Metrika
https://doi.org/10.1007/s00184-020-00786-y

General flation models for count data

Dankmar Böhning[1] · Helen E. Ogden[1]

Received: 16 January 2020

Abstract
The paper discusses very general extensions to existing inflation models for discrete random variables, allowing an arbitrary set of points in the sample space to be either inflated or deflated relative to a baseline distribution. The term flation is introduced to cover either inflation or deflation of counts. Examples include one-inflated count models where the baseline distribution is zero-truncated and count models for data with a few unusual large values. The main result is that inference about the baseline distribution can be based solely on the truncated distribution which arises when the entire set of flation points is truncated. A major application of this result relates to estimating the size of a hidden target population, and examples are provided to illustrate our findings.

Keywords One-inflation · Multiple inflation · Zero-truncation with one-inflation · Multiple truncation

1 Introduction

Zero-inflation models, in particular zero-inflated Poisson models, are very popular in many areas of science and applications that deal with count data modelling. Since the original work of Lambert (1992), contributions have been manifold [see Wagh and Kamalja (2018) for a recent review]. The central idea of a zero-inflated model is to put an extra point-mass at zero, where the counts larger than zero are modelled by a simple count distribution, called the *baseline distribution*. Most applications consider the Poisson as baseline distribution, but more recently other baseline distributions have been considered, including the negative-binomial and binomial distributions.

✉ Helen E. Ogden
H.E.Ogden@soton.ac.uk

Dankmar Böhning
D.A.Bohning@soton.ac.uk

[1] Southampton Statistical Sciences Research Institute, School of Mathematical Sciences, University of Southampton, Southampton, UK

Published online: 10 July 2020

Estimation of parameters is typically based upon the EM algorithm [see Meng (1997) for an early work and McLachlan and Krishnan (2007) for the general concept]. Very little consideration has been given to cases in which inflation occurs at count locations other than zero. A rare exception is Godwin (2017), who considers inflation at counts of one for zero-truncated count models that arise in the capture-recapture setting. Furthermore, very little work has been done for cases where the weight attached to a specific count is allowed to be negative. Dietz and Böhning (2000) consider such a scenario for zero-modified Poisson models.

Here, we define a more general class of models for count data, where an arbitrary set of counts are allowed to have either inflated or deflated probabilities relative to a baseline model. We use the term "flation" to allow for either inflation or deflation of values. This includes the cases of zero- and one-inflation as special cases. We derive a simple closed-form expression for the maximum likelihood estimator of the flation parameters, and show that the profile likelihood for the parameters of the underlying count model is identical to the likelihood of the model after truncating the flated (i.e. inflated or deflated) counts.

The paper is organised as follows. Section 2 introduces and discusses the model, and Sect. 3 illustrates the problem in two substantially different examples. Section 4 contains the main results, and Sect. 5 describes how information criteria may be used to choose between candidate models. Section 6 describes the use of our methods for zero-truncated count modelling and Horvitz–Thompson estimation of population size. Section 7 shows how our results can be applied in the examples. The paper ends with a brief discussion.

2 A general flation model

Let Y be any discrete random variable, taking values in some set $\mathcal{Y}$. For count data, we typically have $\mathcal{Y} = \{0, 1, 2, \ldots\}$, but the model may be defined in the more general case.

We start with a baseline model for the probability mass function $p(y;\theta)$, for each $y \in \mathcal{Y}$, depending on unknown baseline parameters $\theta \in \Theta \subseteq \mathbb{R}^p$.

To define a general flation model, suppose that there is specified subset of d values $\mathcal{Y}^+ = \{y_1^+, \ldots, y_d^+\} \subseteq \mathcal{Y}$ which are allowed to have probabilities different from that assumed under the baseline model. The flation model has probability mass function

$$p_+(y;\theta,\pi) = \begin{cases} \pi_i + \pi_0\, p(y;\theta) & \text{if } y = y_i^+ \\ \pi_0\, p(y;\theta) & \text{if } y \notin \mathcal{Y}^+, \end{cases} \tag{1}$$

where $\pi = (\pi_0, \pi_1, \ldots, \pi_d)$, and $\sum_{i=0}^{d} \pi_i = 1$. We must also restrict π to ensure that $p_+(y;\theta,\pi) \geq 0$ for all $y \in \mathcal{Y}$, which means that $\pi_0 \geq 0$ and $\pi_i \geq -\pi_0 p(y_i^+;\theta)$. We call π the flation parameters.

Since we do not insist that $\pi_i \geq 0$, as well as modelling inflation of certain values, the model may also be used to model situations in which some values exhibit *deflation*, in that the true probability for that value is lower than under the baseline model.

Table 1 Frequencies of the number of times perpetrators of a domestic violence incident were identified in the Netherlands in 2009 (van der Heijden et al. 2014)

Number of	0	1	2	3	4	5	6	7	8	9
Frequency	–	15,169	1957	393	99	28	8	6	1	1

3 Examples

We illustrate the occurrence of inflation with two examples.

3.1 Domestic violence in the Netherlands

It is known that domestic violence is largely a hidden activity and many incidents remain unreported (Summers and Hoffman 2002) and it is of interest to estimate the size of the hidden activity. We consider domestic violence data provided by van der Heijden et al. (2014), which record the number of times perpetrators were identified in a domestic violence incident in the Netherlands during the year 2009, excluding the police region for The Hague. There were 15,169 perpetrators identified as being involved in a domestic violence incident exactly once, 1957 exactly twice, and so forth. In total, there were 17,662 different perpetrators identified.

If we are willing to assume a Poisson model, $p(y;\theta) = \exp(-\theta)\theta^y/y!$, then exploiting the fact that $p(0;\theta) = \frac{1}{2}p(1;\theta)^2/p(2;\theta)$ leads to the estimate $\hat{f_0} = \frac{1}{2}f_1^2/f_2$ where f_x denotes the frequency of perpetrators with x identifications. For the data of Table 1, we find the value of 58,789 hidden perpetrators. This estimator was developed for use in ecology (Chao 1984, 1987, 1989) and it is now one of the most popular estimators in ecology. This is due to the fact it has a lower bound property which means that under heterogeneity the estimator provides a lower bound (in expectation) for the true size of the population of interest.

This estimator depends on the baseline distribution and will change with its form. If the baseline distribution follows a geometric distribution $p(y;\theta) = \theta(1-\theta)^y$, for $y = 0, 1, \ldots$ then $\hat{f_0} = f_1^2/f_2$, with a value of 117,577 for the data of Table 1.

Figure 1 shows a plot of the estimated CDF from a zero-truncated Poisson model, including a 95% confidence interval for the CDF, with the empirical CDF overlaid. The zero-truncated Poisson distribution underestimates the number of ones, and fits the data poorly, so an alternative model is required. One possible cause of this one-inflation is that there may be a behavioural change as a result of identification: many perpetrators may refrain from this activity after the first incidence. We will determine whether the data are consistent with such a behavioural change, and study the implications for estimation of the number of unobserved perpetrators in the population.

3.2 Studies using counts of falls as outcome

In fall prevention trials, in particular with patients with Parkinson's disease, the number of falls is typically zero, one, or two. In Table 2, a typical distributions of fall counts

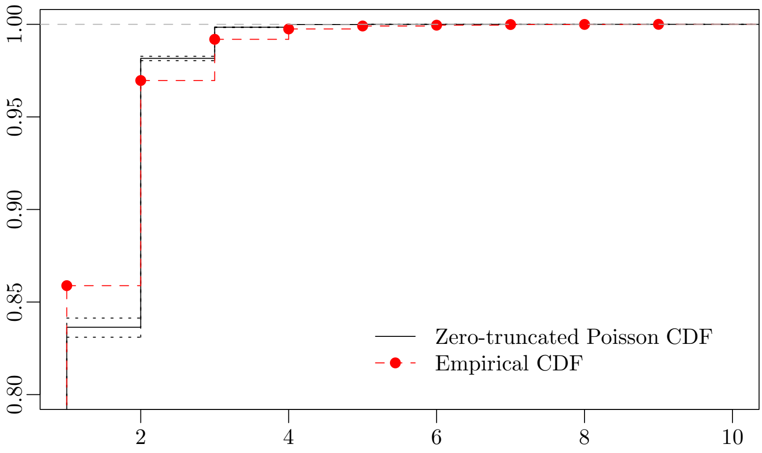


Fig. 1 The estimated CDF from a zero-truncated Poisson model for the domestic violence data, including a 95% confidence interval for the CDF, with the empirical CDF overlaid. The y-axis is zoomed in to the region (0.8, 1)

Table 2 The distribution of fall counts in a study on Parkinson's disease (Ashburn et al. 2007)

Number of falls	0	1	2	3	4	5	6	7	8	9	11	15	19	499
Frequency	50	30	14	9	6	4	3	4	1	2	3	1	1	1

in a fall prevention study is presented (Ashburn et al. 2007). In this data set, 75% of all patients fell at most 3 times, with the remainder of the patients falling at most 19 times—except one patient who had a very large number of falls, 499 in total. This observation would be difficult to catch even for a long-tailed distribution. However, it is also clear that it could have significant impact on the mean structure if ignored. Here it seems more reasonable to consider a flated distribution which puts small mass at the large fall.

4 Inference about the model parameters

Suppose we have data $y_1, \ldots, y_n$, assumed to be samples from model (1). The log-likelihood for the baseline and flation parameters is

$$\begin{aligned}\log L(\theta, \pi) &= \sum_{i=1}^{n} \log p_+(y_i; \theta) \\ &= \sum_{y_i \notin \mathcal{Y}^+} \{\log \pi_0 + \log p(y_i; \theta)\} + \sum_{j=1}^{d} \sum_{y_i = y_j^+} \log\{\pi_j + \pi_0\, p(y_j^+; \theta)\}\end{aligned}$$

$$= f_0 \log \pi_0 + \sum_{y_i \notin \mathcal{Y}^+} \log p(y_i; \theta) + \sum_{j=1}^{d} f_j \log\{\pi_j + \pi_0 \, p_j(\theta)\},$$

where we write $f_j = |\{i : y_i = y_j^+\}|$ for the number of times y_j^+ appears in the sample, $p_j(\theta) = p(y_j^+; \theta)$ for the probability of y_j^+ under the baseline model, and $f_0 = n - \sum_{j=1}^{d} f_j$ for the number of non-flated values in the sample.

Next, we aim to find the profile log-likelihood for θ. To do this, we first find $\hat{\pi}(\theta)$ to maximise the log-likelihood for each fixed θ, subject to the constraint $\sum_{i=0}^{d} \pi_i = 1$.

Proposition 1 *For each fixed* θ

$$\hat{\pi}_0(\theta) = \frac{f_0}{np_0(\theta)}, \textit{ and } \hat{\pi}_j(\theta) = \frac{f_j p_0(\theta) - f_0 p_j(\theta)}{np_0(\theta)}, \textit{ for } j = 1, \ldots, d,$$

where $p_0(\theta) = 1 - \sum_{j=1}^{d} p_j(\theta)$ *is the total probability assigned to all non-flated values under the baseline model.*

The proof is given in "Appendix 1". By Proposition 1, the profile log-likelihood for θ is

$$\begin{aligned} \log L_p(\theta) &= \log L(\theta, \hat{\pi}(\theta)) && (2) \\ &= \sum_{y_i \notin \mathcal{Y}^+} \left\{ \log \frac{f_0}{np_0(\theta)} + \log p(y_i; \theta) \right\} + \sum_{j=1}^{d} f_j \log \frac{f_j}{n} \\ &= \sum_{y_i \notin \mathcal{Y}^+} \{\log p(y_i; \theta) - \log p_0(\theta)\} + \sum_{j=0}^{d} f_j \log \frac{f_j}{n}. && (3) \end{aligned}$$

We may maximise this profile log-likelihood to estimate θ as $\hat{\theta} = \arg\max \log L(\theta, \hat{\pi}(\theta))$. The maximum likelihood estimator for π is $\hat{\pi} = \hat{\pi}(\hat{\theta})$.

We could also use this profile log-likelihood to test hypotheses or find confidence intervals for θ.

Remark 1 As a consequence of Proposition 1, for each $\theta \in \Theta$, the fitted probability of observing each $y_i^+ \in \mathcal{Y}^+$ is

$$\begin{aligned} p_+(y_i^+; \theta, \hat{\pi}(\theta)) &= \hat{\pi}_i(\theta) + \hat{\pi}_0(\theta) p(y_i^+, \theta) \\ &= \frac{f_i p_0(\theta) - f_0 p_i(\theta)}{np_0(\theta)} + \frac{f_0}{np_0(\theta)} p_i(\theta) \\ &= \frac{f_i}{n}, \end{aligned}$$

so the fitted probability for any point in the flation set is equal the corresponding observed proportion.

Remark 2 If Y follows (1), for each $y \notin \mathcal{Y}^+$, we can find the flation-truncated distribution

$$\begin{aligned} p^{\text{trunc}}(y;\theta) &= P(Y = y | Y \notin \mathcal{Y}^+) \\ &= \frac{p^+(y;\theta)}{\sum_{z \notin \mathcal{Y}^+} p^+(z;\theta)} \\ &= \frac{\pi_0\, p(y;\theta)}{\sum_{z \notin \mathcal{Y}^+} \pi_0\, p(z;\theta)} \\ &= \frac{p(y;\theta)}{p_0(\theta)} \end{aligned} \tag{4}$$

We could consider the log-likelihood based only those observations y_{trunc} not in $\mathcal{Y}^+$, assumed to be samples from the flation-truncated distribution (4). This truncated log-likelihood is

$$\log L(\theta; y_{\text{trunc}}) = \sum_{y_i \notin \mathcal{Y}^+} \log p^{\text{trunc}}(y_i;\theta) = \sum_{y_i \notin \mathcal{Y}^+} \{\log p(y_i;\theta) - \log p_0(\theta)\}\,. \tag{5}$$

The profile log-likelihood (3) for θ under the flation model is the truncated log-likelihood plus a term not involving θ, so all inference on θ is equivalent to that obtained by truncating $\mathcal{Y}^+$.

5 Model choice

In Sect. 4, we described how to conduct inference on the model parameters, assuming a particular baseline model and that $\mathcal{Y}^+$ is a fixed set. In reality, we may want to compare several candidate baseline models and choices of the flation set $\mathcal{Y}^+$.

We could use standard information criteria to do this. For a given candidate model, consisting of a specified baseline model and flation set, suppose that the dimension of the baseline parameter θ is p, and the dimension of the flation set is d. Suppose that the profile log-likelihood for this model is $\log L_p(\theta)$, given by (3), and write $\hat{\ell} = \log L_p(\hat{\theta})$ for the maximised log-likelihood under the model.

The Akaike information criterion (AIC) is

$$\text{AIC} = 2(p + d) - 2\hat{\ell},$$

and the Bayesian information criterion (BIC) is

$$\text{BIC} = (p + d)\log(n) - 2\hat{\ell}.$$

We will use these information criteria to compare various possible models for our example data sets.

Here $\hat{\ell}$ is the maximised log-likelihood under the flation model, which is different from the maximised truncated log-likelihood (5). While for a fixed model, inference about the model parameters may be conducted by truncating the flation set, we must use the full flation likelihood, involving all observations, to choose between models.

We have focussed here on selection criteria such as AIC and BIC as these allow a general approach to model assessment. However, other methods are also possible, including likelihood ratio testing for specific models. For example, in investigating inflation models (all π_i s non-negative) it might be of interest to test the null-hypothesis $H_0 : \pi_0 = 1$ (no inflation present) against the alternative $H_1 : \pi_0 \in (0, 1)$. Particular caution is required in this case, as the null hypothesis is part of the boundary of the alternative and the likelihood ratio test has a non-standard distribution under H_0. For details see Self and Liang (1987).

6 Estimating the number of unobserved zeroes

6.1 Horvitz–Thompson estimation

We consider now the special case $\mathcal{Y} = \{0, 1, 2, \ldots\}$, with some appropriate count distribution as the baseline model. We assume that, as a consequence of the sampling model, zeros are not observed, so that the observational baseline model becomes $p(y;\theta)/[1 - p(y_0;\theta)]$ for $y_0 = 0$ and $y = 1, 2, \ldots$. This corresponds to the situation of Example 3.1. In order to conduct inference on θ, we can simply consider the baseline distribution truncated at both 0 and the flation set $\mathcal{Y}^+$, equivalent to including 0 in $\mathcal{Y}^+$ in (5).

Recall that $p_0(\theta)$ denotes the probability assigned to all non-flated values under the baseline model and f_0 is the frequency of all non-flated values in $\mathcal{Y}$. We are interested in estimating the frequency f_{y_0} associated with the truncated value y_0 as this is the hidden part of the target population. This can be simply achieved by a Horvitz–Thompson-type estimator

$$\widehat{f_{y_0}}(\theta) = \frac{f_0}{p_0(\theta)} p(y_0;\theta). \tag{6}$$

Assuming that the hidden counts belong to the baseline model, $\widehat{f_{y_0}}(\theta)$ is an unbiased estimator of the true frequency of hidden counts $Np(y_0;\theta)$, where N is the size of the target population. An unbiased estimator of the population size N itself is then

$$\hat{N}(\theta) = n + \widehat{f_{y_0}}(\theta),$$

where n is the size of the entire observed sample. In practice, the unknown value of θ needs to be replaced by its maximum likelihood estimate from the truncated sample (which will exclude y_0 and potentially other values in $\mathcal{Y}$).

6.2 Standard errors

We are interested in providing a standard error for $\widehat{f_{y_0}}(\hat{\theta})$ assuming that θ is estimated (say by using maximum likelihood). We use the technique of conditional moments to find that

$$\begin{aligned}&\mathrm{Var}\left(\widehat{f_{y_0}}(\hat{\theta})\right)\\&= E\left[\mathrm{Var}\left(\frac{f_0}{p_0(\hat{\theta})}p(y_0;\hat{\theta})|f_0\right)\right]+\mathrm{Var}\left[E\left(\frac{f_0}{p_0(\hat{\theta})}p(y_0;\hat{\theta})|f_0\right)\right].\end{aligned} \tag{7}$$

The first term in (7) can be estimated as

$$f_0^2\mathrm{Var}\left[p(y_0;\hat{\theta})/p_0(\hat{\theta})\right]=f_0^2\mathrm{Var}\left(T(\hat{\theta})\right)\approx f_0^2T'(\hat{\theta})^2\mathrm{Var}(\hat{\theta}), \tag{8}$$

where we have used the $\delta-$method for the variance of the transformation $T(\theta)=p(y_0;\theta)/p_0(\theta)$. This is the most general form and more concrete results will depend on the structure of $T(.)$. We assume here that an estimate of $\mathrm{Var}(\hat{\theta})$ is readily available from standard results, e.g. by using the Fisher information if θ is estimated by maximum likelihood.

The second term in (7) can be estimated as

$$\frac{p(y_0;\hat{\theta})^2}{p_0(\hat{\theta})^2}\mathrm{Var}(f_0)=\frac{p(y_0;\hat{\theta})^2}{p_0(\hat{\theta})^2}Np_0(\hat{\theta})(1-p_0(\hat{\theta}))\approx\frac{p(y_0;\hat{\theta})^2}{p_0(\hat{\theta})^2}f_0(1-p_0(\hat{\theta})). \tag{9}$$

The second term allows a very general form of result and does not depend on the specific form of $T(.)$.

To illustrate these findings let us consider the geometric distribution with probability mass function $p(y;\theta)=\theta(1-\theta)^y$ for $y=0,1,\ldots,$. First, we take the case of zero-truncation with one-flation. As the zero-one truncated is again a geometric distribution, we can find the maximum likelihood estimator as $\hat{\theta}=f_0/(f_0+S)$ where $S=\sum_{x=0}^{m-2}xg_x$ where m is the largest observed count and $g_{x-2}=f_x$ for $x=2,\ldots,m$, $f_0=f_2+\cdots+f_m$. Furthermore, an estimate of $\mathrm{Var}(\hat{\theta})$ can be found as the negative inverse observed Fisher information as $\left(\frac{f_0}{\hat{\theta}^2}+\frac{S}{(1-\hat{\theta})^2}\right)^{-1}$.

Hence, we are ready to calculate (8) as

$$f_0^2\frac{(1+\hat{\theta})^2}{(1-\hat{\theta})^6}\mathrm{Var}(\hat{\theta})$$

and (9) as

$$\frac{\hat{\theta}^2}{[1-\hat{\theta}-\hat{\theta}(1-\hat{\theta})]^2}[\hat{\theta}+\hat{\theta}(1-\hat{\theta})]f_0.$$

We now look at the zero-truncation case without one-flation. Again the zero-truncated is a geometric distribution, so that we can find the maximum likelihood estimator as $\hat{\theta}=f_0/(f_0+S)$ where $S=\sum_{x=0}^{m-1}xg_x$ where m is the largest observed count and $g_{x-1} = f_x$ for $x = 1, \dots m,\ f_0 = f_1 + \cdots + f_m$. Furthermore, an estimate of $\mathrm{Var}(\hat{\theta})$ can be found as the negative inverse observed Fisher information as $\left(\frac{f_0}{\hat{\theta}^2} + \frac{S}{(1-\hat{\theta})^2}\right)^{-1}$. We can calculate again (8) as

$$f_0^2 \frac{1}{(1-\hat{\theta})^4} \mathrm{Var}(\hat{\theta})$$

and (9) as

$$\frac{\hat{\theta}^3}{[1-\hat{\theta}]^2} f_0.$$

This completes the illustration. We will use the standard errors to provide confidence intervals for the estimated frequency of hidden members of the target population, under the assumption that these hidden members belong to the baseline model.

7 Application to the examples

7.1 Domestic violence in the Netherlands

It is reasonable to assume that among the unobserved perpetrators of domestic violence, there is a sub-population who would exhibit a behavioural change if they were caught, and refrain from further violence. Under this assumption, it is not possible to estimate the size of this sub-population from the data, as they would never be observed more than once. Instead, we may use the methods described in Sect. 6 to estimate the number of unobserved perpetrators who would not change their behaviour after being caught. The size of this second sub-population of offenders can also be used as a lower bound for the size of the entire unobserved population of perpetrators. We investigate how this lower bound compares with an estimate of the size of the unobserved population made under the assumption that identification of perpetrators does not lead to a behavioural change.

We consider the geometric distribution $p(y; \theta) = (1-\theta)^y \theta$ with $\theta \in (0, 1)$ as baseline model. The log-likelihood for the zero-one truncated model is

$$\log L(y; \theta) = \sum_{y=2}^{m} f_y \log[(1-\theta)^{y-2}\theta].$$

Here m is the largest observed count ($m = 9$). The maximum likelihood estimate is given as $\hat{\theta} = \sum_{y \geq 2} f_y / [\sum_{y \geq 2} y f_y - \sum_{y \geq 2} f_y]$ which we use as a plug-in estimate in (6). Fitting a zero-one-truncated model leads to a maximum likelihood estimate of $\hat{\theta} = 0.77$ with an associated Horvitz–Thompson-type estimate of the number of hidden perpetrators in the second sub-population of $\widehat{f_{y0}}(\hat{\theta}) = 35{,}832$. This seems like a large number but is considerably lower than Chao's 'lower' bound estimate (using

Table 3 AIC and BIC for various models for the domestic violence data with estimates of the hidden population size

Base	Flation	AIC	BIC	$\widehat{f_{y_0}}(\hat{\theta})$ (95%CI)
Poisson	None	18,445.78	18,453.56	42,539 (40,771–44,307)
Geometric	None	18,043.26	18,051.04	96,191 (92,360–100,021)
Poisson	1	17,955.37	17,970.93	5979 (5027–6930)
Geometric	1	17,890.16	17,905.72	35,832 (30,481–41,183)

the geometric as baseline) of 117,577. The estimate $\widehat{f_{y_0}}(\hat{\theta}) = 96{,}191$ under the zero-truncated geometric model is similar to Chao's estimate as it also uses the frequency of ones.

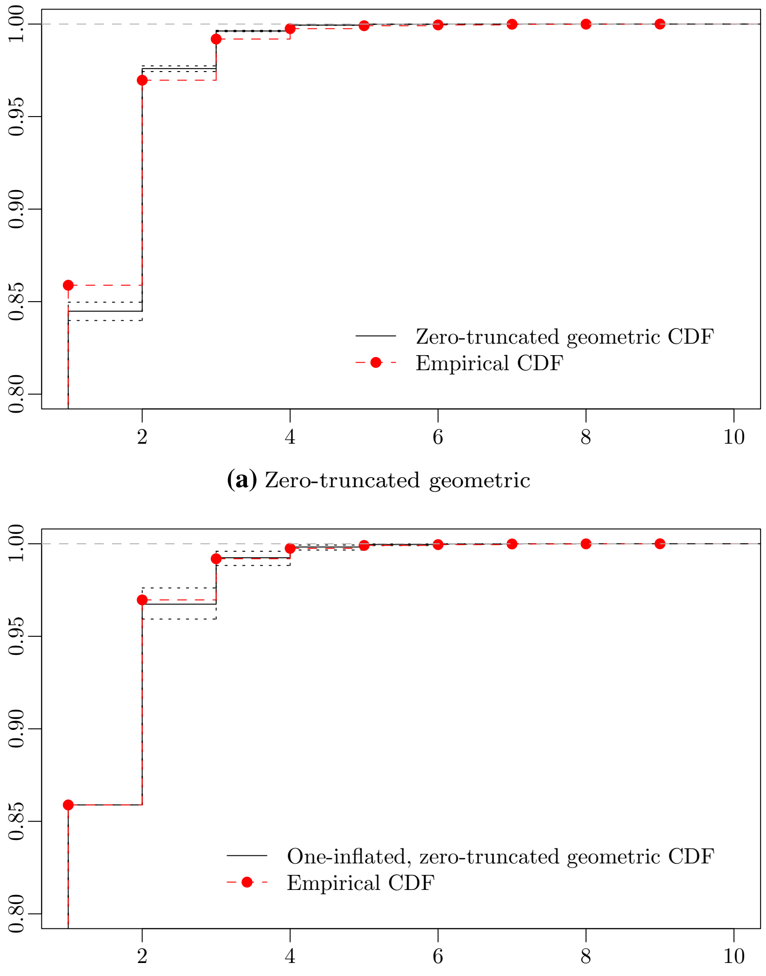
Table 3 compares Poisson and geometric models for this data, with either no flation or one-flation. Both AIC and BIC lead us to prefer the geometric model with one-flation. The last column of Table 3 provides an estimate of the size of the second hidden sub-population of perpetrators, including a 95% confidence interval. The lower end of this confidence interval may be viewed as a lower bound for the entire hidden population of perpetrators. This lower bound is substantially smaller than both Chao's lower bound and the Horvitz–Thompson estimate under the geometric model without flation. This demonstrates that ignoring the possible impact of the observation mechanism on future incidents may lead to an over-estimation of the size of a hidden population.

Figure 2a shows a plot of the estimated CDF assuming a zero-truncated geometric model, including a 95% confidence interval for the CDF, with the empirical CDF overlaid. The zero-truncated geometric distribution provides a better fit to the data than the zero-truncated Poisson distribution from Fig. 1, but still underestimates the number of ones. Figure 2b shows a similar plot but with the a one-flated zero-truncated geometric model. This provides a good fit to the data, with the empirical CDF matching the estimated CDF closely, and contained within the limits of the confidence bands. The data are consistent with the theory that a sub-population of perpetrators refrain from domestic violence after first identification.

7.2 Studies using counts of falls as outcome

In this case we find for the ordinary Poisson model a maximum likelihood estimate of 5.95 with 95% confidence interval of 5.54–6.38 whereas the flated Poisson model with flation parameter $y_0 = 499$ has a maximum likelihood estimate of 2.09 with 95% confidence interval of 1.86–2.36. This shows the large impact of high fallers in this study and the potential for flated effect estimates if large observations are ignored.

Table 4 compares Poisson, geometric and negative Binomial models for this data, with either no flation, flation at 499. Both AIC and BIC lead us to prefer the negative Binomial with flation at 499. We can conclude that the fit of the model can be largely improved here by using a simple one-flation model for the large observation and potential biasing effects can thus be avoided.



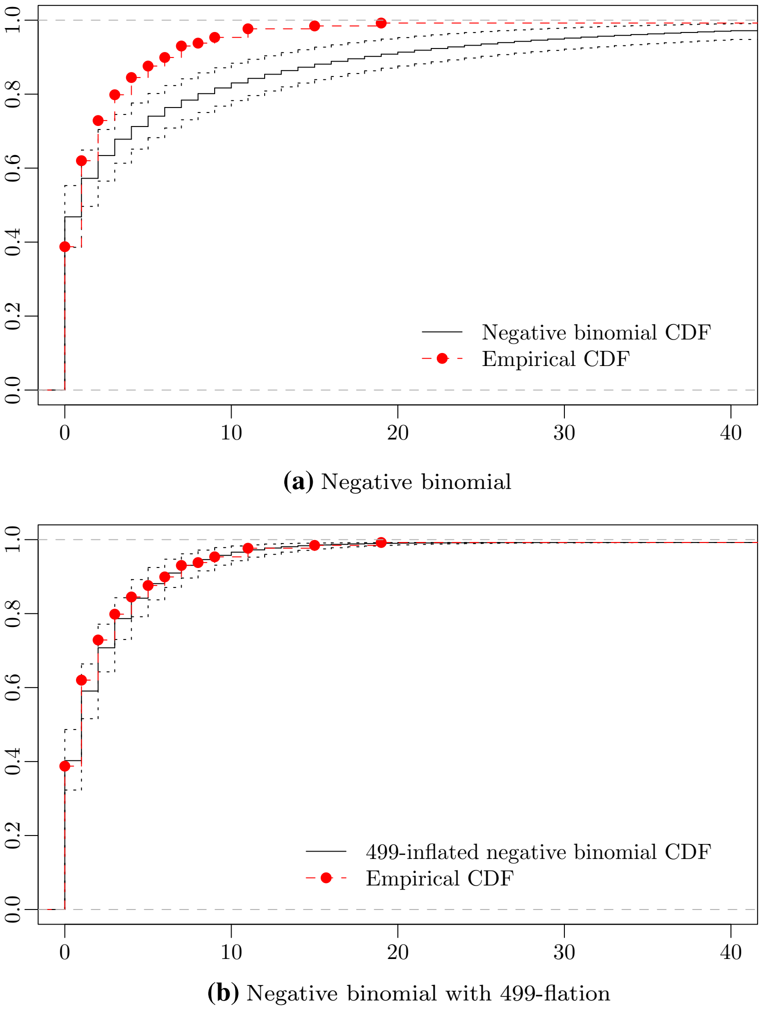
(a) Zero-truncated geometric

(b) One-flated zero-truncated geometric

Fig. 2 The estimated CDF from two models for the domestic violence data, including a 95% confidence interval for the CDF, with the empirical CDF overlaid. The y-axis is zoomed in to the region (0.8, 1)

Table 4 AIC and BIC for various models for the falls data

Base	Flation	AIC	BIC
Poisson	None	4559.29	4562.15
Geometric	None	740.50	743.36
Neg. Bin.	None	595.56	601.28
Poisson	499	703.32	709.03
Geometric	499	514.10	519.82
Neg. Bin.	499	509.11	517.69



(a) Negative binomial

(b) Negative binomial with 499-flation

Fig. 3 The estimated CDF from two models for the falls data, including a 95% confidence interval for the CDF, with the empirical CDF overlaid

Figure 3a shows a plot of the estimated CDF assuming a negative binomial model, including a 95% confidence interval for the CDF, with the empirical CDF overlaid. The negative binomial is a more flexible distribution than the Poisson, but fails to fit the data well. Figure 3b shows a similar plot but with the 499-flated negative binomial model. We observe that this provides a good fit to the data, with the empirical CDF matching the estimated CDF closely, and contained within the limits of the confidence bands.

Instead of allowing the count of 499 to be flated, it is also possible to consider flation for a large range of counts in this example, for instance all counts from 20 to 1000. In this case, the fitted models under 20–1000 flation are very close to those with

Table 5 The distributions used to generate data for the simulation study

Model	Baseline distribution	Flation	Flation parameters	N
P	Poisson (0.347)	None	N/A	60,201
G	geometric (0.845)	None	N/A	113,819
P1	Poisson (0.793)	1	$\pi_1 = 0.780$	19,616
G1	geometric (0.769)	1	$\pi_1 = 0.828$	20,347

N samples were simulated from each of these distributions, then the zeroes were removed

499 flation, both in terms of the estimates of the baseline parameters and of the fitted model probabilities.

8 Simulation study

To validate our approach, we conduct a simulation study based on the application to the domestic violence data from Sect. 7.1. We simulate data from each of the four models considered there: a Poisson model (which we refer to as model P), a geometric model (G) a Poisson model with one-flation (P1), and a geometric model with one-flation (G1). In each case, we use the maximum likelihood estimates from the domestic violence data as parameter values in our simulation. Each simulated dataset is constructed by taking N samples from the distribution, then removing all zero counts. For each model, we choose this total population size N so that the expected number of non-zero counts is 17,662, to match the number of perpetrators identified at least once in the domestic violence data. To do this, we choose N to be the closest integer to $17{,}662/\left[1 - P(Y = 0)\right]$, where Y represents the distribution being simulated from, before zero-truncation occurs. For example, for model P1, with probability $\pi_1 = 0.780$, we take $Y = 1$, and with probability $1-\pi_1$ we take $Y \sim \text{Poisson}(0.793)$. In this case, $P(Y = 0) = (1-0.780)\exp(-0.793) = 0.100$, so we choose $N \approx 17{,}662/0.9$. Table 5 gives the parameters used to generate data from each of the four models.

We simulate 10,000 datasets from each model. We fit each of the four candidate models to each simulated dataset, and compute the AIC and BIC for each model. Table 6 shows the proportion of times each model was selected by AIC and by BIC. In this example, BIC outperforms AIC. When we simulate from either model without one-flation, BIC nearly always chooses the correct model. When we simulate from either model with one-flation, AIC and BIC choose the same model as one another. Both criteria always reject the models without one-flation, but are not always able to distinguish between models P1 and G1, which are both subject to one-flation.

We also consider the quality of our estimates of the total population size, N. Recall that we fix N to the values given in Table 5. For each simulated dataset and model, we get an estimate of N. N is estimated either assuming that the form of the true model is known, or that the form of the model must first be selected by AIC or BIC. For the ith simulated dataset, we calculate an estimate of N for each of the four candidate models, then let $\hat{N}_i$ be the estimate of N from the selected model. We compute the

Table 6 The proportion of times each model was selected in the simulation study, using either AIC or BIC

(a) AIC					(b) BIC				
True model	Selected model				True model	Selected model			
	P	G	P1	G1		P	G	P1	G1
P	0.766	0.000	0.133	0.101	P	0.997	0.000	0.001	0.002
G	0.000	0.795	0.067	0.137	G	0.000	0.994	0.004	0.002
P1	0.000	0.000	0.889	0.111	P1	0.000	0.000	0.889	0.111
G1	0.000	0.000	0.338	0.662	G1	0.000	0.000	0.338	0.662

Table 7 The relative bias and relative RMSE of estimates of the total population size, *N*

(a) Relative bias				(b) Relative RMSE			
True model	Model selection method			True model	Model selection method		
	True model	AIC	BIC		True model	AIC	BIC
P	0.001	0.304	0.006	P	0.016	0.961	0.143
G	0.001	−0.045	−0.002	G	0.018	0.187	0.049
P1	0.001	0.058	0.058	P1	0.015	0.173	0.173
G1	0.008	−0.033	−0.033	G1	0.042	0.071	0.071

bias and root mean squared error as

$$\text{bias} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{N}_i - N, \quad \text{RMSE} = \left[\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} (\hat{N}_i - N)^2 \right]^{1/2},$$

where $n_{\text{sim}} = 10{,}000$ is the number of simulated datasets. To make results more comparable across scenarios, we standardise by finding the relative bias and relative RMSE, as

$$\text{relative bias} = \frac{\text{bias}}{N}, \quad \text{relative RMSE} = \frac{\text{RMSE}}{N}.$$

Table 7 gives the relative bias and relative RMSE of $\hat{N}$ for each case.

Assuming the model is known, we also calculated standard errors for the estimates of N for each of the models, using the process described in Sect. 6.2. For each simulated dataset, we used these standard errors to find a 95% confidence interval for N. In our 10,000 simulations, the empirical coverage of the confidence interval for N was 94% for all models, very close to the nominal 95% level.

9 Discussion

Our paper contains the striking result that inflation models (and flation models more generally) can be fitted simply by truncating the respective flation points. The estimates for the flation weights can then be found in closed form. We allow the flation weights to be positive or negative, allowing either inflation or deflation. In practice, we see usually that flation weights are positive.

An open question remains of how to generalise the result for available covariate information. Of course, the result can be generalised in a stratified way, but the challenge is to develop a full modelling approach that incorporates the covariates.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest. We are grateful to a referee for helpful comments.

Proof of results

Proof of Proposition 1 To find $\hat{\pi}(\theta)$, we define the Lagrangian

$$\begin{aligned} g(\theta; \pi, \lambda) &= \log L(\theta, \pi) - \lambda\Big(\sum_{j=0}^{d} \pi_j - 1\Big) \\ &= f_0 \log \pi_0 + \sum_{y_i \notin \mathcal{Y}^+} \log p(y_i; \theta) \\ &\quad + \sum_{j=1}^{d} f_j \log\{\pi_j + \pi_0 p_j(\theta)\} - \lambda\Big(\sum_{j=0}^{d} \pi_j - 1\Big) \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial g}{\partial \pi_0} &= \frac{f_0}{\pi_0} + \sum_{j=1}^{d} \frac{f_j p_j(\theta)}{\pi_j + \pi_0 p_j(\theta)} - \lambda, \\ \frac{\partial g}{\partial \pi_j} &= \frac{f_j}{\pi_j + \pi_0 p_j(\theta)} - \lambda \end{aligned}$$

for $j = 1, \ldots, d$, and

$$\frac{\partial g}{\partial \lambda} = \sum_{j=0}^{d} \pi_j - 1.$$

Choosing $\hat{\pi}(\theta)$ to solve $\frac{\partial g}{\partial \pi_j}|_{\hat{\pi}(\theta)} = 0$ and $\frac{\partial g}{\partial \lambda}|_{\hat{\pi}(\theta)} = 0$ gives

$$\hat{\pi}_0(\theta) = \frac{f_0}{\lambda p_0(\theta)}$$

and

$$\hat{\pi}_j(\theta) = \frac{1}{\lambda}\left(f_j - \frac{f_0 p_j(\theta)}{p_0(\theta)}\right).$$

Using the constraint $\sum_{j=0}^{d} \hat{\pi}_j(\theta) = 1$ we find $\lambda = n$, and the result follows.

As π needs to meet the restriction that $p_+(y; \theta, \pi) \geq 0$ for all $y \in \mathcal{Y}$, which is equivalent to $\pi_0 \geq 0$ and $\pi_i \geq -\pi_0 p(y_i^+; \theta)$ for $i = 1, \ldots, d$. So, let us define $\varPi = \{\pi = (\pi_0, \pi_1, \ldots, \pi_d | \pi_0 \geq 0$ and $\pi_i \geq -\pi_0 p(y_i^+; \theta)$ for $i = 1, \ldots, d\}$ and show that $\varPi$ is convex. We consider $\pi^{(1)}, \pi^{(2)} \in \varPi$ and $\alpha \in (0, 1)$. We need to show that

$$\alpha \pi^{(1)} + (1 - \alpha)\pi^{(2)} \in \varPi.$$

Clearly, if $\pi_0^{(1)} \pi_0^{(2)}$ are both non-negative, so is $\alpha \pi_0^{(1)} + (1 - \alpha)\pi_0^{(2)}$. Next, we need to show that

$$\alpha \pi_i^{(1)} + (1 - \alpha)\pi_i^{(2)} \geq -[\alpha \pi_0^{(1)} + (1 - \alpha)\pi_0^{(2)}] p(y_i^+; \theta)$$

for $i = 1, \ldots, d$. This follows from the fact that $\alpha \pi_i^{(1)} \geq -\alpha \pi_0^{(1)} p(y_i^+; \theta)$ and $(1 - \alpha)\pi_i^{(2)} \geq -(1 - \alpha)\pi_0^{(2)} p(y_i^+; \theta)$, as $\pi \in \varPi$.

Furthermore, a direct argument, using that log(.) is strictly concave, shows that $\log L(\theta, \pi)$ is strictly concave in π for fixed θ. Indeed, if $\pi^{(1)}, \pi^{(2)} > 0$ with $\mathbf{1}^T \pi^{(i)} = 1$ for $i = 1, 2$, $\mathbf{1}$ the vector of ones and $\alpha \in (0, 1)$, then

$$\log L(\theta, (1 - \alpha)\pi^{(1)} + \alpha \pi^{(1)}) \geq (1 - \alpha) \log L(\theta, \pi^{(1)}) + \alpha \log L(\theta, \pi^{(2)})$$

with equality only if $\pi^{(1)} = \pi^{(2)}$. Hence $\hat{\pi}$ is indeed the unique profile maximum likelihood estimate. This ends the proof. □

References

Ashburn A, Fazakarley L, Ballinger C, Pickering R, McLellan LD, Fitton C (2007) A randomised controlled trial of a home based exercise programme to reduce the risk of falling among people with Parkinson's disease. J Neurol Neurosurg Psychiatry 78(7):678–684. https://doi.org/10.1136/jnnp.2006.099333

Chao A (1984) Nonparametric estimation of the number of classes in a population. Scand J Stat 11(4):265–270

Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43(4):783–791

Chao A (1989) Estimating population size for sparse data in capture-recapture experiments. Biometrics 45(2):427–438

Dietz E, Böhning D (2000) On estimation of the Poisson parameter in zero-modified Poisson models. Comput Stat Data Anal 34:441–459. https://doi.org/10.1016/S0167-9473(99)00111-5

Godwin RT (2017) One-inflation and unobserved heterogeneity in population size estimation. Biom J 59(1):79–93. https://doi.org/10.1002/bimj.201600063

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14

McLachlan G, Krishnan T (2007) The EM algorithm and extensions, 2nd edn. Wiley, Hoboken

Meng XL (1997) The EM algorithm and medical studies: a historical link. Stat Methods Med Res 6:3–23

Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82(398):605–610. https://doi.org/10.1080/01621459.1987.10478472

Summers RW, Hoffman AM (2002) Domestic violence: a global view. Greenwood Press, Westport

van der Heijden PGM, Cruyff M, Böhning D (2014) Capture recapture to estimate criminal populations. In: Bruinsma GJN, Weisburd DL (eds) Encyclopedia of criminology and criminal justice. Springer, Berlin, pp 267–278

Wagh YS, Kamalja KK (2018) Zero-inflated models and estimation in zero-inflated Poisson distribution. Commun Stat Simul Comput 47(8):2248–2265. https://doi.org/10.1080/03610918.2017.1341526

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.