Generalized Poisson Distribution

Bo Yang
boyang@knights.ucf.edu

May 7, 2013

Abstract

Generalized Poisson Distribution to model and bin metagenomic species.

1 Methods

$X = (x_1, x_2, \dots, x_n)$: observed data. x_i is the unique occurences of the $i-$th $l-$tuple in all reads, and l is a fixed number. n is the number of unique k-mers.
$L = (l_1, l_2, \dots, l_n)$: observed data. l_i is the occurrences of x_i, i.e. the total number of a k-mer.
$\Theta = (\alpha_1, \alpha_2, \dots, \alpha_m, \lambda_{jk})$: parameters, where $j = \{1, 2, \dots, m\}$ $k = \{1, 2\}$. And

$\alpha_1, \alpha_2, \dots, \alpha_m$: the probability that a k-mer is from a Generalized Poisson Distribution(GPD). m is the number of different GPDs.

λ_{jk}: parameters for generalized poisson distribution,

$$p_j(x_i) \triangleq P(x_i|\lambda_{j1}, \lambda_{j2}) = \frac{\lambda_{j1}(\lambda_{j1} + x_i\lambda_{j2})^{x_i-1} e^{-(\lambda_{j1}+x_i\lambda_{j2})}}{x_i!} \tag{1}$$

where $\lambda_{j1} > 0$, $0 < \lambda_{j2} < 1$.
$Y = \{y_{ij}\}$: missing data, where $i = \{1, 2, \dots, n\}$,$j = \{1, 2, \dots, m\}$, and
$y_{ij} = 1$ if x_i is from the $j-$th Generalized Poisson Distribution.
$y_{ij} = 0$ if x_i is not from the $j-$th Generalized Poisson Distribution.
The likelihood function is

$$L(X, Y, L|\Theta) = P_Y(X, Y, L|\Theta) = \prod_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\alpha_j p_j(x_i) \tag{2}$$

$$\log L(X, Y, L|\Theta) = \sum_{i=1}^{n}\log\sum_{j=1}^{m} y_{ij}\alpha_j p_j(x_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log\alpha_j p_j(x_i) \tag{3}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}[\log\alpha_j + \log\lambda_{j1} + (x_i - 1)\log(\lambda_{j1} + x_i\lambda_{j2}) - (\lambda_{j1} + x_i\lambda_{j2}) - \log(x_i!)]$$

Since y_{ij} is missing, we try to estimate it by its mean:

$$E(y_{ij}) = P(y_{ij} = 1) = \frac{\alpha_j p_j(x_i)}{\sum_{k=1}^{m}\alpha_k p_k(x_i)} = z_{ij} \tag{4}$$

Note: z_{ij} depends on the current parameters, which we assume to be $\Theta^{(t-1)}$. Correspondingly, we assume z_{ij} under the current parameters are $z_{ij}^{(t-1)}$.

So we have the missing data y_{ij} replaced by its expression in $\log L(X, Y, L|\Theta)$. We define

$$Q\left(\Theta^{(t)}|\Theta^{(t-1)}\right) = \sum_{i=1}^{n} l_i \sum_{j=1}^{m} z_{ij}^{(t-1)}\log\alpha_j p_j(x_i) \tag{5}$$

In other words, $Q\left(\Theta^{(t)}|\Theta^{(t-1)}\right)$ is the log likelihood function with the missing data y_{ij} integrated out under the current parameters $\Theta^{(t-1)}$. We now want to estimate the new parameter $\Theta^{(t)}$ by maximal likelihood estimation. So we calculate

$$\frac{\partial Q\left(\Theta^{(t)}|\Theta^{(t-1)}\right)}{\partial\alpha_j}, \frac{\partial Q\left(\Theta^{(t)}|\Theta^{(t-1)}\right)}{\partial\lambda_{j1}}, \frac{\partial Q\left(\Theta^{(t)}|\Theta^{(t-1)}\right)}{\partial\lambda_{j2}} \tag{6}$$

and so we have

$$\begin{aligned}
\alpha_j^{(t)} &= \frac{1}{N}\sum_{i=1}^{n} z_{ij}^{(t-1)} l_i \\
\frac{\partial Q}{\partial \lambda_{j1}} &= \sum_{i=1}^{n} l_i z_{ij}^{(t-1)} \left(\frac{1}{\lambda_{j1}} + \frac{x_i - 1}{\lambda_{j1} + x_i\lambda_{j2}} - 1\right) = 0 \\
\frac{\partial Q}{\partial \lambda_{j2}} &= \sum_{i=1}^{n} l_i z_{ij}^{(t-1)} \left(\frac{x_i\left(x_i - 1\right)}{\lambda_{j1} + x_i\lambda_{j2}} - x_i\right) = 0
\end{aligned} \tag{7}$$

where $N = \sum_{i=1}^{n} l_i$.

In order to calculate λ_{j1} and λ_{j2}, we will resort to Newton's method.

$$\lambda_{j1} = \frac{\sum_{i=1}^{n} l_i z_{ij}^{(t-1)} x_i}{\sum_{i=1}^{n} l_i z_{ij}^{(t-1)}} \left(1 - \lambda_{j2}\right) = w\left(1 - \lambda_{j2}\right) \tag{8}$$

where $w = \frac{\sum_{i=1}^{n} l_i z_{ij}^{(t-1)} x_i}{\sum_{i=1}^{n} l_i z_{ij}^{(t-1)}}$.

$$f\left(\lambda_{j2}\right) = \sum_{i=1}^{n} l_i z_{ij}^{(t-1)} \left(\frac{x_i(x_i - 1)}{w + (x_i - w)\lambda_{j2}} - x_i\right) = 0 \tag{9}$$

and

$$f^{'}\left(\lambda_{j2}\right) = -\sum_{i=1}^{n} l_i z_{ij}^{(t-1)} \frac{x_i(x_i - 1)(x_i - w)}{\left[w + (x_i - w)\lambda_{j2}\right]^2} \tag{10}$$

According to Newton's Method,

$$\lambda_{j2}^{(k+1)} = \lambda_{j2}^{(k)} - \frac{f\left(\lambda_{j2}\right)}{f'\left(\lambda_{j2}\right)} \tag{11}$$

and the stop criteria is

$$\left|\frac{\lambda_{j2}^{(k+1)} - \lambda_{j2}^{(k)}}{\lambda_{j2}^{(k+1)}}\right| < \varepsilon \tag{12}$$

Another way to calculate λ_{j1} and λ_{j2} is
Define vector

$$g = f\left(\lambda_{j1}, \lambda_{j2}\right) = \begin{bmatrix} f_1\left(\lambda_{j1}, \lambda_{j2}\right) \\ f_2\left(\lambda_{j1}, \lambda_{j2}\right) \end{bmatrix} = \begin{bmatrix} \frac{\partial Q}{\partial \lambda_{j1}} \\ \frac{\partial Q}{\partial \lambda_{j2}} \end{bmatrix} \tag{13}$$

so that

$$H = Df\left(\lambda_{j1}, \lambda_{j2}\right) = \begin{bmatrix} \frac{\partial f_1}{\partial \lambda_{j1}} & \frac{\partial f_1}{\partial \lambda_{j2}} \\ \frac{\partial f_2}{\partial \lambda_{j1}} & \frac{\partial f_2}{\partial \lambda_{j2}} \end{bmatrix} \tag{14}$$

$$= \begin{bmatrix} -\sum_{i=1}^{n} z_{ij}^{(t-1)} \left(\frac{1}{\lambda_{j1}^2} + \frac{x_i - 1}{\left(\lambda_{j1} + x_i\lambda_{j2}\right)^2}\right) & -\sum_{i=1}^{n} z_{ij}^{(t-1)} \frac{x_i(x_i - 1)}{\left(\lambda_{j1} + x_i\lambda_{j2}\right)^2} \\ -\sum_{i=1}^{n} z_{ij}^{(t-1)} \frac{x_i(x_i - 1)}{\left(\lambda_{j1} + x_i\lambda_{j2}\right)^2} & -\sum_{i=1}^{n} z_{ij}^{(t-1)} \frac{x_i^2(x_i - 1)}{\left(\lambda_{j1} + x_i\lambda_{j2}\right)^2} \end{bmatrix} \tag{15}$$

Assume $\Delta = H^{-1}g$, so we have

$$\begin{bmatrix} \lambda_{j1} \\ \lambda_{j2} \end{bmatrix}^{(t+1)} = \begin{bmatrix} \lambda_{j1} \\ \lambda_{j2} \end{bmatrix}^{(t)} - \Delta \tag{16}$$

Convergence criteria for Newton's Method and EM algorithm:

$$\left|\frac{\sqrt{\left(\lambda_{j1}^{(t+1)} - \lambda_{j1}^{(t)}\right)^2 + \left(\lambda_{j2}^{(t+1)} - \lambda_{j2}^{(t)}\right)^2}}{\sqrt{\lambda_{j1}^{(t)^2} + \lambda_{j2}^{(t)^2}}}\right| < \varepsilon \tag{17}$$

where $\varepsilon = 0.001$ or 0.0001.

Computing probability by logarithm

Since

$$\log p_j(x_i) = \log \lambda_{j1} + (x_i - 1)\log(\lambda_{j1} + x_i\lambda_{j2}) - (\lambda_{j1} + x_i\lambda_{j2}) - \log(x_i!) \tag{18}$$

and $\log(x_i!) = \sum_{k=1}^{x_i} \log k$,

$$p_j(x_i) = \exp\left(\log \lambda_{j1} + (x_i - 1)\log(\lambda_{j1} + x_i\lambda_{j2}) - (\lambda_{j1} + x_i\lambda_{j2}) - \sum_{k=1}^{x_i} \log k\right) \tag{19}$$

Once the EM algorithm converges, we can estimate the probability of a read assigned to a bin, based on its l-tuples binning result as,

$$P(r_k \in s_j) = \frac{\prod_{w_i \in r_k} P(y_{ij} = 1)}{\sum_{s_j \in S}\left(\prod_{w_i \in r_k} P(y_{ij} = 1)\right)} = \frac{\prod_{i=0}^{n} z_{ij}}{\sum_{j=0}^{m}\left(\prod_{i=0}^{n} z_{ij}\right)} \tag{20}$$

where r_k is a given read, w_i is the l-tuples that belong to r_k, and s_j is a bin. A read will be assigned to the bin with the highest probability among all bins.