Hidden Markov Models and Animal Movement

Owen Gould

July 2024

1 Introduction

To best model the movement patterns of the elephants at Kruger National Park (KNP), we will attempt to fit a statistical model to existing elephant movement data from the area. We employ the modeling framework as outlined by Cannon et al., which consists of

- **choosing** the form of the model;
- **fitting** the model to the data;
- **assessing** the model fit; and
- **using** the model to answer the desired research question [3].

The entire process is outlined, in detail, in the subsections below.

1.1 Choosing the Model Form

Biological intuition tells us that the particular behavioral state of an animal (e.g., foraging, migrating, or resting) influences its movement patterns over time. Therefore, while a random walk might be a good baseline to model elephant movement patterns, we might suspect that a single model will not be appropriate for all time or under all conditions. This leads us to the **Hidden Markov Model (HMM)**, which allows us to introduce additional flexibility to our model by allowing it to switch between different random walks over time, each with a different set of parameters.

In the HMM, we have a series of discrete time steps, as well as two concurrent sequences. The first is the sequence of “hidden” states, in the sense that they are unknown to the observer. The other is the observed states, each of which is a probabilistic function of the corresponding hidden state at that time step. As the name also implies, the sequence of hidden states is assumed to be a Markov process. That is, the next state in the sequence depends only on the current state; the process is “memoryless.” For an illustration of this model, see Figure 1.

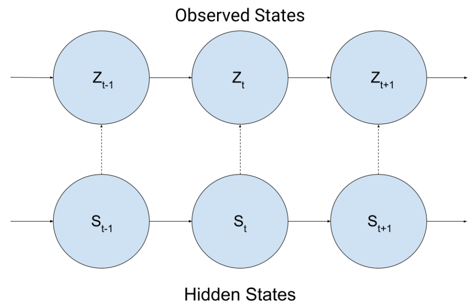

Figure 1: A Diagram of the Hidden Markov Model

Figure 2 depicts a simple version of this state-switching model. In it, we have two states S_1, S_2. We denote the probability of moving from State i to State j via $P(i, j)$. The HMM is therefore known as a "doubly stochastic process", since the hidden states are random, as are the observed states produced by them.

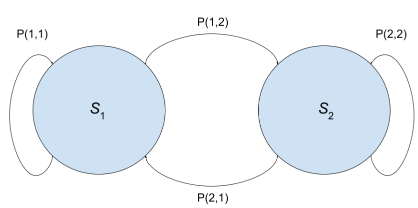

Figure 2: A Two-State State-Switching Model

To apply this model to the context of animal movement, suppose we have a series of observed movement patterns of the form $Z_1, \ldots, Z_T$, where $Z_i = \begin{bmatrix} l_i \\ \phi_i \end{bmatrix}$. Here, l_i denotes the step length at time i (the euclidean distance between the angle between between the animal's position at time step i and time step $i - 1$) and ϕ_i denotes the turning angle at time i (the angle between between the animal's position at time step i and time step $i - 1$).

We then assume that there are unobservable states $S_1, \ldots, S_T$ which we treat as a proxy for the animal's behavioral state. We define S_i via

$$S_i \sim \begin{bmatrix} \text{Gamma}(\mu_i, \sigma_i) \\ \text{Von Mises}(\mu_i, \kappa_i) \end{bmatrix}.$$

In other words, the step length at each time interval i is randomly sampled from a Gamma distribution with mean μ_i and standard deviation σ_i, whereas the turning angle at each time interval i is randomly sampled from a Von Mises distribution with mean μ_i and concentration κ_i. (The Gamma distribution is chosen for step lengths because it is a continuous distribution with a support set limited to the positive real numbers, and the Von Mises distribution is chosen for turning angles because it is the circular analogue of the normal distribution).

For the purposes of this investigation, we wish to optimize the model parameters in a way that best explains the observed states and can be used to infer future observable states. In particular, we wish to segment the behavior of elephants into the appropriate number of states and best estimate the parameters for the probability distributions for each state.

1.2 Fitting the Model

In this subsection, we fit a Hidden Markov Model using the R statistical software. In particular, we use the `moveHMM` package, as designed by Théo Michelot [7]. Its workflow summarized in Figure 3.

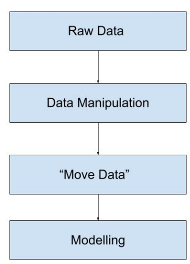

Figure 3: A Summary of the moveHMM workflow

The first step is to import the raw GPS data, which came from Slowtow 2019 [8]. This data consists of time series measurements of the latitude and longitude of 14 elephant herds in Kruger National Park, which were collected in roughly 30-minute intervals over a period of two years. However, because there GPS collars could not always establish a connection every half hour, there are missing values that create gaps in the data.

To illustrate this problem, consider the following simultaneous plots of AM-107 and AM-307 (Figures 4-6). The first is a Mercator projection plot of latitude and longitude, the second is a plot of latitude versus time, and the third is a plot of longitude versus time. In each plot, a single point corresponds to a single observation and a "path" geometry is used to connect the points. A straight line with no points along it, therefore, corresponds to a gap in the time series data. These gaps, some of which are hours long, violate the assumptions of the models used in the `moveHMM` package.

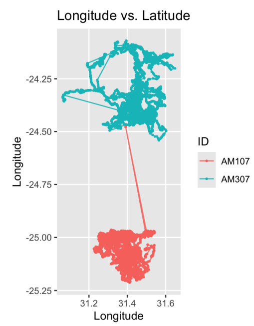

Figure 4: Longitude vs. Latitude for AM-107 and AM-307

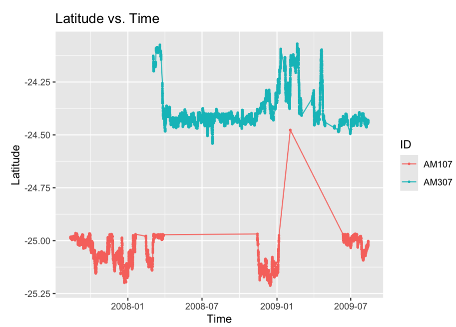

Figure 5: Latitude vs. Time for AM-107 and AM-307

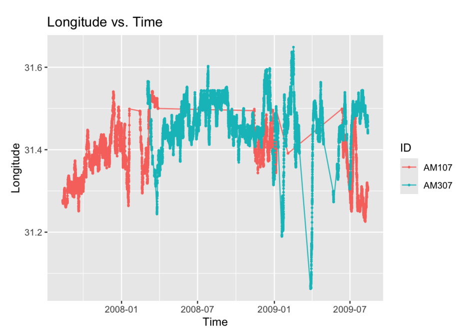

Figure 6: Longitude vs. Time for AM-107 and AM-307

To resolve this issue, we first subdivide any walk with a gap longer than two hours using a utility function designed by Théo Michelot [6]. Then, we fill in a timestamp for every 30 minute interval and impute a "missing" (`NA`) value if data is not available. The `moveHMM` package can handle these missing values, so long as they occur in regular intervals [9].

From there, we use the `moveHMM` function `prepData`, which uses the longitude and latitude coordinates at each timestamp and computes the the step length (the Euclidean distance) and the turning angle between each successive observation. A visualization of how the step length and turning angle are determined from a sequence of position data is shown in Figure 7.

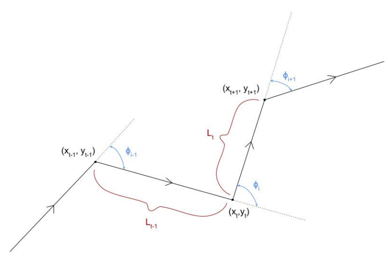

Figure 7: GPS Position Data to HMM Move Data

Before we can fit the model, we must decide beforehand how many states we want our model to use. This decision is often motivated by a combination of biological intuition and model selection techniques [9]. Prior research has established a framework that separates animal behavior into two states: an "encamped" state and an "exploratory" state. In the encamped state, which may correspond to an activity like foraging, step lengths are generally low and turning angles are generally to be high. In the exploratory state, which may correspond to an activity like migration, step lengths are generally low and turning angles are generally low (that is, there is directional persistence over time) [4]. A separate study on elephants has shown that elephants exhibit a similar segmentation of their behavior [2]. Accordingly, we expect there to be two true behavioral states among the elephants at Kruger National Park.

We must also supply an initial guess for the parameters of the probability distributions comprising each state. This is because the `moveHMM` package fits the model using maximum likelihood estimation. If the initial parameter values supplied are inappropriate, optimization may not return the most accurate parameters or may fail entirely [5]. To illustrate this problem, imagine the following scenario: you are a hiker, and your goal is to find the highest peak in a given mountain range. However, it is very foggy, so you cannot rely on your eyesight. Instead, your strategy is to continue hiking as long as you are moving uphill. If you notice you are going downhill, you must have passed the peak, so you go back the way you came until you are no longer moving downhill. This strategy may help you reach *a peak* (i.e., a local likelihood maximum), but it does not guarantee that you reach the *highest peak* (i.e., the global likelihood maximum). Whether you reach the highest peak is entirely dependent on your starting position (i.e., your set of initial guesses for the parameters of each probability distribution) [4]. To illustrate this problem, refer to Figure 8.

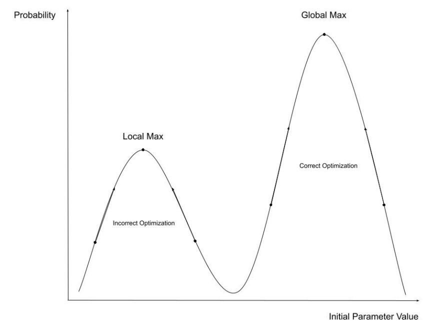

Figure 8: A Diagram Illustrating the Importance of Choosing the Right Starting Parameters

To arrive at an educated guess, we examine histograms of the step lengths and turning angles, then make an educated guess for plausible means μ_i for each state. We also compare the histogram of turning angles to a theoretical Von Mises Distribution for various values of the concentrations to arrive at an educated guess of concentrations κ_i for each state. After that, we iterate over a series of possible parameter values, fitting a model to each set, and comparing the maximum likelihood estimate to our original model to verify that the correct optimization did occur.

We then fit the HMM to the movement data. The parameters for the step length (in kilometers) and turning angle (in radians), rounded to three decimal places, is summarized in two tables below.

	State 1	State 2
Mean	0.107	0.623
Standard Deviation	0.106	0.625

Table 1: Step Length Parameters

	State 1	State 2
Mean	-0.012	-0.007
Concentration	0.802	1.849

Table 2: Turning Angle Parameters

We can see that the average step length is larger in State 2 than in State 1. We can also see that the mean turning angle is smaller in State 2 than State 1. Furthermore, the higher concentration in State 2 indicates that there is less variability in the turning angles (i.e., there is greater directional persistence). These facts allow us to conclude that States 1 and 2 correspond to an "encamped state" and an "exploratory state," respectively. It is worth noting that maximum likelihood estimation resulted in a two state model that exhibited such a distinction that comported with our biological intuition and prior research. It would have been entirely possible, had there been no significant difference between the two states, that maximum likelihood estimation would have produced a two-state model with similar sets of parameters for the two states.

We can visualize the difference between these two states by plotting separate histograms of step length and turning angle in the original movement data and overlaying the estimated density plot of each state (See Figures 9 and 10).

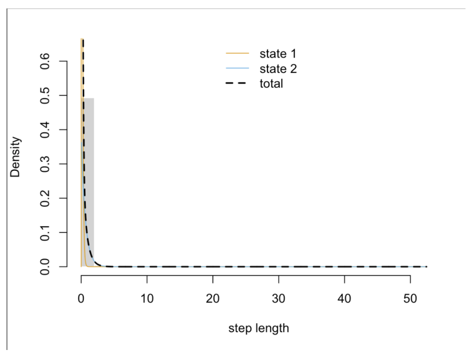

Figure 9: Histogram of Step Lengths with Densities Curves of Each State

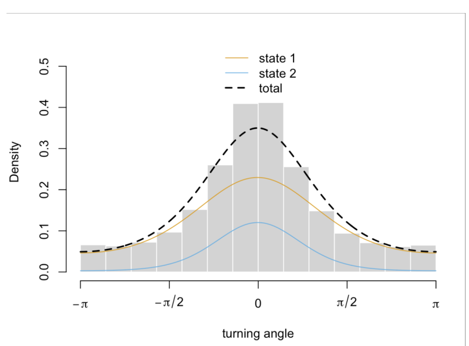

Figure 10: Histogram of Turning Angles with Densities Curves of Each State

Although the step lengths are overwhelmingly concentrated in a single bin, close inspection reveals a greater density of State 2 (the exploratory state) across higher step lengths than State 1 (the encamped state). Examining the histogram of turning angles shows a higher concentration of angles close to π and $-\pi$ and an overall greater variability in turning angle for State 1, the encamped state. (State 1 also appears to have a higher concentration of turning angles around 0 than State 2 does. This is because, the model estimated elephants to be in the encamped state roughly two-thirds of the time.)

While there appears to be a practical significance between the exploratory and encamped states in our model, we must also verify that the distinction between these two states is *statistically significant*. To do this, we also fit a one-state model (i.e., a model built on the assumption that elephants have consistent behavioral patterns) and compare the AIC of it with the AIC of our two-state model. The AIC, or Akaike Information Criteria, quantifies the goodness of a fit by estimating its prediction error while balancing fit with simplicity. After all, increasingly complicated models could be built, but they may only be capturing noise in the data rather than real trends. The lower the AIC value, the greater the level of fit.

Model	AIC
One-State	494565.8
Two-State	445312.3

Table 3: AIC for One and Two-State Models

As is evident in the Table 3, the two-state model has a lower AIC value. We therefore have statistically significant evidence that elephants in Kruger National Park exhibit two distinct behavioral states (exploratory and encamped) rather than just one.

1.3 Assessing the Model

To assess the validity of the assumptions underlying our Hidden Markov Model, we reflect on the nature of our data and our model. We also create and examine a set of diagnostic plots to further assess these assumptions, as well as the overall model fit.

One assumption of the moveHMM package that was not addressed in the previous subsection was that there be little to no measurement error in the data. As mentioned previously, the elephants were tracked using GPS collars. In the context of animal movement, this is generally considered a reliable enough method of data collection that we have no concerns about our data meeting this requirement [4].

An important assumption of the HMM is the Markov assumption. That is, the next state in the chain depends only on the current state, and no others [1].

Yet another assumption underlying the model is independence: that the observed state at any given time depends only on the contemporary behavioral state, and that, conditioned on the current state, the step length and turning angle are independent of each other. The independence assumption is met by our division of elephant movement patterns into two states [10].

We now use R to generate a series of diagnostic plots for both step length and turning angle, as shown in Figure 11. The R output contains, in descending order for each variable, the time series plot, the qq-plot, and the autocorrelation (ACF) plot.

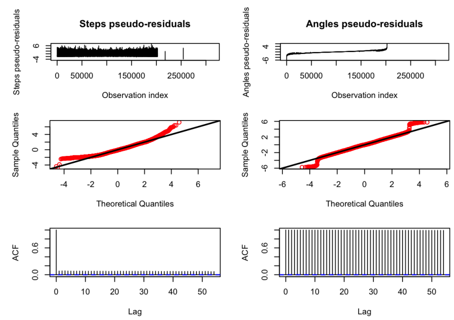

Figure 11: Diagnostic Plots for the Two-State Model

The first diagnostic plot is the qq-plot. It plots the quantiles of the normal pseudo-residuals against the standard normal distribution. If the normal pseudo-residuals fall along the line $y = x$, then they are approximately normally distributed, indicating that the model is a good fit. In each qq-plot, the points deviate significantly from this line at the tails; the plot of angle pseudo-residuals even exhibits a "jumping" behavior at each end. This gives us reason to be concerned about the overall model fit and to take less stock in the precise parameter estimates.

The second diagnostic plot is the ACF plot. It visualizes the correlation between a given time series and a lagged version of itself. Since the black vertical lines along the x-axis fall outside of the interval around 0 created by two dashed horizontal lines, we have high autocorrelation, which is a bad sign for our model.

This means that the 30-minute intervals in our model are merely snapshots of a larger journey and too granular a measurement to produce effective observations at each time step index. In this context, we have reason to doubt whether elephants' behavioral states are switching independently every 30-minutes; they may exhibit greater long-term planning with regards to their movement.

1.4 Using the Model

Although our current Hidden Markov Model is flawed, it is still our best data-driven estimate of the underlying movement patterns of elephants in KNP. This model then form a component of the "lifelike model" in the next section. This model will better simulate the movement patterns of the elephants at KNP in relation to their environment by allowing us to investigate the impact of other factors that are suspected to influence elephant behavior but were not measured in the original data. (The lifelike model's incorporation of a footfall measurement and elephants bias away from locations they have already visited will also introduce a degree of memory, intending to offset the problem of autocorrelation discussed in the previous subsection.) Note, however, for the sake of simplicity, the behavioral state of the elephants do no switch at each time step. Instead, the encamped state is used for the dry season and the exploratory state is used for the wet season.

2 Acknowledgments

Théo Michelot's vignettes for the `moveHMM` package, as well as a webinar he did for Ecological Forecasting, was immensely helpful in guiding me through the data manipulation and model fitting in R. The work he did on a smaller subset of the elephant movement data became a useful building block for the rest of my work. I am also grateful for Grinnell College's Professor Wells' guidance in interpreting the model diagnostics and drawing conclusions from them.

References

[1] Mariette Awad and Rahul Khanna. "Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers". In: Berkeley, CA: Apress, 2015. Chap. Hidden Markov Model, pp. 81–104. ISBN: 978-1-4302-5990-9. DOI: 10.1007/978-1-4302-5990-9_5. URL: https://doi.org/10.1007/978-1-4302-5990-9_5.

[2] Christopher Beirne et al. "African forest elephant movements depend on time scale and individual behavior". en. In: *Sci. Rep.* 11.1 (June 2021), p. 12634.

[3] A. Cannon et al. *STAT2: Modeling with Regression and ANOVA*. W. H. Freeman, 2018. ISBN: 9781319067502. URL: https://books.google.com/books?id=SG15DwAAQBAJ.

[4] Ecological Forecasting. *Statistical Methods Series: Movement Ecology*. Youtube. 2022. URL: https://www.youtube.com/watch?v=WELTpbB5BuU.

[5] Théo Michelot Roland Langrock. *A short guide to choosing initial parameter values for the estimation in moveHMM*. 2023.

[6] Théo Michelot. *Utility Functions*. https://github.com/eco4cast/Statistical-Methods-Seminar-Series/tree/main/michelot_movement/code. 2022.

[7] Théo Michelot, Roland Langrock, and Toby A. Patterson. "moveHMM: an R package for the statistical modelling of animal movement data using hidden Markov models". In: *Methods in Ecology and Evolution* (2016). DOI: 10.1111/2041-210X.12578. URL: https://doi.org/10.1111/2041-210X.12578.

[8] Rob Slotow, Maria Thaker, and Abi Tamim Vanak. *Data from: Fine-scale tracking of ambient temperature and movement reveals shuttling behavior of elephants to water*. 2019.

[9] Roland Langrock Toby Patterson Théo Michelot. *moveHMM: an R package for the analysis of animal movement data*. 2023.

[10] Roland Langrock Walter Zucchini Iain L. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R*. 2. Taylor Francis Group, 2016.