

Mixture Models

Models for Socio-Environmental Data

Chris Che-Castaldo, Mary B. Collins, N. Thompson Hobbs

May 29, 2019



Course topics

Principles

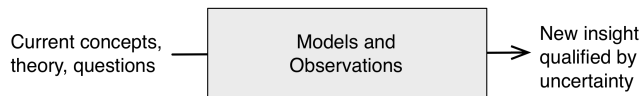
- Laws of probability
- Distribution theory
- Moment matching
- Bayes' theorem
- Conjugacy

Implementation and inference

- MCMC
- JAGS
- Inference from single and multiple models
- Model checking

Hierarchical models

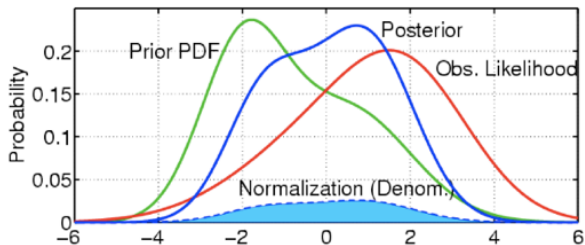
- Introduction
- Multi-level regression
- Mixture and occupancy
- State-space
- Spatial



Flow of ideas

- ▶ Mixture models in general
- ▶ Zero-inflation as a useful example of mixture models
- ▶ State-space examples of zero-inflation
 - ▶ Occupancy
 - ▶ Sensitivity-specificity
 - ▶ Capture-mark-recapture

Remember these bumpy distributions?



Mixture models

Mixture Distribution:

- ▶ $[y|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, p] = p[y|\boldsymbol{\theta}_1]_1 + (1-p)[y|\boldsymbol{\theta}_2]_2$
 - ▶ $0 \leq p \leq 1$.
 - ▶ $[y|\boldsymbol{\theta}_1]_1$ and $[y|\boldsymbol{\theta}_2]_2$ integrate to 1.

K-Mixture Distribution:

- ▶ $[y|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \mathbf{p}] = \sum_{k=1}^K p_k [y|\boldsymbol{\theta}_k]_k$
 - ▶ $p_k \geq 0$ for all k .
 - ▶ $\sum_{k=1}^K p_k = 1$.
 - ▶ all $[y|\boldsymbol{\theta}_k]_k$ integrate to 1.

Mixture models

Mixture Distribution:

- ▶ $[y|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, p] = p[y|\boldsymbol{\theta}_1]_1 + (1-p)[y|\boldsymbol{\theta}_2]_2$
 - ▶ $0 \leq p \leq 1$.
 - ▶ $[y|\boldsymbol{\theta}_1]_1$ and $[y|\boldsymbol{\theta}_2]_2$ integrate to 1.

K-Mixture Distribution:

- ▶ $[y|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \mathbf{p}] = \sum_{k=1}^K p_k [y|\boldsymbol{\theta}_k]_k$
 - ▶ $p_k \geq 0$ for all k .
 - ▶ $\sum_{k=1}^K p_k = 1$.
 - ▶ all $[y|\boldsymbol{\theta}_k]_k$ integrate to 1.

Mixture models

Mixture Distribution:

- ▶ $[y|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, p] = p[y|\boldsymbol{\theta}_1]_1 + (1 - p)[y|\boldsymbol{\theta}_2]_2$
 - ▶ $0 \leq p \leq 1$.
 - ▶ $[y|\boldsymbol{\theta}_1]_1$ and $[y|\boldsymbol{\theta}_2]_2$ integrate to 1.

K-Mixture Distribution:

- ▶ $[y|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \mathbf{p}] = \sum_{k=1}^K p_k [y|\boldsymbol{\theta}_k]_k$
 - ▶ $p_k \geq 0$ for all k .
 - ▶ $\sum_{k=1}^K p_k = 1$.
 - ▶ all $[y|\boldsymbol{\theta}_k]_k$ integrate to 1.

Mixture models

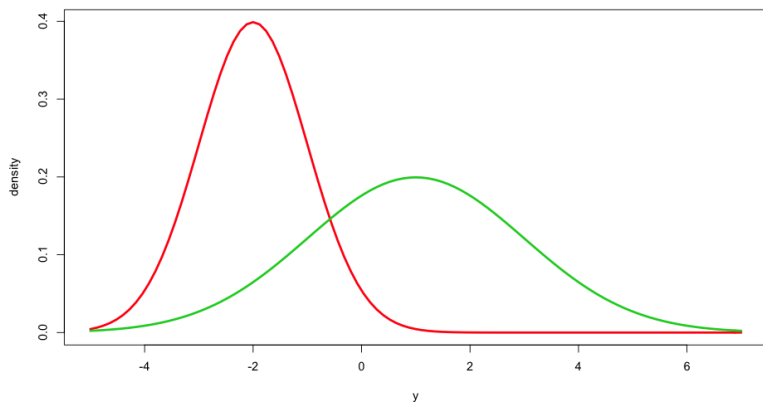
Mixture Distribution:

- ▶ $[y|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, p] = p[y|\boldsymbol{\theta}_1]_1 + (1-p)[y|\boldsymbol{\theta}_2]_2$
 - ▶ $0 \leq p \leq 1$.
 - ▶ $[y|\boldsymbol{\theta}_1]_1$ and $[y|\boldsymbol{\theta}_2]_2$ integrate to 1.

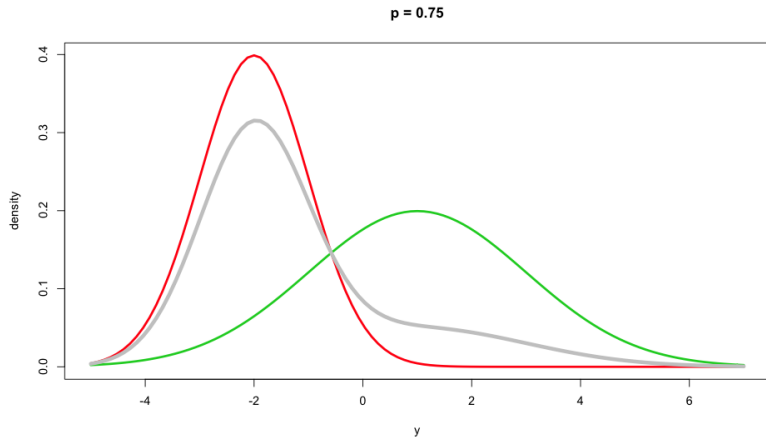
K-Mixture Distribution:

- ▶ $[y|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \mathbf{p}] = \sum_{k=1}^K p_k [y|\boldsymbol{\theta}_k]_k$
 - ▶ $p_k \geq 0$ for all k .
 - ▶ $\sum_{k=1}^K p_k = 1$.
 - ▶ all $[y|\boldsymbol{\theta}_k]_k$ integrate to 1.

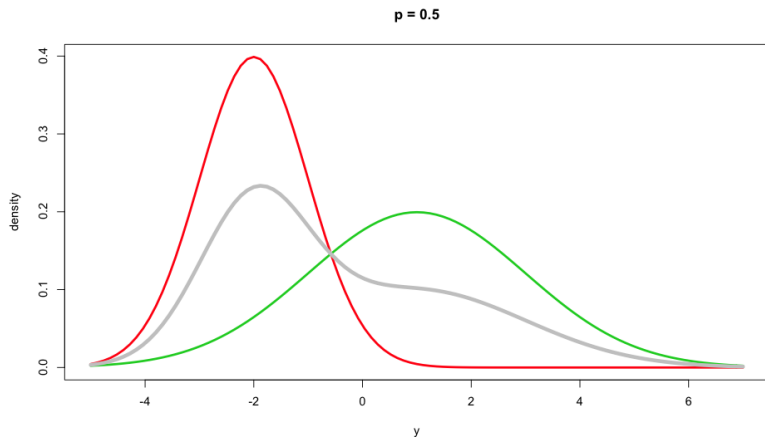
Mixture models



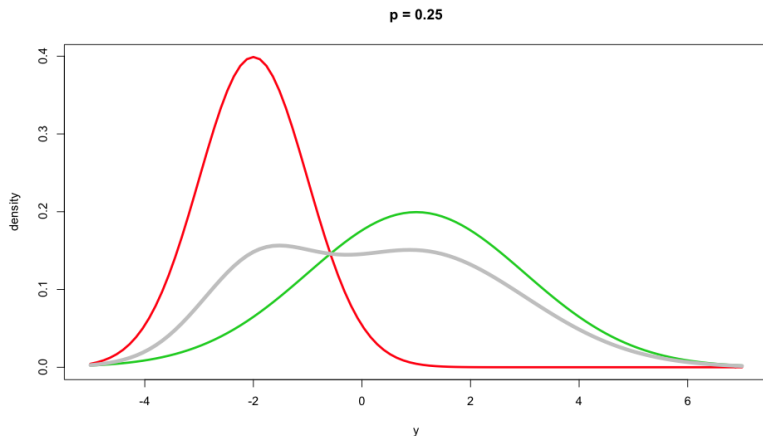
Introduction



Mixture models



Mixture models



Example: Darwin's Finches Model (Hendry et al. 2006):

$$y_i \sim p \cdot \text{normal}(\mu_1, \sigma^2) + (1 - p) \cdot \text{normal}(\mu_2, \sigma^2)$$

- ▶ $i = 1, \dots, n$
- ▶ $\mu_1 \neq \mu_2$
- ▶ $0 < p < 1$



Example: Darwin's Finches

Use latent (auxiliary) variables to make the mixture model hierarchical:

$$y_i \sim \begin{cases} \text{normal}(\mu_1, \sigma^2) & \text{if } z_i = 1 \\ \text{normal}(\mu_2, \sigma^2) & \text{if } z_i = 0 \end{cases}$$

where,

$$z_i \sim \text{Bernoulli}(p)$$

and,

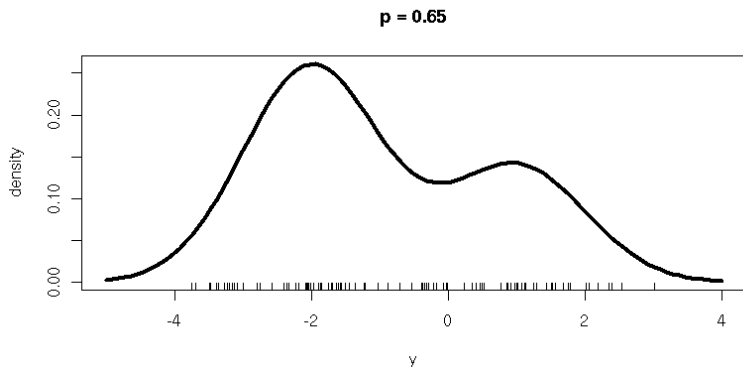
$$p \sim \text{uniform}(0, 1)$$

$$\mu_1 \sim \text{normal}(0, 100)$$

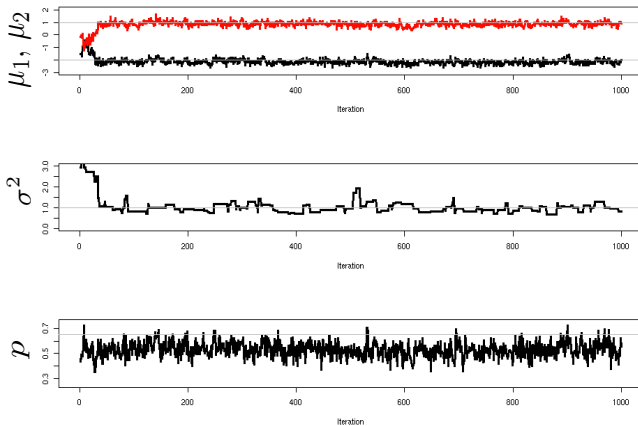
$$\mu_2 \sim \text{normal}(0, 100)$$

$$\sigma^2 \sim \text{inverse gamma}(.01, .01)$$

Data analysis finch model



MCMC output



Coding mixture models

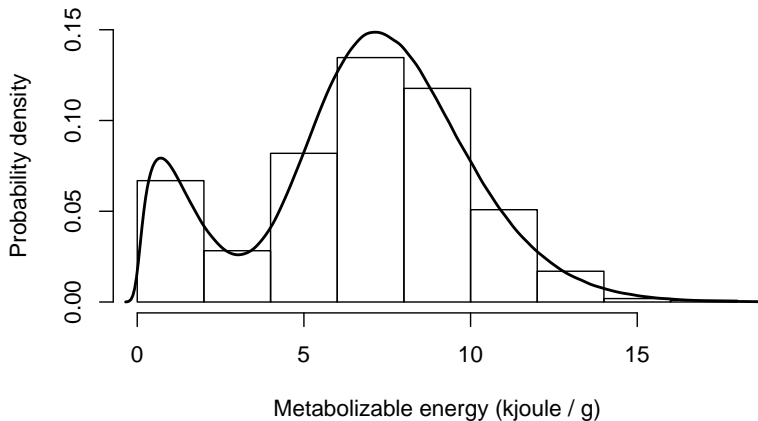
- Specify appropriate priors for mean and variance of two distributions to reflect their shapes and location. Use JAGS truncation function to assure that samples from one distribution are always greater than from the other.

```
mu[1] ~ dgamma(7/100,4/100) # vague priors specifying mean of 7/4
mu[2] ~ dgamma(49/500, 7/500) T(mu[1],) #mu[2] will always exceed mu[1]
```

- Specify appropriate initial values.
- Specify 0 to 1 prior on mixture parameter p .
- Write a likelihood that mixes the 2 distributions using Bernoulli draws from the distribution of p . Something like:

```
z[i] ~ dbern(p)
mu.mix[i] <- z[i] * mu[1] + (1-z[i]) * mu[2]
sigma.mix[i] <- z[i] * sigma[1] + (1-z[i]) * sigma[2]
y[i] ~ dnorm(mu.mix[i], sigma.mix[i]^-2)
```

Model fit



Martin et al. 2005

Ecology Letters, (2005) 8: 1235–1246

doi: 10.1111/j.1461-0248.2005.00826.x

**REVIEWS AND
SYNTHESES****Zero tolerance ecology: improving ecological inference by modelling the source of zero observations**

Tara G. Martin,^{1*} Brendan A. Wintle,² Jonathan R. Rhodes,³ Petra M. Kuhnert,⁴ Scott A. Field,⁵ Samantha J. Low-Choy,⁶ Andrew J. Tyre^{7†} and Hugh P. Possingham¹

Abstract

A common feature of ecological data sets is their tendency to contain many zero values. Statistical inference based on such data are likely to be inefficient or wrong unless careful thought is given to how these zeros arose and how best to model them. In this paper, we propose a framework for understanding how zero-inflated data sets originate and deciding how best to model them. We define and classify the different kinds of zeros that occur in ecological data and describe how they arise: either from 'true zero' or 'false zero' observations. After reviewing recent developments in modelling zero-inflated data sets, we use practical examples to demonstrate how failing to account for the source of zero inflation can reduce our ability to detect relationships in ecological data and at worst lead to incorrect inference. The adoption of methods that explicitly model the sources of zero observations will sharpen insights and improve the robustness of ecological analyses.

Keywords

Bayesian inference, detectability, excess zeros, false negative, mixture model, observation error, sampling error, zero-inflated binomial, zero-inflated Poisson, zero inflation.

Ecology Letters (2005) 8: 1235–1246

Zero-inflation as mixture model

Imagine that you sampled many plots along a coastline, counting the number of species of mussels within each plot. In essence there are two sources of zeros. Some zeros arise because the plot was placed areas that are not mussel habitat, while other zeros occur in plots placed in mussel habitat but that contain no mussels as a result of sampling variation. The Poisson distribution offers a logical choice for modeling the distribution of counts in mussel habitat, but it cannot portray the zeros that arise because plots were placed in areas where mussels never live.

Zero-inflation as mixture model

$$y_i \sim \begin{cases} 0 & w_i = 1 \\ \text{Poisson}(\lambda) & w_i = 0 \end{cases}$$

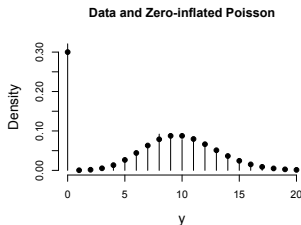
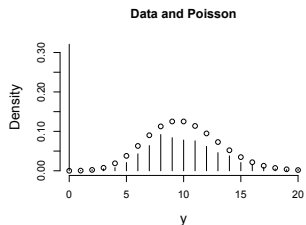
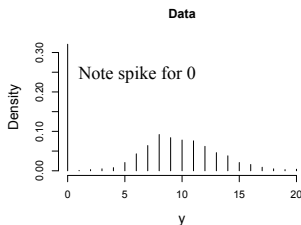
$$\begin{aligned} y_i &\sim \text{Poisson}(y_i | \lambda(1 - w_i)) \cdot \text{Bernoulli}(w_i | \phi) \text{beta}(\phi | 1, 1) \\ &\times \text{gamma}(\lambda | .01, .01), \end{aligned}$$

where λ is the average number of mussels per plot; ϕ is the probability that the plot is outside of mussel habitat, and w_i is an indicator variable for non-habitat.

Bayesian network (DAG)

Draw it.

Zero-inflation as mixture model



Zero-inflation example

Zero-inflated models

Poisson:

$$[\lambda, \alpha, \beta, \mathbf{w}, \phi | \mathbf{y}] \propto \prod_{i=1}^n \text{Poisson}(y_i | \lambda(1 - w_i) \text{Bernoulli}(w_i | \phi) \\ \times \text{beta}(\phi | \alpha, \beta) [\lambda] [\alpha] [\beta]$$

Binomial:

$$[p, \alpha, \beta, \mathbf{w}, \phi | \mathbf{y}] \propto \prod_{j=1}^J \text{binomial}(y_j | n_j, p(1 - w_j) \text{Bernoulli}(w_j | \phi) \\ \times \text{beta}(\phi | \alpha, \beta) [p] [\alpha] [\beta]$$

In both cases ϕ is the probability of a zero that is not accounted for by sampling variation alone. Negative binomial and multinomial random variables can also be modeled this way.

What are state space models?

$$\begin{aligned} &[y_i | \boldsymbol{\theta}_d, z_i] \\ &[z_i | \boldsymbol{\theta}_p, z_i] \end{aligned}$$

The idea is simple. We have a stochastic model of an unobserved, true state z_i and a stochastic model that relates our observations (y_i) to the true state.

$$\begin{aligned} &[\mathbf{z}, \boldsymbol{\theta}_{process}, \boldsymbol{\theta}_{data} | \mathbf{y}] \propto \\ &\prod_{i=1}^n [y_i | \boldsymbol{\theta}_{data}, z_i] [z_i | \boldsymbol{\theta}_{process}, z_i] [\boldsymbol{\theta}_{process}, \boldsymbol{\theta}_{data}] \end{aligned}$$

Occupancy models as an application of zero inflation

	Visit, t		
Site, i	1	2	3
1	1	1	0
2	0	0	0
3	0	0	1
.	0	0	0
.	1	0	0
n	0	0	0

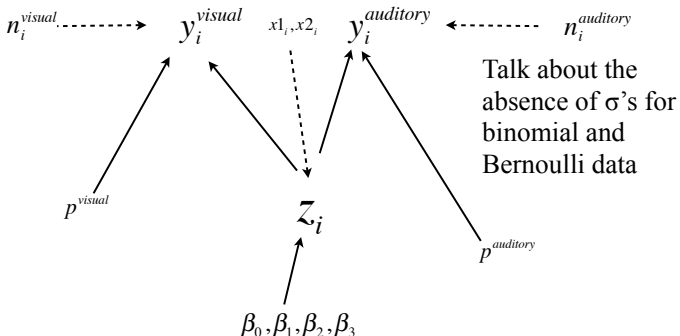
Occupancy models as an application of zero inflation

- ▶ We want to understand controls on presence or absence of individuals within discrete categories.
- ▶ We have a problem interpreting zeros.
 - ▶ The individual is truly absent.
 - ▶ The individual is present but undetected.
- ▶ We want to use a model to explain spatial or temporal variation in presence or absence.
- ▶ We need to model uncertainty arising from two sources:
 - ▶ The failure of the model to portray the process
 - ▶ The error in our observations arising because we fail to perfectly observe presence or absence

Exercise: Courtesy of McCarthy 2007: Box 5.9

- ▶ Kristen Parris studied controls on the distribution of tree frogs in the riparian zone of streams on the east coast of sub-tropical Australia.
- ▶ Multiple surveys were conducted at 64 sites using 2 observation methods, visual searches at night and auditory searches for responses to taped calls.
- ▶ We assume the presence /absence of frogs at a site does not change during the multiple surveys. (Frogs don't fly.) This is known as the *closure* assumption
- ▶ Presence / absence of frogs was modeled as a function of 1) stream size (measured as annual volume of rainfall in the catchment above the site) 2) the presence or absence of palms at the site (an indicator of mesic or xeric conditions) and 3) the interaction between palms and stream size.
- ▶ Data are the total number of times frogs were detected using each method for each site and the number of surveys for each site.

Bayesian network (DAG)



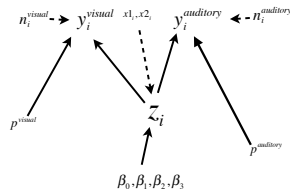
Posterior and joint distributions:

$$[z, \beta, p^{auditory}, p^{visual} | \mathbf{y}^{auditory}, \mathbf{y}^{visual}] \propto$$

$$\prod_{i=1}^{64} \text{binomial}(y_i^{auditory} | n_i^{auditory}, z_i \cdot p^{auditory}) \text{binomial}(y_i^{visual} | n_i^{visual}, z_i \cdot p^{visual}) \times$$

$$\text{Bernoulli}(z_i | \text{invlogit}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)) \times$$

$$\prod_{j=0}^3 \text{normal}(\beta_j | 0, .0001) \text{uniform}(p^{visual} | 0, 1) \text{uniform}(p^{auditory} | 0, 1)$$

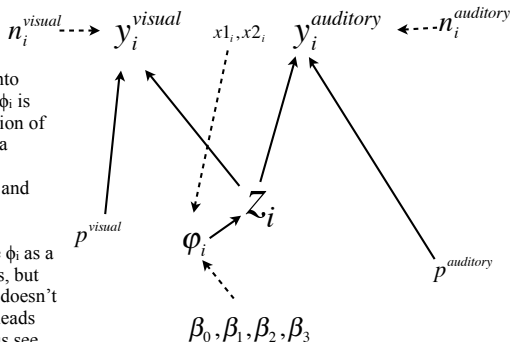


A common error:

Important

Here is where you can run into trouble. (I did.) Notice that ϕ_i is simply a calculation, a function of the β 's and the x 's. It is not a stochastic quantity.

The relationship between z_i and the β 's and the x 's form the stochastic quantity. You are certainly free to estimate the ϕ_i as a function of random variables, but putting them in the diagram doesn't make sense in terms of the heads and tails of arrows helping us see the conditioning. If you want them in the diagram, use a different type of arrow indicating a calculated quantity, as I did here.



Code

```

model
{
  b0 ~ dnorm(0, 1.0E-6) # uninformative priors for the variables
  b[1] ~ dnorm(0, 1.0E-6)
  b[2] ~ dnorm(0, 1.0E-6)
  b[3] ~ dnorm(0, 1.0E-6)
  p.visual ~ dunif(0, 1) # detection probabilities when the species is present
  p.auditory ~ dunif(0, 1)
  mLnCV <- mean(LnCV[]) # average catchment volume
  for (i in 1:64) # for each of the 64 sites
  {
    phi[i] <- ilogit(b0 + b[1]*(LnCV[i] - mLnCV) + b[2]*palms[i] + b[3]*(LnCV[i] -
mLnCV)*palms[i]) # probability of presence using centered data
    z[i] ~ dbern(phi[i]) # actual, latent presence at the site, 0 or 1
    y.visual[i] ~ dbin(p.visual*z[i], n.visual[i]) # number eye detections with
    y.auditory[i] ~ dbin(p.auditory*z[i], n.auditory[i]) # number of ear detec-
tions
  }
  # predicted relationships--derived quantities; note that predic-
tions are done using centered data. There is no back-transform.
  for (i in 1:20)
  {
    LVol[i] <- 2 + 3*i/20 # covers the range of stream sizes
    logit(predpalms[i]) <- b0 + (b[1] + b[3])*(LVol[i] - mLnCV) + b[2]
    logit(prednopalmes[i]) <- b0 + b[1]*(LVol[i] - mLnCV)
  }
}

```

Implementation tricks

Refer to code above

- ▶ Initialize all latent quantities i.e., all $z[i] = 1$. The $\phi[i]$ don't need to be initialized because they are derived quantities.
- ▶ You may need to bound $\phi[i]$, i.e.:

```
phi[i] <- max(min(ilogit(a + b[1]*(LnCV[i] -
mLnCV) + b[2]*palms[i] + b[3]*(LnCV[i] -
mLnCV))*palms[i], .0000001), .999999)
```

Sensitivity - specificity

- ▶ Sensitivity: The probability that a truly infected individual gives a positive test result ϕ
- ▶ Specificity: The probability that an uninfected individual gives a negative test result. $(1 - \rho)$
- ▶ Test result $y_i = 1$ if positive for disease, 0 otherwise.

Sensitivity - specificity

$$y_i \sim \begin{cases} \text{Bernoulli}(\pi) & z_i = 1 \\ \text{Bernoulli}(\rho) & z_i = 0 \end{cases}$$

$$y_i \sim \text{Bernoulli}(z_i \phi) + \text{Bernoulli}((1 - z_i)\rho)$$

$$z_i \sim \text{Bernoulli}(\beta_0 + \mathbf{x}_i \boldsymbol{\beta})$$

$$\begin{aligned} [\boldsymbol{\beta}, \mathbf{z}, \phi, \rho \mid \mathbf{y}] &\propto \prod_{i=1}^n [y_i \mid z_i \phi + (1 - z_i)\rho] \\ &\quad \times [z_i \mid \beta_0 + \mathbf{x}_i \boldsymbol{\beta}] \\ &\quad \times \text{priors} \end{aligned}$$

Will require informed priors on ϕ and ρ . They are not identifiable from test results alone.

Capture-mark-recapture

	Capture occasion, t										
Individual, i	1	2	3	T
1	1	1	0	1	0	0	0	0	0	0	0
2	1	0	1	1	1	0	0	0	0	0	0
3	0	0	1	1	1	0	0	0	0	0	0
.											
.											
M	0	0	0	0	0	0	0	1	1	1	0

- ▶ True state $z_i = 1$ if animal alive, 0 otherwise.
- ▶ Leftmost 1 is time of first capture.
- ▶ Population must be closed (no emigration) during sampling occasions.

Capture-mark-recapture

$$\begin{aligned}
 y_{i,t} &\sim \text{Bernoulli}(z_{i,t}p_t) \\
 z_{i,t} &\sim \text{Bernoulli}(z_{i,t-1}\phi_t) \\
 [\mathbf{z}, \boldsymbol{\phi}, \mathbf{p} | \mathbf{y}] &\propto \prod_{i=1}^M \prod_{t=1}^T [y_{i,t} | z_{i,t}p_t][z_{i,t} | z_{i,t-1}\phi_t] \\
 &\times \text{priors}
 \end{aligned}$$

Can substitute inverse logit models of covariates for ϕ_t and / or p_t .

Data augmentation

Augment the observed data with large number $n_z \gg N$ of all zero capture histories.

	Capture occasion, t										
Individual, i	1	2	3	T
1	1	1	0	1	0	0	0	0	0	0	0
2	1	0	1	1	1	0	0	0	0	0	0
3	0	0	1	1	1	0	0	0	0	0	0
.											
.											
M	0	0	0	0	0	0	0	1	1	1	0
.	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0
$M + n_z$	0	0	0	0	0	0	0	0	0	0	0

Jolly-Seber model with data augmentation

$$y_{i,t} \sim \text{Bernoulli}(z_t p_t)$$

$$z_{i,t} \sim \text{Bernoulli}(z_{i,t-1} \phi_t + (1 - z_{i,t-1}) \gamma_t)$$

$$z_{1,t} \sim \text{Bernoulli}(\gamma_1)$$

$$\begin{aligned}
 [\phi, \boldsymbol{\gamma}, \mathbf{p}, \mathbf{Z} \mid \mathbf{Y}] &\propto \underbrace{\prod_{i=1}^{M+n_z}}_{\text{initial conditions}} \underbrace{\prod_{t=2}^T}_{\text{observation}} \underbrace{[y_{i,t} | z_{i,t} p]_t}_{\text{survival}} \underbrace{[z_{i,t} | z_{i,t-1} \phi]_t}_{\text{addition}} + \prod_{k=1}^t (1 - z_{i,k}) \gamma_t \\
 &\times \underbrace{[z_{1,t} | \gamma_1]}_{\text{initial conditions}} [\phi][\gamma_t][p_t].
 \end{aligned}$$

- ▶ n_z = number of augmented data records
- ▶ γ_i = probability of inclusion. Does not have biological interpretation.
- ▶ Approximate marginal posterior distribution of population size N as a derived quantity $N = \sum_{i=1}^{M+n_z} z_i$.
- ▶ Does not require closure assumption.

Data augmentation

- ▶ Imagine for simplicity that we sample and mark a population and observe it on two subsequent occasions. There are four possible capture histories for each individual: $(1,1,1)$; $(1,0,1)$, $(1,1,0)$, and $(0,0,0)$
- ▶ $(0,0,0)$ *represents an individual that is a member of the population but that was never marked or observed*. Estimating the total population size requires us to find the number of these *all zero* capture histories. The sum of the “not observed” capture histories and M is the true population size.
- ▶ Data augmentation works because the model of the augmented dataset is a zero-inflated version of either a binomial or a multinomial model. There is enough information in the observed data to estimate the probability of the all zero case.

For details on data augmentation see

1. J. A. Royle. Analysis of capture-recapture models with individual covariates using data augmentation. *Biometrics*, 65(1):267–274, 2009.
2. J. A. Royle and R. M. Dorazio. Parameter-expanded data augmentation for bayesian analysis of capture- recapture models. 152(Supplement 2):521–537, 2012.
3. J. A. Royle, R. M. Dorazio, and W. A. Link. Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics*, 16(1):67–85, 2007.

Multi-state models

We notate the true state of the individual at time t using a three element vector $\mathbf{z}_{i,t} = (1, 0, 0)'$ if it is alive at time t ; $\mathbf{z}_{i,t} = (0, 1, 0)'$ if it is dead at time t , and $\mathbf{z}_{i,t} = (0, 0, 1)'$ if it is not in the study. Transitions among states are governed by the matrix \mathbf{S} :

$$\mathbf{S} = \begin{pmatrix} \phi & 0 & \gamma_t \\ 1 - \phi & 1 & 0 \\ 0 & 0 & 1 - \gamma_t \end{pmatrix}$$

Multi-state models

We define an observation matrix as

$$\mathbf{D} = \begin{pmatrix} p & 0 & 0 \\ 0 & q & 0 \\ 1-p & 1-q & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (1)$$

where p is the probability of observing animal i at time t conditional on it being alive and q is the probability of observing animal i at time t conditional on it being dead.

Multi-state models

For a three state model:

$$\mathbf{y}_{i,t} \mid \mathbf{z}_{i,t}, \mathbf{D} \sim \text{multinomial}(1, \mathbf{D}\mathbf{z}_{i,t}) \quad (2)$$

$$\mathbf{z}_{i,t} \mid \mathbf{z}_{i,t-1}, \mathbf{S} \sim \text{multinomial}(1, \mathbf{S}\mathbf{z}_{i,t-1}) \quad (3)$$

$$\mathbf{z}_{i,1} \mid \gamma_1 \sim \text{multinomial}(1, (\gamma_1, 0, 1 - \gamma_1)') \quad (4)$$

$$\begin{aligned} [\phi, \gamma, p, q, \mathbf{Z} \mid \mathbf{Y}] &\propto \prod_{i=1}^M \prod_{t=2}^T [\mathbf{y}_{i,t} \mid \mathbf{D}\mathbf{z}_{i,t}] [\mathbf{z}_{i,t} \mid \mathbf{S}\mathbf{z}_{i,t-1}] [\mathbf{z}_{i,1} \mid \gamma_1] \quad (5) \\ &\times [\phi] [\gamma_t] [p] [q]. \quad (6) \end{aligned}$$

Further study

1. J. A. Royle and R. M. Dorazio. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations, and communities. Academic Press, London, UK, 2008. NA A hierarchical perspective. Academic Press, Waltham, MA, USA, 2012.
2. R. King. A review of Bayesian state-space modelling of capture-recapture-recovery data. Interface Focus, 2(2):190–204, 2012.
3. M. A. McCarthy. Bayesian methods for ecology. Cambridge University Press, Cambridge, UK, 2007.