CONABIO - Septiembre 2019

An Introduction to Hidden Markov Models

Vianey Leos(-)Barajas

Depts. of Forestry Envir Res and Statistics
North Carolina State University

Table of contents

1. Introduction

2. Data Examples

3. hidden Markov models

4. Extending the basic HMM

5. Inference in HMMs

6. Overview + Markov-switching processes

Introduction

Where do we start?

- Claim:
  - **False:** Hidden Markov models objectively classify behaviors.
  - **True (for the most part):** Hidden Markov models can distinguish between discernible signals and capture temporal dependence.

Where do we start?

- Claim:
  - **False:** Hidden Markov models objectively classify behaviors.
  - **True (for the most part):** Hidden Markov models can distinguish between discernible signals and capture temporal dependence.
- Reality: No model is magic, but **every model tells a story**.

Where do we start?

- Claim:
  - **False:** Hidden Markov models objectively classify behaviors.
  - **True (for the most part):** Hidden Markov models can distinguish between discernible signals and capture temporal dependence.
- Reality: No model is magic, but **every model tells a story**.
- What's important:

Where do we start?

- Claim:
  - **False:** Hidden Markov models objectively classify behaviors.
  - **True (for the most part):** Hidden Markov models can distinguish between discernible signals and capture temporal dependence.
- Reality: No model is magic, but **every model tells a story**.
- What's important:
  - Embracing your domain expertise.
  - You know your animals.
  - Constructing a model while thinking about what your animal actually does.

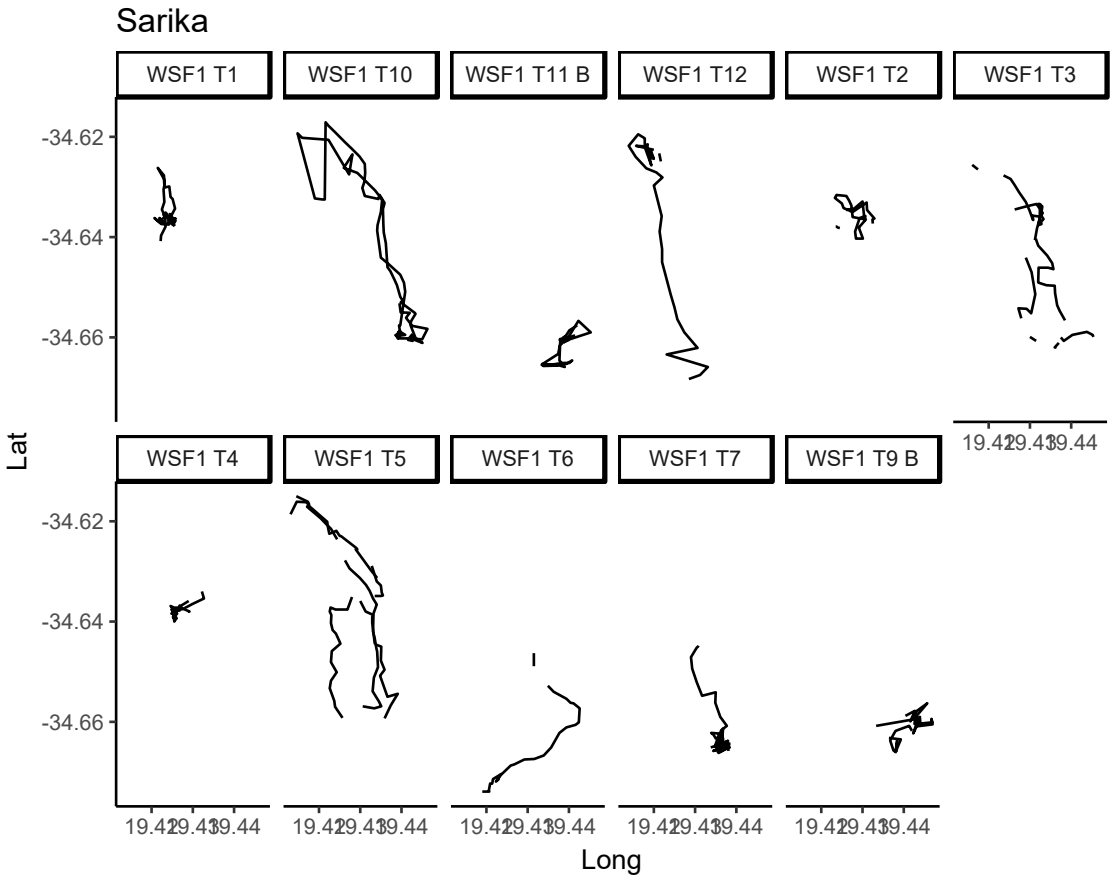
No Model is Magic

Really. Not even hidden Markov models.

Data Examples

Example 1: White Sharks Active Tracking Data

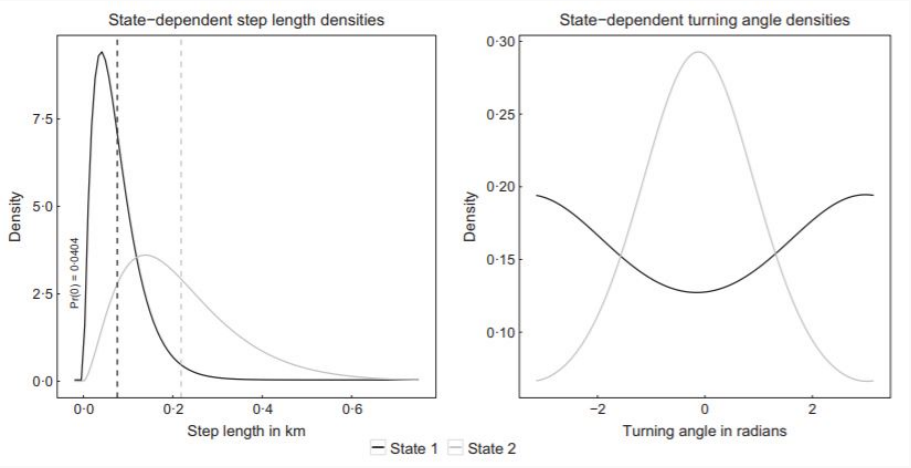
Position every 5 minutes (filled with missing data when needed):

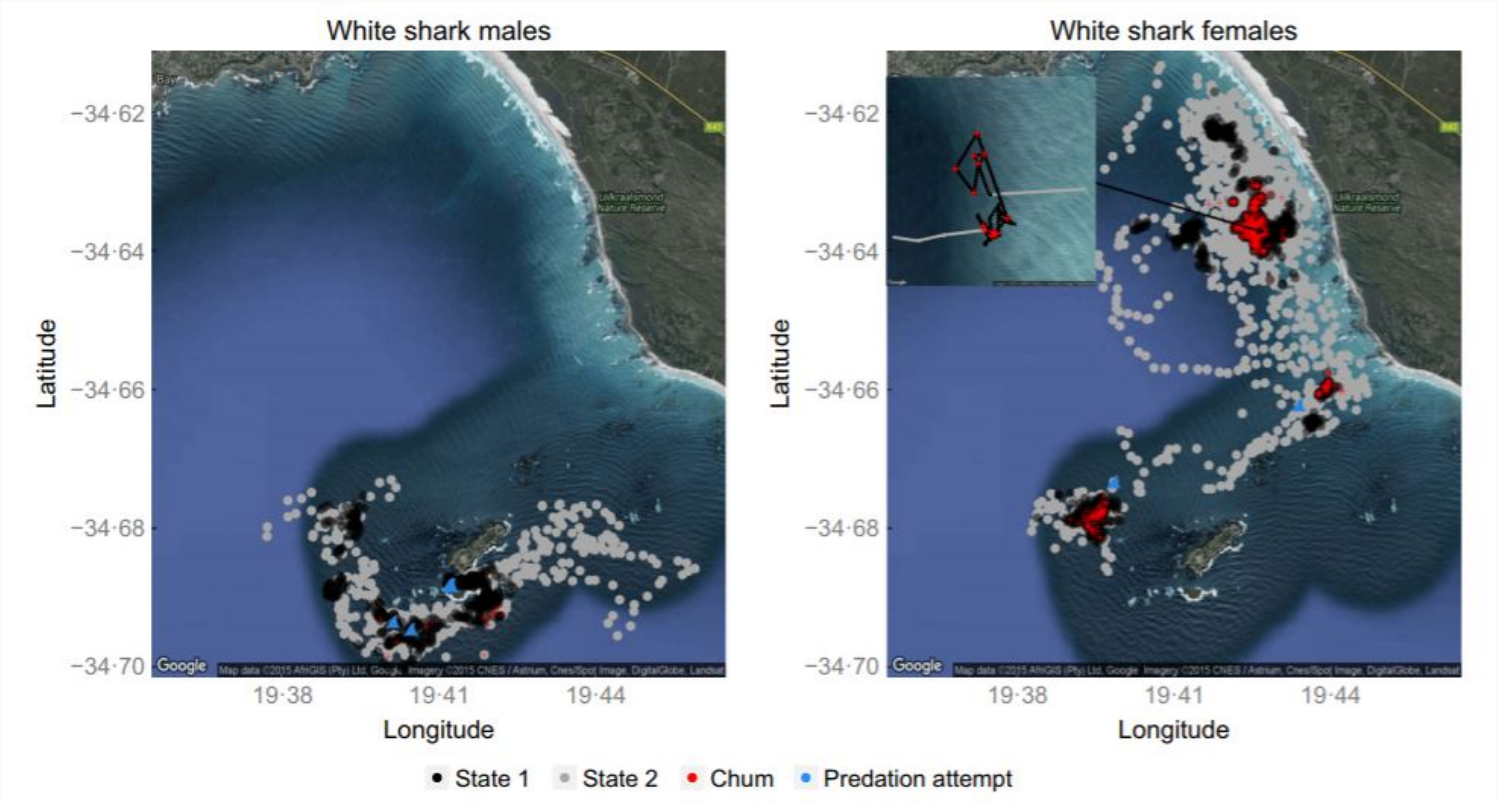


Visually: Some directed, traveling behavior, other times she remained in the same area.

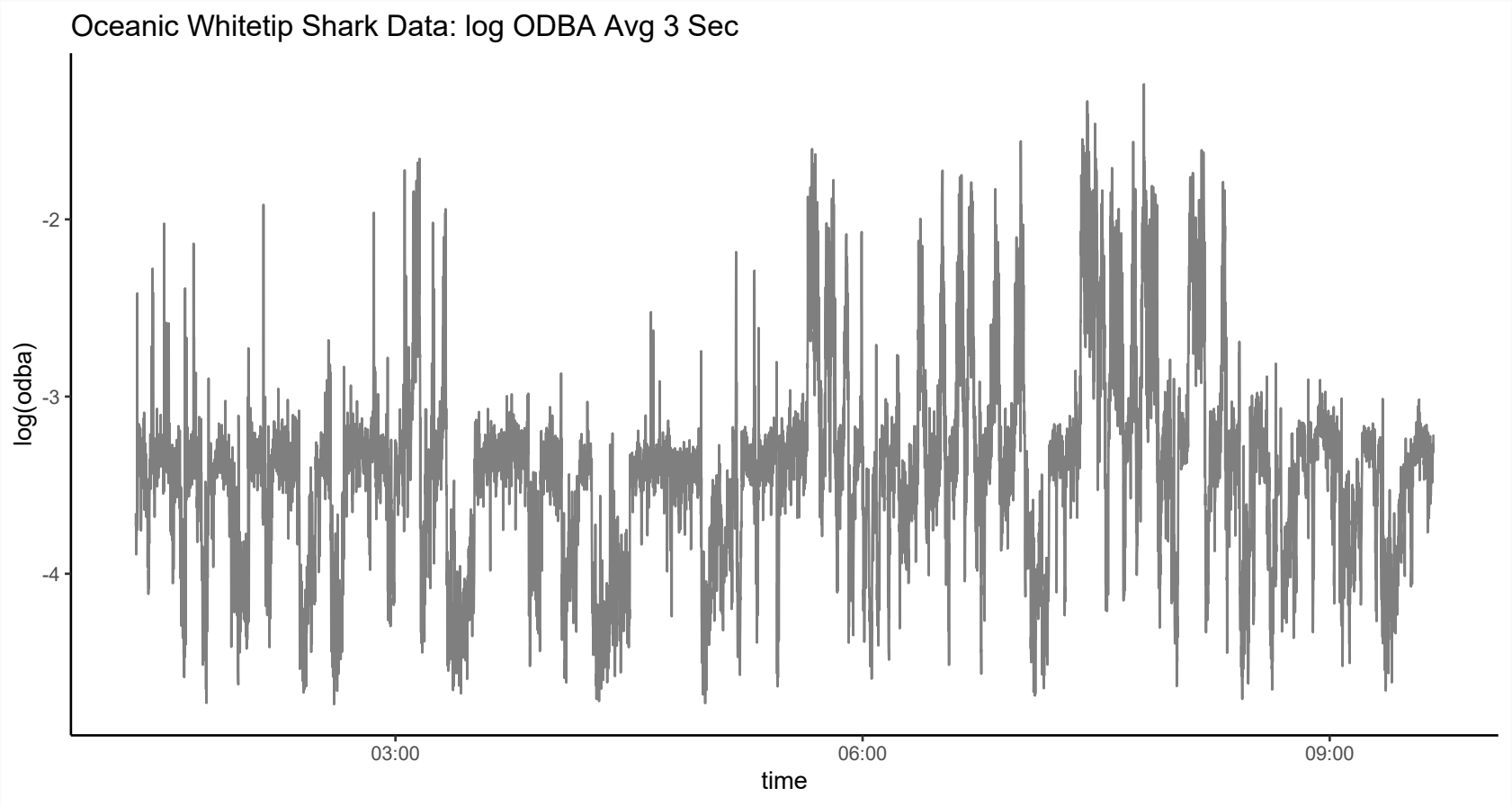
Example 1: White Sharks Active Tracking Data (cont.)

HMM Results:



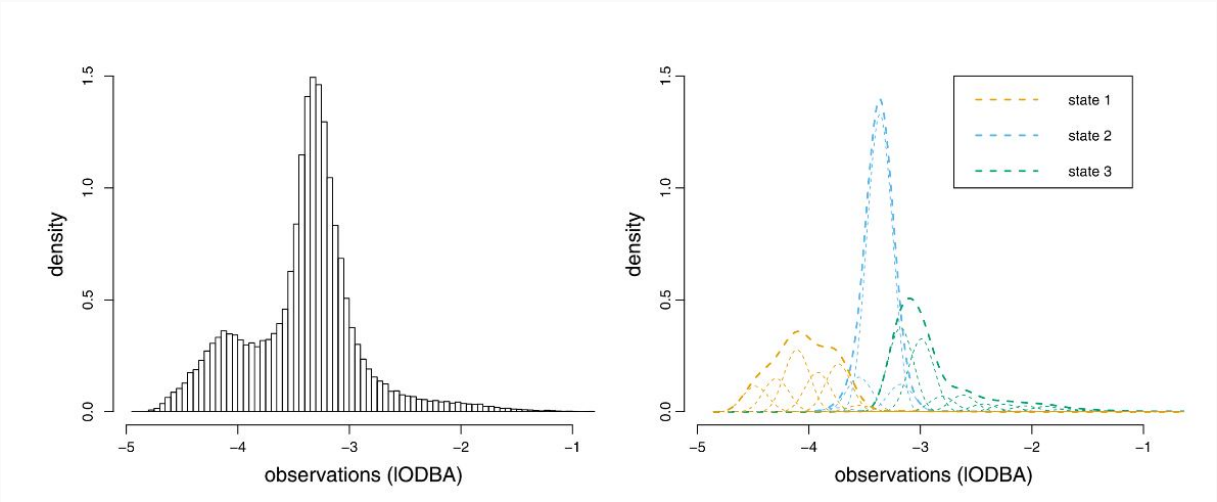


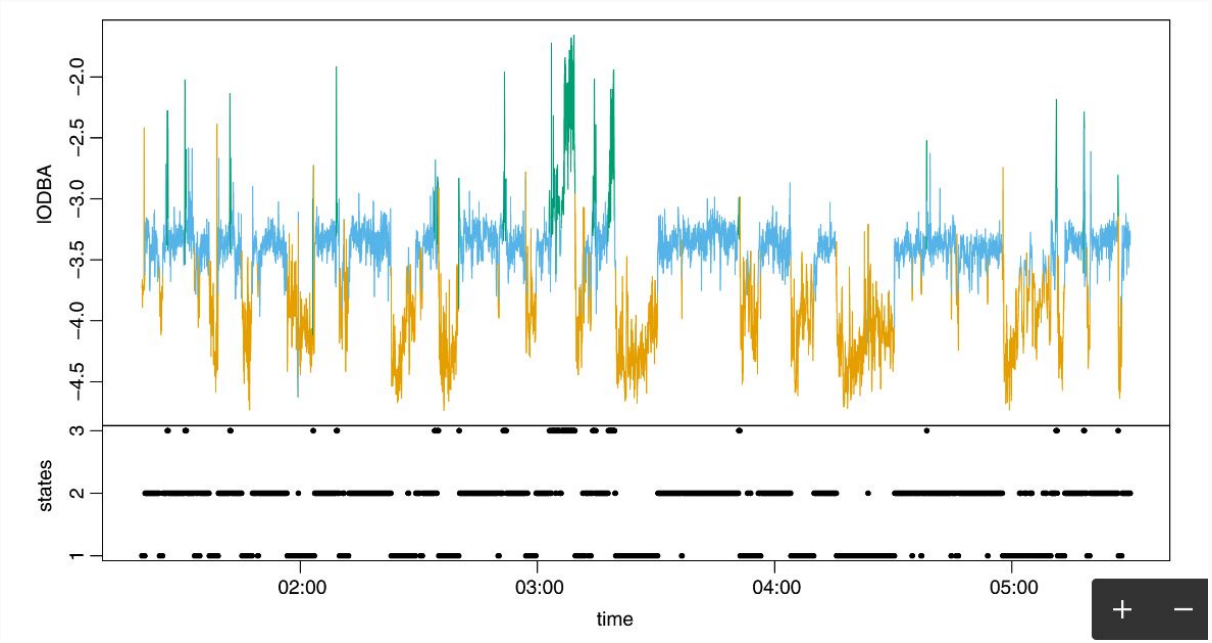
Example 2: Oceanic Whitetip Shark Acceleration Data



Example 2: Oceanic Whitetip Shark Acceleration Data (cont.)

HMM Results:





Data requirements

Data Processing is the Most Important Step

Biological Question of Interest $\rightarrow$ Quantification $\rightarrow$ Data Features

Requirements:

- Observations collected at regular temporal scales (can be filled with missing data)

hidden Markov models

Observations

Processes: Observation – $\{\mathbf{Y}_t\}_{t=1}^{T}$

Figure 1: Observations over time.

Observations and States

Processes: Observation – $\{\mathbf{Y}_t\}_{t=1}^{T}$ State – $\{\mathbf{S}_t\}_{t=1}^{T}$

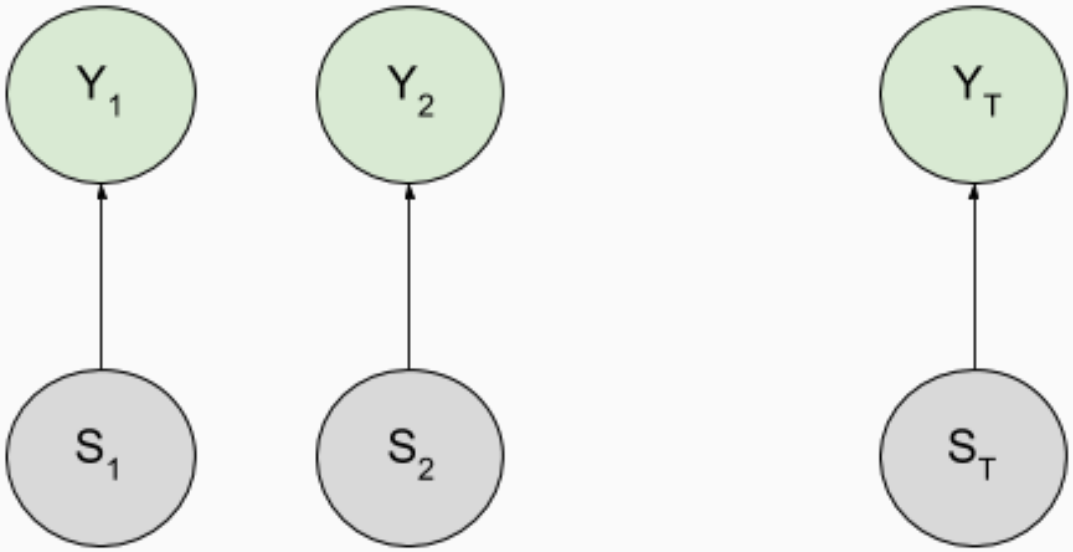


Figure 2: Observation and state (behavior) process.

Finite mixture model, assuming independence. At each point in time, S_t takes on one of N possible values.

hidden Markov models

Processes: Observation – $\{\mathbf{Y}_t\}_{t=1}^{T}$ State – $\{\mathbf{S}_t\}_{t=1}^{T}$

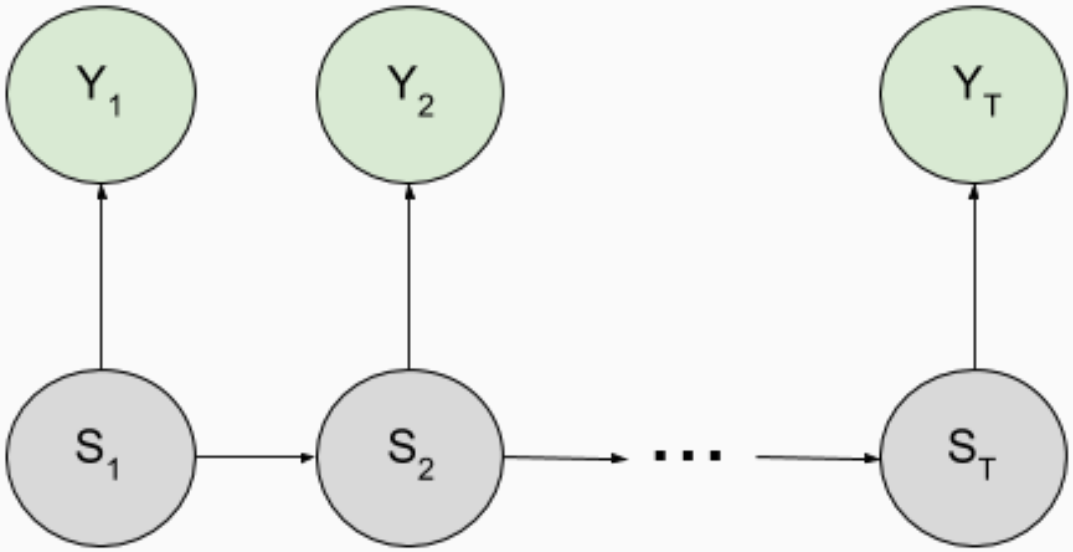


Figure 3: Graphical dependence structure implied by a hidden Markov model.

HMMs (cont.)

A basic finite-state, discrete-time HMM is fully specified by four components:

Mathematical Description

- Number of states, N

Biological Interpretation

- Number of movement patterns (proxies for behavior)

HMMs (cont.)

A basic finite-state, discrete-time HMM is fully specified by four components:

Mathematical Description

- Number of states, N
- State-dependent distributions: $\{f(\mathbf{y}_t|S_t = n)\}_{n=1}^{N}$

Biological Interpretation

- Number of movement patterns (proxies for behavior)
- Movement pattern distributions

HMMs (cont.)

A basic finite-state, discrete-time HMM is fully specified by four components:

Mathematical Description

- Number of states, N
- State-dependent distributions: $\{f(\mathbf{y}_t|S_t = n)\}_{n=1}^{N}$
- Transition probability matrix (t.p.m.), $\mathbf{\Gamma}$

Biological Interpretation

- Number of movement patterns (proxies for behavior)
- Movement pattern distributions
- How does the animal switch across behaviors over time?

HMMs (cont.)

A basic finite-state, discrete-time HMM is fully specified by four components:

Mathematical Description

- Number of states, N
- State-dependent distributions: $\{f(\mathbf{y}_t|S_t = n)\}_{n=1}^{N}$
- Transition probability matrix (t.p.m.), $\mathbf{\Gamma}$
- Initial state distribution, $\boldsymbol{\delta}$

Biological Interpretation

- Number of movement patterns (proxies for behavior)
- Movement pattern distributions
- How does the animal switch across behaviors over time?
- What is the animal doing the first time we see it?

Number of States & State-Dependent Distributions

Number of States - number of movement patterns:

- Chosen *a priori* to fit the model. Part of the exploratory search.
- Helps when chosen (approximately) using domain expertise.

State-dependent (movement patterns) distributions:

Key idea: there are multiple signals that stem from the animal exhibiting a given behavior. We group these signals into behaviors through the use of probability distribution (or mass) functions.

For example, we could have

$f(y_t|S_t = n) \sim Normal(\mu_n, \sigma_n)$ or
$f(y_t|S_t = n) \sim Gamma(\mu_n, \sigma_n)$ or
(any other valid distribution)

Transition Probability Matrix

A **Markov chain**[1] is a stochastic process $\{S_t, t = 1, 2, \ldots\}$, i.e. a sequence of random variables $S_1, S_2, \ldots$, such that:

- $S_t \in \{1, \ldots, N\}$ for all t (i.e. there are N so-called "states")
- the **Markov property** holds:

$$\Pr(S_{t+1} = s_{t+1} \,|\, S_t = s_t, \ldots S_1 = s_1) = \Pr(S_{t+1} = s_{t+1} \,|\, S_t = s_t)$$

The **transition probability matrix** (t.p.m.) is given by

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & \ddots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NN} \end{pmatrix}$$

There are two important restrictions on the parameters:

- $\gamma_{ij} \in [0, 1]$ for all i, j
- $\sum_{j=1}^{N} \gamma_{ij} = 1$ for all i

[1] more precisely: a discrete-time, finite-state Markov chain

State duration implied by a Markov chain

How long do we spend in a state?

Let D_n be a random variable connected with the amount of time that we spend in a state *before switching*. Then, the implied distribution for D_n when assuming the Markov property is given as,

$$Pr(D_n = d_n) \sim Geom(1 - \lambda_{nn})^2$$

such that $E(D_n) = 1/(1 - \lambda_{nn})$.

[2] What's the story here? Do the calculations make sense? Work out together.

Markov chains — stationary distribution

Suppose we are dealing with a homogeneous and "well-behaved"[3] Markov chain.

What if we observe this Markov chain at a random point in time, after it has already been running for some time? What state will we find it in?

$\rightarrow$ the probabilities of encountering the process in any of the N states is then given by the **stationary distribution** — this is the vector $\boldsymbol{\delta} \in \mathbb{R}^N$ such that

$$\boldsymbol{\delta}\boldsymbol{\Gamma} = \boldsymbol{\delta} \text{ subject to } \sum_{i=1}^{N} \delta_i = 1$$

[3] it needs to be irreducible and aperiodic — properties that are pretty much always met in practice!

HMMs

- an N-state HMM is a (doubly) stochastic process in discrete time, with
  - an unobservable **state process** $S_1, S_2, \ldots, S_T$ taking values in $\{1, \ldots, N\}$,
  - an observed **state-dependent process** $Y_1, Y_2, \ldots, Y_T$,
- such that
  - $f(y_t \mid s_1, \ldots, s_t, y_1, \ldots, y_{t-1}) = f(y_t \mid s_t)$ (conditional independence assumption)
  - $f(s_t \mid s_1, \ldots, s_{t-1}) = f(s_t \mid s_{t-1})$ (Markov property)

The HMM Story So far...

- We begin at time $t = 1$, the first observation of the animal. The initial distribution assigns probabilities to the animal exhibiting one of the N states (movement patterns/behaviors) during this period.
- Given whatever it's doing at S_1, it produces an observation y_1 from the distribution $f(y_1|S_1 = s_1)$.
- What the animal does during S_2 depends on what it was doing at S_1.
- Given whatever it's doing at S_2, it produces an observation y_2 from the distribution $f(y_2|S_2 = s_2)$.
- What the animal does during S_3 depends on what it was doing at S_2.
- (and so on)

Another slightly different version...

- We begin at time $t = 1$, the first observation of the animal. The initial distribution assigns probabilities to the animal exhibiting one of the N states (movement patterns/behaviors) during this period.
- Given whatever it's doing at S_1, it produces an observation y_1 from the distribution $f(y_1|S_1 = s_1)$.
- We expect that the animal will remain in the same state for a given amount of time, as given by the state duration distribution. Once it 'finishes' its time, it switches to another state with the appropriate probabilities
- If we remain in state n for d_n time points, we draw d_n observations from $f(y_{t-d_n+1:t}|S_{t-d_n+1:t} = n)$.

Simulating Data from an HMM

RMarkdown file.

Extending the basic HMM

Multivariate time series

HMMs are often used to model **multivariate time series**,
$\mathbf{X}_t = (X_{t1}, \ldots, X_{tK})$

$$\begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1K} \end{pmatrix}, \begin{pmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2K} \end{pmatrix}, \ldots, \begin{pmatrix} X_{T1} \\ X_{T2} \\ \vdots \\ X_{TK} \end{pmatrix}.$$

In many scenarios it can be assumed that the K variables are **driven by the same underlying state process** S_t

Dependence assumptions in multivariate HMMs

Two types of conditional independence assumptions can be made:

- **longitudinal conditional independence** — conditional on the states, the vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$ are mutually independent[4]

  select class of multivariate distributions, with some density $f(y_{t1}, \ldots, y_{tK})$

- **contemporaneous conditional independence** — conditional on the states, all components Y_{tk}, $t = 1, \ldots, T$, $k = 1, \ldots, K$ are independent

  $\rightarrow$ multivariate state-dep. distrib. is a product of K univariate distributions
  $\rightarrow$ select classes of univariate distributions, with densities $f(y_{t1}), \ldots, f(y_{tK})$

[4] however, at any time t, the components the vector $\mathbf{Y}_t$ may be correlated!

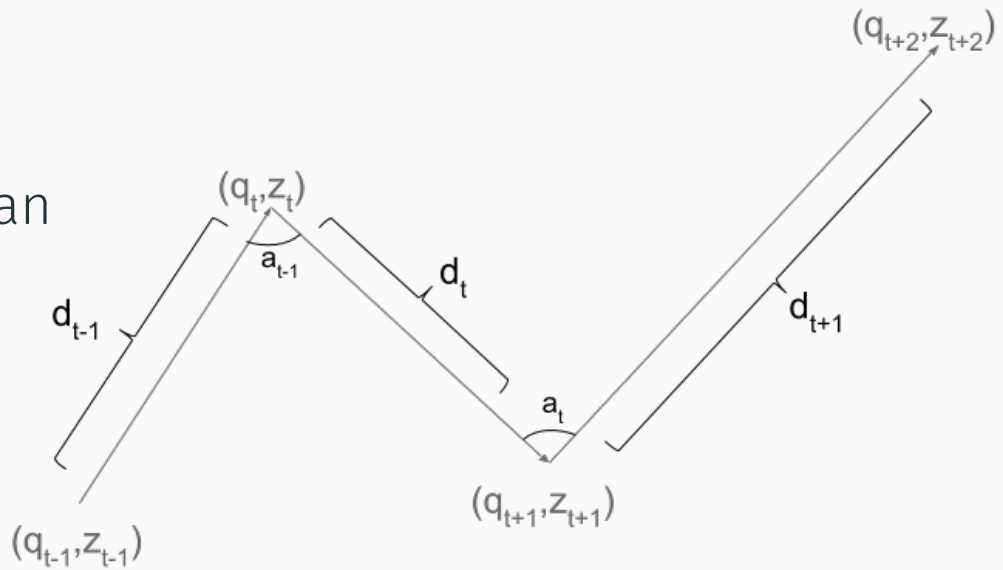
Modeling Positional Data

Coordinates: $\{q_t, z_t\}_{t=1}^{T+1}$

Transform positions into:

- **step lengths** $\{d_t\}_{t=1}^{T}$ (euclidean distance between two points)
- **turning angles** $\{a_t\}_{t=1}^{T}$ (using three consecutive positions)

Let $\mathbf{y}_t = \{d_t, a_t\}_{t=1}^{T}$



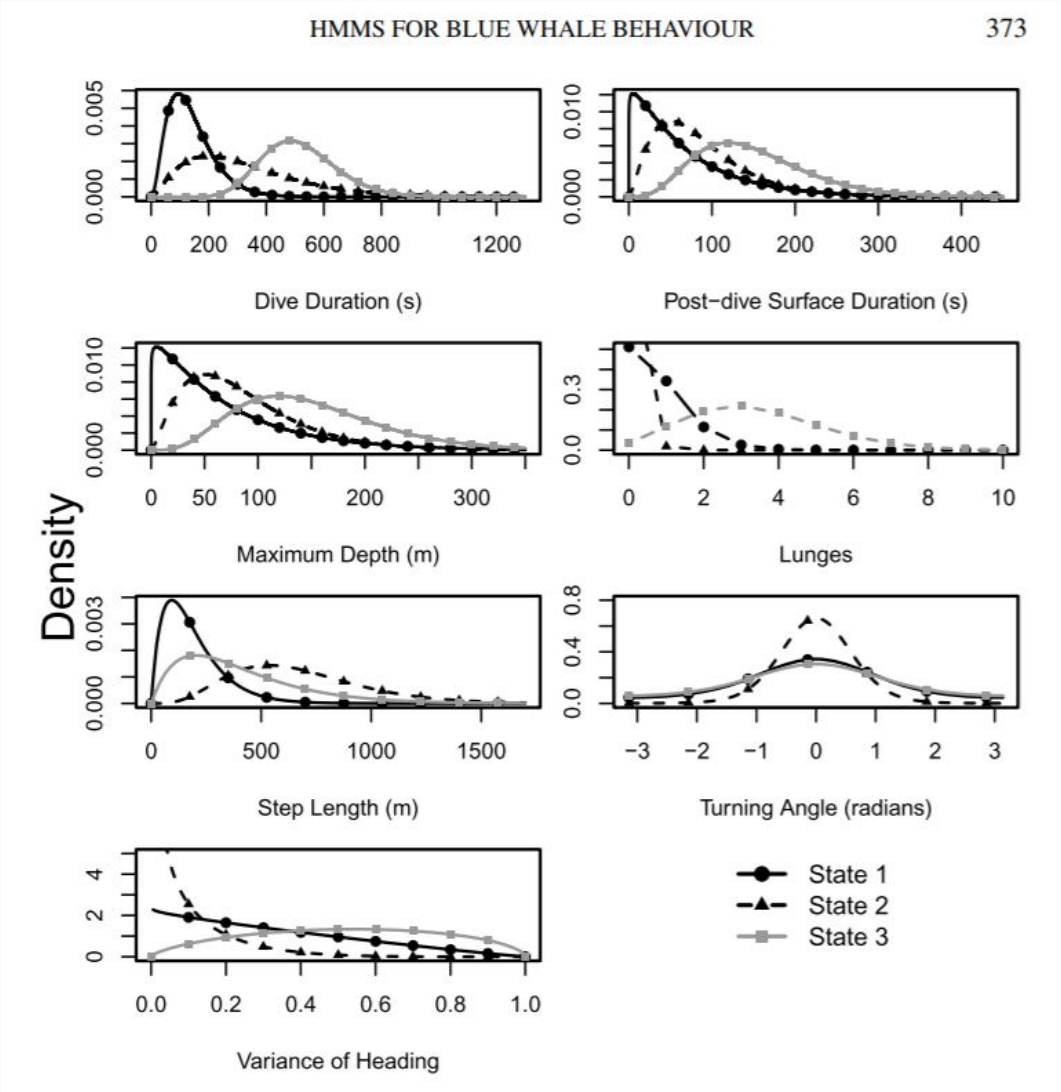
Key Idea:
Observed movements are a result of the underlying behavior that the animal is exhibiting.

HMM for Positional Data

What are the four components of an HMM and how do we construct an HMM for positional data?

Blue Whale Example

Paper by DeRuiter et al (2017): *A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure.*



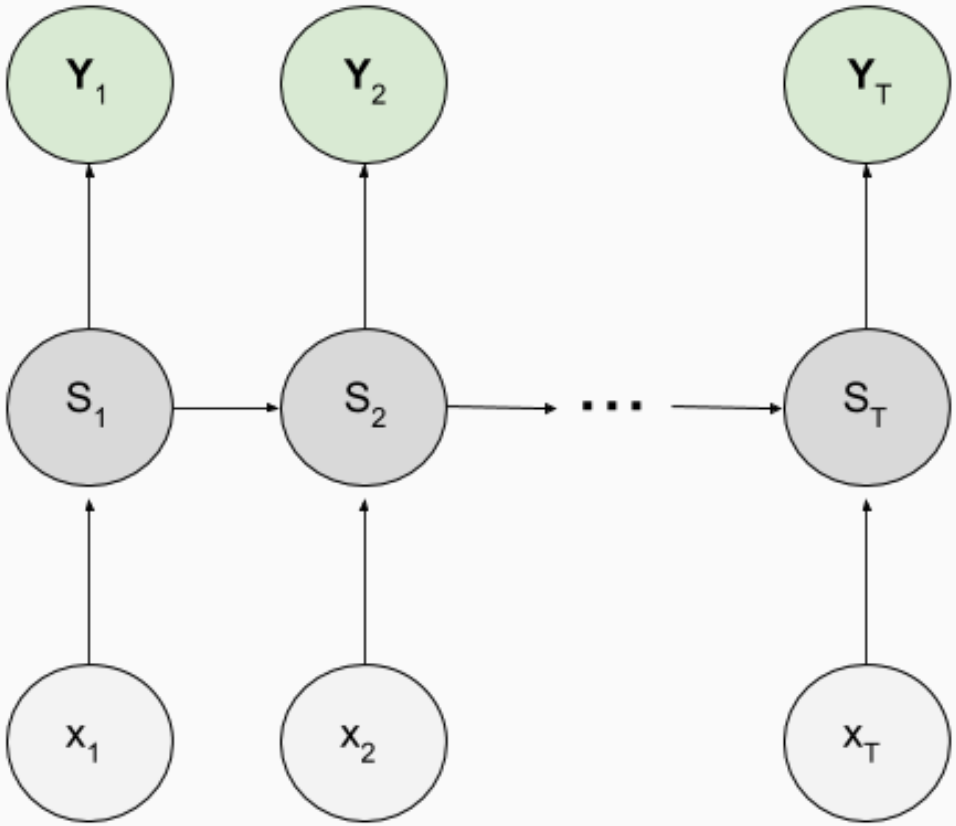
HMMs with Covariates in State Process

MULTINOMIAL LOGISTIC REGRESSION:

$$\gamma_{ij}^t(x_t) = \frac{\exp(\rho_{ij}^t)}{\sum_{j=1}^N \exp(\rho_{ij}^t)}$$

$$\rho_{ij}^t = \begin{cases} \tau_0^{(ij)} + \tau_1^{(ij)} x_t & \text{if } i \neq j; \\ 0 & \text{o.w.} \end{cases}$$

for $i, j \in \{1, \ldots, N\}$



Inference in HMMs

Likelihood(s)

We need to know what the likelihood of the HMM is in order to do inference. There are two likelihoods of interest in the HMM literature. First, the ***complete-data likelihood*** is the joint distribution of the observation and state process, written as

$$f(\mathsf{Y}, \mathsf{S}) = \delta_{S_1} f_{S_1}(y_1) \prod_{t=2}^{T} \gamma_{S_{t-1}, S_t} f_{S_t}(y_t)$$

Very popular to use when doing inference, as the expression is straightforward, easy to interpret. Can conduct inference by alternating estimation of the state and observation process, *à la* expectation-maximization.

Likelihood(s)

The other likelihood in the HMM literature, just referred to as the likelihood, perhaps to some the *marginal distribution*, is the distribution of the joint observation process *only*.

$$\begin{aligned}\mathcal{L} = f(\mathsf{Y}) &= \sum_{S_1=1}^{N}\sum_{S_2=1}^{N}\cdots\sum_{S_T=1}^{N} f(\mathsf{Y},\mathsf{S}) \\ &= \sum_{S_1=1}^{N}\sum_{S_2=1}^{N}\cdots\sum_{S_T=1}^{N} \delta_{S_1} f_{S_1}(y_1)\prod_{t=2}^{T}\gamma_{S_{t-1},S_t} f_{S_t}(y_t)\end{aligned}$$

The number of sums we have to calculate: N^T.

Likelihood(s)

However, we can also write the likelihood as a matrix product. Let $\mathbf{B}(\mathbf{y}_t)$ be an $N \times N$ diagonal matrix with entries $\mathrm{B}_{nn}(\mathbf{y}_t) = f(\mathbf{y}_t | S_t = n)$ for $n = 1, \ldots, N$, then we can express the likelihood of an individual time series as a matrix product[5],

$$\mathcal{L} = \boldsymbol{\delta}(x_1)^\top \mathbf{B}(\mathbf{y}_1) \prod_{t=2}^{T} \boldsymbol{\Gamma}(x_t) \mathbf{B}(\mathbf{y}_t) \mathbf{1}$$

with $\mathbf{1}$ = (1, ..., 1) denoting a vector of length N.

[5] the only joint distribution I've ever seen expressed as a matrix product

Likelihoods Comparison

Which likelihood form to use?

1. The complete-data likelihood is easy, nice to look at, straightforward, but...very deceiving. Conducting inference for the observation and state process requires a lot of calculations depending on what approach is taken. It's also slower because we *think* we need to care about the observed state process.

2. The likelihood is not easy to look at, looks complex. However, there is no need to care about the state process, and marginalization actually concentrates the likelihood around the parameter values (a good property!). But, what do we even use to do inference with this?

Dynamic programming techniques

Recall that for the likelihood, we either need to conduct N^T sums or evaluate a matrix product (numerically unstable given we multiply lots of probabilities).

In HMMs, we'll turn to dynamic programming techniques for

- likelihood evaluation
- state decoding
- construction of various distributional forms

Forward Algorithm

For the sequence of observations $\{y_t\}_{t=1}^T$ we can evaluate $\mathcal{L}$ via the *forward algorithm*[6]. We define the sequence of **forward variables**, $\{\boldsymbol{\alpha}_t\}_{t=1}^T$, starting at time t=1,

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{B}(\mathbf{y}_1), \quad \text{where} \quad \alpha_1(n) = \Pr(S_1 = n, \mathbf{y}_1)$$

Then,

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{B}(\mathbf{y}_t), \quad \text{where} \quad \alpha_t(n) = \Pr(S_t = n, \mathbf{y}_1, \ldots, \mathbf{y}_t)$$

Then, the likelihood is obtained by summing over $\boldsymbol{\alpha}_T$,

$$\mathcal{L} = \sum_{n=1}^{N} \alpha_T(n).$$

[6] presented here without scaling – can modify to avoid numerical underflow

Inference

Given we can evaluate the likelihood with the forward algorithm, this means we can now do inference!

It leaves us with two routes:

- **Numerical maximization**
- **Bayesian inference**

Keep in mind, the likelihood is not available in closed form so we need to use an algorithm that performs well. From here on out, we'll do Bayesian inference in the software **Stan**.

The Underlying State Process

What about the underlying state process, $\mathbf{S}$? We have no need (*at all*) to conduct inference for the state and observation process jointly. Using the forward algorithm, we can conduct inference on the parameters only (highly recommended).

How do we estimate the underlying state process then?

State decoding in HMMs — overview

Given a fitted model, it is often of interest to **decode the hidden states** underlying the observed time series.

Local decoding:

- consider $\Pr(S_t = i | y_1, \ldots, y_T)$
- most probable state at time t is maximum of the above over $i = 1, \ldots, N$

(looks at each time point in isolation)

Global decoding:

- consider $\Pr(S_1 = i_1, \ldots, S_T = i_T | y_1, \ldots, y_T)$
- most probable state sequence is maximum of the above over $(i_1, \ldots, i_T) \in \{1, \ldots, N\}^T$

(looks at the sequence as a whole)

Usually the outcome is either identical or at least very similar.

Backward Algorithm

For local state decoding, along with the forward variables, we need the backward variables. We define the sequence of **backward variables**, $\{\boldsymbol{\beta}_t\}_{n=1}^{N}$. For $t = T$, $\boldsymbol{\beta}_t = \mathbf{1}$, for $\mathbf{1}$ a vector of ones of size N.

For $t \neq T$, we have

$$\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\mathbf{B}(\mathbf{y}_{t+1})\boldsymbol{\beta}_{t+1},$$

where,

$$\beta_t(n) = \Pr(S_t = n|\mathbf{y}_t, \ldots, \mathbf{y}_t)$$

Forward-Backward Algorithm

Local state decoding: $\Pr(S_t = n|\mathbf{y})$, for $n \in \{1, \ldots, N\}$

Using the *forward* and *backward* variables, we can evaluate the state probabilities:

$$\Pr(S_t = n|\mathbf{y}) = \frac{\alpha_t(n)\beta_t(n)}{\sum_{n=1}^{N} \alpha_t(n)\beta_t(n)}$$

The Viterbi algorithm

Define:

$\xi_1(i) = f(S_1 = i, y_1)$

$\xi_t(i) = \max_{i_1,\ldots,i_{t-1}} f(S_1 = i_1, \ldots, S_{t-1} = i_{t-1}, S_t = i, y_1, \ldots, y_t)$

(effectively the highest possible prob. of sequences ending in state i at t)

Recursive scheme:

$$\xi_t(j) = \big(\max_{i=1,\ldots,N} \big(\xi_{t-1}(i)\gamma_{ij}\big)\big) \cdot f(y_t|S_t = j)$$

After calculating the $\xi_t(i)$ as above, we backtrack the optimal state sequence:

- $i_T^* = \operatorname{argmax}_{i=1,\ldots,N} \xi_T(i)$
- $i_t^* = \operatorname{argmax}_{i=1,\ldots,N} \big(\xi_t(i)\gamma_{i,i_{t+1}^*}\big)$ (for $t = T-1, T-2, \ldots, 1$)

$\rightarrow (i_1^*, \ldots, i_T^*)$ is the most likely state sequence

Model checking in HMMs

Main options to check if a fitted HMM is adequate:

1. graphical comparison of marginal distribution under fitted HMM and empirical distribution, to check adequacy of state-dep. distributions

2. simulate data from the fitted model, then compare the patterns found in the simulated data with those of the real data (patterns to look for: marginal distribution, autocorrelation, etc.)

3. a residual analysis

Overview + Markov-switching processes

Why did we want to fit an HMM?

There are two main features of the data that an HMM captures:

- **the marginal distribution of the observation at time** t
- **the autocorrelation structure**

In animal movement, we never validate our model by how well it captures 'actual behavior'. So any paper out there that claims that HMMs are tools for automatic identification of animal behavior is very wrong and very misleading.

Oftentimes, with the right data, the right domain expertise, we can capture simple, biologically relevant patterns in the movement data that can serve as good proxies for some very general behaviors of interest (active/not active – area restricted search/directed traveling).

Markov-switching processes

Composed of four items:

- Number of states, N
- State-dependent distributions: $\{f(\mathbf{y}_t|S_t = n)\}_{n=1}^N$
- Transition probability matrix (t.p.m.), $\mathbf{\Gamma}$
- Initial state distribution, $\boldsymbol{\delta}$

For a Markov-switching simple linear regression model, the state-dependent distributions are written as:

$$f(y_t|S_t = n) \sim N(\beta_{0,n} + \beta_{1,n}x_t, \sigma_n^2)$$

The state-dependent parameters are the collection of $\{\boldsymbol{\beta}_n\}_{n=1}^N$ and $\{\sigma_n\}_{n=1}^N$.

Markov-switching processes

Wait...that looks like an HMM....(YES)

A very important and practical reason to recognize the same general structural forms is so that we can use the same 'HMM machinery' to conduct inference.

We can then,

- marginalize over the latent states
- conduct inference using the likelihood directly
- do Bayesian inference in Stan
- state decoding

Stan code

In the github repository, https://github.com/vianeylb/CONABIO_AME2019, there are several R and Stan scripts to simulate data from various models and conduct inference.

Time to play with code!