Bayesian Mark-Recapture State-Space Model

Nick Cummings
Jennifer Hoeting
Colorado State University

October, 2010

State-Space

We propose to model the mark-recapture data using a state-space model. The disease state of each animal for each year is represented as a vector:

$$
\mathbf{y}_{i,t} =
\begin{bmatrix}
1 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}
$$

Each cell in the vector represents each of 7 states:

1. susceptible
2. infected
3. died susceptible
4. died infected
5. out of study
6. dead but infection status unknown
7. unobserved

So animal i, above, is observed susceptible at time t.

States over Time

To represent observations of disease state of the animal over time (five years), we use a matrix:

$$y_i = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Each column is a year. In this example, animal i is

- susceptible in year 1 (when it enters the study)
- survives to year 2, still susceptible
- contracts CWD in year 3
- survives to year 4
- dies in year 5

Absence from Study

An example of a deer that is essentially left-censored:

$$
\mathbf{y}_i =
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0
\end{bmatrix}
$$

The above animal is

- ▶ first captured in year 2 susceptible
- ▶ is unobserved in year 3
- ▶ is observed infected in year 4
- ▶ and dies in year 5.

Alternate Notation

To consider this a different way, we can write the history of deer i as a vector, where the number in cell t denotes the state of deer in year t:

$$\mathbf{y}_i = \begin{bmatrix} 5 \\ 1 \\ 7 \\ 2 \\ 4 \end{bmatrix}$$

The above animal has the same observed history as the previous deer.

True State (Hidden Data)

For each observed y_i matrix (or vector), there are one or more underlying possibilities for the hidden, "true" data. We denote the hidden data as $x_{i,t}$:

$$\mathsf{x}_{i,t} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

For the hidden data, there are only 5 possible states:

1. susceptible
2. infected
3. died susceptible
4. died infected
5. out of study

So animal i, above, is susceptible at time t.

Example: All Data Observed

If the true state deer i was observed for each year it was in the study, there is only one possibility for the hidden data:

$$\mathbf{y}_i = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{x}_i = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Example: Unobserved Years

If, however, deer i contains years unobserved, there are multiple possibilities (not of equal probability) for the hidden data:

$$
y_i = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}
$$

$$
x_i = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}
$$

$$
x_i = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}
$$

Multinomial Data Model

Each year of the observed data is modeled as a multinomial realization of a matrix of (unknown) observation probabilities multiplied by the hidden data vector for that year.

$$y_{t,i} \sim Multinomial(1, Dx_{t,i})$$

where D=

$$
\begin{bmatrix}
p_s & 0 & 0 & 0 & 0 \\
0 & p_i(1 - y_{t-1,i,2}) + y_{t-1,i,2} & 0 & 0 & 0 \\
0 & 0 & p_d & 0 & 0 \\
0 & 0 & 0 & p_d(1 - y_{t-1,i,2}) + y_{t-1,i,2} & 0 \\
0 & 0 & 0 & 0 & 0 \\
1 - p_s & (1 - p_i)(1 - y_{t-1,i,2}) & 1 - p_d & (1 - p_d)(1 - y_{t-1,i,2}) & 0
\end{bmatrix}
$$

for $i = 1, \ldots, n$ (n = total number of deer)
for $t = T_{i,1} + 1, \ldots, T$ ($T_{i,1}$ = first year animal i is in study, T = total number of study years)

▶ p_s is the probability of observing a susceptible animal

▶ p_i is the probability of observing an infected animal

▶ p_d is the probability of observing the CWD status of a deceased deer

Multinomial Process Model

Each year of the hidden data is modeled as a multinomial realization of a matrix of (unknown) probabilities multiplied by the previous year's state vector.

$$x_{t,i} \sim Multinomial(1, Mx_{t-1,i})$$

where M= $\begin{bmatrix} \phi_s(1-\phi_t) & 0 & 0 & 0 & 0 \\ \phi_i\phi_t & \phi_i & 0 & 0 & 0 \\ (1-\phi_s)(1-\phi_t) & 0 & 0 & 0 & 0 \\ \phi_t(1-\phi_i) & 1-\phi_i & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$

for $i = 1, \ldots, n$ (n = total number of deer)

for $t = T_{i,1} + 1, \ldots, T$ ($T_{i,1}$ = first year animal i is in study, T = total number of study years)

▶ ϕ_t is the probability that a susceptible becomes infected

▶ ϕ_s is the probability that a susceptible survives

▶ ϕ_i is the probability that an infected survives

Priors

We will use MCMC to obtain credible intervals estimating ϕ_t, ϕ_i, and the other parameters.
To determine whether our sample size will likely yield satisfactorily narrow intervals, we currently place uniform (uninformative) priors on all parameters:

$\phi_t \sim \mathit{Unif}(0, 1)$

But we may ultimately use previous studies to inform our prior distributions.

Incorporating Covariates

Furthermore, the current model assumes only the state data; that is, the yearly disease and mortality status of each deer. In future, we will incorporate covariates such as age, genetic relationship, spatial data, etc., using a logit transform:

$$\log \left(\frac{\phi_{tu,j}}{1-\phi_{tu,j}} \right) = \beta_0 + \beta_1 s_{1,u,j} + \cdots + \beta_k s_{k,u,j} + \epsilon_j$$

where, for example,

- $s_{1,u,j}$ is the age of deer j at year u
- $s_{2,j}$ is an indicator variable denoting whether deer j is in a familial group with a known infected deer
- etc.

Hidden Data Simulation

Using reasonable estimates for each of the parameters and
140 total deer captured in the first year alone, we obtain the
following averages for the simulated hidden data:

		Year				
		1	2	3	4	5
States	Susceptible	0.97	0.66	0.43	0.29	0.21
	Infected	0.03	0.04	0.04	0.04	0.03
	Died Susceptible	0.00	0.29	0.19	0.12	0.07
	Died Infected	0.00	0.02	0.04	0.01	0.01
	Out of Study	0.00	0.00	0.31	0.54	0.67

Observed Data Simulation

From the simulated hidden data, the following averages for the observed data are simulated:

		Year				
		1	2	3	4	5
	Susceptible	0.97	0.56	0.37	0.29	0.19
	Infected	0.03	0.04	0.04	0.04	0.03
States	Died Susceptible	0.00	0.14	0.11	0.08	0.04
	Died Infected	0.00	0.01	0.03	0.01	0.01
	Out of Study	0.00	0.00	0.31	0.54	0.67
	Dead (Status Unknown)	0.00	0.15	0.09	0.04	0.03
	Unobserved	0.00	0.09	0.06	0.00	0.03

Parameter Estimation Results

Again using reasonable parameter values for simulation, we generate realistic data for the following starting numbers of captured deer:

	Year Entering Study			
New Deer Capture	140	50	50	50
Infected Deer	4	2	1	2

Parameter Estimation Results

Using MCMC, we obtain the following results for the parameters of primary interest:

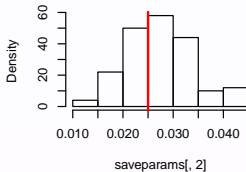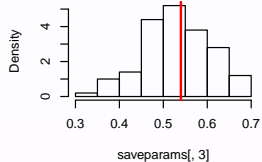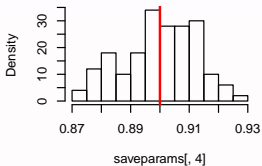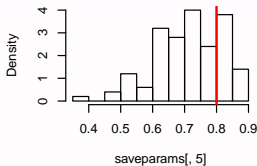| Parameter	True Value	Mean	Std. Dev.	95% CI
ϕ_t	.05	0.05	0.01	(0.03,0.08)
ϕ_i	.50	0.50	0.07	(0.37,0.64)

Posterior Distribution of phi.t
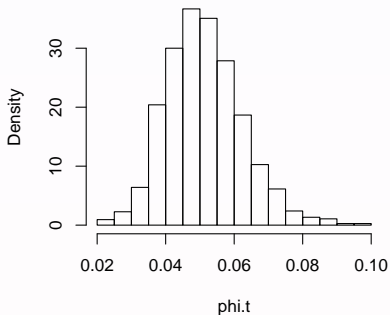
Posterior Distribution of phi.i
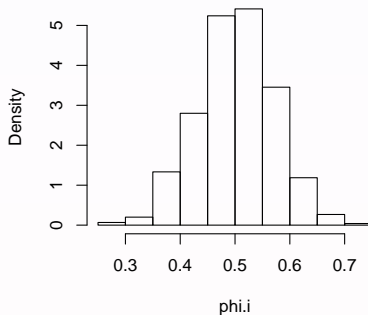
Parameter Estimation (More Realistic Parameters)

Results using values that are likely more accurate:

Parameter	True Value	Mean	Std. Dev.
ϕ_s	.85	0.86	0.01
ϕ_t	.025	0.025	0.01
ϕ_i	.54	0.52	0.08
p_s	.9	0.90	0.01
p_i	.8	0.66	0.17
p_d	.6	0.58	0.05

Posterior Distribution of phi.s
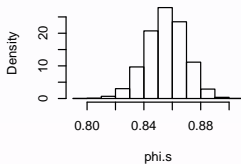
Posterior Distribution of phi.t
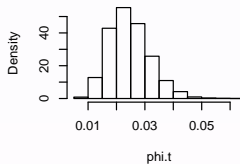
Posterior Distribution of phi.i
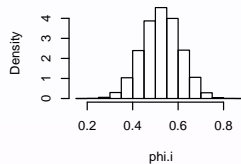
Posterior Distribution of p.s
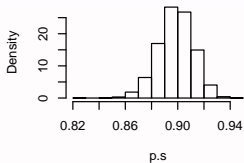
Posterior Distribution of p.i

Posterior Distribution of p.d

Observed Data
○○○○

Hidden Data
○○○

Model
○○

Parameter Estimation
○○

Preliminary Results
○○○○○○○●