Generally Altered, Inflated, Truncated and Deflated Regression

Thomas Yee and Chenchen Ma

University of Auckland and Peking University

29 Nov 2023 @ Bay of Islands

t.yee@auckland.ac.nz
http://www.stat.auckland.ac.nz/~yee

Outline of this document

1. Recall. . .
    - Zero-inflated Examples
2. Further Motivation: Heaped Data
    - What is Heaped Data?
    - Example(s)
3. The GAITD Combo Model
    - The Main Idea
    - All the Extensions
    - GAITD Combo PMF
    - The GT–Expansion Method
    - Moments and CDF†
4. Estimation
    - Fisher Scoring and IRLS†
    - Upper Tail Truncation†
    - Initial Values†
    - VGAM Software
5. Examples
    - Example 1. GTE Method and Heaping on Sleep Duration

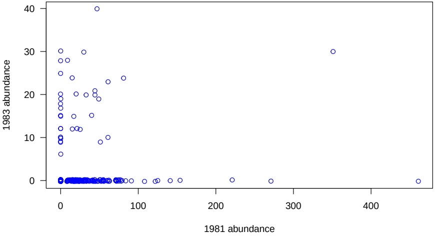
- Example 2. Smoking Duration

6. Closing Comments
   - Advice†
   - Some Future Work
   - Summary and Concluding Remarks

7. References

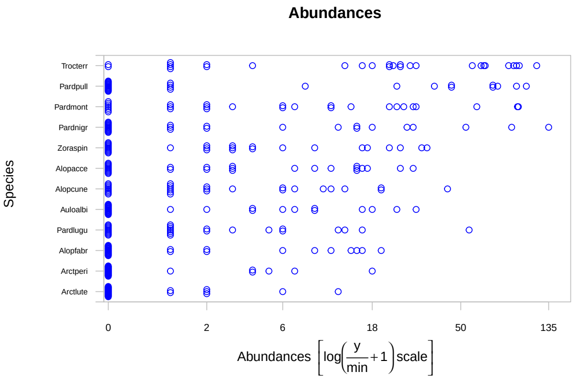
Recall. . .

Zero-inflated Examples



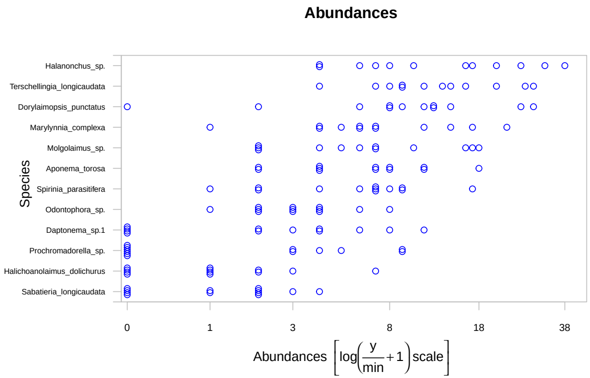
```
> data(tikus, package = "mvabund")
> tikusdat <- mvabund(tikus$abund)
> prop.table(table(tikusdat == 0))


 FALSE   TRUE
0.1164 0.8836
```



```
> data(spider, package = "mvabund")
> spiddat <- as.mvabund(spider$abund)
> prop.table(table(spiddat == 0))


 FALSE   TRUE
0.5417 0.4583
```



```
> data(solberg, package = "mvabund")
> solbdat <- as.mvabund(solberg$abund)
> prop.table(table(solbdat == 0))


 FALSE   TRUE
0.3884 0.6116
```

Further Motivation: Heaped Data

What is Heaped Data?

Heaping is a very common aberration in *retrospective self-reported survey* data, e.g., "**How many cigarettes did you smoke last week?**", "**At what age did you quit smoking?**"

- Often characterized by an excess of multiples of 10 or 5 upon rounding, e.g., 20 for a pack of cigarettes.
- Its effect on inference is usually unpredictable, e.g., the direction of bias depending on multiple heap locations.

Heaping is a form of *measurement error*.

Some real examples:

- self-reported smoking rates (e.g., Lewis-Esquerre et al. 2005, Wang and Heitjan 2008, Wang et al. 2012),
- age reported in multiples of 5 (e.g., Stockwell and Wicks 1974),
- duration of breastfeeding (e.g., Singh et al. 1994),
- household total expenditure (e.g., Browning et al. 2003),
- number of drug partners (e.g., Roberts and Brewer 2001),
- age at menopause (e.g., Crawford et al. 2002).

Also known as *digit preference* and *heaped data*, Crawford et al. (2015): "Inference for heaped data is an important statistical problem."

Heaping is reviewed in Wang et al. (2012). To date, methods and software for heaping have been largely inadequate.

There seems only one R package that deals with heaped data: Kernelheaping, and this is not for estimation.

Example(s)

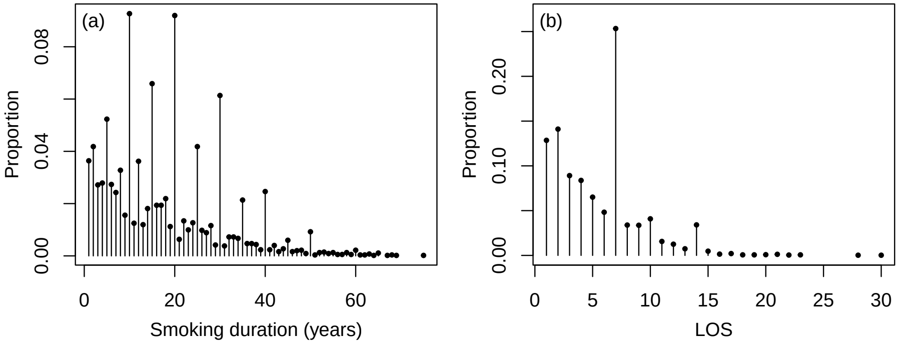


Figure: The raw data are capped and come from VGAMdata. (a) Spikeplot of smoking duration from 5525 current or past smokers in a large cross-sectional study, from `xs.nz`. (b) Flamingo Hotel length of stay proportions, from `flamingo`.

The GAITD Combo Model

The Main Idea

Combine alteration (hurdle), inflation, deflation and truncation into a single model.

Precedence of the operators:

1. Truncation first,
2. alteration next,
3. inflation and deflation last.

Allow *parametric* and *nonparametric* alteration, inflation and deflation. Hence 7 types of *special* values.

Wish to offer maximum flexibility $\Longrightarrow$ more usefulness.

All the Extensions

Let $\mathcal{R}$ be the *support* of the parent (base) distribution, e.g., $\{0, 1, \ldots\}$ for the Poisson. Extend previous work in three directions:

(I) Any subset of the support can be altered, inflated, deflated or truncated, cf. treating only the singleton $\{0\}$ as special. The first three are denoted $\mathcal{A}$, $\mathcal{I}$, and $\mathcal{D}$ with finite cardinality, but $|\mathcal{T}| = \infty$ permitted.

(II) Rather than allowing only one of $\mathcal{A}$, $\mathcal{I}$ $\mathcal{D}$ and $\mathcal{T}$, the operators are combined into a single model and are allowed to operate concurrently. The $\mathcal{A}$, $\mathcal{I}$, $\mathcal{D}$ and $\mathcal{T}$ are mutually disjoint.

(III) Utilizing (I) and (II) on $\mathcal{A}$ and $\mathcal{I}$, parametric and nonparametric forms are spawned. These are further combined into a single model, called the GAITD combo; $\mathcal{S} = \{\mathcal{A}_p,\ \mathcal{A}_{np}, \mathcal{I}_p,\ \mathcal{I}_{np}, \mathcal{D}_p,\ \mathcal{D}_{np}, \mathcal{T}\}$.

(IV) Although we develop (I)–(II) mainly for 1- and 2-parameter count parent distributions (Poisson, logarithmic, zeta, NB) our work is envisaged for continuous distributions.

Why General Truncation?†

1. In the lower and upper tails is obvious.
2. Also, e.g., asking for a favourite number between 0 and 20 say, it is likely that *tetraphobia* in East Asian culture and *triskaidekaphobia* in Western culture would show truncation or deficits of 4s and 13s respectively.

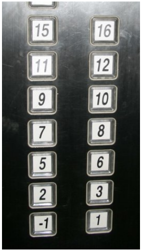


Figure: Triskaidekaphobia and other nasties. An elevator in an apartment building in Shanghai. Which floor number(s) are missing? Sources: Wiki and T. Jin.

GAITD Combo PMF

$$\Pr(Y_* = y; \boldsymbol{\theta}_\pi, \omega_p, \boldsymbol{\theta}_\alpha, \phi_p, \boldsymbol{\theta}_\iota, \psi_p, \boldsymbol{\theta}_\delta, \boldsymbol{\omega}_{np}, \boldsymbol{\phi}_{np}, \boldsymbol{\psi}_{np}) =$$

$$\begin{cases} 0, & y \in \mathcal{T}, \\ \omega_p \, f_\alpha(y) \,/ \sum\limits_{u \in \mathcal{A}_p} f_\alpha(u), & y \in \mathcal{A}_p, \\ \omega_s, & y = a_s \in \mathcal{A}_{np}, \;\; s = 1, \ldots, |\mathcal{A}_{np}|, \\ \Delta \, f_\pi(y) + \phi_p \, f_\iota(y) \,/ \sum\limits_{u \in \mathcal{I}_p} f_\iota(u), & y \in \mathcal{I}_p, \\ \Delta \, f_\pi(y) + \phi_s, & y = i_s \in \mathcal{I}_{np}, \;\; s = 1, \ldots, |\mathcal{I}_{np}|, \\ \Delta \, f_\pi(y) - \psi_p \, f_\delta(y) \,/ \sum\limits_{u \in \mathcal{D}_p} f_\delta(u), & y \in \mathcal{D}_p, \\ \Delta \, f_\pi(y) - \psi_s, & y = d_s \in \mathcal{D}_{np}, \;\; s = 1, \ldots, |\mathcal{D}_{np}|, \\ \Delta \, f_\pi(y), & y \in \mathcal{R} \backslash \mathcal{S}, \end{cases} \tag{1}$$

where the normalizing constant is

$$\Delta \quad = \quad \frac{1 - \omega_p - \phi_p + \psi_p - \sum\limits_{u=1}^{|\mathcal{A}_{np}|} \omega_u - \sum\limits_{u=1}^{|\mathcal{I}_{np}|} \phi_u + \sum\limits_{u=1}^{|\mathcal{D}_{np}|} \psi_u}{1 - \sum\limits_{a \in \{\mathcal{A}_p, \; \mathcal{A}_{np}\}} f_\pi(a) - \sum\limits_{t \in \mathcal{T}} f_\pi(t)}. \tag{2}$$

A 7-component mixture model with *nested* support & multinomial logit model!

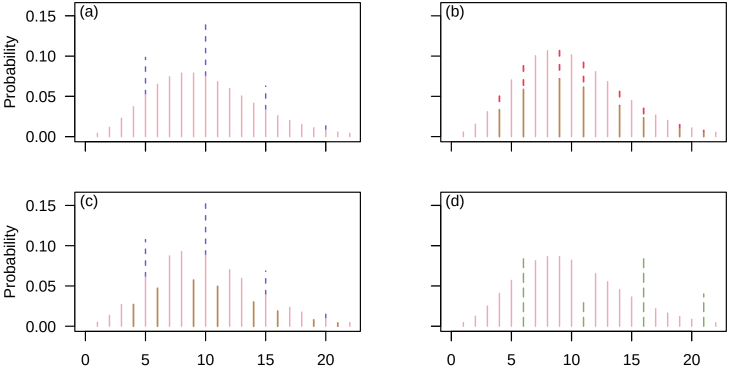


Figure: Heaped and/or seeped data—idealized forms in (a)–(c). Here, $\mathcal{I}_p = \{5, 10, 15, 20\}$, $\mathcal{D}_p = \{4, 6, 9, 11, 14, 16, 19, 21\}$, $\mathcal{T} = \{0\}$ with $\omega_p = 0.15$, $\psi_p = 0.15$, $\mu_\pi = 10$, $k_\pi = 10$ so that $f_\pi = f_\iota = f_\delta =$ NB(10, 10) PMF. (a) GIT–NB–NB; (b) GTD–NB–NB with the dip probabilities shown; (c) GITD-NB-NB-NB combines them together; (d) GAT-NB-MLM($\boldsymbol{\omega}_{np} = (0.09, 0.03, 0.09, 0.04)^T$). Colours: see p.36.

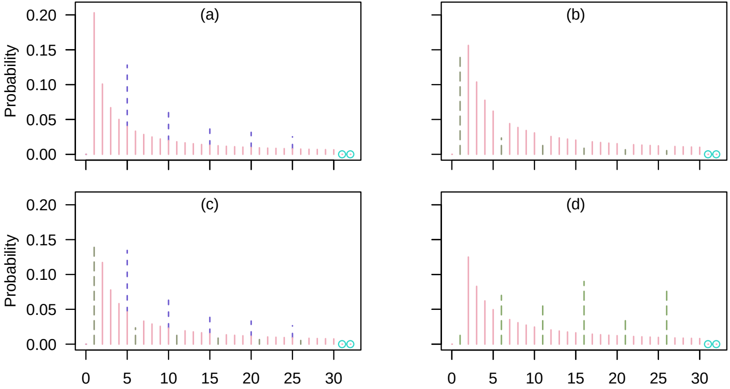


Figure: Heaped and/or seeped data—idealized forms in (a)–(c). Here, $\mathcal{I}_p = \{0, 5, 10, 15, 20, 25\}$, $\mathcal{A}_p$ or $\mathcal{A}_{np} = \{1, 6, 11, 16, 21, 26\}$, $\mathcal{T} = \{31, 32, \ldots\}$ with $\phi_p = 0.2$, $\omega_p = 0.2$, $\theta_\pi = 0.01$ is the shape parameter. (a) GIT–zeta–zeta; (b) GAT–zeta–zeta; (c) GAIT-zeta-zeta-zeta combines them together; (d) GAT-zeta-MLM($\boldsymbol{\omega}_{np} = (0.02, 0.07, 0.06, 0.09, 0.04, 0.08)^T$). Colours: see p.36.

Seven Modes

GAITD regression can handle bimodality, trimodality,. . . up to *7 modes*!

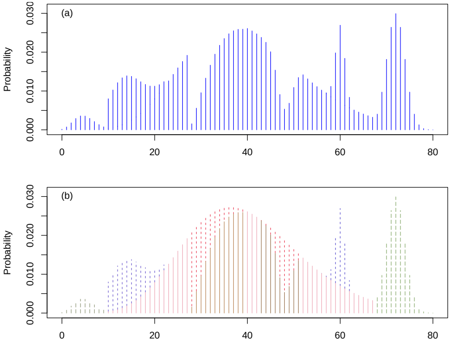


Figure: A GAITD–NB distribution with seven modes; (a) overall masked PMF; (b) PMF decomposed by the special values using colour and various line types, e.g., the dip probabilities appear in reddish dashed lines.

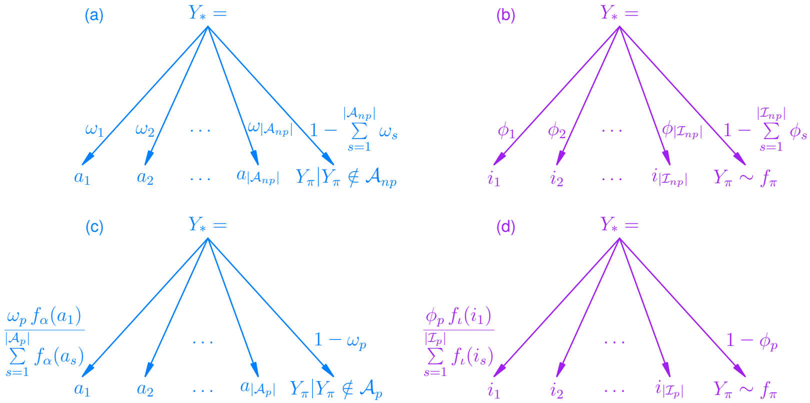

(a)
$Y_* =$
ω_1
ω_2
$\cdots$
$\omega_{|\mathcal{A}_{np}|}$
$1 - \sum_{s=1}^{|\mathcal{A}_{np}|} \omega_s$
a_1
a_2
$\cdots$
$a_{|\mathcal{A}_{np}|}$
$Y_\pi | Y_\pi \notin \mathcal{A}_{np}$
(b)
$Y_* =$
ϕ_1
ϕ_2
$\cdots$
$\phi_{|\mathcal{I}_{np}|}$
$1 - \sum_{s=1}^{|\mathcal{I}_{np}|} \phi_s$
i_1
i_2
$\cdots$
$i_{|\mathcal{I}_{np}|}$
$Y_\pi \sim f_\pi$
(c)
$Y_* =$
$\frac{\omega_p \, f_\alpha(a_1)}{\sum_{s=1}^{|\mathcal{A}_p|} f_\alpha(a_s)}$
$\cdots$
$1 - \omega_p$
a_1
a_2
$\cdots$
$a_{|\mathcal{A}_p|}$
$Y_\pi | Y_\pi \notin \mathcal{A}_p$
(d)
$Y_* =$
$\frac{\phi_p \, f_\iota(i_1)}{\sum_{s=1}^{|\mathcal{I}_p|} f_\iota(i_s)}$
$\cdots$
$1 - \phi_p$
i_1
i_2
$\cdots$
$i_{|\mathcal{I}_p|}$
$Y_\pi \sim f_\pi$

Figure: (a) GA–f_π–MLM, (b) GI–f_π–MLM, (c) GA–f_π–f_α, (d) GI–f_π–f_ι, where Y_π corresponds to the parent distribution and Y_* is the response of interest.

Notation and Nomenclature†

- Subscripts *p* for parametric, *np* for nonparametric.
- Operators and sets:
  - Alteration $\mathcal{A}_p$, $\mathcal{A}_{np}$
  - Inflation $\mathcal{I}_p$, $\mathcal{I}_{np}$
  - Truncation $\mathcal{T}_p$
  - Deflation $\mathcal{D}_p$, $\mathcal{D}_{np}$
- Parametric distributions:
  - Parent f_π
  - Altered (mixture) f_α
  - Inflated (mixture) f_ι
  - Deflated (mixture) f_δ
- Call ω_s a *heaping probability* if $\omega_s > f_\pi(a_s)$, else a *seeping probability*.
- Use $\omega_{\lceil a \rceil} = \{\omega_i : a_i = a\}$ so that $\sum_{a \in \mathcal{A}} a \cdot \omega_{\lceil a \rceil}$ equals $\sum_{i=1}^{L} a_i \cdot \omega_i$.
- $\mathcal{R}$ is the support, $\mathcal{S}$ are the special values.
- $\mathrm{I}(\cdot)$ is the indicator function, $\mathcal{A}$ has cardinality $|\mathcal{A}|$.

Answerable Questions

- *generally-altered regression* explains why observations are there,
- *generally-inflated regression* accounts for why they are there *in excess*, and
- *generally-deflated regression* explains why observations are *not* there.

Associated terms:

- underrepresentation versus overrrepresentation,
- misreported versus actual data.

Some Parent Distributions

t

Distribution	$\mathcal{R}$	Mean	PMF $f(y)$	Parameter space
Poisson	$0(1)\infty$	μ	$e^{-\theta}\,\theta^y/y!$	$0 < \theta$
Logarithmic	$1(1)\infty$	$\dfrac{\kappa\,\theta}{1-\theta}$	$\kappa\,\theta^y/y$	$0 < \theta < 1$
Zeta	$1(1)\infty$	$\dfrac{\zeta(\theta)}{\zeta(\theta+1)}$ if $1 < \theta$	$[y^{\theta+1}\cdot\zeta(\theta+1)]^{-1}$	$0 < \theta$
Neg. Binom.	$0(1)\infty$	μ	$\binom{y+k-1}{y}\dfrac{\mu^y k^k}{(\mu+k)^{y+k}}$	$0 < \mu$, $0 < k$

Table: GAITD distributions implemented in VGAM. Here, $\kappa = [-\log(1-\theta)]^{-1}$.

Note:

- In VGAMextra `logffMlink()` and `zetaffMlink()` enable $\eta_1 = \log E(Y)$.

Identifiability†

Avoiding degeneracy, the constraints on the parameter space needed for the PMF to be identifiable are

$$
\begin{aligned}
&0 < \phi_s \text{ for } s = 1, \ldots, |\mathcal{I}_{np}|, && 0 < \phi_p, \\
&0 < \psi_s \text{ for } s = 1, \ldots, |\mathcal{D}_{np}|, && 0 < \psi_p, \\
&0 < \omega_s \text{ for } s = 1, \ldots, |\mathcal{A}_{np}|, && 0 < \omega_p, \\
&\phi_p - \psi_p + \omega_p + \sum_{s=1}^{|\mathcal{I}_{np}|} \phi_s - \sum_{s=1}^{|\mathcal{D}_{np}|} \psi_s + \sum_{s=1}^{|\mathcal{A}_{np}|} \omega_s < 1, && |\mathcal{R} \backslash \mathcal{S}| > 0.
\end{aligned} \tag{3}
$$

- The last condition ensures that $\mathcal{R}$ cannot be inflated, deflated, altered or truncated, and guarantees that $\dim(\boldsymbol{\theta}_\pi) \geq 1$; and if $\widetilde{\mathcal{R}}$ is the support of the sample (i.e., set of all Y values) then the sample versions of the above must hold too. Practically however, $|\mathcal{R} \backslash \mathcal{S}| > 1$ to avoid a trivial regression.
- $\omega_p = 0$ is not permitted since otherwise $\mathcal{A}_p$ could be subsumed into $\mathcal{T}$, and a similar argument holds for $\phi_p = 0$, $\omega_s = 0$ and $\phi_s = 0$.

- The ZAP can arise in two ways: either $\mathcal{A}_p = \{0\}$ or $\mathcal{A}_{np} = \{0\}$; and likewise $\mathcal{I}_p = \{0\}$ or $\mathcal{I}_{np} = \{0\}$ for the ZIP. To ensure the parameters are identifiable one can further enforce

$$|\mathcal{A}_p| \neq 1, \qquad |\mathcal{I}_p| \neq 1 \;\text{ and }\; |\mathcal{D}_p| \neq 1. \tag{4}$$

Note: altered values are effectively disconnected from the rest of the data because they are only loosely coupled by a MLM. The remainder of the data are used to estimate the parameters $\boldsymbol{\theta}$.

The GT–Expansion Method

GAITD regression can easily handle underdispersion, e.g.,

- Multiply Y by 2 and truncate the odd values in between,
- Multiply Y by 3 and truncate all but multiples of 3 in between, etc.

Suggestion: transform to equidispersion or mild overdispersion.

It is a special case of the inverse location–scale transformation

$$Y_{**} \;=\; \nu + mY_*, \quad m \in \mathbb{Z}^+, \quad \nu \in \mathbb{Z}, \tag{5}$$

where usually the multiplier $m > 1$ is small and ν nonnegative.

Alternative example: Poisson f_π, if $\nu = 0$ then $\widehat{m} = \overline{y_*}/s^2_{y_*}$ is the moment estimator. Could round this or choose an integer $> \widehat{m}$.

See the Example 1 on p.39.

The GTE method might make the Conway–Maxwell distribution less important. Implementations of such include COMPoissonReg (Sellers et al. 2019) and mpcmp (Fung et al. 2020) will face competition!

Moments and CDF†

The kth moment is $\mathrm{E}[Y_*^k] =$

$$\frac{\omega_p \sum_{a\in\mathcal{A}_p} a^k \, f_\alpha(a)}{\sum_{a\in\mathcal{A}_p} f_\alpha(a)} + \frac{\phi_p \sum_{i\in\mathcal{I}_p} i^k \, f_\iota(i)}{\sum_{i\in\mathcal{I}_p} f_\iota(i)} - \frac{\psi_p \sum_{d\in\mathcal{D}_p} d^k \, f_\delta(d)}{\sum_{d\in\mathcal{D}_p} f_\delta(d)} + \sum_{a\in\mathcal{A}_{np}} a^k \, \omega_{\lceil a\rceil} + \sum_{i\in\mathcal{I}_{np}} i^k \, \phi_{\lceil i\rceil} - \sum_{d\in\mathcal{D}_{np}} d^k \, \psi_{\lceil d\rceil} + \Delta\cdot\left\{\mathrm{E}[Y_\pi^k] - \sum_{t\in\mathcal{T}} t^k \, f_\pi(t) - \sum_{a\in\{\mathcal{A}_p,\,\mathcal{A}_{np}\}} a^k \, f_\pi(a)\right\}. \tag{6}$$

The mean is returned as the `fitted()` values.

Let F_* be the GAIT cumulative distribution function (CDF) and F_π the CDF of the parent distribution. Then $F_*(y) =$

$$\omega_p \frac{\sum_{a\in\mathcal{A}_p} \mathrm{I}(a \le y)\, f_\alpha(a)}{\sum_{a\in\mathcal{A}_p} f_\alpha(a)} + \phi_p \frac{\sum_{i\in\mathcal{I}_p} \mathrm{I}(i \le y)\, f_\iota(i)}{\sum_{i\in\mathcal{I}_p} f_\iota(i)} - \psi_p \frac{\sum_{d\in\mathcal{D}_p} \mathrm{I}(d \le y)\, f_\delta(d)}{\sum_{d\in\mathcal{D}_p} f_\delta(d)} + \sum_{a\in\mathcal{A}_{np}} \mathrm{I}(a \le y)\, \omega_{\lceil a\rceil} + \sum_{i\in\mathcal{I}_{np}} \mathrm{I}(i \le y)\, \phi_{\lceil i\rceil} - \sum_{d\in\mathcal{D}_{np}} \mathrm{I}(d \le y)\, \psi_{\lceil d\rceil} + \Delta\cdot\left\{F_\pi(y) - \sum_{t\in\mathcal{T}} \mathrm{I}(t \le y)\, f_\pi(t) - \sum_{a\in\{\mathcal{A}_p,\,\mathcal{A}_{np}\}} \mathrm{I}(a \le y)\, f_\pi(a)\right\}. \tag{7}$$

Returned by, e.g., `pgaitdpois()`.

Estimation

Fisher Scoring and IRLS†

For the full GAITD combo model, $\mathcal{S} = \{\mathcal{A}_p, \mathcal{A}_{np}, \mathcal{I}_p, \mathcal{I}_{np}, \mathcal{T}, \mathcal{D}_p, \mathcal{D}_{np}\}$ and

$$\begin{aligned}
\ell &= \sum_{a \in \mathcal{A}_p} \mathrm{I}(y = a) \cdot \log\left[\omega_p\, A_\alpha(y)\right] + \sum_{a \in \mathcal{A}_{np}} \mathrm{I}(y = a) \cdot \log\, \omega_{\lceil a \rceil} + \\
&\quad \sum_{s=1}^{|\mathcal{I}_{np}|} \mathrm{I}(y = i_s) \cdot \log \Delta_s(y) + \sum_{i \in \mathcal{I}_p} \mathrm{I}(y = i) \cdot \log \Delta_p(y) + \\
&\quad \sum_{s=1}^{|\mathcal{D}_{np}|} \mathrm{I}(y = d_s) \cdot \log \Delta_s^-(y) + \sum_{d \in \mathcal{D}_p} \mathrm{I}(y = d) \cdot \log \Delta_p^-(y) + \mathrm{I}(y \notin \mathcal{S}) \cdot \log\left[\Upsilon B_\pi(y)\right], \qquad (8)
\end{aligned}$$

$$\begin{aligned}
\Upsilon &= 1 - \omega_p - \phi_p + \psi_p - \sum_{u=1}^{|\mathcal{A}_{np}|} \omega_u - \sum_{u=1}^{|\mathcal{I}_{np}|} \phi_u + \sum_{u=1}^{|\mathcal{D}_{np}|} \psi_u, \qquad (9) \\
\Delta &= \Upsilon / K_\pi, \qquad (10) \\
\Delta_v(y) &= \Upsilon\, B_\pi(y) + \phi_v \qquad y \in \mathcal{I}_{np}, \quad v = 1, \ldots, |\mathcal{I}_{np}|, \\
\Delta_p(y) &= \Upsilon\, B_\pi(y) + \phi_p\, A_\iota(y), \qquad y \in \mathcal{I}_p,
\end{aligned}$$

$$
\begin{aligned}
\Delta_v^-(y) &= \Upsilon\, B_\pi(y) - \psi_v \qquad y \in \mathcal{D}_{np}, \quad v = 1, \ldots, |\mathcal{D}_{np}|, \\
\Delta_p^-(y) &= \Upsilon\, B_\pi(y) - \psi_p\, A_\delta(y), \qquad y \in K_p, \\
A_\alpha(y) &= f_\alpha(y) / \sum_{u \in \mathcal{A}_p} f_\alpha(u) \;=\; f_\alpha(y)/K_\alpha, \qquad y \in \mathcal{A}_p, \\
A_\iota(y) &= f_\iota(y) / \sum_{u \in \mathcal{I}_p} f_\iota(u) \;=\; f_\iota(y)/K_\iota, \qquad y \in \mathcal{I}_p, \\
A_\delta(y) &= f_\delta(y) / \sum_{u \in \mathcal{D}_p} f_\delta(u) \;=\; f_\delta(y)/K_\delta, \qquad y \in \mathcal{D}_p, \\
K_\pi &= 1 - \sum_{a \in \{\mathcal{A}_{np},\, \mathcal{A}_{np}\}} f_\pi(a) - \sum_{t \in \mathcal{T}} f_\pi(t), \\
K_\alpha &= \sum_{a \in \mathcal{A}_p} f_\alpha(a), \qquad K_\iota \;=\; \sum_{i \in \mathcal{I}_p} f_\iota(i), \qquad K_\delta \;=\; \sum_{d \in \mathcal{D}_p} f_\delta(d), \\
A'_\alpha(y) &= f'_\alpha(y)/K_\alpha - f_\alpha(y)\, K'_\alpha / K^2_\alpha, \\
A''_\alpha(y) &= f''_\alpha(y)/K_\alpha - 2\, f'_\alpha(y)\, K'_\alpha / K^2_\alpha - f_\alpha(y)\, K''_\alpha / K^2_\alpha + 2\, f_\alpha(y)\, \left(K'_\alpha\right)^2 / K^3_\alpha, \\
B_\pi(y) &= f_\pi(y)/K_\pi, \qquad y \notin \{\mathcal{A}_{np}, \mathcal{A}_p, \mathcal{T}\}, \\
B'_\pi(y) &= \frac{f'_\pi(y)}{K_\pi} - \frac{f_\pi(y)\, K'_\pi}{K^2_\pi},
\end{aligned}
\tag{11}
$$

$$B''_{\pi}(y) \quad = \quad \frac{f''_{\pi}(y)}{K_{\pi}} - \frac{2\, f'_{\pi}(y)\, K'_{\pi}}{K_{\pi}^2} - \frac{f_{\pi}(y)\, K''_{\pi}}{K_{\pi}^2} + \frac{2\, f_{\pi}(y)\, \left(K'_{\pi}\right)^2}{K_{\pi}^3}.$$

Then

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta_j} \quad &= \quad \sum_{a \in \mathcal{A}_p} \mathrm{I}(y = a) \cdot \frac{A'_{\alpha}(y)}{A_{\alpha}(y)} + \mathrm{I}(y \notin \mathcal{S}) \cdot \frac{B'_{\pi}(y)}{B_{\pi}(y)} + \\
& \quad \sum_{i \in \mathcal{I}_p} \mathrm{I}(y = i) \cdot \frac{\Upsilon\, B'_{\pi}(y) + \phi_p\, A'_{\iota}(y)}{\Delta_p(y)} + \sum_{s=1}^{|\mathcal{I}_{np}|} \mathrm{I}(y = i_s) \cdot \frac{\Upsilon\, B'_{\pi}(i_s)}{\Delta_s(i_s)} + \\
& \quad \sum_{d \in \mathcal{D}_p} \mathrm{I}(y = d) \cdot \frac{\Upsilon\, B'_{\pi}(y) - \psi_p\, A'_{\delta}(y)}{\Delta_p^{-}(y)} + \sum_{s=1}^{|\mathcal{D}_{np}|} \mathrm{I}(y = d_s) \cdot \frac{\Upsilon\, B'_{\pi}(d_s)}{\Delta_s^{-}(d_s)}, \qquad j = \pi, \alpha, \iota, \delta, \\
\frac{\partial \ell}{\partial \phi_v} \quad &= \quad \sum_{s=1}^{|\mathcal{I}_{np}|} \mathrm{I}(y = i_s) \cdot \frac{\mathrm{I}(s = v) - B_{\pi}(y)}{\Delta_s(y)} - \sum_{s=1}^{|\mathcal{I}_{p}|} \mathrm{I}(y = i_s) \cdot \frac{B_{\pi}(y)}{\Delta_p(y)} - \\
& \quad \sum_{s=1}^{|\mathcal{D}_{np}|} \mathrm{I}(y = d_s) \cdot \frac{B_{\pi}(y)}{\Delta_s^{-}(y)} - \sum_{s=1}^{|\mathcal{D}_{p}|} \mathrm{I}(y = d_s) \cdot \frac{B_{\pi}(y)}{\Delta_p^{-}(y)} - \frac{\mathrm{I}(y \notin \mathcal{S})}{\Upsilon}, \qquad v = 1, \ldots, |\mathcal{I}_{np}|,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \psi_v} \;=\;& -\sum_{s=1}^{|\mathcal{D}_{np}|} \mathrm{I}(y = d_s) \cdot \frac{\mathrm{I}(s = v) - B_\pi(y)}{\Delta_s^-(y)} + \sum_{s=1}^{|\mathcal{D}_p|} \mathrm{I}(y = d_s) \cdot \frac{B_\pi(y)}{\Delta_p^-(y)} + \\
& \sum_{i \in \mathcal{I}_p} \mathrm{I}(y = i) \cdot \frac{B_\pi(y)}{\Delta_p(y)} + \sum_{s=1}^{|\mathcal{I}_{np}|} \frac{\mathrm{I}(y = i_s)\, B_\pi(y)}{\Delta_s(y)} + \frac{\mathrm{I}(y \notin \mathcal{S})}{\Upsilon}, \qquad v = 1, \ldots, |\mathcal{D}_{np}|, \\
\frac{\partial \ell}{\partial \omega_v} \;=\;& \frac{\mathrm{I}(y = a_v)}{\omega_v} - \sum_{s=1}^{|\mathcal{I}_{np}|} \frac{\mathrm{I}(y = i_s) \cdot B_\pi(y)}{\Delta_s(y)} - \sum_{s=1}^{|\mathcal{I}_p|} \mathrm{I}(y = i_s) \cdot \frac{B_\pi(y)}{\Delta_p(y)} - \\
& \sum_{d \in \mathcal{D}_p} \mathrm{I}(y = d) \cdot \frac{B_\pi(y)}{\Delta_p^-(y)} - \sum_{s=1}^{|\mathcal{D}_{np}|} \frac{\mathrm{I}(y = d_s)\, B_\pi(y)}{\Delta_s^-(y)} - \frac{\mathrm{I}(y \notin \mathcal{S})}{\Upsilon}, \qquad v = 1, \ldots, |\mathcal{A}_{np}|, \\
\frac{\partial \ell}{\partial \omega_p} \;=\;& \sum_{a \in \mathcal{A}_p} \frac{\mathrm{I}(y = a)}{\omega_p} - \sum_{i \in \mathcal{I}_p} \frac{\mathrm{I}(y = i)\, B_\pi(y)}{\Delta_p(y)} - \sum_{s=1}^{|\mathcal{I}_{np}|} \frac{\mathrm{I}(y = i_s)\, B_\pi(y)}{\Delta_s(y)} - \\
& \sum_{d \in \mathcal{D}_p} \mathrm{I}(y = d) \cdot \frac{B_\pi(y)}{\Delta_p^-(y)} - \sum_{s=1}^{|\mathcal{D}_{np}|} \frac{\mathrm{I}(y = d_s)\, B_\pi(y)}{\Delta_s^-(y)} - \frac{\mathrm{I}(y \notin \mathcal{S})}{\Upsilon},
\end{aligned}$$

$$
\begin{aligned}
-\,\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \theta_j^2}\right) \quad = \quad & \sum_{i \in \mathcal{I}_p} \frac{1}{\Delta_p(i)} \left\{\Upsilon\, B_\pi'(i) + \phi_p\, A_\iota'(i)\right\}^2 - \sum_{i \in \mathcal{I}_p} \left\{\Upsilon\, B_\pi''(i) + \phi_p\, A_\iota''(i)\right\} + \\
& \sum_{s=1}^{|\mathcal{I}_{np}|} \frac{[\Upsilon\, B_\pi'(i_s)]^2}{\Delta_s(i_s)} - \sum_{s=1}^{|\mathcal{I}_{np}|} \Upsilon\, B_\pi''(i_s) + \omega_p \left[\frac{1}{K_\alpha} \sum_{a \in \mathcal{A}_p} \frac{[f_\alpha'(a)]^2}{f_\alpha(a)} - \left(\frac{K_\alpha'}{K_\alpha}\right)^2 \right] + \\
& \sum_{d \in \mathcal{D}_p} \frac{1}{\Delta_p^-(d)} \left\{\Upsilon\, B_\pi'(d) - \psi_p\, A_\delta'(d)\right\}^2 - \sum_{d \in \mathcal{D}_p} \left\{\Upsilon\, B_\pi''(d) - \psi_p\, A_\delta''(d)\right\} + \\
& \sum_{s=1}^{|\mathcal{D}_{np}|} \frac{[\Upsilon\, B_\pi'(d_s)]^2}{\Delta_s^-(d_s)} - \sum_{s=1}^{|\mathcal{D}_{np}|} \Upsilon\, B_\pi''(d_s) + \\
& \Pr(y \notin \mathcal{S}) \cdot \left\{ -\,\mathrm{E}_c\, \frac{f_\pi''(y)}{f_\pi(y)} + \mathrm{E}_c \left(\frac{f_\pi'(y)}{f_\pi(y)}\right)^2 + \frac{K_\pi''}{K_\pi} - \left(\frac{K_\pi'}{K_\pi}\right)^2 \right\}, \\
\mathrm{E}\left(\frac{-\partial^2 \ell}{\partial \phi_u\, \partial \phi_v}\right) \quad = \quad & \sum_{s=1}^{|\mathcal{I}_{np}|} \frac{[\mathrm{I}(s=u) - B_\pi(i_s)]\,[\mathrm{I}(s=v) - B_\pi(i_s)]}{\Delta_s(i_s)} + \sum_{s=1}^{|\mathcal{I}_p|} \frac{B_\pi^2(i_s)}{\Delta_p(i_s)} + \\
& \sum_{s=1}^{|\mathcal{D}_{np}|} \frac{B_\pi^2(d_s)}{\Delta_s^-(d_s)} + \sum_{s=1}^{|\mathcal{D}_p|} \frac{B_\pi^2(d_s)}{\Delta_p^-(d_s)} + \frac{\Pr(y \notin \mathcal{S})}{\Upsilon^2}, \qquad \text{etc., etc.}
\end{aligned}
$$

Some notes:

1. It is readily shown that $\Pr(y \notin \mathcal{S}) =$

$$\begin{aligned}
&\Upsilon\left(1 - K_\pi^{-1} \cdot \left[\sum_{i\in\{\mathcal{I}_{np},\,\mathcal{I}_p\}} f_\pi(i) + \sum_{d\in\{\mathcal{D}_{np},\,\mathcal{D}_p\}} f_\pi(d)\right]\right) \\
= \quad & 1 - \omega_p - \sum_{s=1}^{|\mathcal{A}_{np}|} \omega_s - \sum_{s=1}^{|\mathcal{I}_{np}|} \Delta_s(i_s) - \sum_{i\in\mathcal{I}_p} \Delta_p(i) - \sum_{s=1}^{|\mathcal{D}_{np}|} \Delta_s^-(d_s) - \sum_{d\in\mathcal{D}_p} \Delta_p^-(d). \qquad (12)
\end{aligned}$$

2. In general, the conditional expectations over $\mathcal{R}\backslash\mathcal{S}$ used is

$$\mathrm{E}\left[g(Y_\pi)|Y_\pi \notin \mathcal{S}\right] = \frac{\mathrm{E}\left[g(Y_\pi)\right] - \sum_{s\in\mathcal{S}} s\, f_\pi(s)}{\left(1 - \sum_{s\in\mathcal{S}} f_\pi(s)\right)} \qquad (13)$$

for some function $g(Y)$, since $\mathcal{S}$ takes on the role of $\mathcal{T}$.

The MLM and all the Etas

Recall the *multinomial logit model* (*MLM*) for probabilities $\boldsymbol{p} = (p_1, \ldots, p_D)^T$ has $g = \text{multilogit}(p_1, \ldots, p_D)$ given by

$$g(p_s) \;=\; \eta_s \;\;=\;\; \log\left\{p_s \Big/ \left(1 - \sum_{u=1}^{D} p_u\right)\right\}, \qquad s = 1, \ldots, D, \tag{14}$$

so that $p_{D+1} = 1 - p_1 - \cdots - p_D$ (reference group). The inverse link (*softmax*) is $p_s = e^{\eta_s} / \sum_{u=1}^{D+1} e^{\eta_u}$ where $\eta_{D+1} \equiv 0$.

Fitted as a VGLM for 1-parameter distributions, $\boldsymbol{\eta}^T =$

$$\Big(g_{\pi}(\theta_{\pi}),\; \log\frac{\omega_p}{\mathcal{N}},\; g_{\alpha}(\theta_{\alpha}),\; \log\frac{\phi_p}{\mathcal{N}},\; g_{\iota}(\theta_{\iota}),\; \log\frac{\psi_p}{\mathcal{N}},\; g_{\delta}(\theta_{\delta}),\; \log\frac{\omega_1}{\mathcal{N}},\; \ldots,\; \log\frac{\omega_{LA}}{\mathcal{N}},\; \log\frac{\phi_1}{\mathcal{N}},\; \ldots,\; \log\frac{\phi_{LI}}{\mathcal{N}},\; \log\frac{\psi_1}{\mathcal{N}},\; \ldots,\; \log\frac{\psi_{LD}}{\mathcal{N}} \Big) \tag{15}$$

where $g_{\cdot}(\cdot)$ are the links, $LA = |\mathcal{A}_{np}|$, $LI = |\mathcal{I}_{np}|$, $LD = |\mathcal{D}_{np}|$ and $\mathcal{N} = 1 - \omega_p - \phi_p - \psi_p - \sum_{u=1}^{|\mathcal{A}_{np}|} \omega_u - \sum_{u=1}^{|\mathcal{I}_{np}|} \phi_u - \sum_{u=1}^{|\mathcal{D}_{np}|} \psi_u$.

Change of variable†

All parameters except for $\boldsymbol{\theta}_\pi$, $\boldsymbol{\theta}_\alpha$, $\boldsymbol{\theta}_\iota$ and $\boldsymbol{\theta}_\delta$ are estimated by a MLM so need to apply a change of variable from the given EIM to $\boldsymbol{\eta}$. Writing $\boldsymbol{p} = (\omega_p,\, \phi_p,\, \omega_1,\, \ldots,$ $\omega_{|\mathcal{A}_{np}|},\, \phi_1,\, \ldots,\, \phi_{|\mathcal{I}_{np}|})^T$, then

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\eta}} &= \left(\mathrm{Diag}(\boldsymbol{p}) - \boldsymbol{p}\,\boldsymbol{p}^T\right) \frac{\partial \ell}{\partial \boldsymbol{p}}, \\
\boldsymbol{\mathcal{I}}_E(\boldsymbol{\eta}) &= \left(\mathrm{Diag}(\boldsymbol{p}) - \boldsymbol{p}\,\boldsymbol{p}^T\right) \mathrm{E}\left(\frac{-\partial^2 \ell}{\partial \boldsymbol{p}\, \partial \boldsymbol{p}^T} \right) \left(\mathrm{Diag}(\boldsymbol{p}) - \boldsymbol{p}\,\boldsymbol{p}^T\right).
\end{aligned}$$

Efficiently compute the latter by exploiting structure in $\mathrm{Diag}(\boldsymbol{p}) - \boldsymbol{p}\,\boldsymbol{p}^T$:

$$[\boldsymbol{\mathcal{I}}_E(\boldsymbol{\eta})]_{u,v} \;=\; p_u\, p_v \left\{ I_{uv} - \sum_{s=1}^{M} p_s\,(I_{us} + I_{vs}) + \sum_{s=1}^{M} p_s^2\, I_{ss} + 2 \sum_{s<t} p_s\, p_t\, I_{st} \right\} \quad (16)$$

(where I_{st} is the (s,t) element of $\boldsymbol{\mathcal{I}}_E(\boldsymbol{p})$)—it involves computing only a single quadratic form $\boldsymbol{p}^T \boldsymbol{\mathcal{I}}_E(\boldsymbol{p})\, \boldsymbol{p}$.

Upper Tail Truncation†

For the GT–Poisson, one can handle truncation past U:

```
vglm(..., gaitdpoisson(truncate = 0:4, max.support = 20))
```

The support of Y is 5(1)20 but `truncate = c(0:4, 21:Inf)` is impractical. To handle `max.support = U` use

$$\sum_{t=U+1}^{\infty} t\, f_{\pi}(t) \;=\; \lambda_{\pi} \left[1 - e^{-\lambda_{\pi}} \sum_{t=0}^{U-1} \frac{\lambda_{\pi}^{t}}{t!} \right], \quad \sum_{t=U+1}^{\infty} t(t-1)\, f_{\pi}(t) \;=\; \lambda_{\pi}^{2} \left[1 - e^{-\lambda_{\pi}} \sum_{t=0}^{U-2} \frac{\lambda_{\pi}^{t}}{t!} \right],$$

$$\sum_{t=U+1}^{\infty} f_{\pi}'(t) \;=\; f_{\pi}(U), \qquad \sum_{t=U+1}^{\infty} f_{\pi}''(t) \;=\; f_{\pi}(U-1) - f_{\pi}(U).$$

Some notes:

- Tractable formulas also exist for the Logarithmic(θ) but not for Zeta(θ) or NB(μ, k).
- No need for `min.support`, of course.

Initial Values†

- Initial values $\omega_p^{(0)}$ and $\omega_s^{(0)}$ easily obtained from the MLEs (sample proportions) assuming intercept-only, but shrinking them towards 0 is better.
- Good starting values $\phi_p^{(0)}$ and $\phi_s^{(0)}$ are difficult (confounded with the scaled f_π). A grid-search works well assumping that the $\phi_s^{(0)}$ are equal.
- Typically for a well-specified model, the Fisher scoring/IRLS algorithm converges within 6–8 iterations, like ordinary GLMs.
- The computations may be subject to numerical problems if the special values are extremely remote in the support, e.g., for Pois($\lambda = 10$), $i_s > 303$ on most machines. If to machine precision $f_\pi(i_s)$ is evaluated as 0 then naive programming will lead to 0/0 being computed. But possible to give warnings and take corrective action against such possibilities.

VGAM Software

VGAM 1.1-9 on CRAN now has the following functions.

t

Distribution	R functions	VGAM family function
GAITD–Pois–Pois–Pois–Pois	[dpqr]gaitdpois()	gaitdpoisson()
GAITD–Log–Log–Log–Log	[dpqr]gaitdlog()	gaitdlog()
GAITD–Zeta–Zeta–Zeta–Zeta	[dpqr]gaitdzeta()	gaitdzeta()
GAITD–NB–NB–NB–NB	[dpqr]gaitdnbinom()	gaitdnbinomial()

Table: Prefix “d” = density, “p” = CDF, “q” = inverse CDF, “r” = random deviates.

t

R function	Comments
`spikeplot(y)`	Spike-plots a data vector `y`. **Colours**: **p**arent is **p**ink, **t**runcated are **t**urquoise hollow circles, **a**ltered probabilities are **a**vocado, **i**nflated probabilities are **i**ndigo, **d**eflated probabilities are **d**eer.
`plotdgaitd(fit)`	Spike-plots the PMF of `fit`.
`dgaitdplot()`	Spike-plots the PMF, given $f.$, $\mathcal{A}_p$, $\mathcal{A}_{np}$, $\mathcal{I}_p$, $\mathcal{I}_{np}$, $\mathcal{T}$, $\mathcal{D}_p$, $\mathcal{D}_{np}$.
`rootogram4(fit)`	Rootograms (e.g., hanging, suspended, etc.) of `fit`.
`goffset()`	For the GTE method: offset matrix.
`Trunc(Range, mux = 2)`	For the GTE method: returns a vector of values from `Range[1]` to `Range[2]` that are not multiples of `mux`.

Table: Supporting functions for GAITD regression.

```
> args(gaitdpoisson)

function (a.mix = NULL, i.mix = NULL, d.mix = NULL, a.mlm = NULL,
    i.mlm = NULL, d.mlm = NULL, truncate = NULL, max.support = Inf,
    zero = c("pobs", "pstr", "pdip"), eq.ap = TRUE, eq.ip = TRUE,
    eq.dp = TRUE, parallel.a = FALSE, parallel.i = FALSE, parallel.d = FALSE,
    llambda.p = "loglink", llambda.a = llambda.p, llambda.i = llambda.p,
    llambda.d = llambda.p, type.fitted = c("mean", "lambdas",
        "pobs.mlm", "pstr.mlm", "pdip.mlm", "pobs.mix", "pstr.mix",
        "pdip.mix", "Pobs.mix", "Pstr.mix", "Pdip.mix", "nonspecial",
        "Numer", "Denom.p", "sum.mlm.i", "sum.mix.i", "sum.mlm.d",
        "sum.mix.d", "ptrunc.p", "cdf.max.s"), gpstr.mix = ppoints(7)/3,
    gpstr.mlm = ppoints(7)/(3 + length(i.mlm)), imethod = 1,
    mux.init = c(0.75, 0.5, 0.75), ilambda.p = NULL, ilambda.a = ilambda.p,
    ilambda.i = ilambda.p, ilambda.d = ilambda.p, ipobs.mix = NULL,
    ipstr.mix = NULL, ipdip.mix = NULL, ipobs.mlm = NULL, ipstr.mlm = NULL,
    ipdip.mlm = NULL, byrow.aid = FALSE, ishrinkage = 0.95, probs.y = 0.35)
NULL
```

Notes:

- `NULL` $= \{\}$. The first arguments are for $\mathcal{A}_p$, $\mathcal{I}_p$, $\mathcal{D}_p$, $\mathcal{A}_{np}$, $\mathcal{I}_{np}$, $\mathcal{D}_{np}$, $\mathcal{T}$.
- `eq.ap`, `eq.ip`, `eq.dp` are logical, e.g., $\lambda_\pi = \lambda_\alpha$ in the GAITD–Poisson.
- `parallel.ap`, `parallel.ip`, `parallel.dp` refer to the MLM, e.g., all ω_s are equal.

Table: Argument `type.fitted` and fitted values for GAITD models. Many of the fitted values are terms in the combo PMF (1). Not all options are applicable for any particular fitted object. **Warning**: these details are subject to future change.

Argument	Fitted value
`"mean"`	μ, $n \times 1$, Eqn. (6) with $k = 1$, the default
`"pobs.mix"`	$\Pr(y \in \mathcal{A}_p)$, ω_p, $n \times 1$
`"pstr.mix"`	ϕ_p, $n \times 1$
`"pdip.mix"`	ψ_p, $n \times 1$
`"pobs.mlm"`	$\Pr(y \in \mathcal{A}_{np})$, ω_s, $n\times$ `length(a.mlm)`
`"pstr.mlm"`	ϕ_s, $n\times$ `length(i.mlm)`
`"pdip.mlm"`	ψ_s, $n\times$ `length(d.mlm)`
`"ptrunc.p"`	$\sum_{t\in\mathcal{T}} f_\pi(t)$ including truncated values in the upper tail $>$ `max.support`
`"cdf.max.s"`	$F_\pi($`max.support`$)$
`"Pobs.mix"`	$\Pr(y \in \mathcal{A}_p)$, $\omega_p\, f_\alpha(y) \,/ \sum_{u\in\mathcal{A}_p} f_\alpha(u)$, $n\times$ `length(a.mix)`
`"Pstr.mix"`	$\phi_p\, f_\iota(y) \,/ \sum_{u\in\mathcal{I}_p} f_\iota(u)$, $n\times$ `length(i.mix)`
`"Pdip.mix"`	$\psi_p\, f_\delta(y) \,/ \sum_{u\in\mathcal{D}_p} f_\delta(u)$, $n\times$ `length(d.mix)`
`"sum.mix.i"`	$\Pr(y \in \mathcal{I}_p)$, $n\times$ `length(i.mix)`, 4th line in (1)
`"sum.mlm.i"`	$\Pr(y \in \mathcal{I}_{np})$, $n\times$ `length(i.mlm)`, 5th line in (1)
`"Numer"`	Numerator, Υ, see (9), $n \times 1$
`"Denom"`	Denominator, used in (1), $1 - \sum_{a\in\{\mathcal{A}_p,\,\mathcal{A}_{np}\}} f_\pi(a) - \sum_{t\in\mathcal{T}} f_\pi(t)$, $n \times 1$
`"nonspecial"`	$\Pr(y \notin \mathcal{S})$, see (12), $n \times 1$

Example 1. GTE Method and Heaping on Sleep Duration

The NZ cross-sectional data `xs.nz` in VGAMdata has a variable called `sleep` giving the self-reported sleep hours to the question *"How many hours do you usually sleep each night?"*

After removing the `NA`s and outliers (2.4%) then $n = 10264$ individuals.

Table: Usual sleep duration in a New Zealand cross-sectional data.

Hours	3	4	5	6	7	8	9	10	11	12
Frequency	16	125	443	1760	3076	3766	891	170	10	7

```
> data("xs.nz", package = "VGAMdata")
> sxs.nz <- subset(xs.nz, !is.na(sleep))  # Remove missing values
> sleep.min <-  3  # Smallest value allowed
> sleep.max <- 12  # Largest  value allowed; Remove outliers
> sxs.nz <- subset(sxs.nz, sleep.min <= sleep & sleep <= sleep.max)
> with(sxs.nz, table(sleep))

sleep
   3    4    5    6    7    8    9   10   11   12
  16  125  443 1760 3076 3766  891  170   10    7

> with(sxs.nz, spikeplot(sleep))
```

See p.46. The data are left-skewed.

The data is strongly underdispersed with respect to the Poisson: the sample mean and variance are 7.297 and 1.286.

Let $y \in \{3, \ldots, 12\}$ so we chose $\mathcal{T} = \{0, 1, 2, 13, 14, \ldots\}$ to account for the absence of such values from the data set.

Firstly we search for the optimal multiplier *m*.

```
org1 <- with(sxs.nz, range(sleep))  # Original data range
m.max <- 8  # Try multipliers 1:m.max
aics <- logliks <- numeric(m.max)
allfits <- vector("list", m.max)
names(allfits) <- names(aics) <-
names(logliks) <- as.character(1:m.max)
for (mux in 1:m.max) {   # seq(m.max)
  fit <- vglm(mux * sleep ~ 1,   # trace = TRUE,
              gaitdpoisson(truncate = c(0:(sleep.min * mux - 1),
                                         Trunc(org1, mux)),
                           max.support = sleep.max * mux,
                           i.mlm = 8 * mux),
              offset = rep(log(mux), nrow(sxs.nz)),
              data = sxs.nz)
  allfits[[mux]] <- fit
  logliks[mux] <- logLik(fit)
  aics[mux] <- AIC(fit)
}
```

The results are

```
> (mux.best <- as.vector(which.max(logliks)))  # 5 is best

[1] 5
```

```
> plot(logliks, col = "blue", type = "b", xlab = "Multiplier")
> abline(v = 5, h = logliks[5], col = 'orange', lty = "dashed")
```

The LHS of the plot is where the Poisson is used to fit underdispersed data, while the RHS is where the transformed data is overdispersed.

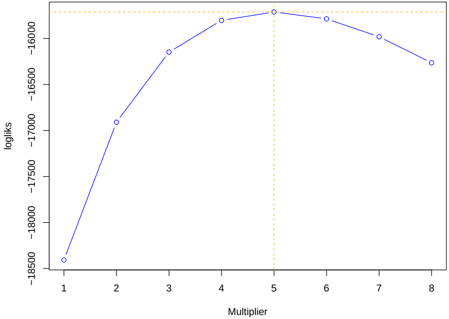


Figure: Log-likelihood ℓ versus multiplier m for several GAITD-Poisson fits according to the GTE method.

Incidentally, in contrast, the method of moments estimator for m is

```
> with(sxs.nz, mean(sleep) / var(sleep))

[1] 5.672
```

We select the best fit and then continue on the analysis.

```
> fit1 <- allfits[[mux.best]]
> with(sxs.nz, spikeplot(mux.best * sleep, col = "red", lwd = 2,
                         xlab = "Expanded response", ylab = "",
                         xlim = c(mux.best * sleep.min,
                                  mux.best * sleep.max)))
> plotdgaitd(fit1, new.plot = FALSE, offset.x = 0.5,
             all.lwd = 2, col.p = "blue")
```

See p.46. The GAITD-Pois fit appears close to the observed proportions.

```
> logLik(fit1)

[1] -15712

> coef(fit1, matrix = TRUE)

            loglink(lambda.p) multilogitlink(pstr.mlm40)
(Intercept)             1.969                     -3.292

> head(fitted(fit1, type = "pstr.mlm"), 2)

         40
[1,] 0.1568
[2,] 0.1568

> KLD(fit1)

[1] 4.573
```

The fitted model indicates that the amount of inflation at 8 is about $\widehat{\phi}_{\lceil 8 \rceil} \approx 0.1568$—almost 1/6 of the entire data set. The Kullback-Leibler divergence is 4.573 relative to an ordinary Poisson.
With regard to the mean sleep duration, we have

```
> head(fitted(fit1), 1) / mux.best

   [,1]
2 7.297

> exp(confint(fit1)[1, ])

 2.5 % 97.5 %
 7.139  7.194
```

That is, the overall GAITD mean is estimated by 7.297 hours whereas an approximate 95% confidence interval for μ_π, the mean sleep of the Poisson parent parameter, is $[7.139, 7.194]$ hours.

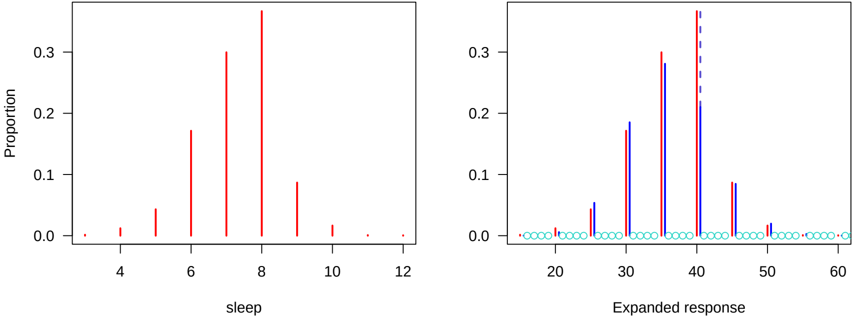


Figure: (a) Spikeplot of `sleep` in `xs.nz` from VGAMdata; (b) GTE method with GAITD-Poisson fit is overlaid in blue—the truncated values are turquoise hollow points.

Examples

Example 2. Smoking Duration

Use `xs.nz` in VGAMdata, a NZ prospective observational study in the 1990s. Have $n = 10,529$ as an approximate random sample of the working population. Let $Y =$ smoking duration (years).

```
xs.nz <- transform(xs.nz, roundsmokeyears = round(smokeyears))
smoke.df <- subset(xs.nz, roundsmokeyears > 0 &
                          !is.na(smokeyears) &
                          !is.na(ethnicity) &
                          !is.na(sex))
smoke.df <- transform(smoke.df, smokeyears = roundsmokeyears)
```

Then $n = 5492$. Let's spikeplot the data.

```
mylwd <- 1.5
myxlab <- "Smoking duration (years)"
with(xs.nz, spikeplot(smokeyears, lwd = mylwd, las = 1, xlab = myxlab))
put.caption("(a)", w.x = c(0.5, 0.5))
myylim <- c(0, 0.11)
with(smoke.df, spikeplot(smokeyears, lwd = mylwd, las = 1,
                         xlab = myxlab, ylim = myylim))
put.caption("(b)", w.x = c(0.5, 0.5))
```

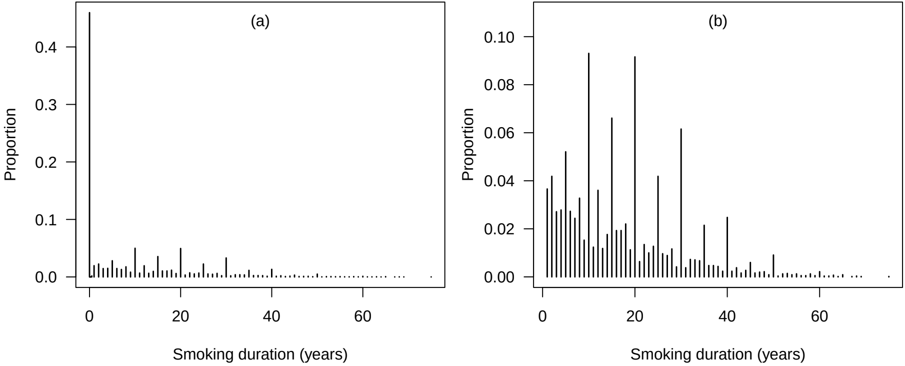


Figure: Spikeplot of `smokeyears` from VGAMdata. (a) From `xs.nz`; (b) From the subset `smoke.df`.

The most pronounced heaped values include 5, 10, 20, 30, 40, 50, 60, as well as 12, 25, 35. A careful examination also suggests that 9, 11, 13, 19, 21, 29, 31 are seeped.

Let's fit an intercept-only GAITD regression. Because the nonsmokers are such a large group it is necessary to model `smoke.df` only—we run out of baseline reserve probability. We choose

$$\begin{aligned} \mathcal{I}_p &= \{5, 10, 20, 30, 40, 50, 60\}, \quad \mathcal{A}_p = \{2, 15, 25, 35, 45\}, \\ \mathcal{I}_{np} &= \{1, 8, 12, 18\}, \quad \mathcal{D}_p = \{9, 11, 13, 19, 21, 29, 31\}, \quad \mathcal{T} = \{0\}. \end{aligned}$$

We use a NB parent to handle overdispersion and we relax the assumptions that the altered and inflated distributions are equal to the parent.

```
i.mix <- c(5, 10, 20, 30, 40, 50, 60)
a.mix <- c(2, 15, 25, 35, 45)
i.mlm <- c(1, 8, 12, 18)
d.mix <- c(9, 11, 13, 19, 21, 29, 31)
tvec <- 0
fit1.sy <-
  vglm(smokeyears ~ 1,
       gaitdnbinomial(i.mix = i.mix, i.mlm = i.mlm, a.mix = a.mix,
                      eq.dp = FALSE,  # This line is good
                      eq.ip = FALSE, eq.ap = FALSE,  # This line is good
                      d.mix = d.mix, truncate = tvec),
       crit = "coef", trace = FALSE, data = smoke.df)
```

The acceptable number of IRLS iterations needed for convergence is suggestive that the model fits the data reasonably well. In fact, changing to `eq.dp = TRUE` decreases the number of iterations to a reasonable number.

The following plot shows a good correspondence between the model and data. To conserve the baseline reserve probability, $\mathcal{I}_p$ was used to model the layer of largest spikes while $\mathcal{A}_p$ for the inner layer.

```
mylwd <- 1.5
with(smoke.df, spikeplot(smokeyears, las = 1, lwd = mylwd,
                         xlim = c(0, 59), xlab = myxlab, ylim = myylim))
plotdgaitd(fit1.sy, new.plot = FALSE, offset.x = 0.33,
           all.lwd = mylwd, deflation = TRUE)
```

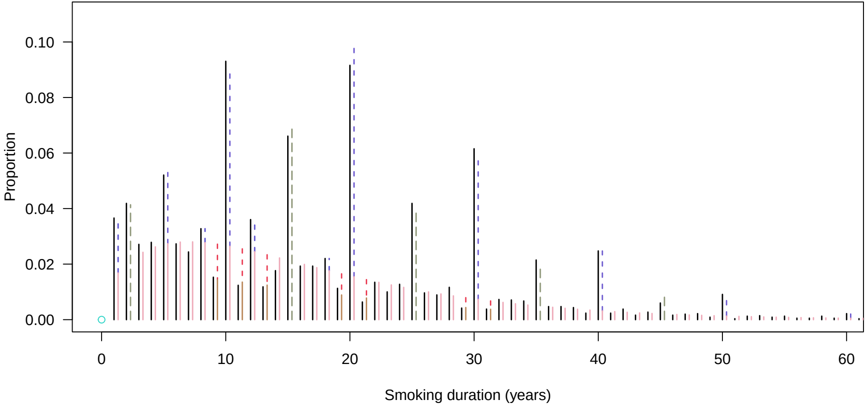


Figure: How the GAITD regression and `smoke.df` compare for `smokeyears`.

```
> t(coef(fit1.sy, matrix = TRUE))

                            (Intercept)
loglink(munb.p)                  2.7806
loglink(size.p)                  0.5920
multilogitlink(pobs.mix)        -0.9801
loglink(munb.a)                  2.9109
loglink(size.a)                  0.7901
multilogitlink(pstr.mix)        -0.6258
loglink(munb.i)                  3.1248
loglink(size.i)                  1.4627
multilogitlink(pdip.mix)        -2.1238
loglink(munb.d)                  2.7571
loglink(size.d)                  0.6099
multilogitlink(pstr.mlm1)       -3.1748
multilogitlink(pstr.mlm8)       -4.5433
multilogitlink(pstr.mlm12)      -3.7095
multilogitlink(pstr.mlm18)      -4.6729
```

And some more output:

```
> head(fitted(fit1.sy, type.fitted = "pobs.mix"), 1)

      [,1]
[1,] 0.1773

> head(fitted(fit1.sy, type.fitted = "Pobs.mix"), 1)

           2      15      25      35       45
[1,] 0.04138 0.06905 0.03979 0.01894 0.008189

> head(fitted(fit1.sy, type.fitted = "pstr.mix"), 1)

      [,1]
[1,] 0.2528

> head(fitted(fit1.sy, type.fitted = "Pstr.mix"), 1)

          5      10      20      30      40       50       60
[1,] 0.0256 0.06338 0.08224 0.04987 0.02156 0.007689 0.002422

> #head(fitted(fit1.sy, type.fitted = "pdip.mix"), 1)
> #head(fitted(fit1.sy, type.fitted = "Pdip.mix"), 1)
> head(fitted(fit1.sy, type.fitted = "nonspecial"), 1)

      [,1]
[1,] 0.2957
```

```
> head(cbind(smoke.df$smokeyears, fitted(fit1.sy, type.fitted = "munbs")))

       munb.p munb.a munb.i munb.d
[1,] 17  16.13  18.37  22.75  15.75
[2,] 12  16.13  18.37  22.75  15.75
[3,]  8  16.13  18.37  22.75  15.75
[4,]  3  16.13  18.37  22.75  15.75
[5,] 17  16.13  18.37  22.75  15.75
[6,]  9  16.13  18.37  22.75  15.75

> #(pdip.hat <- c(fitted(fit1.sy, type.fitted = "pdip.mix"))[1])
> #(pns.hat <- c(mean(fitted(fit1.sy, type.fitted = "nonspecial"))))
> (pheapseep <- Pheapseep(fit1.sy))  # \Xi

[1] 0.4108
```

That is, for the subset of current or ex-smokers, we can say that approximately 41.1% of the data can be said to be heaped.

```
> rootogram4(fit1.sy, max = 100, main = "", xlim = c(0, 60),
             style = "hanging", col = "red", fill = "lightgreen")
```

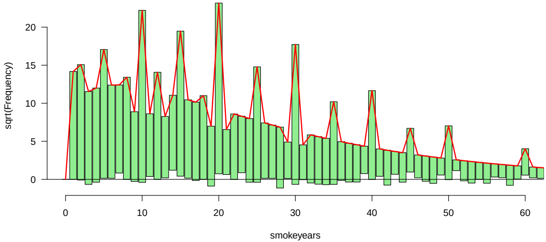


Figure: Hanging rootogram of the intercept-only GAITD regression `fit1.sy`.

The response residuals have no systematic lack-of-fit. Conclusion: the model is an acceptable fit.

Adding Covariates†

How do things change when adjusting for sex and ethnicity? Here, we set eq.dp = TRUE because the estimates look similar and for numerical stability.

```
> fit2.sy <-
    vglm(smokeyears ~ sex + ethnicity,
         gaitdnbinomial(i.mix = i.mix, i.mlm = i.mlm, a.mix = a.mix,
  #                     eq.dp = FALSE,
                        eq.ip = FALSE, eq.ap = FALSE,  # This line is good
                        d.mix = d.mix, truncate = tvec),
         etastart = predict(fit1.sy),  # Improved initial values
         crit = "coef", trace = FALSE, data = smoke.df)
```

Some output:

```
> round(t(coef(fit2.sy, matrix = TRUE)), 3)

                          (Intercept)  sexM ethnicityMaori ethnicityPolynesian ethnicityOther
loglink(munb.p)                 2.736 0.124         -0.262              -0.238         -0.217
loglink(size.p)                 0.631 0.000          0.000               0.000          0.000
multilogitlink(pobs.mix)       -0.976 0.000          0.000               0.000          0.000
loglink(munb.a)                 2.839 0.134         -0.185              -0.315         -0.450
loglink(size.a)                 0.806 0.000          0.000               0.000          0.000
multilogitlink(pstr.mix)       -0.622 0.000          0.000               0.000          0.000
loglink(munb.i)                 3.118 0.041         -0.137              -0.189         -0.395
loglink(size.i)                 1.476 0.000          0.000               0.000          0.000
multilogitlink(pdip.mix)       -2.110 0.000          0.000               0.000          0.000
loglink(munb.d)                 2.736 0.124         -0.262              -0.238         -0.217
loglink(size.d)                 0.631 0.000          0.000               0.000          0.000
multilogitlink(pstr.mlm1)      -3.112 0.000          0.000               0.000          0.000
multilogitlink(pstr.mlm8)      -4.508 0.000          0.000               0.000          0.000
multilogitlink(pstr.mlm12)     -3.716 0.000          0.000               0.000          0.000
multilogitlink(pstr.mlm18)     -4.660 0.000          0.000               0.000          0.000
```

The data suggest the following:

- Europeans smoke longer than the other three ethnicities. In fact, there appears little difference between the three.
- Males smoke longer in general. However, there does not seem to be a difference between males and females in the inflated values (spikes).

Closing Comments

Advice†

1. Firstly spikeplot the response and study it well!
2. Select the special values carefully, e.g., don't inflate seeped values! If unsure, alter them.
3. Don't overfit the model... will it generalize? For example, Warton (2005) concluded that it was rarely necessary for 0-inflation when NB $\widehat{\mu} \approx 0$. Amateurs inflate too much.
4. Monitor convergence; set `trace = TRUE`. Be wary if it takes more than 10 iterations to convergence.
5. Fit an *intercept-only* model first and then add covariates. Use simpler models for initial values, e.g., `etastart`.
6. If necessary, input better initial values than the self-starting ones.

Some Future Work

1. Develop GAITD regression for continuous distributions, e.g., normal and gamma. Called a *concrete* (**con***tinuous–dis***crete**) distribution.

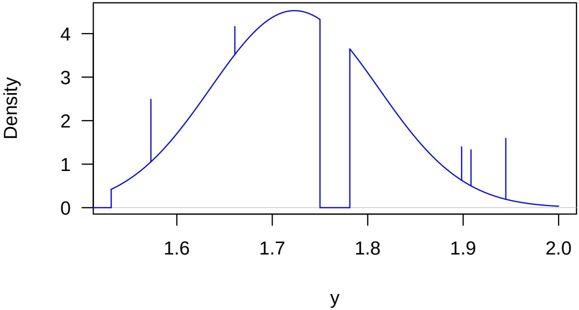


Figure: Hypothetical data from a GAITD normal distribution.

Summary and Concluding Remarks

1. The GAITD combo is easy-to-understand and offers much flexibility. Potentially useful for
   - spiked and dipped data,
   - truncated data,
   - underdispersed, equidispersed and overdispersed data,
   - heaped and seeped data (measurement error).

   Integrated and with parametric and nonparametric subcomponents.
2. Software is available.
3. Q: could the GAITD-NB become the Swiss army knife of count distributions?
4. Full details in Yee and Ma (2024).
5. Suggestions with real-life problems are welcome.
6. There is still a lot more work to be done. . .

Thanks for your attention!

References

- Crawford, F. W., Weiss, R. E., Suchard, M. A., 2015. Sex, lies and self-reported counts: Bayesian mixture models for heaping in longitudinal count data via birth-death processes. Ann. Appl. Stat. 9 (2), 572–596.

- Fung, T., Alwan, A., Wishart, J., Huang, A., 2020. mpcmp: Mean-Parametrized Conway-Maxwell Poisson (COM-Poisson) Regression. R package version 0.3.6.
URL `https://CRAN.R-project.org/package=mpcmp`

- Haslett, J., Parnell, A., Hinde, J., Moral, R. A., 2022. Modelling excess zeros in count data: a new perspective on modelling approaches. Inter. Statist. Rev. 90 (In press).

- Lewis-Esquerre, J. M., Colby, S. M., O'Leary Tevyaw, T., Eaton, C. A., Kahler, C. W., Monti, P. M., 2005. Validation of the timeline follow-back in the assessment of adolescent smoking. Drug Alcoh. Depend. 79 (1), 33–43.

- MacMahon, S., Norton, R., Jackson, R., Mackie, M., Cheng, A., Vander Hoorn, S., Milne, A., McCulloch, A., 1995. Fletcher Challenge-University of Auckland Heart & Health Study: Design and baseline findings. N. Z. Med. J. 108, 499–502.

- Miranda, V., Yee, T. W., 2023. VGAMextra: Additions and Extensions of the 'VGAM' Package. R package version 0.0-6.
URL `https://CRAN.R-project.org/package=VGAMextra`

- Sellers, K., Lotze, T., Raim, A., 2022. COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression. R package version 0.8.0.
URL `https://CRAN.R-project.org/package=COMPoissonReg`

- Wang, H., Heitjan, D. F., 2008. Modeling heaping in self-reported cigarette counts. Statistics in Medicine 27 (19), 3789–3804.

- Wang, H., Shiffman, S., Griffith, S. D., Heitjan, D. F., 2012. Truth and memory: Linking instantaneous and retrospective self-reported cigarette consumption. Ann. Appl. Stat. 6 (4), 1689–1706.

- Warton, D. I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16, 275–289.
- Yee, T. W., 2023. VGAM: Vector Generalized Linear and Additive Models. R package version 1.1-9.
  URL `https://CRAN.R-project.org/package=VGAM`
- Yee, T. W., Gray, J., 2023. VGAMdata: Data Supporting the 'VGAM' Package. R package version 1.1-9.
  URL `https://CRAN.R-project.org/package=VGAMdata`
- Yee, T. W., Ma, C., 2024. Generally altered, inflated, truncated and deflated regression. Statist. Sci. 39.
- Yee, T. W., Ma, C., Frigau, L., 2023. Heaping and seeping, GAITD regression and doubly-constrained reduced rank vector generalized linear models, in smoking studies. In preparation .