Mixture Models

Bayesian Models for Ecologists

DEB 2042028

Overview

1. Mixture distributions
   - Example: Darwin's Finches
2. Zero-inflated models
   - Martin et al. (2005)
   - Zero-inflated Poisson Regression
3. Occupancy Models

Introduction

Two-component Mixture Distribution:

$$[\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, p] = p[\mathbf{y}|\boldsymbol{\theta}_1]_1 + (1-p)[\mathbf{y}|\boldsymbol{\theta}_2]_2$$

- $0 \leq p \leq 1$
- $[\mathbf{y}|\boldsymbol{\theta}_1]_1$ and $[\mathbf{y}|\boldsymbol{\theta}_2]_2$ integrate to 1

Introduction

Two-component Mixture Distribution:

$$[\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, p] = p[\mathbf{y}|\boldsymbol{\theta}_1]_1 + (1-p)[\mathbf{y}|\boldsymbol{\theta}_2]_2$$

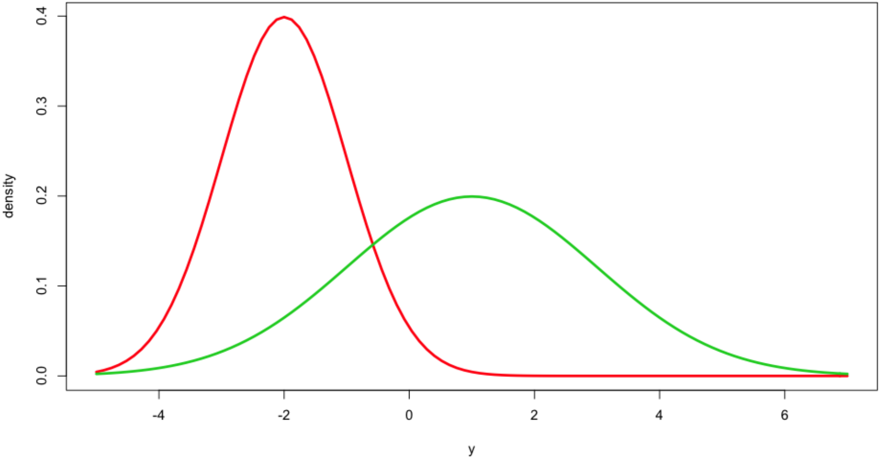
- $0 \leq p \leq 1$
- $[\mathbf{y}|\boldsymbol{\theta}_1]_1$ and $[\mathbf{y}|\boldsymbol{\theta}_2]_2$ integrate to 1

K-Mixture Distribution:

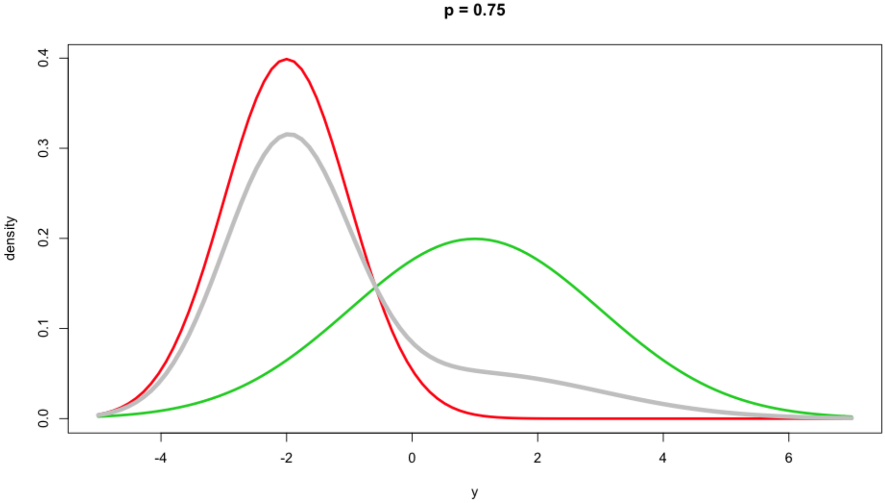
$$[\mathbf{y}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \mathbf{p}] = \sum_{k=1}^{K} p_k [\mathbf{y}|\boldsymbol{\theta}_k]_k$$

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

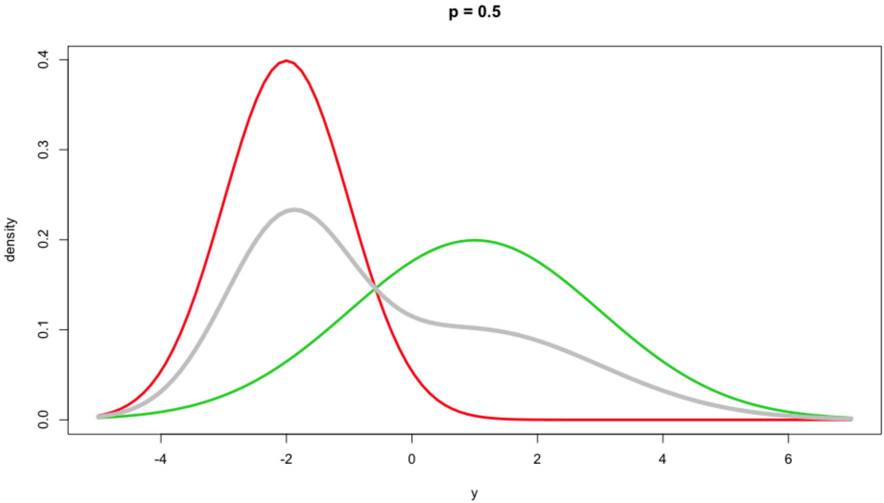
Introduction



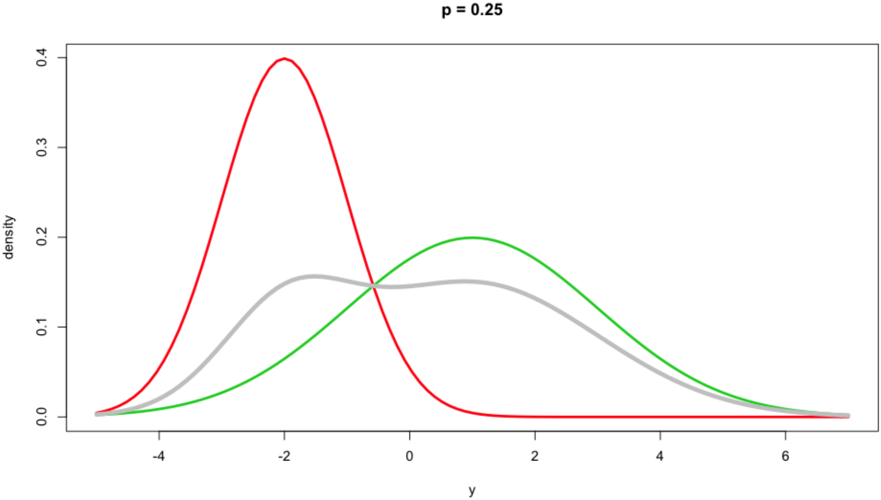
Introduction



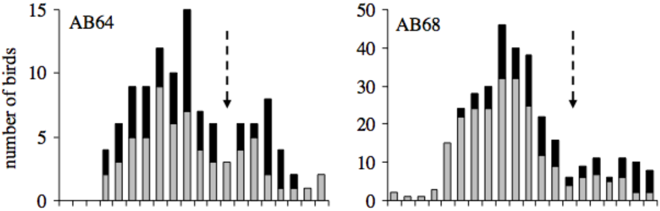
Introduction



Introduction



Example: Darwin's Finches



Example: Darwin's Finches

Model (Hendry et al. 2006):

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

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

Example: Darwin's Finches

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

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

where

$$z_i \sim \mathrm{Bern}(p)$$

$$\begin{aligned} p &\sim ??? \\ \mu_1 &\sim ??? \\ \mu_2 &\sim ??? \\ \sigma^2 &\sim ??? \end{aligned}$$

Example: Darwin's Finches

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

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

where

$$z_i \sim \mathrm{Bern}(p)$$

$$\begin{aligned} p &\sim \mathrm{Beta}(\alpha, \beta) \\ \mu_1 &\sim \mathrm{N}(\mu_0, \sigma_0^2) \\ \mu_2 &\sim \mathrm{N}(\mu_0, \sigma_0^2) \\ \sigma^2 &\sim \mathrm{IG}(r, q) \end{aligned}$$

Mixture Model DAG

Implementation in JAGS

$$[\mu_1, \mu_2, \sigma^2, \mathbf{z}, p|\mathbf{y}] \propto \left(\prod_{i=1}^{n} [y_i|\mu_1, \sigma^2]^{z_i} [y_i|\mu_2, \sigma^2]^{1-z_i} [z_i|p] \right) [p][\mu_1][\mu_2][\sigma^2]$$

```
 model{
    mu1 ~ dnorm(mu10,tau10)
    mu2 ~ dnorm(mu20,tau20)
    tau ~ dgamma(.01,.01)
    p ~ dbeta(1,1)

    for(i in 1:n){
      y[i] ~ dnorm(z[i]*mu1+(1-z[i])*mu2,tau)
    }
    for(i in 1:n){
      z[i] ~ dbin(p,1)
    }
}
```

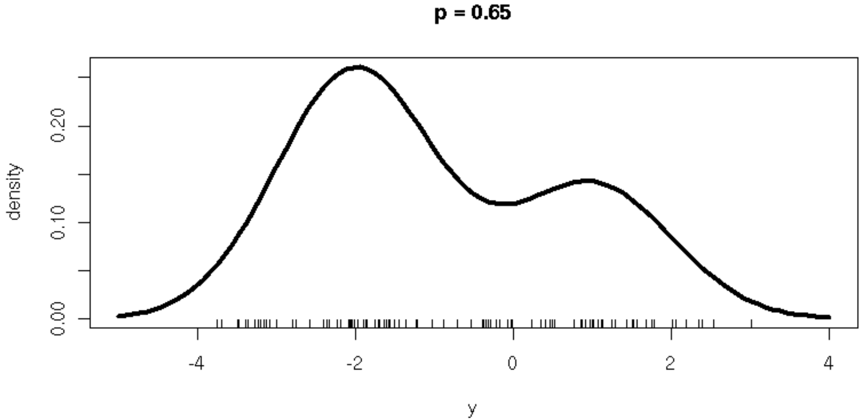
Implementation in JAGS

```
for(i in 1:n){
   y[i] ~ dnorm(z[i]*mu1+(1-z[i])*mu2,tau)
}
```

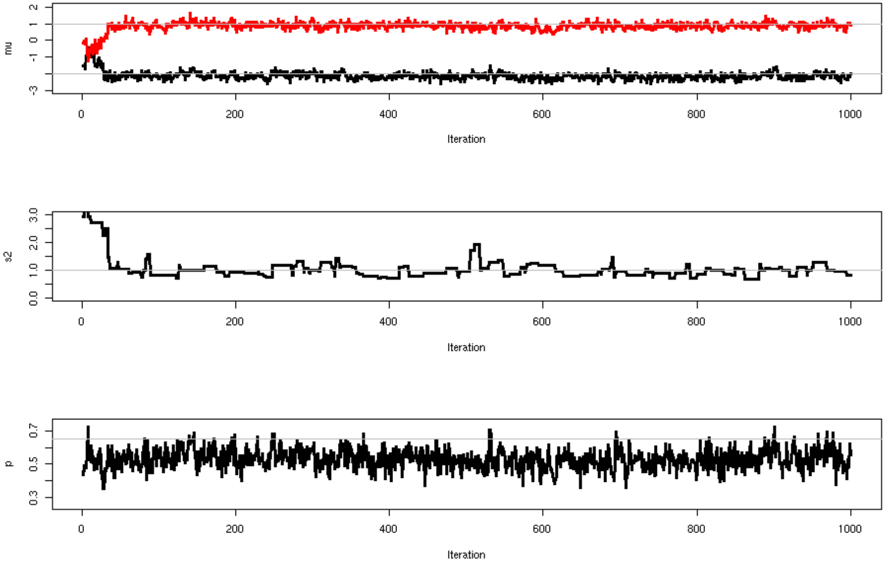
Equivalently, compute mixture parameters in a separate line. . .

```
for(i in 1:n){
   mu.mix[i] <- z[i]*mu1+(1-z[i])*mu2
   y[i] ~ dnorm(mu.mix[i],tau)
}
```

Data Analysis: Darwin's Finches



Data Analysis: Darwin’s Finches



Implementation in JAGS

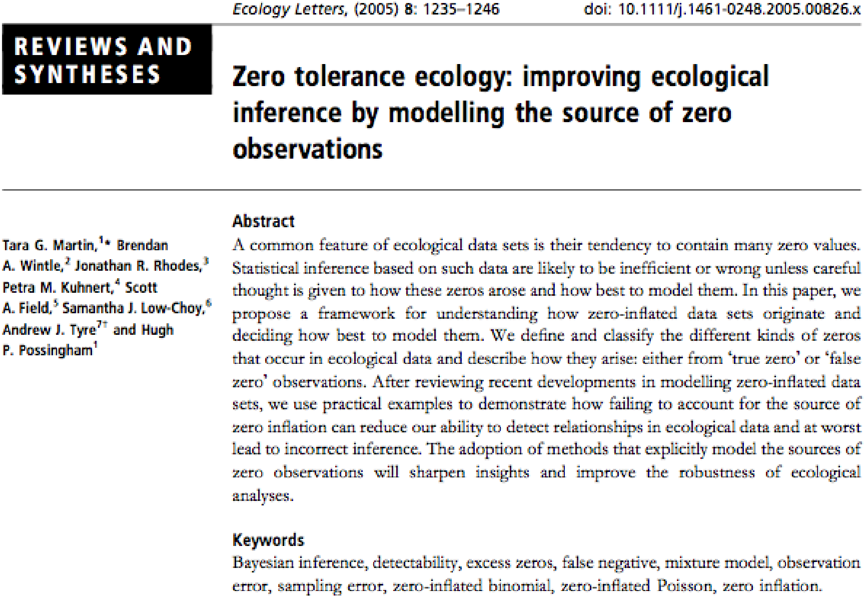
Making inference on the individual mixture components can be complicated by the fact that the labeling of “component 1” and “component 2” is arbitrary. One way to resolve this is to label them based on their moments, e.g. mean value.

This can be accomplished by requiring $mu_2 > \mu_1$. This is enforced in JAGS using the truncation function:

```
 model{
    mu1 ~ dnorm(mu10,tau10)
    mu2 ~ dnorm(mu20,tau20) T(mu1,). # mu2 will be greater than mu1
    tau ~ dgamma(.01,.01)
    p ~ dbeta(1,1)

    ...
}
```

Zero-inflated mdoels: 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. Possingham[1]

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.

Zero-inflation as a 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-inflated models: Martin et al. (2005)

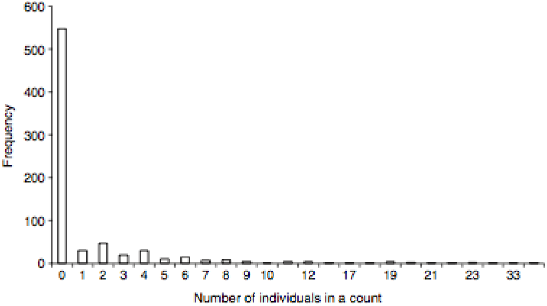


Figure 1 Example of a typical zero-inflated data set. Frequency of counts for 31 bird species across eight sites and three grazing treatments ($n = 744$) from Martin *et al.* (2005). Over 70% of the data set is represented by zero counts, which is more than expected if a Poisson distribution is assumed for the species' abundances.

Poisson regression model

Suppose we start by modeling the data with a Poisson regression model.

$$y_i \sim \text{Poison}(\lambda_i)$$

$$\log(\lambda_i) = \mathbf{x}_i' \boldsymbol{\beta}$$

$$\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Poisson regression model - DAG

Zero-inflated Poisson regression model

$$y_i \sim \begin{cases} 0 & , \; z_i = 0 \\ \text{Poison}(\lambda_i) & , \; z_i = 1 \end{cases}$$

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

$$\log(\lambda_i) = \mathbf{x}_i' \beta$$

$$\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p \sim \text{Beta}(\alpha_1, \alpha_2)$$

Zero-inflated Poisson regression model - DAG

Implementation in JAGS

```
model{
    beta ~ dmnorm(mu0,Tau0)
    p ~ dbeta(1,1)

    for(i in 1:n){
      y[i] ~ dpois(z[i]*lam[i])
    }
    for(i in 1:n){
      z[i] ~ dbin(p,1)
      log(lam[i]) <- X[i,1:pp]%*%beta
    }
}
```

Other applications of mixture models

1. Occupancy models
2. Capture-recapture models
3. Serology models with imperfect tests (sensitivity/specificity)