# GAM Zero-Inflated N-Mixture Model

Yunyi SHEN

January 2, 2019

## 1 Model Setting

### 1.1 Combined Distribution of Latent Population Size and Imperfect Detections

Distribution of counting data follows N-Mixture Models in every sample period number of individuals encountered follows binomial distribution given population size at site i. Assume population size follows Poisson distribution. Data's distribution is given by total probability theorem:

$$\begin{aligned} P(\vec{d}) &= \sum_n P(d|n,\theta)P(n|\theta) \\ &= \sum_n Binom(d|n,p)Pois(n|\lambda) \end{aligned} \tag{1}$$

### 1.2 Zero Inflation

Zero inflate is common in ecology data. For instance sites not occupied can cause zero inflation. One common way to deal with this is to use zero inflated distribution to model data. Here we generally follow occupancy modeling, assume for probability $psi$ a site were occupied and $1-\psi$ it is empty and cause zero inflation. Thus the full distribution of data $\vec{d}$ is given by:

$$\begin{array}{cc} \sum_n Binom(d|n,p)Pois(n|\lambda) & \psi \\ 0 & 1-\psi \end{array} \tag{2}$$

### 1.3 Proof the Distribution of Latent n and Detections d is in Exponential Family

*Proof.* First, set $(n,\vec{d})$ is the latent population and detection vectors at site i. I intend to prove that the distribution of this vector is with in exponential family witch has form:

$$f(y|\theta) = b(y)exp(\eta^T T(y) - a(\eta)) \tag{3}$$

Assume detections given latent population n were binomial distributed with parameter $p_j$, thus:

$$\begin{aligned} P(\vec{d}|n) &= \prod_{j=1}^{w} \binom{n}{d_j} p_j^{d_j}(1-p_j)^{n-d_j} \\ &= \prod_{j=1}^{w} \binom{n}{d_j} exp(\sum_{j=1}^{w} d_j log\frac{p_j}{1-p_j} + n\sum_{j=1}^{w} log(1-p_j)) \end{aligned} \tag{4}$$

which showed that binomial distribution given n belongs to exponential family.

Then assume latent population was Poisson distributed with rate $\lambda$

$$\begin{aligned} P(n|\theta) &= e^{-\lambda}\frac{\lambda^n}{n!} \\ &= \frac{1}{n!}exp(nlog(\lambda)-\lambda) \end{aligned} \tag{5}$$

Thus the total probability contains the latent and detections is given by:

$$\begin{aligned} P(n,\vec{d}|\theta) &= P(d|n,\theta)P(n|\theta) \\ &= \frac{1}{n!}\prod_{j=1}^{w}\binom{n}{d_j}exp(\sum_{j=1}^{w}d_jlog\frac{p_j}{1-p_j}+n\sum_{j=1}^{w}log(1-p_j))exp(nlog(\lambda)-\lambda) \\ &= \frac{1}{n!}\prod_{j=1}^{w}\binom{n}{d_j}exp[\sum_{j=1}^{w}d_jlog\frac{p_j}{1-p_j}+n(\sum_{j=1}^{w}log(1-p_j)+log(\lambda))-\lambda] \\ &= \frac{1}{n!}\prod_{j=1}^{w}\binom{n}{d_j}exp(\eta^T(n,\vec{d})-\lambda) \end{aligned} \tag{6}$$

in which

$$\eta^T = (\sum_{j=1}^{w}log(1-p_j)+log(\lambda), log\frac{\vec{p}}{1-\vec{p}})$$

$$T(y) = y$$

$\lambda$ can be calculated using $\eta$ since it contains all $p_j$ and $\lambda$ itself. Later on, we note this function as

$$f(n,d|\theta)$$

□

# 2 Model Estimation

## 2.1 EM Algorithm to Deal with the Missing Population Size and Occupancy

Since we are missing observation of Latent Population Size $n$ and Occupancy status $z$, we can not directly obtain the MLE for GAMs. Thus here we proposed an EM algorithm to deal with missing observations.

Instead of maximize log likelihood directly EM algorithm maximize the lower bound of the log likelihood every iteration. In each iteration, this algorithm contains two steps, Expectation(E) step and Maximization(M) step. In the E step, the algorithm will take expected value of log likelihood under the posterior distribution of missing observations, and in M step, maximize this expected value. The algorithm will iterate until estimation converges.

### 2.1.1 E-step

We first write down the total probability assuming knowing occupancy status $z$ and latent population size $n$:

$$\begin{aligned} P(\vec{d}, n, z) &= [\psi Pois(n|\lambda) \prod_{j=1}^{w} Bin(d_j|n, p_j)]^z (I_{\vec{d}=0}(1-\psi))^{1-z} \\ &= (\psi f(n, \vec{d}|\theta))^z (I_{\vec{d}=0}(1-\psi))^{1-z} \end{aligned} \tag{7}$$

In E step of EM algorithm, we need $P(n, z|\vec{d})$, which need to sum all n up, here, we truncated it with some large N.

$$\begin{aligned} P(n, z|\vec{d}) &= \frac{P(n, z, \vec{d})}{\sum_{n=max(d_j)}^{zN} \sum_{z=0}^{1} P(n, z, \vec{d})} \\ &= \frac{(\psi f(n, \vec{d}|\theta))^z (I_{\vec{d}=0}(1-\psi))^{1-z}}{\psi \sum_{n=max(d_j)}^{N} f(n, d|\theta) + (1-\psi) I_{\vec{d}=0}} \end{aligned} \tag{8}$$

Take expectation of logL of total likelihood under $r^{th}$ $\theta$ given by summing every z and n up:

$$\mathbb{E}(l_p|\vec{d}, \theta^{[r]}) = \sum_{sites} \frac{\sum_n^N \psi^{[r]} f(n, \vec{d}|\theta^{[r]}) log[\psi f(n, \vec{d}|\theta)] + (1-\psi^{[r]}) I_{\vec{d}=0} log[(1-\psi) I_{\vec{d}=0}]}{\sum_n^N \psi^{[r]} f(n, \vec{d}|\theta^{[r]}) + (1-\psi^{[r]}) I_{\vec{d}=0}} \tag{9}$$

This involves 3 of GAMs to be maximized during M step.

Later on, we note the a normalizing constant given $\theta^{[r]}$ to be:

$$\sum_n^N \psi^{[r]} f(n, \vec{d}|\theta^{[r]}) + (1-\psi^{[r]}) I_{\vec{d}=0} = Z^{[r]}$$

and another constant:

$$\sum_n^N f(n, \vec{d}|\theta^{[r]}) = g^{[r]}$$

and note:

$$f(n, \vec{d}|\theta^{[r]}) = f_{\vec{d}}^{[r]}(n)$$

### 2.1.2 M step, RILS Algorithm

# 3 Model Estimation

We get derivative of $E_l$ and $\psi_i$ at site i

$$\frac{\partial E_l}{\partial \psi_i} = \frac{\psi_i^{[r]} g_i^{[r]}}{\psi_i^{[r]} g_i^{[r]} + (1-\psi_i^{[r]}) I_{\vec{d}_i=0}} - \psi_i \tag{10}$$

Involves the first weighted GAM regarding occupancy rate $\psi$

Then get derivative of $f_{\vec{d}}(n)$ at site i

$$\frac{\partial E_l}{\partial f_{\vec{d}_i}(n)} = \frac{\psi_i^{[r]}}{Z^{[r]}} \sum_{n=0}^{N} \frac{f_{\vec{d}_i}^{[r]}(n)}{f_{\vec{d}_i}(n)} \tag{11}$$

Now we can get the total derivative of the expected value of logL:

$$\begin{aligned}
\frac{\partial E_l}{\partial \beta_k} &= \sum_i \frac{\partial E_l}{\partial f_{\vec{d_i}}(n)} \frac{\partial f_{\vec{d_i}}(n)}{\partial \beta_k} + \sum_i \frac{\partial E_l}{\partial \psi_i}\frac{\partial \psi_i}{\partial \beta_k} \\
&= \sum_i \frac{\psi_i^{[r]}}{Z^{[r]}} \sum_{n=0}^{N} f_{\vec{d_i}}^{[r]}(n) \frac{\partial log(f_{\vec{d_i}}(n))}{\partial \beta_k} + \sum_i \frac{\partial E_l}{\partial \psi_i}\frac{\partial \psi_i}{\partial \beta_k} \\
&= \sum_{n=0}^{N} \sum_i \frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d_i}}^{[r]}(n) \frac{\partial log(f_{\vec{d_i}}(n))}{\partial \beta_k} + \sum_i \frac{\partial E_l}{\partial \psi_i}\frac{\partial \psi_i}{\partial \beta_k}
\end{aligned} \tag{12}$$

Further, we derive the setting of RILS for n and p:

$$\begin{aligned}
&\sum_{n=0}^{N} \sum_i \frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d_i}}^{[r]}(n) \frac{\partial log(f_{\vec{d_i}}(n))}{\partial \beta_k} \\
=&\sum_{n=0}^{N} \sum_i [\frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d_i}}^{[r]}(n) \frac{\partial logPois(n,\lambda_i)}{\partial \beta_k} + \sum_{j=1}^{w} \frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d_i}}^{[r]}(n) \frac{\partial logBin(p_{ij},n)}{\partial \beta_k}]
\end{aligned} \tag{13}$$

Note that:

$$w_i^r(n) = \frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d_i}}^{[r]}(n)$$

First deal with $\lambda$ which is easier:

$$\begin{aligned}
\sum_{n=0}^{N} \sum_i w_i^{[r]}(n) \frac{\partial logPois(n,\lambda_i)}{\partial \beta_k} &= \sum_i \frac{\sum_{n=0}^{N} w_i^{[r]}(n)(n-\mu_{\lambda i})}{\phi V(\mu_{\lambda i})} \frac{\partial \mu_{\lambda i}}{\partial \beta_k} \\
&= \sum_i [\sum_{n=0}^{N} w_i^{[r]}(n)] (\frac{\sum_{n=0}^{N} w_i^{[r]}(n) n}{\sum_{n=0}^{N} w_i^{[r]}(n)} - \mu_{\lambda i}) \frac{1}{\phi V(\mu_{\lambda i})} \frac{\partial \mu_{\lambda i}}{\partial \beta_k}
\end{aligned} \tag{14}$$

Which is a single quasi-Poisson GAM with weight $\sum_{n=0}^{N} w_i^{[r]}(n)$ and pseudo data as weighted average of every possible n.

Then deal with single $p_{ij}$:

$$\begin{aligned}
\frac{\partial logBin(p_{ij},n)}{\partial \beta_k} &= \frac{d_{ij} - np_{ij}}{np_{ij}q_{ij}} \frac{\partial (np_{ij})}{\partial \beta_k} \\
&= \frac{d_{ij} - np_{ij}}{p_{ij}q_{ij}} \frac{\partial (p_{ij})}{\partial \beta_k} \\
&= \frac{d_{ij} - n\mu_{p_{ij}}}{V(\mu_{p_{ij}})} \frac{\partial (\mu_{p_{ij}})}{\partial \beta_k}
\end{aligned} \tag{15}$$

Now it is related to a quasi-Binomial regression.

$$\begin{aligned}
\sum_n w_i^{[r]}(n) \frac{\partial logBin(p_{ij},n)}{\partial \beta_k} &= \sum_n w_i^{[r]}(n) \frac{d_{ij} - n\mu_{p_{ij}}}{V(\mu_{p_{ij}})} \frac{\partial (\mu_{p_{ij}})}{\partial \beta_k} \\
&= \frac{1}{V(\mu_{p_{ij}})} \frac{\partial (\mu_{p_{ij}})}{\partial \beta_k} \sum_n w_i^{[r]}(n)(d_{ij} - n\mu_{p_{ij}}) \\
&= \frac{1}{V(\mu_{p_{ij}})} \frac{\partial (\mu_{p_{ij}})}{\partial \beta_k} \sum_{n=0}^{N} n w_i^{[r]}(n) (\frac{\sum_{n=0}^{N} w_i^{[r]}(n)}{\sum_{n=0}^{N} n w_i^{[r]}(n)} d_{ij} - \mu_{p_{ij}})
\end{aligned} \tag{16}$$

Single $p_{ij}$ GAM uses pseudo-data $\frac{\sum_{n=0}^{N} w_i^{[r]}(n)}{\sum_{n=0}^{N} n w_i^{[r]}(n)} d_{ij}$ and weight $\sum_{n=0}^{N} n w_i^{[r]}(n)$ to fit a quasi-Binomial GAM.

Each iteration's M-step will fit total 3 weighted GAMs and use the modified PIRLS algorithm proposed by Hai Liu and Kung-Sik Chan 2009.

# 4 Model Selection

Follow Hai Liu and Kung-Sik Chan 2009, we use BIC and its Laplacian approximation for model selection.