Flexible Regression and Smoothing
Discrete Distributions

Bob Rigby    Mikis Stasinopoulos

Graz University of Technology, Austria, November 2016
Discrete Distributions
- Count distributions
- Families modelling the variance-mean relationship
- Examples
  - A stylometric application
  - The fish species data
- Binomial response variables
- Example
  - The hospital stay data
The three major problems encountered when modelling count data using the Poisson distribution.

- overdispersion
- excess (or shortage) of zero values
- long tails (rare events)
### Discrete distribution modelling

<table>
<thead>
<tr>
<th>Par.</th>
<th>Modelling</th>
<th>Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Location</td>
<td>PO</td>
</tr>
<tr>
<td>2</td>
<td>Location and scale</td>
<td>NBI, NBII, PIG</td>
</tr>
<tr>
<td>2</td>
<td>Location and zero probability</td>
<td>ZALG, ZAP, ZIP, ZIP2</td>
</tr>
<tr>
<td>3</td>
<td>Location, scale and skewness</td>
<td>DEL, SI, SICHEL</td>
</tr>
<tr>
<td>3</td>
<td>Location, scale and zero probability</td>
<td>ZANBI, ZINBI, ZIPIG</td>
</tr>
</tbody>
</table>
Different count data distributions

- **ZIP(10,0.5)**
- **NB(5,1)**
- **NBtr(2.77,2.61)**
- **PIG(5,1)**
Zero inflated distribution, $Y \sim \text{ZID}$ is given by

- $Y = 0$ with probability $p$
- $Y \sim \text{D}$ with probability $1 - p$.

Hence

$$P(Y = y) = \begin{cases} 
  p + (1 - p)P(Y_1 = 0) & \text{if } y = 0 \\
  (1 - p)P(Y_1 = y) & \text{if } y = 1, 2, 3, \ldots 
\end{cases}$$

where $Y_1 \sim \text{D}$.
Zero adjusted distribution, \( Y \sim \text{ZAD} \) is given by
\[ Y = 0 \text{ with probability } p \]
\[ Y \sim \text{Dtr} \text{ with probability } 1 - p, \]
where \( \text{Dtr} \) is a truncated distribution, \( D \) truncated at zero.

Hence
\[
P(Y = y) = \begin{cases} 
p & \text{if } y = 0 \\ (1 - p) \frac{P(Y_1 = y)}{1 - P(Y_1 = 0)} & \text{if } y = 1, 2, 3, \ldots \end{cases}
\]

where \( Y_1 \sim D \).

\[ (1) \]
Zero altered negative binomial type I, ZANBI

ZANBI( \( \mu = 5 \), \( \sigma = 1 \), \( \nu = 0.1667 \) )

Zero altered negative binomial type I, ZANBI

ZANBI( \( \mu = 4 \), \( \sigma = 1.5 \), \( \nu = 0.09159 \) )

Zero altered negative binomial type I, ZANBI

ZANBI( \( \mu = 3 \), \( \sigma = 2.33 \), \( \nu = 0.01695 \) )

Zero altered negative binomial type I, ZANBI

ZANBI( \( \mu = 2.8 \), \( \sigma = 2.57 \), \( \nu = 0.00213 \) )
Different (overdispersed) count data approaches

(a) Ad-hoc solutions
   (i) quasi-likelihood (QL), Extended QL
   (ii) Efron’s Double Exponential
   (iii) pseudo-likelihood (PL)

(b) Discretized continuous distributions
   for example if $F_W(w)$ is the cdf a continuous random variable $W$
   defined in $\mathbb{R}^+$ then $f_Y(y) = F_W(y + 1) - F_W(y)$

(c) Random effect at the observation level solutions.
   $f_Y(y) = \int f(y|\gamma)f_{\gamma}(\gamma)d\gamma$. 
(c) Random effect at the observation level

(i) when an an explicit continuous mixture distribution, \( f_Y(y) \), exists.

(ii) when a continuous mixture distribution, \( f_Y(y) \), is not explicit but is approximated by integrating out the random effect using approximations, e.g. Gaussian quadrature or Laplace approximation.

(iii) when a 'non-parametric' mixture (effectively a finite mixture) is assumed for the response variable.
Random effect at the observation level case (i)

(i) Explicit continuous mixture distribution

\[ f_Y(y) = \int_{\text{discrete}} f(y|\gamma) f_{\gamma}(\gamma) \ d\gamma \]

- \( Y \sim NBI(\mu, \sigma) \)
- \( Y|\gamma \sim PO(\gamma\mu) \)
- \( \gamma \sim GA(1, \sigma^{1/2}) \)
(ii) Non-explicit continuous mixture distribution

\[ f_Y(y) = \int_{\text{discrete}} f(y|\gamma) \cdot f_\gamma(\gamma) \cdot d\gamma \]

- \( Y \sim PO - Normal(\mu, \sigma) \)
- \( Y|\gamma \sim PO(\gamma \mu) \)
- \( \log(\gamma) \sim NO(1, \sigma) \)
(iii) Non-parametric mixture distribution

\[
    f_Y(y) = \sum_{k=1}^{K} f(y|\gamma_k) p(\gamma = \gamma_k)
\]

where NPFM(2) equals Non-Parametric Finite Mixture with 2 point probabilities.

- \( Y \sim PO - NPFM(\mu, \sigma) \)
- \( Y|\gamma \sim PO(\gamma \mu) \)
- \( \log(\gamma) \sim NPFM(2) \)
### Explicit continuous mixture distribution

<table>
<thead>
<tr>
<th>Distributions</th>
<th>R Name</th>
<th>mixing distribution for $\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>$\text{PO}(\mu)$</td>
<td>-</td>
</tr>
<tr>
<td>Neg. bin. I</td>
<td>$\text{NBI}(\mu, \sigma)$</td>
<td>$\text{GA}(1, \sigma^\frac{1}{2})$</td>
</tr>
<tr>
<td>Neg. bin. II</td>
<td>$\text{NBII}(\mu, \sigma)$</td>
<td>$\text{GA}(1, \sigma^\frac{1}{2}/\mu)$</td>
</tr>
<tr>
<td>Poisson IG</td>
<td>$\text{PIG}(\mu, \sigma)$</td>
<td>$\text{IG}(1, \sigma^\frac{1}{2})$</td>
</tr>
<tr>
<td>Sichel</td>
<td>$\text{SICHEL}(\mu, \sigma, \nu)$</td>
<td>$\text{GIG}(1, \sigma^\frac{1}{2}, \nu)$</td>
</tr>
<tr>
<td>Delaporte</td>
<td>$\text{DEL}(\mu, \sigma, \nu)$</td>
<td>$\text{SG}(1, \sigma^\frac{1}{2}, \nu)$</td>
</tr>
<tr>
<td>Zero inflated Poisson</td>
<td>$\text{ZIP}(\mu, \sigma)$</td>
<td>$\text{BI}(1, 1 - \sigma)$</td>
</tr>
<tr>
<td>Zero inflated Poisson 2</td>
<td>$\text{ZIP2}(\mu, \sigma)$</td>
<td>$(1 - \sigma)^{-1}\text{BI}(1, 1 - \sigma)$</td>
</tr>
<tr>
<td>Zero inflated neg. bin.</td>
<td>$\text{ZINBI}(\mu, \sigma, \nu)$</td>
<td>zero inflated gamma</td>
</tr>
<tr>
<td>Poisson-Tweedie</td>
<td>-</td>
<td>Tweedie family</td>
</tr>
</tbody>
</table>
## Table: Discrete gamlss family distributions for count data

<table>
<thead>
<tr>
<th>R Name</th>
<th>params</th>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO((\mu))</td>
<td>1</td>
<td>(\mu)</td>
<td>(\mu)</td>
</tr>
<tr>
<td>NBI((\mu, \sigma))</td>
<td>2</td>
<td>(\mu)</td>
<td>(\mu + \sigma \mu^2)</td>
</tr>
<tr>
<td>NBII((\mu, \sigma))</td>
<td>2</td>
<td>(\mu)</td>
<td>(\mu + \sigma \mu)</td>
</tr>
<tr>
<td>PIG((\mu, \sigma))</td>
<td>2</td>
<td>(\mu)</td>
<td>(\mu + \sigma \mu^2)</td>
</tr>
<tr>
<td>SICHEL((\mu, \sigma, \nu))</td>
<td>3</td>
<td>(\mu)</td>
<td>(\mu + h(\sigma, \nu) \mu^2)</td>
</tr>
<tr>
<td>DEL((\mu, \sigma, \nu))</td>
<td>3</td>
<td>(\mu)</td>
<td>(\mu + \sigma (1 - \nu)^2 \mu^2)</td>
</tr>
<tr>
<td>ZIP((\mu, \sigma))</td>
<td>2</td>
<td>((1 - \sigma) \mu)</td>
<td>((1 - \sigma) \mu + \sigma (1 - \sigma) \mu^2)</td>
</tr>
<tr>
<td>ZIP2((\mu, \sigma))</td>
<td>2</td>
<td>(\mu)</td>
<td>(\mu + \frac{\sigma}{(1 - \sigma)} \mu^2)</td>
</tr>
</tbody>
</table>
Families modelling the variance-mean relationship

\[ V[Y] = \mu + \mu^2 V[\gamma] \] where \( V[\gamma] = \nu(\sigma, \nu, \tau) \) is a function of the parameters of the mixing distribution \( f_\gamma(\gamma) \).

Alternative variance-mean relationship can be obtained by reparametrization.

i.e NB type I \( V[Y] = \mu + \sigma \mu^2 \).

If \( \sigma = \sigma_1/\mu \) then
\[ V[Y] = (1 + \sigma_1)\mu \] (negative binomial type II)

\( \sigma = \sigma_1\mu \) then \( V[Y] = \mu + \sigma_1\mu^3 \).

More generally \( \sigma = \sigma_1\mu^{2-\nu} \) giving \( V(Y) = \mu + \sigma_1\mu^\nu \)
Comparison of the marginal distributions using a (ratio moment) diagram of their skewness and kurtosis
A stylometric application

Data summary:

R data file: stylo in package `gamlss.data` of dimensions $64 \times 2$

source: Dr Mario Corina-Borja

variables

word : is the number of times a word appears in a single text

freq : the frequency of the number of times a word appears in a text

purpose: to demonstrate the fitting of a truncated discrete dist.

conclusion the truncated SICHEL distributions fits best
library(gamlss.tr)
data(stylo)
plot(freq ~ word, data = stylo, type = "h", xlim = + c(0, 22), xlab = "no of times", ylab = + "frequencies", col = "blue")
The stylometric data
A stylometric application

```r
> library(gamlss.tr)
> gen.trun(par = 0, family = PO, type = "left")
A truncated family of distributions from PO has been generated
and saved under the names:
dPOtr pPOtr qPOtr rPOtr POtr
The type of truncation is left and the truncation parameter is 0
> gen.trun(par = 0, family = NBII, type = "left")
...
> gen.trun(par = 0, family = DEL, type = "left")
...
> gen.trun(par = 0, family = SICHEL, type = "left",
+ delta = 0.001)
...
A stylometric application

```r
> mPO <- gamlss(word ~ 1, weights = freq, data = stylo, + family = POtr, trace = FALSE)
> mNBII <- gamlss(word ~ 1, weights = freq, data = stylo, + family = NBIItr, n.cyc = 50, trace = FALSE)
> mDEL <- gamlss(word ~ 1, weights = freq, data = stylo, + family = DELtr, n.cyc = 50, trace = FALSE)
> mSI <- gamlss(word ~ 1, weights = freq, data = stylo, + family = SICHELtr, n.cyc = 50, trace = FALSE)
> GAIC(mPO, mNBII, mDEL, mSI)

   df    AIC
mSI 3 5148.454
mDEL 3 5160.581
mNBII 2 5311.627
mPO 1 9207.459
```
The stylometric data

(b) Poisson

(c) negative binomial II

(c) Delaporte

(d) Sichel
The fish species data

**Data summary:** the fish species data

**R data file:** species in package gamlss.data of dimensions $70 \times 2$

**variables**

- **fish**: the number of different species in 70 lakes in the world
- **lake**: the lake area
The fish species data
The fish species data

There are several questions that need to be answered.

- How does the mean of $y$ depend on $x$?
- Is $y$ overdispersed Poisson?
- How does the variance $y$ depend on its mean?
- What is the distribution of $y$ given $x$?
- Do the scale and shape parameters of the distribution of $y$ depend on $x$?
Overdispersed count data approaches

Table: Comparison of models for the fish species data

<table>
<thead>
<tr>
<th>Model</th>
<th>$f_Y(y)$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\nu$</th>
<th>$DEV$</th>
<th>$df$</th>
<th>$AIC$</th>
<th>$SBC$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PO</td>
<td>$x &lt; 2 &gt;$</td>
<td>-</td>
<td>-</td>
<td>1849.3</td>
<td>3</td>
<td>1855.3</td>
<td>1862.0</td>
</tr>
<tr>
<td>2</td>
<td>NBI</td>
<td>$x$</td>
<td>1</td>
<td>-</td>
<td>619.8</td>
<td>3</td>
<td>625.8</td>
<td>632.6</td>
</tr>
<tr>
<td>3</td>
<td>NBI</td>
<td>$x &lt; 2 &gt;$</td>
<td>1</td>
<td>-</td>
<td>614.3</td>
<td>4</td>
<td>622.3</td>
<td>631.3</td>
</tr>
<tr>
<td>4</td>
<td>NBI</td>
<td>cs$(x, 3)$</td>
<td>1</td>
<td>-</td>
<td>611.9</td>
<td>6</td>
<td>623.9</td>
<td>637.4</td>
</tr>
<tr>
<td>5</td>
<td>NBI</td>
<td>$x &lt; 2 &gt;$</td>
<td>$x$</td>
<td>-</td>
<td>605.0</td>
<td>5</td>
<td>615.0</td>
<td>626.2</td>
</tr>
<tr>
<td>6</td>
<td>NBI-fam</td>
<td>$x &lt; 2 &gt;$</td>
<td>1</td>
<td>1</td>
<td>606.0</td>
<td>5</td>
<td>616.0</td>
<td>627.3</td>
</tr>
<tr>
<td>7</td>
<td>NBI-fam</td>
<td>$x &lt; 2 &gt;$</td>
<td>$x$</td>
<td>1</td>
<td>604.9</td>
<td>6</td>
<td>616.9</td>
<td>630.4</td>
</tr>
</tbody>
</table>
## Overdispersed count data approaches

<table>
<thead>
<tr>
<th>Model</th>
<th>$f_Y(y)$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\nu$</th>
<th>DEV</th>
<th>df</th>
<th>AIC</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>PIG</td>
<td>$&lt; 2$</td>
<td>1</td>
<td>-</td>
<td>613.3</td>
<td>4</td>
<td>621.3</td>
<td>630.3</td>
</tr>
<tr>
<td>9</td>
<td>SI</td>
<td>$&lt; 2$</td>
<td>1</td>
<td>$x$</td>
<td>597.7</td>
<td>6</td>
<td>609.7</td>
<td>623.2</td>
</tr>
<tr>
<td>10</td>
<td>DEL</td>
<td>$&lt; 2$</td>
<td>1</td>
<td>$x$</td>
<td>600.6</td>
<td>6</td>
<td>612.6</td>
<td>626.1</td>
</tr>
<tr>
<td>11</td>
<td>DEL</td>
<td>$&lt; 2$</td>
<td>-</td>
<td>$x$</td>
<td>600.6</td>
<td>5</td>
<td>610.6</td>
<td>621.9</td>
</tr>
<tr>
<td>12</td>
<td>PO-Normal</td>
<td>$&lt; 2$</td>
<td>1</td>
<td>-</td>
<td>615.2</td>
<td>4</td>
<td>623.2</td>
<td>632.2</td>
</tr>
<tr>
<td>13</td>
<td>NBI-Normal</td>
<td>$&lt; 2$</td>
<td>$x$</td>
<td>1</td>
<td>603.7</td>
<td>6</td>
<td>615.7</td>
<td>629.2</td>
</tr>
<tr>
<td>14</td>
<td>PO-NPFM(5)</td>
<td>$&lt; 2$</td>
<td>-</td>
<td>-</td>
<td>601.9</td>
<td>13</td>
<td>627.9</td>
<td>657.2</td>
</tr>
<tr>
<td>15</td>
<td>NB-NPFM(2)</td>
<td>$&lt; 2$</td>
<td>1</td>
<td>-</td>
<td>611.9</td>
<td>6</td>
<td>623.9</td>
<td>637.4</td>
</tr>
<tr>
<td>16</td>
<td>doublePO</td>
<td>$&lt; 2$</td>
<td>$x$</td>
<td>-</td>
<td>616.4</td>
<td>5</td>
<td>626.4</td>
<td>637.6</td>
</tr>
<tr>
<td>17</td>
<td>IGdisc</td>
<td>$&lt; 2$</td>
<td>1</td>
<td>-</td>
<td>603.3</td>
<td>4</td>
<td>611.3</td>
<td>620.3</td>
</tr>
</tbody>
</table>
Fitted mean of the Sichel distribution
Fitted Sichel distributions for observations (a) 40 and (b) 67
There are only two distributions here

- binomial
- beta binomial
The hospital stay data

Data summary:

- R data file: aep in package **gamlss** of dimensions 1383 × 8
- source: Gange *et al.* (1996)

**variables**

- los: total number of days
- noinap: number of inappropriate days patient stay in hospital
- loglos: the log of los/10
- sex: the gender of patient
- ward: type of ward in the hospital (medical, surgical or other)
- year: 1988 or 1990
- age: age of the patient subtracted from 55
- y: the response variable, a matrix with columns (noinap, los-noinap)
The hospital stay data
The hospital stay data

```r
> mI <- gamlss(y ~ ward + year + loglos, sigma.fo = ~year, + family = BB, data = aep)
> mII <- gamlss(y ~ ward + year + loglos, sigma.fo = ~year + + ward, family = BB, data = aep)
> mIII <- gamlss(y ~ ward + year + cs(loglos, 1), + sigma.fo = ~year + ward, family = BB, data = aep)
> mIV <- gamlss(y ~ ward + year + cs(loglos, 1) + cs(age, 1), + sigma.fo = ~year + ward, family = BB, data = aep)
> GAIC(mI, mII, mIII, mIV, k = 0)

   df         AIC
mIV 12.00010 4454.362
mIII 10.00045 4459.427
mII  9.00000 4483.020
mI   7.00000 4519.441
```
### The hospital stay data

<table>
<thead>
<tr>
<th>Models</th>
<th>Links</th>
<th>Terms</th>
<th>GD (AIC) [SBC]</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>logit($\mu$) log($\sigma$)</td>
<td>1+ward+loglos+year 1+year</td>
<td>4519.4 (4533.4) [4570.1]</td>
</tr>
<tr>
<td>II</td>
<td>logit($\mu$) log($\sigma$)</td>
<td>1+ward+loglos+year 1+year+ward</td>
<td>4483.0 (4501.0) [4548.1]</td>
</tr>
<tr>
<td>III</td>
<td>logit($\mu$) log($\sigma$)</td>
<td>1+ward+cs(loglos,1)+year 1+year+ward</td>
<td>4459.4 (4479.4) [4531.8]</td>
</tr>
<tr>
<td>IV</td>
<td>logit($\mu$) log($\sigma$)</td>
<td>1+ward+cs(loglos,1)+year+cs(age,1) 1+year+ward</td>
<td>4454.4 (4478.4) [4541.2]</td>
</tr>
</tbody>
</table>
The hospital stay data

```r
> op <- par(mfrow = c(2, 2))
> term.plot(mIV, se = T)
> par(op)
> op <- par(mfrow = c(2, 1))
> term.plot(mIV, "sigma", se = T)
> par(op)
> rqres.plot(mIV)
```
The hospital stay data: fitted model for $\mu$
The hospital stay data: fitted model for $\sigma$
The hospital stay data: normalised randomised quantile residuals
END

for more information see

www.gamlss.org