The Hidden Markov Model

The State Variables

The *telprep* package uses a hidden Markov model (HMM) to estimate the survival and detection probability of fish, evaluate the faultiness of mortality sensors, and determine the most likely path of survival states for each fish. In building the mathematical framework of the HMM, we start by letting $S_i(t)$ be a state variable with $S_i(t) \in \{1, 2\}$ for $i \in \{1, 2, \cdots, n\}$ and $t \in \mathbb{R}$. Here, i is used to index individual fish whereas t is defined at the time of the detection period (or flight grouping). Here, $S_i(t) = 1$ is defined as the event that the i^{th} fish is alive at time t and $S_i(t) = 2$, that the fish has expired at or before time t.

A Continuous-Time Markov Process

In modeling the time-homogeneous instantaneous rate of state transition, transition intensities are defined as $q_{irs} = \lim_{\delta t \to 0} P(S_i(t + \delta t) = s | S_i(t) = r)$. Here, $q_{irs}(t)$ represents the instantaneous risk that fish i moves from state r to s at time t. When it is assumed that the transition intensities do not depend upon individual fish, the index i can be dropped. The transition intensities can be represented as a stochastic matrix, called a transition intensity matrix. The transition intensity matrix Q is defined as $Q_{rs} = q_{rs}$.

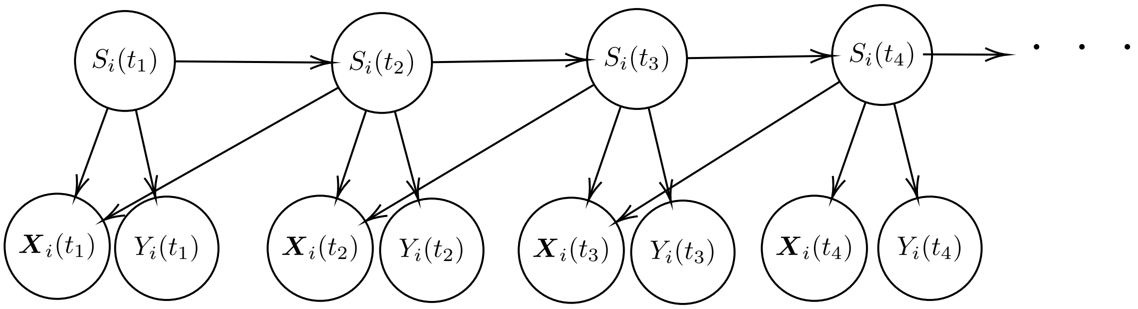
Later on, the likelihood will be formulated in terms of a transition matrix $P(t)$. In defining the transition matrix for a time-homogeneous continuous-time Markov process under the assumption that the transition intensities do not depend upon individual fish, the transition probabilities are first defined as $p_{rs}(t) = P(S(t + u) = s | S(u) = r)$. That is, $p_{rs}(t)$ represents the probability of being in state s at time $t + u$ given that the state at time u is r. The transition matrix is now defined as $P(t)_{rs} = p_{rs}(t)$. For time-homogeneous processes, the relationship between the transition and transition-intensity matrix is given by $P(t) = \mathrm{Exp}(tQ)$ where the function Exp is the matrix exponential.

The Observation Variables

Let $\mathbf{X}_i(t) = \{X_{1i}(t), X_{2i}(t)\}$ and $Y_i(t)$ be detection variables for $t \in \{t_1, t_2, \cdots, t_m\} = T \subset \mathbb{R}$ where m is the number of detection periods and t_i is the time of the i^{th} detection period (i.e. the flight time). Here, $X_{1i}(t)$ and $X_{2i}(t)$ are used to represent the detection longitude and latitude of the i^{th} fish at time t if the fish was detected during the detection period. It is assumed that $\mathbf{X}_i(t) \in R \cup \{0\}$ where R is the spatial domain of the river system and $\mathbf{X}_i(t) = 0$ is the event that the i^{th} fish is undetected at time t. $Y_i(t)$ is used to represent the mortality status of the tag of the i^{th} fish. It is assumed that $Y_i(t) \in \{1, 2, 3\}$ where $Y_i(t) = 1$ is the event that the i^{th} fish is undetected during at time t; $Y_i(t) = 2$ that the fish is detected and that a mortality signal is given by the tag; and $Y_i(t) = 3$, that the fish is detected and no mortality signal is given by the tag.

The Likelihood

We assume the conditional dependencies of $S_i(t)$, $\mathbf{X}_i(t)$ and $Y_i(t)$ respect the following trellis diagram:



With this assumption, the joint distribution can be factored as

$$p(\mathbf{x}, \mathbf{y}, \mathbf{s}) = \prod_{i=1}^{n} p[s_i(t_1)]p[\mathbf{x}_i(t_1)|s_i(t_1)]p[y_i(t_1)|s_i(t_1)] \prod_{j=2}^{m} p[s_i(t_j)|s_i(t_{j-1})]p[\mathbf{x}_i(t_j), \mathbf{x}_i(t_{j-1})|s_i(t_j)]p[y_i(t_j)|s_i(t_j)]$$

The emission probabilities $p[\mathbf{x}_i(t_j), \mathbf{x}_i(t_{j-1})|s_i(t_j)]$ are assumed to depend upon the Euclidean distance between $\mathbf{x}_i(t_j)$ and $\mathbf{x}_i(t_{j-1})$. More precisely, it is assumed that $p[\mathbf{x}_i(t_j), \mathbf{x}_i(t_{j-1})|s_i(t_j)] = p[z_i(t_j)|s_i(t_j)]$ where the random variable $Z_i(t_j)$ is defined as

$$Z_i(t_j) = \begin{cases} 1 \quad \text{if } \mathbf{X}_i(t_j) \text{ or } \mathbf{X}_i(t_{j-1}) = 0 \\ 2 \quad \text{if } \mathbf{X}_i(t_j) \text{ and } \mathbf{X}_i(t_{(j-1)}) \neq 0 \text{ and } \ d(\mathbf{X}_i(t_j), \mathbf{X}_i(t_{(j-1)})) \leq t^* \\ 3 \quad \text{if } \mathbf{X}_{ij} \text{ and } \mathbf{X}_{i(j-1)} \neq 0 \text{ and } \ d(\mathbf{X}_i(t_j), \mathbf{X}_i(t_{j-1})) > t^* \end{cases}$$

where $d(\mathbf{x}_i(t_j), \mathbf{x}_i(t_{(j-1)}))$ is the Euclidean distance between $\mathbf{x}_i(t_j)$ and $\mathbf{x}_i(t_{(j-1)}))$ and t^* is a predefined threshold distance. A concrete interpretation of t^* will be given later on. With the assumption that $p[\mathbf{x}_i(t_j), \mathbf{x}_i t_{j-1} | s_i(t_j)] = p[z_i(t_j)|s_i(t_j)]$, the Markov assumption that the future evolution only depends on the current state is satisfied.

Model Parameters

In parameterizing the model, the transition matrix is defined as

$$P(t) = \begin{pmatrix} p_{11}(t) & p_{12}(t) \\ 0 & 1 \end{pmatrix}.$$

Here, $p_{11}(t)$ is the survival rate and $p_{12}(t)$ is the mortality rate over t units of time. The second row of the transition matrix expresses the mathematically deep notion that dead fish remain dead.

An emission array E is used to parameterize the emission probabilities (i.e. $p[z_i(t_j)|s_i(t_j)]$ and $p[y_i(t_j)|s_i(t_j)]$). Here, $E(t)_{1kl} = p[z_i(t) = l|s_i(t) = k]$ is defined as

$$E(t)_{1kl} = \begin{pmatrix} g_{111} & g_{112} & g_{113} \\ g_{121} & g_{122} & 0 \end{pmatrix}$$

and $E(t)_{2kl} = p[y_i(t) = l|s_i(t) = k]$ as

$$E(t)_{2kl} = \begin{pmatrix} g_{211} & g_{212} & g_{213} \\ g_{221} & g_{222} & g_{223} \end{pmatrix}.$$

Here, it is assumed that fish that move more than t^* units between consecutive detection periods are alive (i.e. $E(t)_{123} = 0$). While this assumption may not be entirely true in practice (expired fish may be transported downstream), the assumption is cause for little concert; expired fish will likely be stationary during in the following detection period. Alternatively, the expired fish will be transported out of the system. It is also assumed that the emission probabilities are not temporally dependent: the detection probability and efficacy of the mortality signals does not change over time.

Finally it is assumed that $(s_i(t)|s_{i(t-1)})$, $(z_i(t)|s_i(t))$, and $(y_i(t)|s_i(t))$ are categorically distributed with the relevant parameters found in the transition and emission matrices (eg: $(s_i(t)|s_{i(t-1)} =$

1) $\sim$ Categorical$(p_{11}(t), p_{12}(t)))$. At this point the likelihood is well defined. Specific parameters of interest are outlined in the table below:

Table 1: Parameters of Interest

Parameter	**Interpretation**
$p_{11}(t)$	probability that a fish survives for the next t units of time
$p_{12}(t)$	probability that a fish expires in the next t units of time
$1 - g_{211}$	detection probability for live fish
$1 - g_{221}$	detection probability for expired fish
g_{212}	probability that a fish is detected with the mortality signal on given that the fish is alive
g_{213}	probability that a fish is detected with the mortality signal off given that the fish is alive
g_{222}	probability that a fish is detected with the mortality signal on given that the fish has expired
g_{223}	probability that a fish is detected with the mortality signal off given that the fish has expired

Parameter Estimation and Survival State Determination

The Baum Welch algorithm is used to estimate the transition intensities and the emission probabilities. Briefly, the algorithm estimates the model parameters by maximizing the marginal likelihood $L(\mathbf{x}, \mathbf{y}; \theta)$. The benefit to using this approach as opposed to MCMC methods is that it is computationally much quicker.

In addition to the parameter estimates, inference be drawn directly from the model states. To determine the most probable sequence of survival states, the Viterbi algorithm was used. More formally, the Viterbi algorithm was used to determine the argmax$(p(\mathbf{s}|\mathbf{x}, \mathbf{y}))$ (called the Viterbi path). Using the Viterbi path histograms of the mortality timing can be constructed to gain insight into the time at which fish are most likely to expire. The plotting function *make_plot* in the *telprep* package couples the Viterbi path with the locations of best detection to generate the survival maps. The R package msm is used to implement the Baum-Welch and Viterbi algorithms.