavior in particular, one way to avoid the use of continuous time models is to use signal processing techniques such as the Fourier transform on the raw data. The advantages of using Fourier analysis within an HMM has been recently demonstrated in the context of describing daily behavioral cycles of marine mammals (**?**). In addition, Fourier analysis has previously been used in the field of animal movement to explain animal behavior (**?**), specifically fluking (**?**) from accelerometer data. Thus, incorporating Fourier analysis of accelerometer data within the structure of an HMM appeared a promising simple approach to account for additional correlation in data that is cyclical in nature.

We consider several models: the classical hierarchical hidden Markov model (HHMM) (5), the CarHMM model of (4), and our new model that blends the two (CarHHMM). The HHMM consists of a coarse-scale process and a fine-scale process, with standard HMMs modelling both the coarse- and fine-scale processes. The CarHMM modifies the usual HMM by modelling additional dependence into the sequence of observed data. In our combination of the HHMM and the CarHHMM models, we keep the usual Markov structure for the coarse-

scale process, but use the CarHMM model for the fine-scale process.

This work investigates how to incorporate fine-scale processes into the larger structure of hierarchical hidden Markov models while maintaining computational efficiency. We describe a general procedure that can be used to extract features from highly structured fine-scale behaviors that otherwise could not be modeled with existing HMM models. In addition, we bridge the gap between the discrete CarHMM and certain continuous-time stochastic process models by showing that the two are equivalent under certain conditions. We then perform a simulation study to compare the performance each existing model with ours in a controlled setting. Finally, we apply our method to dive data collected from a Northern resident killer whale off the coast of British Columbia, Canada.

2. MODELS AND PARAMETER ESTIMATION

In this section we first remind the reader of the definition of a hidden Markov model (HMM). We then review three variations on the HMM: the CarHMM, HHMM, and HMM-DFT. We also briefly describe a continuous-time model and its connection with the CarHMM. Finally, we generalize the hierarchical structure of these models and describe how they can be flexibly altered and combined to form new models.

2.1. The hidden Markov model (HMM)

A *hidden Markov model* is comprised of a sequence of unobserved states X_t, $t = 1, \ldots, T$, and an associated sequence of possibly high-dimensional obser-

vations Y_t, $t = 1, \ldots, T$, The Y_t's are often referred to as "emissions" and the index t typically refers to time. The X_t's form a Markov chain and take possible values $1, \ldots, N$. Their distribution is governed by the distribution of the initial state X_1 and the $N \times N$ transition probability matrix Γ where $\Gamma_{ij} = \Pr(X_{t+1} = j | X_t = i)$, for $t = 1, \ldots, T-1$, and $i, j = 1, \ldots, N$. We assume that X_1 follows the chain's stationary distribution, which is denoted by $\delta \in \mathbb{R}^N$, with ith component $\delta_i = \Pr\{X_1 = i\}$, $i = 1, \ldots, N$. Recall that a Markov chain's stationary distribution is determined by its probability transition matrix via $\delta = \delta\Gamma$ and $\sum_{i=1}^{N} \delta_i = 1$. The distribution of an emission Y_t depends only on the corresponding state X_t and no other observations or hidden states: $p\left(y_t | \{X_1, \ldots, X_T\}, \{Y_1, \ldots, Y_T\}/\{Y_t\}\right) = p(y_t | X_t)$. These conditional distributions are governed by state-dependent parameters. If $X_t = i$, then the state-dependent parameter is $\theta^{(i)}$ and we denote the conditional distribution of Y_t given $X_t = i$ by its conditional density or probability mass function, denoted $f^{(i)}(\cdot; \theta^{(i)})$, or sometimes $f^{(i)}(\cdot)$. (Fig 1a) represents the dependence structure of an HMM.

To find the maximum likelihood estimates of the parameters Γ and $\Theta \equiv (\theta^{(1)}, \ldots, \theta^{(N)})$, let $y = (y_1, \ldots, y_T)$ be the observed emissions. We write the likelihood $\mathcal{L}_{\text{HMM}}$ in a specific form, using the well-known *forward algorithm* (**?**) as follows:

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

where $\mathbf{1}_N$ is an N-dimensional column vector of ones and $P(y_t; \Theta)$ is an $N \times N$ diagonal matrix with iith entry $f^{(i)}(y_t; \theta^{(i)})$.

We will always parameterize transition probability matrices to remove the constraints that the entries of the matrix are non-negative and that the rows sum to 1. We do this by parameterizing the $N \times N$ transition probability matrix Γ using $\eta \in \mathbb{R}^{N \times N}$ (5):

$$\Gamma_{ij} = \frac{\exp(\eta_{ij})}{\sum_{k=1}^{N} \exp(\eta_{ik})},$$

setting η_{ii} to zero, $i = 1, \ldots, N$, for identifiability. Then $\mathcal{L}_{\text{HMM}}(y; \Theta, \Gamma)$ can be maximized using any numerical optimizer. For simplicity, we will continue to use Γ in our notation, suppressing the reparameterization in terms of η.

2.2. The conditionally auto-regressive hidden Markov model (CarHMM)

A key assumption of an HMM is *conditional independence* between observations given the state sequence. Therefore, traditional HMMs do not hold when the observations exhibit certain forms of significant correlation in time.

The CarHMM, or *conditionally auto-regressive hidden Markov model*, introduced by (4), explicitly models auto-correlation within an HMM. Like a traditional HMM, A CarHMM is made up of a Markov chain of unobserved states $X_1, \ldots, X_T$ that can take on values $1, \ldots, N$, with transition probability matrix Γ and initial distribution δ equal to the stationary distribution of Γ. Unlike a traditional HMM, the CarHMM assumes that the distribution of Y_t, conditional on

$X_1, \ldots, X_T$ and $\{Y_1, \ldots, Y_{t-1}\}$, depends on *both* X_t *and* Y_{t-1}. The first emission Y_1 is assumed to be fixed as an initial value which does not depend upon X_1. (Fig 1b) shows the structure of a CarHMM.

We denote the conditional distribution of Y_t given $Y_{t-1} = y_{t-1}$ and $X_t = i$ as $f^{(i)}(\cdot|y_{t-1}; \theta^{(i)})$ or simply $f^{(i)}(\cdot|y_{t-1})$. As an example, we could assume that this conditional distribution is Normal with parameters $\theta^{(i)} = \{\mu^{(i)}, \sigma^{(i)}, \phi^{(i)}\}$ where:

$$\mathbb{E}(Y_t|Y_{t-1} = y_{t-1}, X_t = i) = \phi^{(i)}\, y_{t-1} \;+\; (1 - \phi^{(i)})\, \mu^{(i)}$$

and

$$\mathbb{V}(Y_t|Y_{t-1} = y_{t-1}, X_t = i) = (\sigma^{(i)})^2.$$

The likelihood for the CarHMM can be easily calculated using the forward algorithm. As previously, let y be the vector of observed emissions. Then

$$\mathcal{L}_{\text{CarHMM}}(y; \Theta, \Gamma) = \delta \prod_{t=2}^{T} \Gamma P(y_t|y_{t-1}; \Theta)\mathbf{1}_N \tag{1}$$

where $P(y_t|y_{t-1}; \Theta)$ is an $N \times N$ diagonal matrix with iith entry equal to $f^{(i)}(y_t|y_{t-1}; \theta^{(i)})$.

2.3. Continuous-time processes

While CarHMMs can model auto-correlation within the observation sequence, they fail when observations are taken at irregular time intervals. Another way to model a stochastic process is to use a stochastic differential equation (SDE). As an example, (6) model the movement of an animal as the solution to the following

stochastic differential equation:

$$dY_t = \beta^{(X_t)}(\gamma^{(X_t)} - Y_t)dt + \omega^{(X_t)}dW_t \tag{2}$$

where X_t follows some stochastic process which defines the hidden behaviour 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. This SDE is referred to as a state-switching Ornstein-Uhlenbeck process. Unlike HMMs, the time index $t \in \mathbb{R}$ exists in continuous time and is not necessarily an integer. If the behavioral state X_t is known and does not change between observations, the solution to (eqn 2) is known to be the following (6):

$$Y_{t+\Delta t}|X_t \sim \mathcal{N}\left((1 - e^{-\beta^{(X_t)}\Delta t})\gamma^{(X_t)} + e^{-\beta^{(X_t)}\Delta t}Y_t, \quad \frac{{\omega^{(X_t)}}^2}{2\beta^{(X_t)}}(1 - e^{-2\beta^{(X_t)}\Delta t})\right) \tag{3}$$

where Δt is the time difference between any two observations Y_t and $Y_{t+\Delta t}$. Most continuous time models are difficult to incorporate into an HHMM and require MCMC methods to fit. However, under certain conditions, the CarHMM is equivalent to a state-switching Ornstein-Uhlenbeck process. This gives new interpretation to the learned parameters of the CarHMM in the context of a continuous-time model.

Theorem 1. *If:*

1. *The hidden behavioural process X from (eq. 2) follows a Markov chain with N possible states and transitions occur at equi-spaced time stamps $(\Delta t, \ldots, (T-1)\Delta t)$, and:*
2. *Observations of the SDE from (eq. 2) are taken at times $(0, \Delta t, \ldots, (T-1)\Delta t)$,*

then the observations Y are equivalent to the output of a conditionally autoregressive hidden Markov model with normal emission distributions and parameters $\Theta = (\theta^{(1)}, \ldots, \theta^{(N)})$; $\theta^{(i)} = \{\mu^{(i)}, \sigma^{(i)}, \phi^{(i)}\}$, where:

$$\mu^{(i)} = \gamma^{(i)}, \qquad \sigma^{(i)} = \sqrt{\frac{\omega^{(i)^2}}{2\beta^{(i)}}(1 - e^{-2\beta^{(i)}\Delta t})}, \qquad \phi^{(i)} = e^{-\beta^{(i)}\Delta t} \tag{4}$$

See the appendix for a proof of Theorem 1.

2.4. The hierarchical hidden Markov model (HHMM)

A *hierarchical hidden Markov model* contains both a coarse-scale process and a fine-scale process, both of which are HMMs. The coarse-scale process is a hidden Markov model as defined previously, where $X_1, \ldots, X_T$ make up an unobserved Markov chain with N possible states and $Y_1, \ldots, Y_T$ are the corresponding observed responses. In the hierarchical setting, each state X_t emits another sequence of fine-scale unobserved states, $X_t^* \equiv (X_{t,1}^*, \ldots, X_{t,T_t^*})$ and a sequence of fine-scale observed emissions, $Y_t^* \equiv (Y_{t,1}^*, \ldots, Y_{t,T_t^*})$. This fine-scale process can be either instead of or in addition to the coarse-scale observations Y_t. The

length of each fine-scale process is denoted as T_t^* and can often be controlled by the practitioner. Conditional on the coarse-scale hidden state X_t, the joint fine-scale process, (X_t^*, Y_t^*), is a hidden Markov model with parameters depending on the value of X_t. Specifically, given that $X_t = i$, X_t^* is a Markov chain with states $1, \ldots, N_t^*$. The distribution of X_t^*, conditional on $X_t = i$, is given by an $N_t^* \times N_t^*$ transition probability matrix $\Gamma^{*(i)}$ and initial probability, denoted by the N_t^*-vector $\delta^{*(i)}$, which we assume is equal to the stationary distribution of the chain. For simplicity, we take $N_t^* \equiv N^*$ although this is not necessary. The distribution of Y_{t,t^*} given $X_{t,t^*} = i^*$ and $X_t = i$ is governed by a parameter $\theta^{(i,i^*)}$ and has density or probability mass function denoted $f^{*(i,i^*)}\left(\cdot; \theta^{(i,i^*)}\right)$ or simply $f^{*(i,i^*)}(\cdot)$. Let $\Theta^{*(i)} = \left(\theta^{(i,1)}, \ldots, \theta^{(i,N^*)}\right)$, the fine-scale emission parameter vector corresponding to $X_t = i$. Given the coarse-scale states, $X_1, \ldots, X_T$, the T fine-scale processes $(X_1^*, Y_1^*), \ldots, (X_T^*, Y_T^*)$, are independent HMMs. Depending upon the process being modeled, it is possible to force certain parameters to be shared across different coarse or fine states. For example, in the killer whale case study in section 4, we force the fine-scale emission parameters to be shared across coarse-scale hidden states (i.e. $\theta^{(1,i^*)} = \ldots = \theta^{(N,i^*)}$ for $i^* = 1, \ldots, N^*$). Figure 1c shows a graphical representation of the dependence structure for an HHMM.

Due to the nested structure of the hierarchical hidden Markov model, the likelihood is easy to calculate via the forward algorithm. Let y be the T-vector of the

observed coarse-scale emissions and y^* be the $(T_1^* + \cdots + T_T^*)$-vector of the observed fine-scale emissions. Let Θ^* denote the collection of all fine-scale emission parameters, $\Theta^{*(i)}$, $i = 1, \ldots, N$, and let Γ^* denote the collection of all fine-scale transition probability matrices, $\Gamma^{*(i)}$, $i = 1, \ldots, N$. The likelihood of the observed data is

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

where $P(y_t, y_t^*; \Theta, \Theta^*, \Gamma^*)$ is an $N \times N$ diagonal matrix with iith entry corresponding to $X_t = i$ and equal to $f^{(i)}(y_t)\mathcal{L}_{\text{HMM}}\left(y_t^*; \Theta^{*(i)}, \Gamma^{*(i)}\right)$.

For more information on specific considerations for HHMMs such as incorporating covariates into the probability transition matrix, state decoding, model selection and model checking, see (1).

2.5. The HMM with discrete Fourier transform (HMM-DFT)

The HMM with discrete Fourier transform, or HMM-DFT, incorporates hierarchical structure into an HMM differently than an HHMM. In particular, the fine-scale process is no longer modeled with an HMM and instead summarized using its Fourier transform. For simplicity, we assume that the length of the fine-scale processes is constant (i.e. that $T_t^* = T^*$), although this need not be the case in general. Suppose that the fine-scale process y_t^* does not switch hidden states, but does exhibit significant periodic behaviour. We then suggest using the discrete

Fourier transform (DFT) on y_t^*:

$$DFT\{y_t^*\}(k) := \hat{y}_t^{*(k)} = \sum_{t^*=1}^{T^*} y_{t,t^*}^* \exp\left(-i\frac{2\pi k}{T^*}(t^*-1)\right) \qquad k = 0, 1, \ldots, T^*-1$$

Summary statistics can then drastically reduce the dimension of $\hat{y}_t^*$. One such example is as follows:

$$z_t^{*(1)} := \mathcal{R}\left(\hat{y}_t^{(0)}\right) \qquad z_t^{*(2)} := \frac{1}{T^*}\sum_{k=1}^{\tilde{\omega}} |\hat{y}_t^{(k)}|^2 \tag{5}$$

$z_t^{*(1)}$ is the average value of y_t^* and $z_t^{*(2)}$ is the squared 2-norm of the component of y_t^* that can be attributed to frequencies between 1 and $\tilde{\omega}$ periods per window length T^*. The maximum frequency $\tilde{\omega}$ is a problem-specific tuning parameter which should be selected with care. These summary statistics are just one possible choice to describe each window; other choices include the dominant frequency and amplitude of y_t^*. Figure (**??**) visually displays the process of transforming y^* into z^*.

Once z_t^* is calculated, it can be treated as an observation of the coarse-scale HMM and incorporated into the emission distribution $f^{(i)}\left(y_t, z_t^*; \theta^{(i)}\right)$, or more succinctly $f^{(i)}\left(y_t, z_t^*\right)$. In total, the likelihood of the hierarchical HMM-DFT is as follows:

$$\mathcal{L}_{\text{HMM-DFT}}(y, z^*; \Theta, \Gamma) = \delta P(y_1, z_1^*; \Theta)\prod_{t=2}^{T}\Gamma P(y_t, z_t^*; \Theta)\mathbf{1}_N \tag{6}$$

where $P(y_t, z_t^*; \Theta)$ is an $N \times N$ diagonal matrix with iith entry equal to $f^{(i)}\left(y_t, z_t^*; \theta^{(i)}\right)$.

It is possible to accommodate for unequal time steps within y_t^* by using the non-uniform discrete Fourier transform (NDFT). We do not describe this method here, but the generalization is straightforward. Refer to (**?**) for details.

2.6. General hierarchical structures

In addition to the DFT and HMM models described above, the fine-scale process Y_t^* can be modeled with *any* parametric model which admits an easy-to-compute likelihood. The state-specific emission distribution $\mathcal{L}_{\text{HMM}}$ from the HHMM likelihood is then replaced by the likelihood of the fine-scale model, $\mathcal{L}_{\text{fine}}(\mathbf{y}_t^*; \Theta^{*(i)})$:

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

where $P(y_t, y_t^*; \Theta, \Theta^*)$ is an $N \times N$ diagonal matrix with iith entry corresponding to $X_t = i$ and equal to $f^{(i)}\left(y_t; \Theta^{(i)}\right) \mathcal{L}_{\text{fine}}\left(y_t^*; \Theta^{*(i)}\right)$. This definition is straightforward to extend to the CarHMM as well:

$$\mathcal{L}_{\text{coarse}}(y, y^*; \Theta, \Theta^*, \Gamma) = \delta \prod_{t=2}^{T} \Gamma P(y_t, y_t^* | y_{t-1}; \Theta, \Theta^*) \mathbf{1}_N$$

where $P(y_t, y_t^* | y_{t-1}; \Theta, \Theta^*)$ is an $N \times N$ diagonal matrix with iith entry corresponding to $X_t = i$ and equal to $f^{(i)}\left(y_t | y_{t-1}; \Theta^{(i)}\right) \mathcal{L}_{\text{fine}}\left(y_t^*; \Theta^{*(i)}\right)$.

Possible candidates for the fine-scale model include any of the models described in the previous subsections (HMM, CarHMM, state-switching OU, HHMM, and HMM-DFT). These HMM models can act as building blocks which can be used to construct increasingly complex hierarchical models based on the HMMs. In the sections that follow, we perform both a simulation study and real-

world case study modelling killer whale dive behaviour using models constructed from these building blocks.

3. SIMULATION STUDY

We base the following simulated study on the acceleration data of a killer whale's dive sequence. The parameters used to generate the data are loosely based on those learned from the case study in Section 4. We fit four separate models to the simulated data and compare their performances against one another.

3.1. Data Simulation

500 separate sequences of 100 killer whale dives were simulated according to an HMM, where the hidden Markov chain X was set as a collection of dive types and the observations Y were the corresponding dive durations (in seconds). Each dive could be one of $N = 2$ dive types, and the duration of dive t, Y_t, followed a gamma distribution whose parameters $\theta^{(i)}$ were dependent on the dive type $X_t = i$. We parameterize the gamma distribution by its mean and variance:

$$\Gamma = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \qquad \delta = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$$

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

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

$$\mathbb{V}(Y_t|X_t = 1) = 25s^2 \qquad \mathbb{V}(Y_t|X_t = 2) = 100s^2$$

Once the dive durations were calculated for all 100 dives, dive t was broken into a sequence of $T_t^* = \lfloor Y_t/2 \rfloor$ two-second segments (the end of the dive sequence was discarded) which made up a second fine-scale hidden Markov model. Each two second segment of the dive was assigned one of $N^* = 2$ behaviours (active swimming or passive gliding) according to a fine-scale Markov chain $X_t^* \equiv \left(X_{t,1}^*, \ldots, X_{t,T_t^*}^*\right)$, and the probability transition matrices for these fine-scale Markov chains were set as

$$\Gamma^{*(1)} = \begin{pmatrix} 0.5 & 0.5 \\ 0.9 & 0.1 \end{pmatrix}, \qquad \Gamma^{*(2)} = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix},$$

where $\Gamma^{*(1)}$ was used for dives where $X_t = 1$ and $\Gamma^{*(2)}$ was used for dives where $X_t = 2$.

Each two second sub-dive window had 100 associated acceleration readings, $Y_{t,t^*}^* \equiv \left(Y_{t,t^*,1}^*, \ldots, Y_{t,t^*,100}^*\right)$. To accurately recreate active swimming versus passive gliding on the fine-scale Markov chain, the DFT of each two second segment $\hat{Y}_{t,t^*}^*$ was simulated such that $Z_{t,t^*}^{*(1)}$ and $Z_{t,t^*}^{*(2)}$ (see Eqn 5) would have the following distributions:

$$\left(Z^{*(1)}_{t,t^*}|Z^{*(1)}_{t,t^*-1}, X^*_{t,t^*}=1\right) \sim \mathcal{N}\left(\phi^{*(1)}Z^{*(1)}_{t,t^*-1}+(1-\phi^{*(1)})\mu^{*(1)}, (\sigma^{*(1)})^2\right)$$

$$\left(Z^{*(1)}_{t,t^*}|Z^{*(1)}_{t,t^*-1}, X^*_{t,t^*}=2\right) \sim \mathcal{N}\left(\phi^{*(2)}Z^{*(1)}_{t,t^*-1}+(1-\phi^{*(2)})\mu^{*(2)}, (\sigma^{*(2)})^2\right)$$

$$\mu^{*(1)}=0.0,\ \ \sigma^{*(1)}=0.05,\ \ \phi^{*(1)}=0.99$$

$$\mu^{*(2)}=0.0,\ \ \sigma^{*(2)}=0.1,\ \ \phi^{*(2)}=0.95$$

$$\left(Z^{*(2)}_{t,t^*}|X^*_{t,t^*}=1\right) \sim \text{Gamma}\left(\alpha^{*(1)}, \beta^{*(1)}\right)$$

$$\left(Z^{*(2)}_{t,t^*}|X^*_{t,t^*}=2\right) \sim \text{Gamma}\left(\alpha^{*(2)}, \beta^{*(2)}\right)$$

$$\alpha^{*(1)}=10.10, \quad \beta^{*(1)}=1.00$$

$$\alpha^{*(2)}=305.94, \quad \beta^{*(2)}=1.00$$

Sub-dive behavior 1 corresponds to passive gliding while sub-dive behaviour 2 corresponds to active swimming with a dominant frequency of $\frac{1}{2}s^{-1}$. Sub-dive behaviors 1 and 2 are the same for both dive types. See the appendix for more details regarding procedure for simulating $\hat{Y}^*$ and Y^* such that Z^* has the preceding distribution. Figure 2 shows the first 5 dives of one simulated data set.

3.2. Model Formulation

The building blocks from the previous section were used to build a well-specified model for this simulated data. Specifically, we used a hierarchical HMM where the sequence of dive durations Y was modeled using a simple HMM, and

the fine-scale process Z^* was modeled using a HMM-DFT with explicit autocorrelation in $Z^{*(1)}$. Naturally, we refer to this model as the *CarHHMM-DFT*.

On the coarse scale, the dive types follow a Markov chain with $N = 2$ possible states and unknown probability transition matrix Γ. Given the dive type, the duration of a dive follows a gamma distribution which depends upon the dive type and unknown parameters $\Theta = \{\{\mu^{(1)}, \sigma^{(1)}\}, \{\mu^{(2)}, \sigma^{(2)}\}\}$.

On the fine scale, the sub-dive behavior of each two-second window comprises a Markov chain with $N^* = 2$ possible states and unknown probability transition matrices $\Gamma^{*(1)}$ and $\Gamma^{*(2)}$, depending upon the dive type. Each two-second window is summarized by the observations $Z^{*(1)}_{t,t^*}$ and $Z^{*(2)}_{t,t^*}$. The distribution of $Z^{*(1)}_{t,t^*}$ is Normal and its parameters depend upon the sub-dive behavior X^*_{t,t^*} and $Z^{*(1)}_{t,t^*-1}$. In particular:

$$\mathbb{E}(Z^{*(1)}_{t,t^*}|Z^{*(1)}_{t,t^*-1} = z, X^*_{t,t^*} = i^*) = \phi_1^{*(i^*)} z + (1 - \phi_1^{*(i^*)})\mu_1^{*(i^*)}$$

$$\mathbb{V}(Z^{*(1)}_{t,t^*}|Z^{*(1)}_{t,t^*-1} = z, X^*_{t,t^*} = i^*) = \left(\sigma_1^{*(i^*)}\right)^2$$

The distribution of $Z^{*(2)}_{t,t^*}$ is gamma and its parameters depend upon only X^*_{t,t^*}:

$$\mathbb{E}(Z^{*(2)}_{t,t^*}|Z^{*(1)}_{t,t^*-1} = z, X^*_{t,t^*} = i^*) = \mu_2^{*(i^*)}$$

$$\mathbb{V}(Z^{*(2)}_{t,t^*}|Z^{*(1)}_{t,t^*-1} = z, X^*_{t,t^*} = i^*) = (\sigma_2^{*(i^*)})^2$$

None of the fine-scale emission distributions depend upon dive type. In total the parameters to learn are

$$\Gamma, \qquad \Gamma^* = \{\Gamma^{*(1)}, \Gamma^{*(2)}\} \qquad \text{(probability transition matrices)}$$

$$\Theta = \{\{\mu^{(1)}, \sigma^{(1)}\}, \{\mu^{(2)}, \sigma^{(2)}\}\} \qquad \text{(coarse-scale emission parameters)}$$

$$\Theta^* = \{\Theta^{*(1)}, \Theta^{*(2)}\} \qquad \text{(fine-scale emission parameters)}$$

$$\Theta^{*(i^*)} = \{\{\mu_1^{*(i^*)}, \sigma_1^{*(i^*)}, \phi_1^{*(i^*)}\}, \{\mu_2^{*(i^*)}, \sigma_2^{*(i^*)}\}\} \qquad (Z^{*(1)} \text{ and } Z^{*(2)} \text{ parameters})$$

The likelihood of this model is still easy to calculate using the forward algorithm, and it can be maximized with respect to the parameters above. See the appendix for details of likelihood evaluation, and Figure **??** shows the corresponding graphical model.

Including the CarHHMM-DFT described above, four different models were fit to the simulated data sets:

1. A **CarHHMM-DFT** as described above.
2. An **HHMM-DFT**, which is similar to the model above, but with no modeled auto-correlation, i.e. $\phi_1^{*(i^*)} = 0$ for $i^* = 1, 2$.
3. A **CarHHMM**, which is similar to the model above, but without $Z^{*(2)}$ as an observation, i.e $\Theta^{*(i^*)} = \{\mu_1^{(i^*)}, \sigma_1^{(i^*)}, \phi_1^{(i^*)}\}$, $i^* = 1, 2$.
4. A **CarHMM-DFT** similar to the CarHHMM-DFT, but with $N = 1$ instead of $N = 2$, i.e. $\Gamma^* = \{\Gamma^{*(1)}\}$ and $\Theta^* = \{\Theta^{*(1)}\}$. This is equivalent to loosing one level of the hierarchical structure.

Each of the last three models leaves out one important aspect of the full CarHHMM-DFT. The CarHMM-DFT lacks a hierarchical structure, the HHMM-DFT is missing auto-correlation within the fine-scale observations, and the CarHHMM does not have access to the Fourier transform sums $(Z^{*(2)})$ as observations. All models were run on the Cedar Compute Canada cluster with 1 CPU and 4 GB of dedicated memory per model.

3.3. Simulation Results

Every model was able to decode the fine-scale hidden states of the process almost perfectly except for the *CarHHMM*, which did not have access to the Fourier modes ($Z^{*(2)}$). This is intuitively clear because the distribution of $Z^{*(2)}$ varies between fine-scale states much more than $Z^{*(1)}$. For the coarse-scale hidden states, the CarHMM lacked a hierarchical structure and could not make any predictions at all. The other three models all achieved an accuracy of approximately 90%, with the CarHHMM slightly more likely to categorize a dive as dive type 2 than the other models. Figure 4 shows the decoded state probabilities for both the fine- and coarse- scales. (Tbl. 1) also lists more details regarding the accuracy and training times of each model.

For the emission distributions of dive duration, estimates of standard error using the observed Fisher information tended to be underestimates due to correlation between $\hat{\mu}$ and $\hat{\sigma}$. In addition, $\hat{\sigma}$ tends to underestimate σ for all models, which is a finding consistent with properties of MLEs, especially because the

sample size for a particular dive type is approximately 50 for each simulation. The CarHMM model in particular severely underestimates σ. A full table of parameter estimates for all models is shown in Table 2.

For acceleration ($Z^{*(1)}$), both the CarHMM-DFT and the CarHHMM-DFT regularly converged to the correct parameters with very little standard error. However, the Fisher standard error regularly overestimated the standard error for $\hat{\mu}$ for both of these models. The HHMM regularly overestimates the variance $Z^{*(1)}$ since it does not incorporate auto-correlation into the emission distribution of $Z^{*(1)}$, and the CarHHMM has large biases in many of its parameter estimates, especially in sub-dive state 2. See Table 3 for a detailed breakdown of parameter values for acceleration emission distributions.

For all models there is no bias in the parameter estimates for the distribution of the Fourier sums ($Z^{*(2)}$). One exception is the CarHHMM, which does not model $Z^{*(2)}$ as an observation. The observed fisher information standard errors are good approximations of the empirical standard error. (Tbl. 4) shows a detailed breakdown of the emission distribution of $Z^{*(2)}$ for all models.

Estimates of the probability transition matrix Γ for all models (except for the CarHMM-DFT) are very accurate, with empirical standard errors of approximately 0.02. This is significantly less than the observed fisher information estimate of standard error of approximately 0.08. For Γ^*, the HHMM-DFT and CarHHMM-DFT both showed practically no bias with standard errors on the or-

der of 10^{-2}. One notable exception is the standard error of $\hat{\Gamma}_{12}^{*(1)}$, whose standard error was on the order of 10^{-4}, which is much lower than the observed Fisher information predicted. The CarHMM-DFT is mis-specified, so its results cannot be easily interpreted. The CarHHMM consistently underestimated the probability of a hidden state change on the sub-dive behavior. Again, one notable exception is $\hat{\Gamma}_{12}^{*(1)}$, which had almost no bias and was remarkably consistent. See (Table 5) for a full list of estimates and standard errors.

I also have plots corresponding to each table. For example, I have included figure 5 as an example below for the emission distribution of the MLEs for dive duration of the CarHMM. Should I use those instead? But there are just SO many plots (4 models * 3 observations * 2 states = 24 plots)

4. KILLER WHALE CASE STUDY

The CarHHMM-DFT was used to analyze dive data from a Northern Resident Killer Whale (NRKW) off the coast of British Columbia, Canada. Acceleration data can be a good proxy for energy expenditure (**?**), but studies suggest that the behavioral state must be taken into account when using accelerometer data (**?**). Therefore, understanding both the behavioral state of the killer whale as well as the distribution of accelerometer data within each behavioral state is import to understand the energetic requirements of killer whales. This, in turn, can assist conservation efforts.

4.1. Data Collection and Preprocessing

The data used in this study was collected on September 2, 2019 from 12:49 pm to 6:06 pm and consists of depth and acceleration in three orthogonal directions. Observations were collected at a rate of 50 Hz. Tagging the killer whale caused anomalous behavior before 1:20 pm and after 6:00 pm, so observations in this time range were ignored. In addition, the tagging technology dropped data between 2:25pm and 2:37pm as well as between 4:07 and 5:07 pm, so any partially observed data within this time range were ignored as well. A killer whale "dive" is considered to be any continuous chunk of data that occurs below 0.5 meters in depth and lasts for at least 10 seconds. Accelerometer and depth data were smoothed by taking a moving average with a window of 1/10th of a second. Data preprocessing was done in part with the *divebomb* package in Python (7). After preprocessing the raw data, a total of 267 dives were observed. A plot of the raw data for all dives and a collection of five selected dives can be seen in Figure 6 and Figure 7, respectively.

4.2. Model Selection

The coarse-scale observations were made up of the collection of dive durations in seconds, and the fine-scale observations were determined from the within-dive acceleration data. The dive durations Y_t were assumed to follow a gamma distribution with unknown parameters $\{\mu, \sigma\}$:

$$\mathbb{E}(Y_t|X_t = i) = \mu^{(i)}$$

$$\mathbb{V}(Y_t|X_t = i) = \left(\sigma^{(i)}\right)^2$$

The acceleration exhibits significant sinusoidal behavior at several points in time (see fig 7), so the fine-scale observations Z^* were made up of the DFT summary statistics of a two-second sliding window. Acceleration data was a 3-dimensional vector, so Z^* was calculated as follows:

$$\mathbf{z}_{t,t^*}^{*(1)} := \mathcal{R}\left(\hat{\mathbf{y}}_{t,t^*}^{*(0)}\right) \qquad z_{t,t^*}^{*(2)} := \frac{1}{100}\sum_{k=1}^{10} ||\hat{\mathbf{y}}_{t,t^*}^{(k)}||^2$$

$\mathbf{z}_{t,t^*}^{*(1)}$ is a 3-dimensional vector while $z_t^{*(2)}$ is a scalar. Summing the first 10 Fourier modes to calculate $z_{t,t^*}^{(2)}$ corresponds to a maximum frequency of $\tilde{\omega} = 5$ Hz. Figure 8 displays a lag plot for all major features and reveals strong auto-correlation within $\mathbf{Z}_{t,t^*}^{*(1)}$ for all dimensions. Auto-correlation was therefore directly modeled into the the distribution of $\mathbf{Z}_{t,t^*}^{*(1)}$, which is assumed to be Normally distributed with the following parameters:

$$\mathbb{E}(\mathbf{Z}_{t,t^*}^{*(1)}|\mathbf{Z}_{t,t^*-1}^{*(1)} = \mathbf{z}, X_{t,t^*}^* = i) = \phi_1^{*(i)}\mathbf{z} + (1 - \phi_1^{*(i)})\mu_1^{*(i)}$$

$$\mathbb{V}(\mathbf{Z}_{t,t^*}^{*(1)}|\mathbf{Z}_{t,t^*-1}^{*(1)} = z, X_{t,t^*}^* = i) = \text{diag}\left[\left(\sigma_1^{*(i)}\right)^2\right]$$

where $\phi_1^{*(i)} \in \mathbb{R}$, $\mu_1^{*(i)} \in \mathbb{R}^3$, and $\sigma_1^{*(i)} \in \mathbb{R}^3$.

While $Z^{*(2)}_{t,t^*}$ also exhibits some auto-correlation, the relationship is less strong, and the biological interpretation of auto-correlation within $Z^{*(2)}_{t,t^*}$ is less clear. Auto-correlation was therefore not incorporated into the emission distribution of $Z^{*(2)}_{t,t^*}$. The distribution of $Z^{*(2)}_{t,t^*}$ was also assumed to be gamma:

$$\mathbb{E}(Z^{*(2)}_{t,t^*}|Z^{*(1)}_{t,t^*-1} = z, X^*_{t,t^*} = i) = \mu_2^{*(i)}$$

$$\mathbb{V}(Z^{*(2)}_{t,t^*}|Z^{*(1)}_{t,t^*-1} = z, X^*_{t,t^*} = i) = \left(\sigma_2^{*(i)}\right)^2$$

$Z^{*(2)}_{t,t^*}$ and $\mathbf{Z}^{*(1)}_{t,t^*}$ were assumed to be independent when conditioned of the sub-dive state X_{t,t^*}.

It is known that information criteria tends to overestimate the number of states in biological processes (**?**), so we instead selected $N = 2$ dive types and $N^* = 3$ sub-dive behaviours heuristically and admittedly somewhat arbitrarily. The absence of principled method to select the number of hidden states is a common issue in statistical ecology, so it is important to use model validation techniques in lieu of information criteria (see section 4.4).

The final model is nearly identical to the one from the simulation study, with the exception that the fine scale Markov chain has three sub-dive behaviors instead of two ($N^* = 3$), and that the observation $\mathbf{z}^{*(1)}_{t,t^*}$ is a vector rather than a scalar. The model is comprised of two levels: a coarse-scale HMM and a fine-scale CarHMM-DFT. The coarse-scale model is comprised of an HMM with no auto-correlation and no DFT with hidden states corresponding to dive types and

observations corresponding to dive durations. The fine-scale model is comprised of a CarHMM-DFT where auto-correlation is modeled into the distribution of $\mathbf{Z}^{*(1)}$ (the average acceleration with a 2-second window), but not $Z^{*(2)}$ (the "wiggliness" of the two second window). Refer to Figure **??** for a full graphical representation of this model.

4.3. Results

Table 6 displays estimates of the emission distribution parameters. Each fitted distribution is also plotted in Figure 9 and Figure 10. While the ecological interpretation of behavioral states is tenuous for HMMs, we hypothesize the following meanings. Dive type 1 corresponds to shorter, shallower dives which serve a variety of purposes, including as rest before dives of type 2, which are deeper and more sustained. Sub-dive behavioural state 1 corresponds to gliding and less overall activity compared to the other behavioral states. The mean of $Z^{*(2)}$ in this state is at least an order of magnitude smaller than sub-dive behavior 2, the variance of $\mathbf{Z}^{*(1)}$ is smaller than sub-dive behavior 2 for every component, and the auto-correlation of $\mathbf{Z}^{*(1)}$ is higher than every other behavioral state. Sub-dive state 3, on the other hand, corresponds to vigorous swimming activity, as the mean of $Z^{*(2)}$ and variance of $\mathbf{Z}^{*(1)}$ for every component are much higher than every other state. The auto-correlation of $\mathbf{Z}^{*(1)}$ is also much lower in this state, implying more variation in acceleration every 2 seconds. Finally, sub-dive state 2 corresponds to a moderate amount of activity, as almost every parameter estimate

is between the other two behavioral states.

The estimated probability transition matrices and associated stationary distributions are

$$\hat{\Gamma} = \begin{pmatrix} 0.849 & 0.151 \\ 0.907 & 0.093 \end{pmatrix}$$

$$\hat{\delta} = \begin{pmatrix} 0.857 & 0.143 \end{pmatrix}$$

$$\hat{\Gamma}^{*(1)} = \begin{pmatrix} 0.724 & 0.276 & 0.000 \\ 0.057 & 0.887 & 0.056 \\ 0.000 & 0.247 & 0.753 \end{pmatrix} \qquad \hat{\Gamma}^{*(2)} = \begin{pmatrix} 0.871 & 0.129 & 0.000 \\ 0.135 & 0.829 & 0.036 \\ 0.000 & 0.246 & 0.754 \end{pmatrix}$$

$$\hat{\delta}^{*(1)} = \begin{pmatrix} 0.143 & 0.698 & 0.159 \end{pmatrix} \qquad \hat{\delta}^{*(2)} = \begin{pmatrix} 0.476 & 0.456 & 0.067 \end{pmatrix}.$$

About 86% of observed dives are short; the whale usually rests for many dives in a row before performing a deep dive. The probability transition matrix $\hat{\Gamma}$ shows that the distribution of a particular dive's type is approximately equal regardless of the previous dive type. The fine-scale probability transition matrices imply that the killer whale is much more likely to be in a less active sub-dive state when performing deep dives than when performing shallow dives- 14% of shallow dives are spent in sub-dive state 1 while 48% of deep dives are spent in sub-dive state 1. Using less active sub-dive states when diving deep could be an energy reduction strategy for these long periods of holding breath- a phenomenon which

has been observed in bottle-nose dolphins (**?**). (Fig 11) shows the decoded dive behavior of 5 selected dives, and (fig 12) and (fig 13) display the probability of each dive type and sub-dive state, respectively.

4.4. Model Validation

Two visual tools were used to evaluate this model: pseudo-residuals and empirical histograms. A pseudo-residual of a particular observation is the marginal CDF of an observation conditioned on all other observations under the learned model (**?**). To easily visualize outliers, the pseudo-residual is often passed through the quantile function of the standard Normal distribution. Mathematically, the pseudo-residual of an observation y_t is $\Phi^{-1}\left(Pr(Y_t < y_t | \{Y_1, \ldots, Y_T\}/\{Y_t\})\right)$, where Φ is the cumulative distribution function of a standard Normal distribution. If the model is correct, then all pseudo-residuals are independent and follow a standard Normal distribution. We find that histograms of the pseudoresiduals of this model mostly support that the model is well-specified. $Z^{*(2)}$ is an exception, as its pseudo-residuals are noticeably right-skewed (fig 14). This implies that the true distribution of $Z^{*(2)}$ may follow a heavier-tailed distribution compared to a gamma distribution such as a power law.

In addition to psuedoresiduals, we plotted histograms of each feature where each observation was weighted by the probability that the whale was in a particular hidden state. This empirical distribution was then plotted over the fitted

probability distribution function of that feature and hidden state. If the model is correct, then the histograms of features should closely resemble their respective fitted probability distribution. Our results mostly support a well-specified model with the exception of $Z^{*(2)}$, which is right-skewed. In addition, $\mathbf{Z}^{*(1)}$ has heavy tails for sub-dive state 3 (fig. 15), indicating the existence of rare events corresponding to very violent thrashing of the killer whale. These outliers are potential subjects for future study.

5. DISCUSSION

We presented a collection of HMM models which can be combined together in increasingly complex hierarchical models to match the complexity of particular problems faced by practitioners. This flexible framework can be used to deal with complicated dependence structures within time-series data.

Traditional HMMs can be used to model a state-switching process with conditionally independent observations and Markovian dynamics within the hidden state. However, many real-world processes are more complicated than this and require more complex models.

The CarHMM generalizes the HMM by explicitly modeling auto-correlation in the emission distributions of the HMM while maintaining the structure needed to evaluate the likelihood using the forward algorithm. In our normal model formulation, we have added only one additional parameter, $\phi^{(i)}$, per possible hidden state (4). Several useful model selection tools such as the lag plot can test if there

is significant auto-correlation within an observation sequence.

Although the CarHMM can incorporate auto-correlation into the structure of an HMM, it can break down when observations are taken at irregular time intervals. A common solution to this issue is to use a continuous-time method such as the state-switching OU processes described by (6). Most continuous time models require relatively slow MCMC algorithms to perform inference, and as a result, are not easily incorporated into the HHMM structure. However, we prove that certain continuous-time methods are equivalent to an CarHMM under certain conditions (see appendix).

For simultaneous observed processes taking place at different time scales, the HHMM as described by (5;1) utilizes hierarchical structures to jointly model both as HMMs. In particular, each hidden state of a coarse-scale HMM is assumed to emit both an observation Y_t as well as another fine-scale HMM with hidden states X_t^* and observations Y_t^*.

Finally, for processes with very high sampling frequencies and/or with intricate fine-scale structure, the HHMM structure developed by (5;1) can be generalized so the fine-scale model can be any model which admits an easy-to-calculate likelihood. For example, if the sampling rate of the fine-scale process is very high, then the fine-scale model can be described by a simple probability distribution over the summary statistics of a moving window of observations. In addition, researchers can recursively stack hierarchical HMM models together

like building blocks to create increasingly complex models. This should be done with care, as it is important to balance the need to effectively capture the process in question with the need to avoiding over-fitting and slow parameter estimation.

One way to temper model complexity is to reduce the dimension of the parameter space by forcing fine-scale states to be shared across the coarse-scale states. Even still, model complexity inevitably grows rapidly as hierarchical structures are stacked on top of each other.

An example of balancing model complexity with fast model fitting is presented in the simulation study. While (fine-scale) sub-dive behavioral states were shared across (coarse-scale) dive types to reduce model complexity for the hierarchical models, by far the fastest model to train ($\tilde{1}5$ minutes) was the CarHMM-DFT. The CarHMM-DFT had no hierarchical component but near-perfect accuracy when decoding sub-dive behavioral states of the simulated whale. This model would be preferred if simultaneous modeling of dive types was not required. However, this model is not sufficient if ecologists wish to understand to joint relationship between dive type and intra-dive behaviour.

The simulation study also shows that the observed Fisher information serves as a suitable approximation for the standard errors of parameter estimates in most cases. One notable exception is that the standard errors of the probability transition matrix estimates ($\hat{\Gamma}$ and $\hat{\Gamma}^*$) tend to be overestimated by the observed Fisher information.

Finally, we used the CarHHMM-DFT to model the behavior of a killer whale off the coast of British Columbia, Canada. The CarHHMM-DFT was able to distinguish three distinct sub-dive behaviors and two dive types simultaneously due to its hierarchical structure. The DFT component proved useful in determining the sub-dive behaviors of the whale, as the mean of the emission distribution of $Z^{*(2)}$ was separated by an order of magnitude between each sub-dive state. Finally, the learned auto-correlation parameter for $\mathbf{Z}^{*(1)}$, ϕ^*, was above 0.5 for every dimension and sub-dive type, providing evidence that the conditionally auto-regressive component of the CarHHMM-DFT resulted in a better fit to the data. The introduction of the parameter ϕ^* also allows $\mathbf{Z}^{*(1)}$ to be interpreted as a state-switching OU process (see appendix).

Because traditional information criteria tend to overestimate the number of states in biological processes (**?**), the number of dive types and sub-dive behaviors was selected in an ad-hoc manner. There does appear to be some heterogeneity within dive types, and future work can be done to determine the optimal number of dive types and within-dive behaviors.

BIBLIOGRAPHY

[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] Sonali Bagchi and Sanjit Mitra. The nonuniform discrete fourier transform. 01 2001.

[3] 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.

[4] 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.

[5] Vianey Leos Barajas, Eric Gangloff, Timo Adam, Roland Langrock, Floris van Beest, Jacob Nabe-Nielsen, and Juan Morales. Multi-scale modeling of animal movement and general behavior data using hidden markov models with hierarchical structures. *Journal of Agricultural Biological and Environmental Statistics*, 02 2017.

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

[7] Alex Nunes. Divebomb, 07 2018.

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

APPENDIX

Proof of Theorem 1. Combining equation (4) with equation (3) gives:

$$\left(Y_{s\Delta t}|X_{(s-1)\Delta t}=i^*\right) \sim \mathcal{N}\left((1-\phi^{(i^*)})\mu^{(i^*)}+\phi^{(i^*)}Y_{(s-1)\Delta t},\ \ \sigma^{2(i^*)}\right),$$
$$s=1,\ldots,T-1$$

If X_t follows a Markov chain with transitions at the observation times, then the behavioural state X_t does not change between observations and we can re-index $Y_{(s-1)\Delta t}=Y'_s$ and $X_{(s-2)\Delta t:(s-1)\Delta t}=X'_s$ for $s=2,\ldots T$, yielding the desired result:

$$(Y'_s|X'_s=i^*) \sim \mathcal{N}\left((1-\phi^{(i^*)})\mu^{(i^*)}+\phi^{(i^*)}Y'_{s-1},\ \ \sigma^{2(i^*)}\right)$$
$$s=2,\ldots,T$$

X' follows a Markov chain, and the distribution of $(Y'_s|X'_s,Y'_{s+1})$ is consistent with that of a CarHMM with normal emission distributions. ■

Description of simulated data.

$\hat{Y}^*_{t,t^*}$ was simulated using the following procedure. The k^{th} Fourier mode of $\hat{Y}^*_{t,t^*}$ is denoted as $\hat{Y}^{*(k)}_{t,t^*}$:

$$(\hat{Y}^{*(0)}_{t,0}|X^*_{t,0}=i^*)\sim\mathcal{N}\left(0,\sigma^{*(i^*)}\right)$$

$$(\hat{Y}^{*(0)}_{t,t^*}|X^*_{t,t^*}=i^*)\sim\mathcal{N}\left(\phi^{*(i^*)}*\hat{Y}^{*(0)}_{t,t^*-1},\sigma^{*(i^*)}\right),\quad t^*=1,2,\ldots,\lfloor Y_t/2\rfloor \tag{1}$$

$$\hat{Y}^{*(k)}_{t,t^*}=a^{(k)}_{t,t^*}i\sqrt{b^{(k)}_{t,t^*}},\qquad k=1,\ldots,49$$

$$a^{(k)}_{t,t^*}\sim\begin{cases}-1 & w.p.\ 1/2\\ 1 & w.p.\ 1/2\end{cases}$$

$$(b^{(k)}_{t,t^*}|X^*_{t,t^*}=1)\sim\text{Gamma}(5/k^2,1)$$

$$(b^{(k)}_{t,t^*}|X^*_{t,t^*}=2)\sim\begin{cases}\text{Gamma}(5/k^2,1), & k\notin\{1,2\}\\ \text{Gamma}(250,1), & k=1\\ \text{Gamma}(50,1), & k=2\end{cases}$$

$$\hat{Y}^{*(50)}_{t,t^*}=0$$

$$\hat{Y}^{*(k)}_{t,t^*}=-\hat{Y}^{*(100-k)}_{t,t^*},\qquad k=51,\ldots,99$$

$Y^*_{t,t^*,1:100}$ was set using the inverse discrete Fourier transform of $\hat{Y}^*_{t,t^*}$:

$$Y^*_{t,t^*,1:100}=IDFT\left(\hat{Y}^*_{t,t^*}\right),\qquad t^*=1,\ldots,\lfloor Y_t/2\rfloor$$

$\hat{Y}^*_{t,t^*}$ is anti-symmetric about $\hat{Y}^{*(50)}_{t,t^*}$ so that its inverse Fourier transform is real-valued. $\hat{Y}^{*(k)}_{t,t^*}$ also decays like $1/k$ so that Y^*_{t,t^*} remains continuous within a 2-second window. Y^*_{t,t^*} is not continuous *between* windows, but the jump discontinuities are not very severe since $\hat{Y}^{*(0)}_{t,t^*}$ and $\hat{Y}^{*(0)}_{t,t^*+1}$ are highly correlated. See (fig. 2) for details.

From here it is straightforward to calculate both $Z^{*(1)}$ and $Z^{*(2)}$. We pick $\tilde{\omega} = 10$ periods per window, or 5 hertz. To find the distribution of $Z^{*(1)}_{t,t^*} = \mathcal{R}\left(\hat{Y}^{(0)}_{t,t^*}\right)$, use (eqn 1):

$$\left(Z^{*(1)}_{t,t^*}|X^*_{t,t^*} = i^*\right) = \left(\mathcal{R}\left(\hat{Y}^{(0)}_{t,t^*}\right)|X^*_{t,t^*} = i^*\right) \sim \mathcal{N}\left(\phi^{*(i^*)} * Z^{*(1)}_{t,t^*-1}, \sigma^{*(i^*)}\right)$$

$Z^{*(2)}_{t,t^*}$ is the sum of Gamma-distributed random variables with the same scale parameter, so the distribution of $Z^{*(2)}_{t,t^*}$ is also a Gamma distribution:

$$Z^{*(2)}_{t,t^*} = \sum_{k=1}^{10} b^{(k)}_{t,t^*}$$

$$\left(Z^{*(2)}_{t,t^*}|X^*_{t,t^*} = 1\right) \sim \text{Gamma}\left(\alpha = \sum_{k=1}^{10} 5/k^2 = 10.10, \beta = 1.00\right)$$

$$\left(Z^{*(2)}_{t,t^*}|X^*_{t,t^*} = 1\right) \sim \text{Gamma}\left(\alpha = 300 + \sum_{k=3}^{10} 5/k^2 = 305.94, \beta = 1.00\right)$$

■

Likelihood of Simulation study model.

The overall likelihood of the CarHHMM-DFT model is as follows:

$$\mathcal{L}_{\text{CarHHMM-DFT}}(y, z^*; \Theta, \Theta^*, \Gamma, \Gamma^*) = \delta P(y_1, z_1^*; \Theta, \Theta^*, \Gamma^*) \prod_{t=2}^{T} \Gamma P(y_t, z_t^*; \Theta, \Theta^*, \Gamma^*) \mathbf{1}_N$$

where:

$$P(y_t, z_t^*; \Theta, \Theta^*, \Gamma^*) = \text{diag}\Big[f^{(1)}(y_t; \Theta^{(i)}) \mathcal{L}_{\text{CarHMM-DFT}} \left(z_t^*; \Theta^{*(1)}, \Gamma^{*(1)} \right), \ldots, f^{(N)}(y_t; \Theta^{(i)}) \mathcal{L}_{\text{CarHMM-DFT}} \left(z_t^*; \Theta^{*(N)}, \Gamma^{*(N)} \right) \Big]$$

$f^{(i)}(y_t; \Theta^{(i)})$ is the emission distribution of the dive duration y_t conditioned on the fact that $X_t = i$.

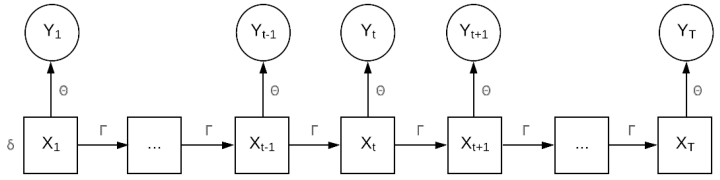
$\mathcal{L}_{\text{CarHMM-DFT}}$ corresponds to the fine-scale chain:

$$\mathcal{L}_{\text{CarHMM-DFT}} \left(z_t^*; \Theta^{*(i)}, \Gamma^{*(i)} \right) = \delta^{*(i)} \prod_{t=2}^{T} \Gamma^{*(i)} P(z_{t,t^*}^* | z_{t,t^*-1}^*; \Theta^{*(i^*)}) \mathbf{1}_N$$

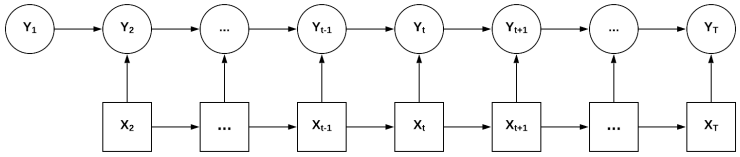
where $P(z_{t,t^*}^* | z_{t,t^*-1}^*; \Theta^{*(i)})$ is an $N^* \times N^*$ diagonal matrix with $i^* i^*$th entry equal to $f^{(i,i^*)}(z_{t,t^*}^* | z_{t,t^*-1}^*; \theta^{*(i,i^*)})$. $f^{(i,i^*)}(z_{t,t^*}^* | z_{t,t^*-1}^*; \theta^{*(i,i^*)})$ is the emission distribution of Z_{t,t^*}^* conditioned on the fact that $X_{t,t^*}^* = i^*$ and Z_{t,t^*-1}^*.

■

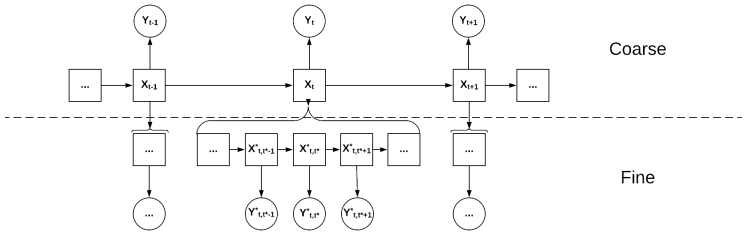
1. FIGURES AND TABLES



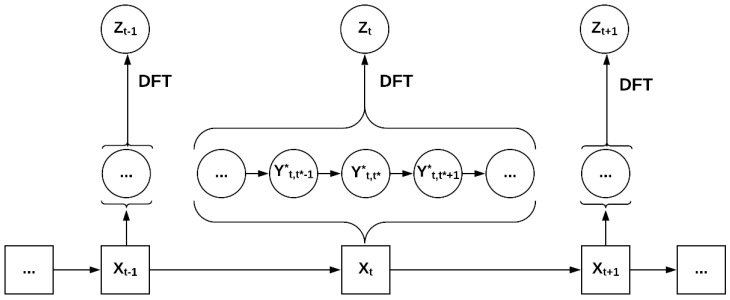
(a) Hidden Markov Model (**HMM**)



(b) Conditionally Auto-regressive HMM (**CarHMM**)



(c) Hierarchical HMM (**HHMM**)



(d) HMM with Discrete Fourier Transform (**HMM-DFT**)

FIGURE 1: : Graphical representations of HMM models

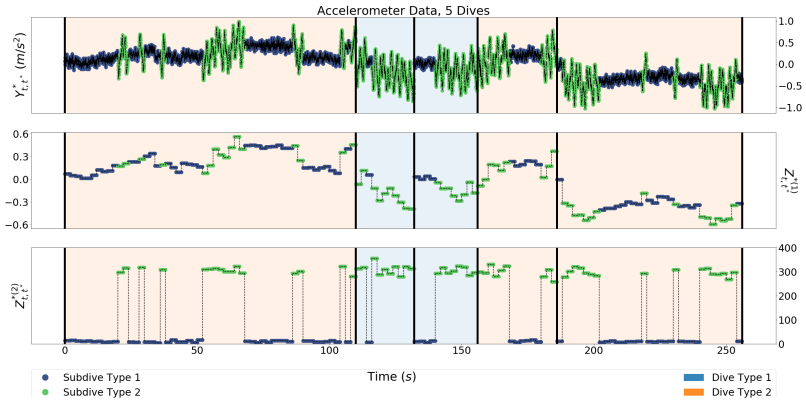


FIGURE 2: : Simulated acceleration data for one dive. The color of the line corresponds to the true fine-scale state of the sub-dive process, while the color of the background corresponds to the true dive type of the simulated whale.

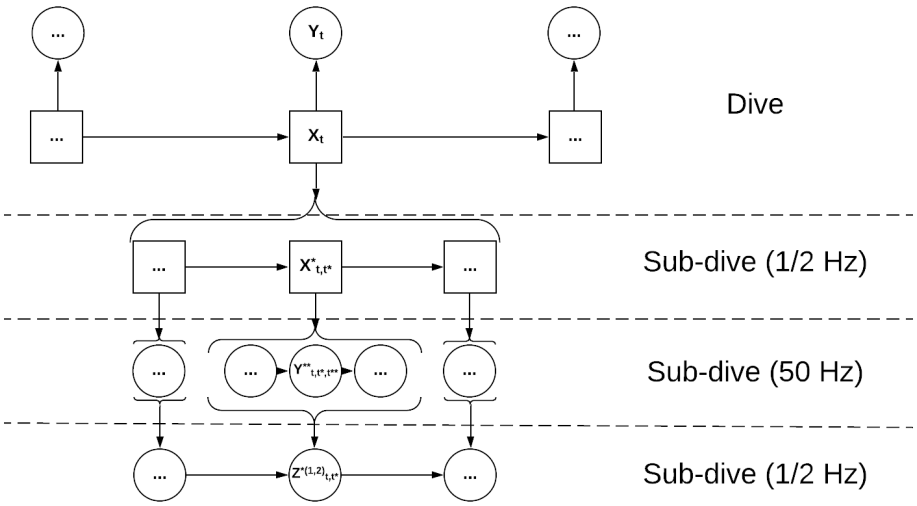
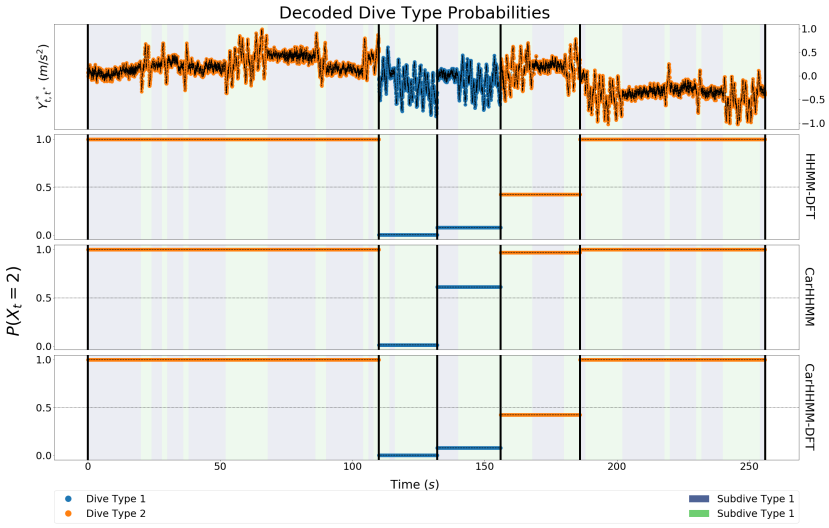
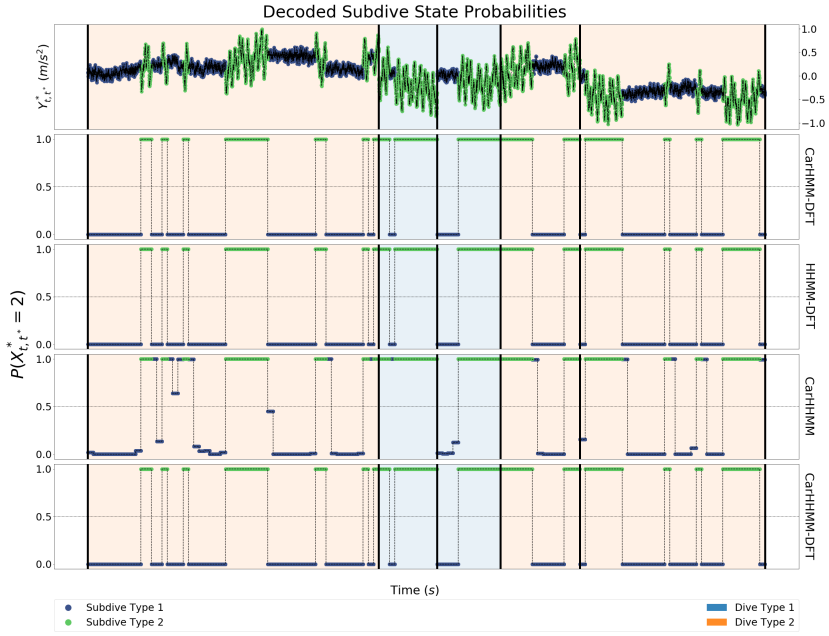


FIGURE 3: : Graphical representation the model used in the simulation and case study.

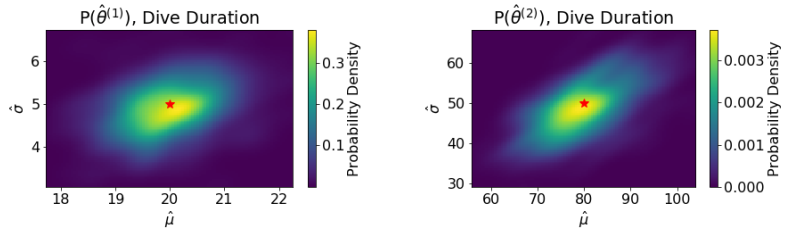


(a) Coarse-scale hidden process



(b) Fine-scale hidden process

FIGURE 4: : Decoded state probabilities of each model



(a) Dive Type 1 (b) Dive Type 2

FIGURE 5: : KDE plot of $\hat{\mu}$ and $\hat{\sigma}$ for the dive duration emission distribution for the CarHMM.

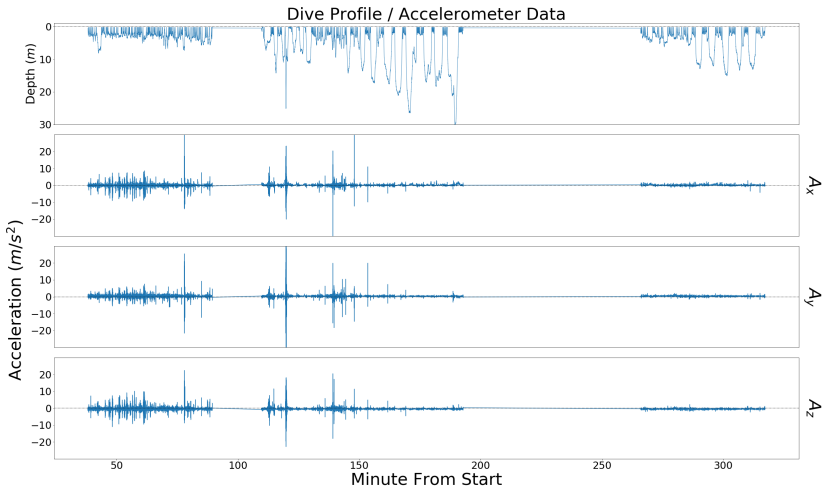


FIGURE 6: : Dive profile and Acceleration data of entire data set

TABLE 1: : Accuracies and run times for all models. All reported values are averages, and $\pm$ refers to the standard deviation.

Model	Training Time (Minutes)	Dive Type	Subdive Type	Dive Accuracy	Subdive Accuracy
CarHMM-DFT	15.74 ± 2.46	Both	Both	————	1.00 ± 0.00
		1	1	————	1.00 ± 0.00
		1	2		1.00 ± 0.00
		2	1	————	1.00 ± 0.00
		2	2		1.00 ± 0.00
HHMM-DFT	82.43 ± 11.48	Both	Both	0.94 ± 0.02	1.00 ± 0.00
		1	1	0.94 ± 0.03	1.00 ± 0.00
		1	2		1.00 ± 0.00
		2	1	0.94 ± 0.03	1.00 ± 0.00
		2	2		1.00 ± 0.00
CarHHMM	70.85 ± 15.89	Both	Both	0.91 ± 0.03	0.89 ± 0.02
		1	1	0.87 ± 0.04	0.44 ± 0.12
		1	2		1.00 ± 0.00
		2	1	0.95 ± 0.03	0.81 ± 0.04
		2	2		1.00 ± 0.00
CarHHMM-DFT	81.22 ± 16.10	Both	Both	0.94 ± 0.02	1.00 ± 0.00
		1	1	0.94 ± 0.03	1.00 ± 0.00
		1	2		1.00 ± 0.00
		2	1	0.94 ± 0.03	1.00 ± 0.00
		2	2		1.00 ± 0.00

TABLE 2: : Estimates and standard errors of parameters for dive duration distribution for all four models. All reported values are averages, except for the Fisher observed standard error, which are medians. $\pm$ refers to the IQR.

Model	Parameter	Dive Type	Estimate	Bias	Empirical SE	Observed Fischer SE
CarHMM-DFT	μ	1	49.72	-0.28	4.78	2.47 ± 0.34
		—	—	—	—	—
	σ	1	39.00	-7.51	5.05	2.50 ± 0.40
		—	—	—	—	—
HHMM-DFT	μ	1	19.99	-0.01	0.75	0.69 ± 0.11
		2	79.85	-0.15	8.05	5.85 ± 1.10
	σ	1	4.90	-0.10	0.61	0.53 ± 0.10
		2	48.74	-1.26	6.50	5.15 ± 1.02
CarHHMM	μ	1	19.91	-0.09	0.77	0.71 ± 0.12
		2	75.80	-4.20	7.72	5.32 ± 0.98
	σ	1	4.73	-0.27	0.59	0.55 ± 0.10
		2	49.48	-0.52	6.26	4.79 ± 0.93
CarHHMM-DFT	μ	1	19.99	-0.01	0.75	0.69 ± 0.12
		2	79.85	-0.15	8.05	5.85 ± 1.10
	σ	1	4.90	-0.10	0.61	0.53 ± 0.10
		2	48.74	-1.26	6.50	5.15 ± 1.02

TABLE 3: : Estimates and standard errors of parameters for $Z^{*(1)}_{t,t^*}$ for all four models. All reported values are averages, except for the Fisher observed standard error, which are medians. $\pm$ refers to the IQR.

Model	Parameter	Subdive Type	Estimate	Bias	Empirical SE	Observed Fischer SE
CarHMM-DFT	μ	1	0.00	0.00	0.00	0.14 ± 0.13
		2	0.00	0.00	0.01	0.06 ± 0.02
	σ	1	0.05	-0.00	0.00	0.00 ± 0.00
		2	0.10	-0.00	0.00	0.00 ± 0.00
	ϕ	1	0.99	-0.00	0.01	0.01 ± 0.00
		2	0.95	-0.00	0.01	0.01 ± 0.00
HHMM-DFT	μ	1	-0.01	-0.01	0.02	0.01 ± 0.00
		2	-0.00	-0.00	0.02	0.01 ± 0.00
	σ	1	0.25	0.20	0.04	0.00 ± 0.00
		2	0.24	0.14	0.03	0.00 ± 0.00
	ϕ	1	——	——	——	——
		2	——	——	——	——
CarHHMM	μ	1	0.00	0.00	0.00	0.08 ± 0.04
		2	-0.01	-0.01	0.01	0.01 ± 0.00
	σ	1	0.05	0.00	0.04	0.00 ± 0.00
		2	0.27	0.17	0.01	0.00 ± 0.00
	ϕ	1	0.97	-0.02	0.10	0.00 ± 0.00
		2	0.49	-0.46	0.05	0.02 ± 0.00
CarHHMM-DFT	μ	1	0.00	0.00	0.00	0.13 ± 0.12
		2	0.00	0.00	0.00	0.06 ± 0.02
	σ	1	0.05	-0.00	0.00	0.00 ± 0.00
		2	0.10	-0.00	0.00	0.00 ± 0.00
	ϕ	1	0.99	-0.00	0.01	0.01 ± 0.00
		2	0.95	-0.00	0.01	0.01 ± 0.00

TABLE 4: : Estimates and standard errors of parameters for $Z^{*(2)}_{t,t^*}$ for all four models. All reported values are averages, except for the Fisher observed standard error, which are medians. $\pm$ refers to the IQR.

Model	Parameter	Subdive Type	Estimate	Bias	Empirical SE	Observed Fischer SE
CarHMM-DFT	μ	1	10.10	-0.00	0.09	0.08 ± 0.01
		2	305.97	0.03	0.54	0.51 ± 0.03
	σ	1	3.18	-0.00	0.07	0.06 ± 0.01
		2	17.46	-0.03	0.37	0.36 ± 0.02
HHMM-DFT	μ	1	10.10	-0.00	0.09	0.08 ± 0.01
		2	305.97	0.03	0.54	0.51 ± 0.03
	σ	1	3.18	-0.00	0.07	0.06 ± 0.01
		2	17.46	-0.03	0.37	0.36 ± 0.02
CarHHMM	μ	1	——	——	——	——
		2	——	——	——	——
	σ	1	——	——	——	——
		2	——	——	——	——
CarHHMM-DFT	μ	1	10.10	-0.00	0.09	0.08 ± 0.01
		2	305.97	0.03	0.54	0.51 ± 0.03
	σ	1	3.18	-0.00	0.07	0.06 ± 0.01
		2	17.46	-0.03	0.37	0.36 ± 0.02

TABLE 5: : Estimates and standard errors of Γ and Γ^* for all four models. All reported values are averages except for the observed fisher SE, which is a median. $\pm$ refers to the IQR.

Model	Parameter	Estimate	Bias	Empirical SE	Observed Fischer SE
HHMM-DFT	Γ_{12}	0.50	-0.00	0.03	0.08 ± 0.01
	Γ_{21}	0.50	-0.00	0.03	0.08 ± 0.01
	$\Gamma_{12}^{*(1)}$	0.50	0.00	0.00	0.07 ± 0.01
	$\Gamma_{21}^{*(1)}$	0.10	0.00	0.02	0.02 ± 0.00
	$\Gamma_{12}^{*(2)}$	0.20	-0.00	0.01	0.01 ± 0.00
	$\Gamma_{21}^{*(2)}$	0.30	-0.00	0.02	0.02 ± 0.00
CarHMM-DFT	Γ_{12}	——	——	——	——
	Γ_{21}	——	——	——	——
	$\Gamma_{12}^{*(1)}$	0.23	——	0.01	0.00 ± 0.00
	$\Gamma_{21}^{*(1)}$	0.23	——	0.02	0.00 ± 0.00
	$\Gamma_{12}^{*(2)}$	——	——	——	——
	$\Gamma_{21}^{*(2)}$	——	——	——	——
CarHHMM	Γ_{12}	0.49	-0.01	0.02	0.08 ± 0.01
	Γ_{21}	0.50	0.00	0.02	0.08 ± 0.01
	$\Gamma_{12}^{*(1)}$	0.50	-0.00	0.00	0.23 ± 0.19
	$\Gamma_{21}^{*(1)}$	0.04	-0.06	0.02	0.01 ± 0.00
	$\Gamma_{12}^{*(2)}$	0.11	-0.09	0.02	0.01 ± 0.00
	$\Gamma_{21}^{*(2)}$	0.11	-0.19	0.03	0.01 ± 0.00
CarHHMM-DFT	Γ_{12}	0.50	-0.00	0.03	0.08 ± 0.01
	Γ_{21}	0.50	-0.00	0.03	0.08 ± 0.01
	$\Gamma_{12}^{*(1)}$	0.50	0.00	0.00	0.07 ± 0.01
	$\Gamma_{21}^{*(1)}$	0.10	0.00	0.02	0.02 ± 0.00
	$\Gamma_{12}^{*(2)}$	0.20	-0.00	0.01	0.01 ± 0.00
	$\Gamma_{21}^{*(2)}$	0.30	-0.00	0.02	0.02 ± 0.00

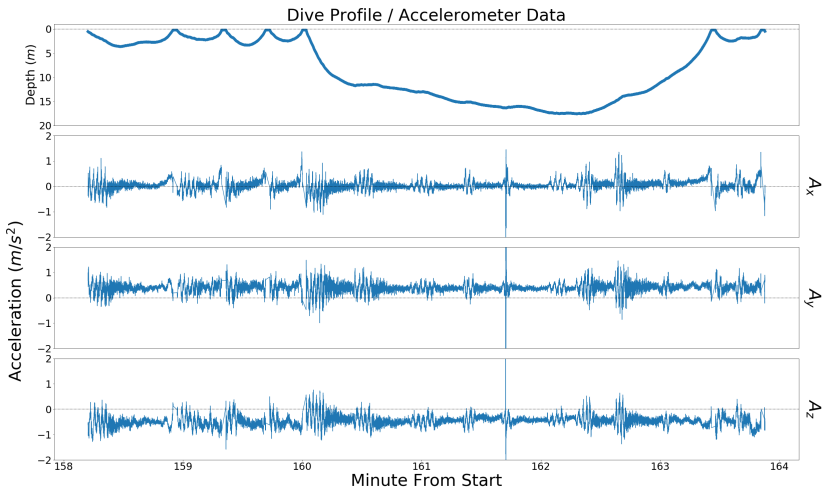


FIGURE 7: : Dive profile and acceleration data for a collection of 5 dives of a killer whale.

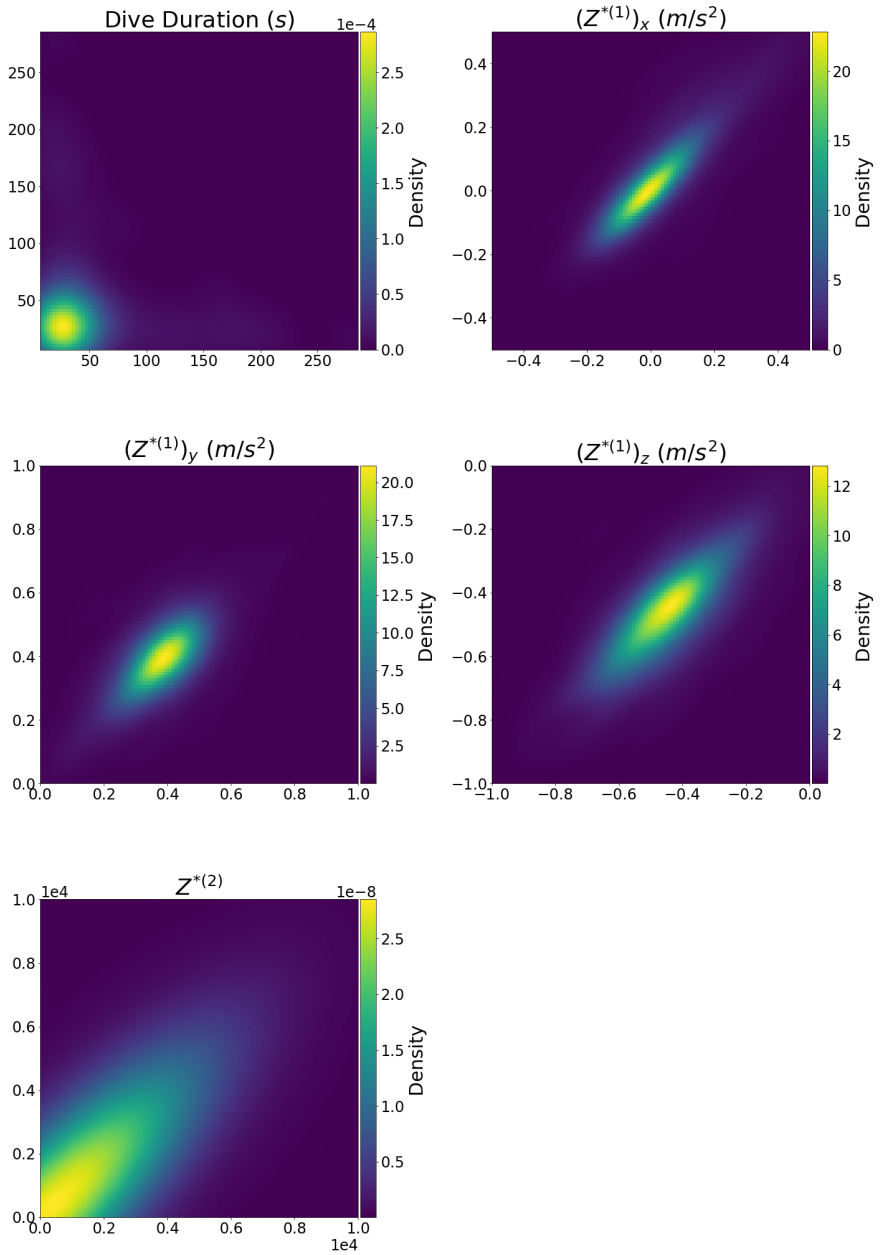


FIGURE 8: : Lag plots of all features on both the fine- and coarse- scale.

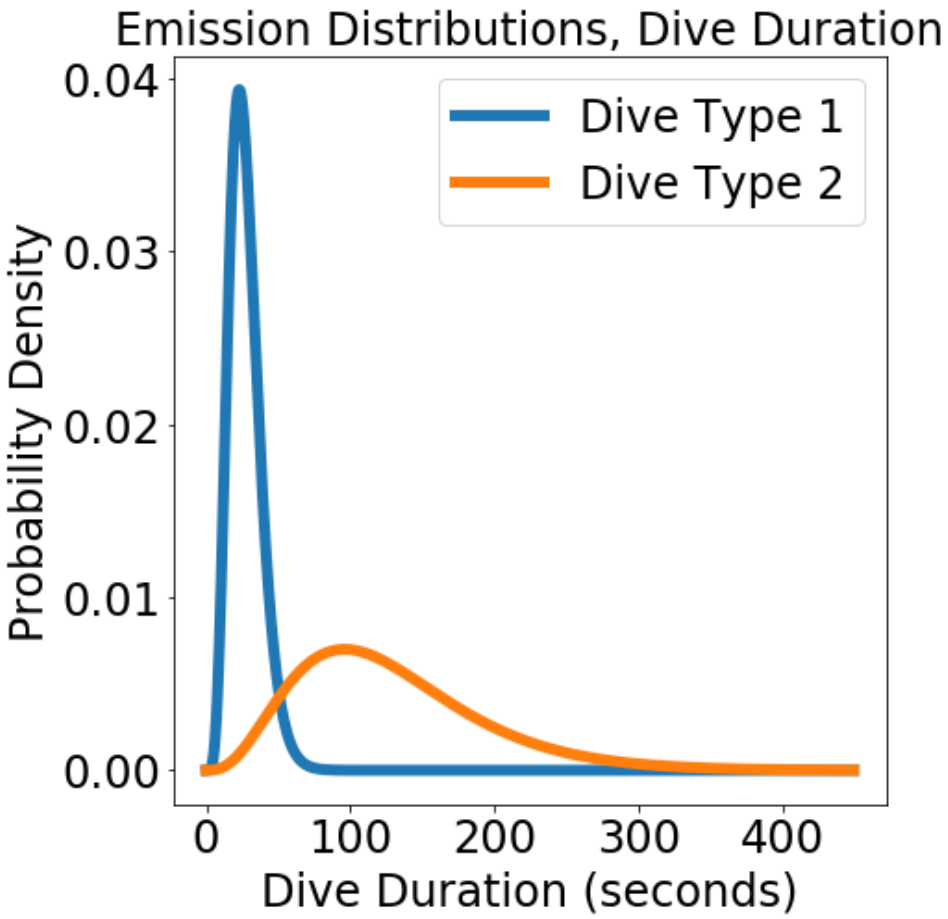


FIGURE 9: : Estimated probability distributions for each coarse-scale observation in each dive type.

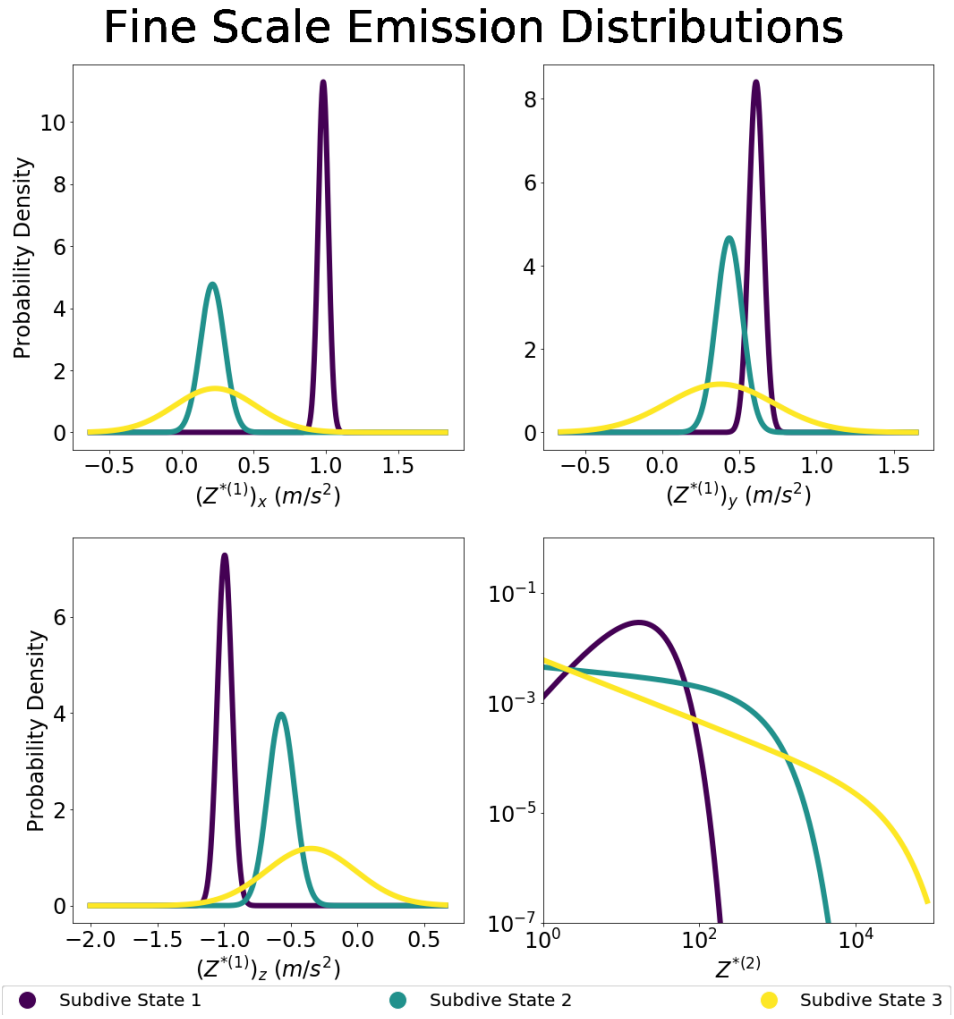


FIGURE 10: : Estimated probability distributions for each fine-scale observation in each behavioral state. Note that the distributions of acceleration do not take auto-correlation into account (see table 6)

TABLE 6: : Estimates and standard errors of emission parameters for killer whale data.

Feature	Dive / Sub-dive Type	Parameter Estimate		
		$\hat{\mu}$	$\hat{\sigma}$	$\hat{\phi}$
Dive Duration (seconds)	1	27.23 ± 0.63	10.89 ± 0.56	—
	2	127.96 ± 11.50	64.13 ± 9.21	—
$Y_x^{*(1)}$	1	0.98 ± 0.07	0.04 ± 0.00	0.99 ± 0.00
	2	0.22 ± 0.01	0.08 ± 0.00	0.87 ± 0.01
	3	0.23 ± 0.03	0.28 ± 0.01	0.62 ± 0.03
$Y_y^{*(1)}$	1	0.61 ± 0.09	0.05 ± 0.00	0.99 ± 0.00
	2	0.43 ± 0.01	0.09 ± 0.00	0.87 ± 0.01
	3	0.38 ± 0.04	0.35 ± 0.01	0.62 ± 0.04
$Y_z^{*(1)}$	1	-1.00 ± 0.11	0.05 ± 0.00	0.99 ± 0.00
	2	-0.57 ± 0.01	0.10 ± 0.00	0.87 ± 0.01
	3	-0.35 ± 0.04	0.34 ± 0.01	0.62 ± 0.04
$Y^{*(2)}$	1	27.16 ± 0.32	16.67 ± 0.32	—
	2	406.98 ± 4.42	438.09 ± 5.49	—
	3	9688.54 ± 221.95	14584.02 ± 358.40	—

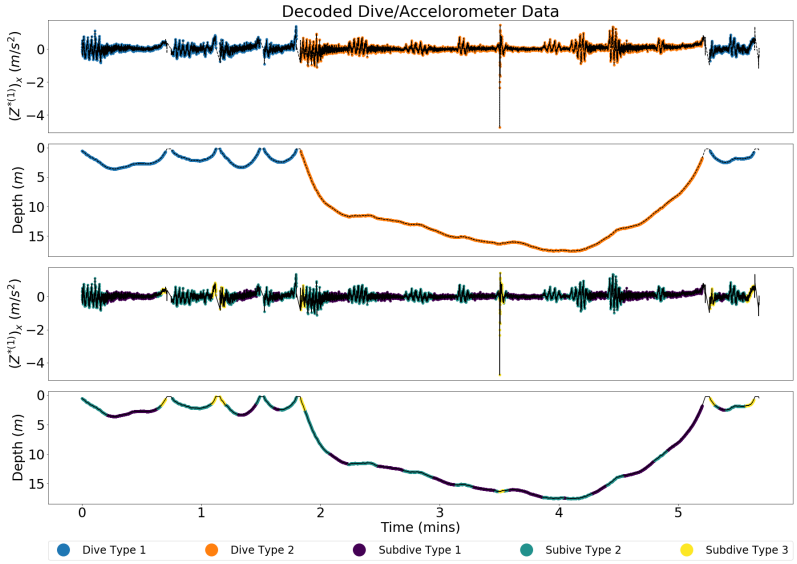


FIGURE 11: : Features of a particular set of killer whale dives and decoded estimates for the intra-dive behavioral states. The color of the plot corresponds to behavioral or dive state with the highest probability.

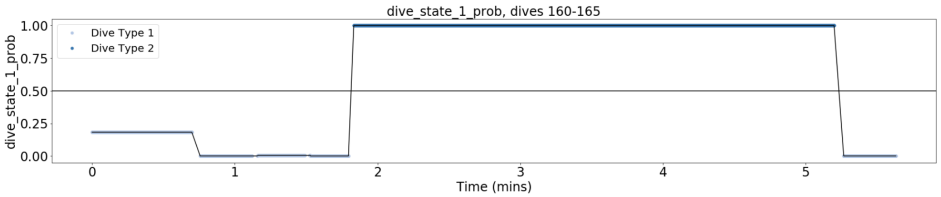
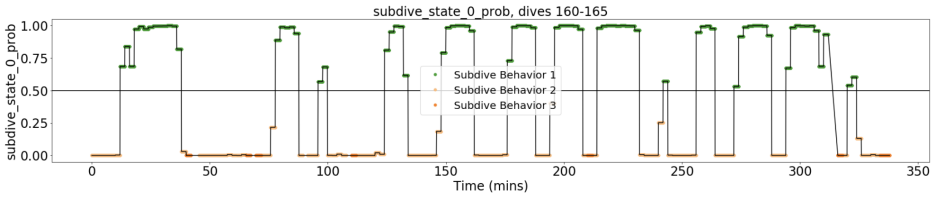
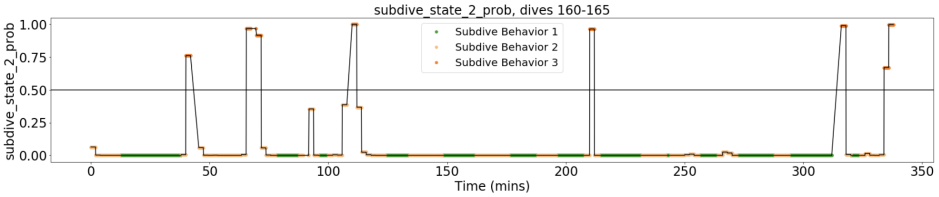


FIGURE 12: : Probabilities of dive types for the set of killer whale dives from (fig 11).



(a) Fine-scale state 1 probabilities



(b) Fine-scale state 3 probabilities

FIGURE 13: : Probabilities of sub-dive types for the set of killer whale dives from (fig 11).

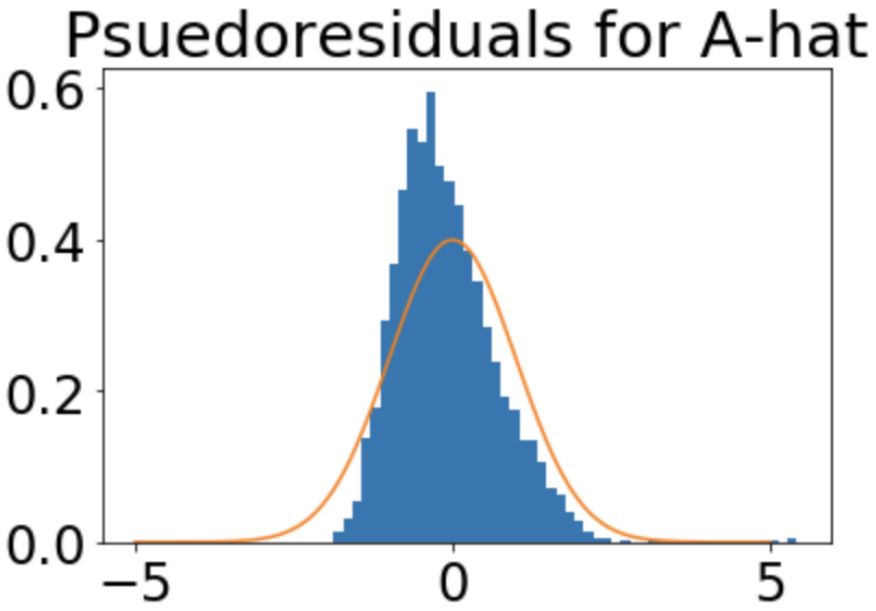


FIGURE 14: : Psuedoresiduals of $Y^{*(2)}$

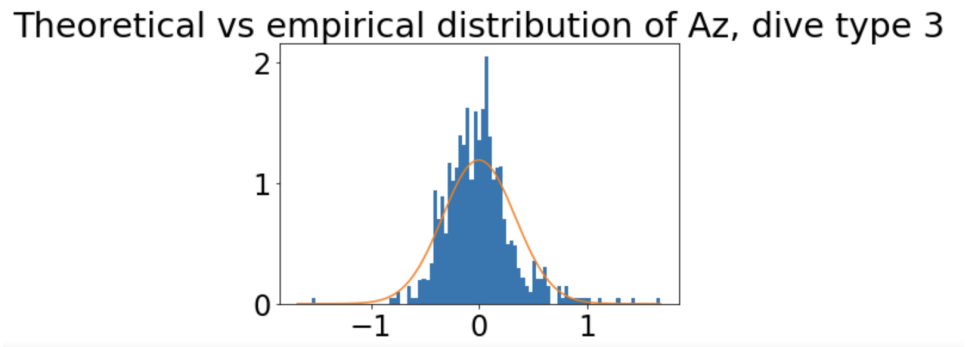


FIGURE 15: : Empirical distribution of $Y_z^{*(1)}$ for sub-dive state 3 plotted over its estimated pdf.