

ZIGAM with Repeated Sampling in a Short Period

Yunyi SHEN

November 16, 2018

1 Model Setting

1.1 Combined Distribution of Latent Population Size and Detections

1.2 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:

$$f(y|\theta) = b(y)\exp(\eta^T T(y) - a(\eta)) \quad (1)$$

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

$$\begin{aligned} P(\vec{d}|n) &= \prod_{j=1}^w \binom{n}{d_j} p_j^{d_j} (1 - p_j)^{n - d_j} \\ &= \prod_{j=1}^w \binom{n}{d_j} \exp\left(\sum_{j=1}^w d_j \log \frac{p_j}{1 - p_j} + n \sum_{j=1}^w \log(1 - p_j)\right) \end{aligned} \quad (2)$$

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

Then assume latent population was Poisson distributed with rate λ

$$\begin{aligned} P(n|\theta) &= e^{-\lambda} \frac{\lambda^n}{n!} \\ &= \frac{1}{n!} \exp(n \log(\lambda) - \lambda) \end{aligned} \quad (3)$$

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

$$\begin{aligned} P(n, \vec{d}|\theta) &= P(d|n, \theta)P(n|\theta) \\ &= \frac{1}{n!} \prod_{j=1}^w \binom{n}{d_j} \exp\left(\sum_{j=1}^w d_j \log \frac{p_j}{1 - p_j} + n \sum_{j=1}^w \log(1 - p_j)\right) \exp(n \log(\lambda) - \lambda) \\ &= \frac{1}{n!} \prod_{j=1}^w \binom{n}{d_j} \exp\left[\sum_{j=1}^w d_j \log \frac{p_j}{1 - p_j} + n\left(\sum_{j=1}^w \log(1 - p_j) + \log(\lambda)\right) - \lambda\right] \\ &= \frac{1}{n!} \prod_{j=1}^w \binom{n}{d_j} \exp(\eta^T (n, \vec{d}) - \lambda) \end{aligned} \quad (4)$$

in which

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

$$T(y) = y$$

λ can be calculated using η since it contains all p_j and λ itself. Later on, we note this function as

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

□

1.3 Zero Inflate

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 :

$$\begin{aligned} P(\vec{d}, n, z) &= [\psi \text{Pois}(n|\lambda) \prod_{j=1}^w \text{Bin}(d_j|n, p_j)]^z (I_{\vec{d}=0}(1 - \psi))^{1-z} \\ &= (\psi f(n, \vec{d}|\theta))^z (I_{\vec{d}=0}(1 - \psi))^{1-z} \end{aligned} \quad (5)$$

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 .

$$\begin{aligned} P(n, z|\vec{d}) &= \frac{P(n, z, \vec{d})}{\sum_{n=\max(d_j)}^N \sum_{z=0}^1 P(n, z, \vec{d})} \\ &= \frac{(\psi f(n, \vec{d}|\theta))^z (I_{\vec{d}=0}(1 - \psi))^{1-z}}{\psi \sum_{n=\max(d_j)}^N f(n, \vec{d}|\theta) + (1 - \psi) I_{\vec{d}=0}} \end{aligned} \quad (6)$$

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

$$\mathbb{E}(l_p|\vec{d}, \theta^{[r]}) = \sum_{sites} \frac{\sum_n \psi^{[r]} f(n, \vec{d}|\theta^{[r]}) \log[\psi f(n, \vec{d}|\theta)] + (1 - \psi^{[r]}) I_{\vec{d}=0} \log[(1 - \psi) I_{\vec{d}=0}]}{\sum_n \psi^{[r]} f(n, \vec{d}|\theta^{[r]}) + (1 - \psi^{[r]}) I_{\vec{d}=0}} \quad (7)$$

This involves large number of GAMs to be maximized during M step, but still can be solved via RILS

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

We get derivative of E_l and ψ_i at site i

$$\frac{\partial E_l}{\partial \psi_i} = \frac{\psi_i^{[r]} g_i^{[r]}}{\psi_i^{[r]} g_i^{[r]} + (1 - \psi_i^{[r]}) I_{\vec{d}_i=0}} - \psi_i \quad (8)$$

Involves the first weighted GAM regarding occupancy rate ψ

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

$$\frac{\partial E_l}{\partial f_{\vec{d}_i}(n)} = \frac{\psi_i^{[r]}}{Z^{[r]}} \sum_{n=0}^N \frac{f_{\vec{d}_i}^{[r]}(n)}{f_{\vec{d}_i}(n)} \quad (9)$$

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

$$\begin{aligned} \frac{\partial E_l}{\partial \beta_k} &= \sum_i \frac{\partial E_l}{\partial f_{\vec{d}_i}(n)} \frac{\partial f_{\vec{d}_i}(n)}{\partial \beta_k} + \sum_i \frac{\partial E_l}{\partial \psi_i} \frac{\partial \psi_i}{\partial \beta_k} \\ &= \sum_i \frac{\psi_i^{[r]}}{Z^{[r]}} \sum_{n=0}^N f_{\vec{d}_i}^{[r]}(n) \frac{\partial \log(f_{\vec{d}_i}(n))}{\partial \beta_k} + \sum_i \frac{\partial E_l}{\partial \psi_i} \frac{\partial \psi_i}{\partial \beta_k} \\ &= \sum_{n=0}^N \sum_i \frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d}_i}^{[r]}(n) \frac{\partial \log(f_{\vec{d}_i}(n))}{\partial \beta_k} + \sum_i \frac{\partial E_l}{\partial \psi_i} \frac{\partial \psi_i}{\partial \beta_k} \end{aligned} \quad (10)$$

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

$$\begin{aligned} &\sum_{n=0}^N \sum_i \frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d}_i}^{[r]}(n) \frac{\partial \log(f_{\vec{d}_i}(n))}{\partial \beta_k} \\ &= \sum_{n=0}^N \sum_i \left[\frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d}_i}^{[r]}(n) \frac{\partial \log \text{Pois}(n, \lambda_i)}{\partial \beta_k} + \sum_{j=1}^w \frac{\psi_i^{[r]}}{Z^{[r]}} f_{\vec{d}_i}^{[r]}(n) \frac{\partial \log \text{Bin}(p_{ij}, n)}{\partial \beta_k} \right] \end{aligned} \quad (11)$$

Note that:

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

First deal with λ which is easier:

$$\begin{aligned} \sum_{n=0}^N \sum_i w_i^{[r]}(n) \frac{\partial \log \text{Pois}(n, \lambda_i)}{\partial \beta_k} &= \sum_i \frac{\sum_{n=0}^N w_i^{[r]}(n)(n - \mu_{\lambda_i})}{\phi V(\mu_{\lambda_i})} \frac{\partial \mu_{\lambda_i}}{\partial \beta_k} \\ &= \sum_i \left[\sum_{n=0}^N w_i^{[r]}(n) \right] \left(\frac{\sum_{n=0}^N w_i^{[r]}(n)n}{\sum_{n=0}^N w_i^{[r]}(n)} - \mu_{\lambda_i} \right) \frac{1}{\phi V(\mu_{\lambda_i})} \frac{\partial \mu_{\lambda_i}}{\partial \beta_k} \end{aligned} \quad (12)$$

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} :

$$\begin{aligned} \frac{\partial \log \text{Bin}(p_{ij}, n)}{\partial \beta_k} &= \frac{d_{ij} - np_{ij}}{np_{ij}q_{ij}} \frac{\partial(np_{ij})}{\partial \beta_k} \\ &= \frac{d_{ij} - \frac{n}{N}Np_{ij}}{Np_{ij}q_{ij}} \frac{\partial(Np_{ij})}{\partial \beta_k} \\ &= \frac{d_{ij} - \frac{n}{N}\mu_{Np_{ij}}}{V(\mu_{Np_{ij}})} \frac{\partial(\mu_{Np_{ij}})}{\partial \beta_k} \end{aligned} \quad (13)$$

Now it is related to a quasi-Binomial regression with sample size N , set by user.

$$\begin{aligned} \sum_n w_i^{[r]}(n) \frac{\partial \log \text{Bin}(p_{ij}, n)}{\partial \beta_k} &= \sum_n w_i^{[r]}(n) \frac{d_{ij} - \frac{n}{N}\mu_{Np_{ij}}}{V(\mu_{Np_{ij}})} \frac{\partial(\mu_{Np_{ij}})}{\partial \beta_k} \\ &= \frac{1}{V(\mu_{Np_{ij}})} \frac{\partial(\mu_{Np_{ij}})}{\partial \beta_k} \sum_n w_i^{[r]}(n) (d_{ij} - \frac{n}{N}\mu_{Np_{ij}}) \\ &= \frac{1}{V(\mu_{Np_{ij}})} \frac{\partial(\mu_{Np_{ij}})}{\partial \beta_k} \frac{\sum_{n=0}^N nw_i^{[r]}(n)}{N} \left(\frac{N \sum_{n=0}^N w_i^{[r]}(n)}{\sum_{n=0}^N nw_i^{[r]}(n)} d_{ij} - \mu_{Np_{ij}} \right) \end{aligned} \quad (14)$$

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

Each iteration's M-step will fit total $w + 2$ weighted GAMs and use the modified IRLS algorithm proposed by Hai Liu and Kung-Sik Chan 2009.