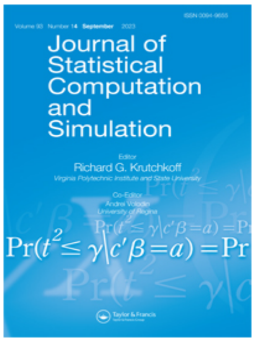
Journal of Statistical Computation and Simulation
Editor
Richard G. Krutchkoff
$\Pr(t^2 \leq \gamma \backslash c'\beta = a) = \Pr$

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: www.tandfonline.com/journals/gscs20

Modelling zero inflated and under-reported count data

Debjit Sengupta & Surupa Roy

To cite this article: Debjit Sengupta & Surupa Roy (2023) Modelling zero inflated and under-reported count data, Journal of Statistical Computation and Simulation, 93:14, 2390-2409, DOI: 10.1080/00949655.2023.2182883

To link to this article: https://doi.org/10.1080/00949655.2023.2182883

Published online: 02 Mar 2023.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2023, VOL. 93, NO. 14, 2390–2409
https://doi.org/10.1080/00949655.2023.2182883

Taylor & Francis
Taylor & Francis Group

Modelling zero inflated and under-reported count data

Debjit Sengupta and Surupa Roy

Department of Statistics, St. Xavier's College, Kolkata, India

ABSTRACT
Poisson distribution is a classic choice for modelling unbounded count data. However, count data arising in various fields of scientific research often have excess zeros and are under-reported. In such situations, Poisson distribution gives a poor fit and Poisson model based inferences lead to biased estimators and inaccurate confidence intervals. In this paper we develop a flexible model which can accommodate excess zeros and undercount. Internal validation data has been used for making likelihood based inferences. The impact of ignoring undercount and excess zeros are studied through extensive simulations. The finite sample behaviour of the estimators are investigated through bootstrap methodology. Finally, a real life data which is supposedly under-reported and known to have excess zeros is analysed.

ARTICLE HISTORY
Received 3 August 2022
Accepted 12 February 2023

KEYWORDS
Excess zero; undercount; surrogate; likelihood; bootstrap method

MATHEMATICS SUBJECT CLASSIFICATIONS
62F10; 62F40; 62P25; 62P20

1. Introduction

Poisson distribution is widely used to model many naturally occurring events varying over time and space. However, in case of excess zeros in the process, Poisson distribution gives a poor fit. In such situations Zero Inflated Poisson distribution (ZIP) becomes useful. Count data with excess zeros are common in a variety of disciplines including medicine [1], public health [2], environmental sciences [3] and agriculture [4]. Some popular examples of zero inflated count process include number of defects in manufactured items [5], number of days a student is suspended in a school year [6], self reported counts of specific high risk behaviours in a given time period [7], number of infected persons per household, when a high proportion of all households is uninfected [7]. The ZIP model is flexible in the sense that it accommodates the over dispersion phenomenon also. Unlike overdispersion arising due to heterogeneity of the data for which a negative binomial model is a good candidate, in A ZIP distribution, overdispersion occurs due to excess zeros in the data.

Very often, in real life, in addition to excess zeros, the data generating mechanism may be such that some occurrences of an event are not counted, thereby, leading to underreporting. Ignoring undercount leads to biased inferenceS. Several authors have developed undercount adjusted Poisson models based on a single count data. However, undercount in ZIP model has not been studied in the literature.

CONTACT Debjit Sengupta ✉ debjits10@gmail.com St. Xavier's College, Kolkata 380009, India

The focus of the current research lies in the joint modelling of excess zeros and undercount, based on a random sample of counts. We thus consider a setup where the true latent counts $Y_1^*, Y_2^*, .., Y_n^*$ are assumed to be a random sample from $ZIP(\lambda, \delta)$, where λ denotes the rate of incidence of an event and δ is the zero inflation parameter. However, in practice, we only observe the surrogate counts (Y) which are such that $Y_i \leq Y_i^*, \forall i$ and the probability of recording a count is π. In the undercount adjusted ZIP model, zeros can occur from three sources (i) zero of the Bernoulli distribution, (ii) zero of the standard Poisson distribution and (iii) zero from the conditional distribution of Y given $Y^* = t$; the probability of which are δ, $exp(-\lambda)$ and $(1-\pi)^t$ respectively. The undercount adjusted ZIP model so developed is over parametrized due to the presence of the nuisance parameter π. For model identifiability, a popular technique is to make use of validation data through a double sampling approach. This technique has been extensively used by researchers [8–17] for modelling error prone Poisson counts. Following similar lines, we have chosen a validation sample as a sub-sample from the original sample. Corresponding to the chosen sub-sample units, the true counts are observed through special efforts. We now motivate our research setting through the following examples.

Example 1.1: Let $Y_1, Y_2, .., Y_n$ denote the self reported counts by n students on the number of days suspended in a school year [6]. The extra zeros occur since some students are never at the risk of suspension. Moreover, self reported suspension data are likely to be underreported.

Example 1.2: Let $Y_1, Y_2, ..Y_n$ denote the reported numbers of violent behaviour by the batterers towards n victims. A ZIP model is found to fit the data well [18]. Further, data on domestic violence is known to be grossly under-reported.

Example 1.3: Let $Y_1, Y_2, .., Y_n$ denote the number of cigarettes smoked daily by a sample of n residents of Ethiopia. For this data, a ZIP model was found appropriate [19]. Moreover, such data is often under-reported due to the recall bias and social desirability bias [20].

In the context of the aforementioned examples, $Y_i^*(\geq Y_i), i = 1, 2, .., r(r < n)$ are the true counts of r randomly chosen students (in Example 1.1) for whom detailed school records on suspension are checked or r randomly selected victims (in Example 1.2) for whom the true counts of domestic violence is observed or r randomly chosen individuals (in Example 1.3) for whom the true number of cigarettes smoked per day is obtained through special efforts. Under the double sampling approach, we find the maximum likelihood estimators (MLE) of λ, δ and π given the data $Y_i, i = 1, 2, \ldots, n$ and $Y_i^*(\geq Y_i), i = 1, 2, \ldots, r$. The finite sample behaviour of the different estimators are investigated using bootstrap resampling techniques.

The article is organized as follows. In Section 2, we define the models and likelihoods. Section 3 describes the moment method of estimation. Section 4 gives detail of a simulation study. Section 5 analyses beverage consumption data which has excess zeros and is supposedly under-reported. Finally we conclude in Section 6.

2. Model and likelihood

Suppose the true count Y^* follows a ZIP distribution given by,

$$P(Y^* = y^*) = \begin{cases} \delta + (1-\delta)\,\mathrm{e}^{-\lambda} & \text{if } y^* = 0 \\ (1-\delta)\dfrac{\mathrm{e}^{-\lambda}\lambda^{y^*}}{y^*!} & \text{if } y^* = 1, 2, \ldots. \end{cases} \tag{1}$$

where δ and λ are the labelling parameters of the distribution. The error prone observed count Y is such that $P(Y \leq Y^*) = 1$. Let U_α denote a Bernoulli random variable such that, $U_\alpha = 1$, if a count is not missed, and 0 otherwise. Also, let π denote the probability of not missing a count. Then the surrogate count is $Y = \sum_{\alpha=1}^{Y^*} U_\alpha$. The conditional distribution of Y given $Y^* = y^*$ is given by,

$$P(Y = y|Y^* = y^*) = \binom{y^*}{y}\pi^y(1-\pi)^{y^*-y}; \quad y = 0, 1, 2, .., y^*. \tag{2}$$

Theorem 2.1: *The marginal distribution of the surrogate count Y is given by,*

$$P(Y = y) = \begin{cases} \delta + (1-\delta)\,\mathrm{e}^{-\lambda\pi} & \text{if } y = 0 \\ (1-\delta)\dfrac{\mathrm{e}^{-\lambda\pi}\{\lambda\pi\}^{y}}{y!} & \text{if } y = 1, 2, \ldots. \end{cases} \tag{3}$$

Proof: The proof is given in Appendix 1. ■

Remark 2.1: From Theorem 2.1, it follows that the distribution of the surrogate count is $ZIP(\lambda\pi, \delta)$. In case $\pi = 1$, $\delta = 0$, it reduces to the standard $P(\lambda)$ distribution. For $\pi = 1, \delta \neq 0$, we get the $ZIP(\lambda, \delta)$ model [21]. For $\delta = 0, \pi < 1$, it reduces to the undercount adjusted Poisson model [22].

Remark 2.2: From the model in (1), $P(Y^* = 0) = \delta + (1-\delta)\,\mathrm{e}^{-\lambda} = P_0^*$ (say). Note that $P_0^* \in [\delta, 1]$. For moderate to large δ, the system will have excess true zeros which will imply excess observed zeros also.

Remark 2.3: From models in (2) and (3) we get $P(Y^* = 0|Y = 0) = (\dfrac{\delta + (1-\delta)\,\mathrm{e}^{-\lambda}}{\delta + (1-\delta)\,\mathrm{e}^{-\lambda\pi}})$ $= P_{0|0}$ (say). Note thatred, for fixed λ and δ, $P_{0|0}$ is an increasing function of π. For fixed λ and π, $P_{0|0}$ is an increasing function of δ. Thus a large number of observed zeros in the system may in fact correspond to true zeros.

Let $Y_1, Y_2, .., Y_n$ denote the independent surrogate responses generated from the distribution in (3). Note that the model in (3) is over parametrized. Hence, based on $Y_1, Y_2, .., Y_n$, one cannot estimate the model parameters (λ, δ) and the nuisance parameter (π) simultaneously. In practice, sometimes, a crude estimate of π is available based on the knowledge of the domain experts. Assuming $\pi = \pi_0$, (a known value), one can use the model in (3), to obtain the MLE of λ and δ. The log likelihood function is given by,

$$l(\lambda,\delta) = \sum_{i=1}^{n} I_i \log\{\delta + (1-\delta)\,\mathrm{e}^{-\lambda\pi_0}\} + \sum_{i=1}^{n}(1-I_i)\{y_i \log\lambda\pi_0 - \lambda\pi_0 + \log(1-\delta) - \log y_i!\}, \tag{4}$$

where $I_i = 1$ if $y_i = 0$ and 0, otherwise. Defining $\psi(\lambda,\delta,\pi) = (1-\delta)\,\mathrm{e}^{-\lambda\pi}$, the score functions for λ and δ are given by,

$$\frac{\partial l(\lambda,\delta)}{\partial\lambda} = -\frac{\pi_0\psi(\lambda,\delta,\pi_0)}{\delta+\psi(\lambda,\delta,\pi_0)}\sum_{i=1}^{n} I_i + \lambda^{-1}\sum_{i=1}^{n}(1-I_i)(y_i - \lambda\pi_0). \tag{5}$$

$$\frac{\partial l(\lambda,\delta)}{\partial\delta} = \frac{1-\mathrm{e}^{-\lambda\pi_0}}{\delta+\psi(\lambda,\delta,\pi_0)}\sum_{i=1}^{n} I_i - (1-\delta)^{-1}\sum_{i=1}^{n}(1-I_i). \tag{6}$$

Equating the score functions in (5) and (6) to zero and solving simultaneously, we can obtain the MLEs of λ and δ which would be designated as the *Benchmark*(BM) estimators. In all future discussions we shall denote these by $\hat{\lambda}_{BM}$ and $\hat{\delta}_{BM}$ respectively. The variance-covariance matrix of $(\hat{\lambda}_{BM}, \hat{\delta}_{BM})$ is given in Appendix 2.

Further, based on the surrogate responses, one can carry out a naive analysis using the model in (1). Let the MLEs of λ and δ so obtained be denoted by $\hat{\lambda}_N$ and $\hat{\delta}_N$ respectively. The related score functions can be obtained from (5) to (6) by setting $\pi_0 = 1$. Similarly, the variance-covariance matrix of $(\hat{\lambda}_N, \hat{\delta}_N)$ can be obtained from Appendix 2 by setting $\pi_0 = 1$.

In case π is unknown, as often is the case in practice, one needs additional data for model identifiability. Here we adopt a double sampling approach of data collection, where an internal validation sample of size r is selected as a sub-sample from the original sample of size n. It is assumed that, in the validation sample, the true counts can be observed without any error. Let S_v and S_{nv} be the sets of validation and non-validation units respectively. Based on the data $\{(Y_i, Y_i^*); i \in S_v\}$ and $\{Y_i; i \in S_{nv}\}$, we obtain the MLEs of λ, δ and π. Throughout, we shall use the notation $f(.)$ to indicate a probability mass function. Now, the log likelihood function can be partitioned as,

$$\begin{aligned} l(\lambda,\delta,\pi) &= \sum_{i\in S_v} f(y_i|y_i^*;\pi) + \sum_{i\in S_v} f(y_i^*;\lambda,\delta) + \sum_{i\in S_{nv}} f(y_i;\lambda,\delta,\pi) \\ &= l_1(\pi) + l_2(\lambda,\delta) + l_3(\lambda,\delta,\pi), \text{(say)}, \end{aligned} \tag{7}$$

where, $l_1(.)$, $l_2(.)$ and $l_3(.)$ are the logarithms of the distributions in (2), (1) and (3) respectively. Writing explicitly,

$$l_1(\pi) = \sum_{i \in S_v} \left\{ y_i \log \pi + (y_i^* - y_i) \log(1-\pi) + \log \binom{y_i^*}{y_i} \right\}. \tag{8}$$

$$\begin{aligned} l_2(\lambda, \delta) &= \sum_{i \in S_v} I_{v,i} \log\{\delta + (1-\delta)\, \mathrm{e}^{-\lambda}\} \\ &\quad + \sum_{i \in S_v} (1 - I_{v,i}) \{y_i^* \log \lambda - \lambda + \log(1-\delta) - \log y_i^*!\}, \end{aligned} \tag{9}$$

where, $\forall i \in S_v$, $I_{v,i} = 1$ if $y_i^* = 0$ and 0, otherwise. Finally,

$$l_3(\lambda, \delta, \pi) = \sum_{i \in S_{nv}} I_{nv,i} \log\{\delta + (1-\delta)\, \mathrm{e}^{-\lambda\pi}\} \tag{10}$$

$$+ \sum_{i \in S_{nv}} (1 - I_{nv,i}) \{y_i \log \lambda\pi - \lambda\pi + \log(1-\delta) - \log y_i!\}, \tag{11}$$

where $\forall i \in S_{nv}$, $I_{nv,i} = 1$ if $y_i = 0$ and 0, otherwise. Writing $\psi(\lambda, \delta, \pi) = (1-\delta)\, \mathrm{e}^{-\lambda\pi}$, the score functions for λ, δ and π denoted respectively by $S(\lambda)$, $S(\delta)$ and $S(\pi)$ are given below.

$$\begin{aligned} S(\lambda) &= \frac{\partial l(\lambda, \delta, \pi)}{\partial \lambda} = \frac{\partial l_2(\lambda, \delta)}{\partial \lambda} + \frac{\partial l_3(\lambda, \delta, \pi)}{\partial \lambda} \\ &= -\frac{\psi(\lambda, \delta, 1)}{\delta + \psi(\lambda, \delta, 1)} \sum_{i \in S_v} I_{v,i} + \lambda^{-1} \sum_{i \in S_v} (1 - I_{v,i})(y_i^* - \lambda) \\ &\quad - \frac{\pi \psi(\lambda, \delta, \pi)}{\delta + \psi(\lambda, \delta, \pi)} \sum_{i \in S_{nv}} I_{nv,i} + \lambda^{-1} \sum_{i \in S_{nv}} (1 - I_{nv,i})(y_i - \lambda\pi). \end{aligned} \tag{12}$$

$$\begin{aligned} S(\delta) &= \frac{\partial l(\lambda, \delta, \pi)}{\partial \delta} = \frac{\partial l_2(\lambda, \delta)}{\partial \delta} + \frac{\partial l_3(\lambda, \delta, \pi)}{\partial \delta} \\ &= \frac{1 - \mathrm{e}^{-\lambda}}{\delta + \psi(\lambda, \delta, 1)} \sum_{i \in S_v} I_{v,i} - (1-\delta)^{-1} \sum_{i \in S_v} (1 - I_{v,i}) \\ &\quad + \frac{1 - \mathrm{e}^{-\lambda\pi}}{\delta + \psi(\lambda, \delta, \pi)} \sum_{i \in S_{nv}} I_{nv,i} - (1-\delta)^{-1} \sum_{i \in S_{nv}} (1 - I_{nv,i}). \end{aligned} \tag{13}$$

$$\begin{aligned} S(\pi) &= \frac{\partial l(\lambda, \delta, \pi)}{\partial \pi} = \frac{\partial l_1(\pi)}{\partial \pi} + \frac{\partial l_3(\lambda, \delta, \pi)}{\partial \pi} \\ &= \sum_{i \in S_v} \frac{(y_i - y_i^* \pi)}{\pi(1-\pi)} - \frac{\lambda \psi(\lambda, \delta, \pi)}{\delta + \psi(\lambda, \delta, \pi)} \sum_{i \in S_{nv}} I_{nv,i} + \pi^{-1} \sum_{i \in S_{nv}} (1 - I_{nv,i})(y_i - \lambda\pi). \end{aligned} \tag{14}$$

Equating the score functions in (12)–(14) to zero and solving simultaneously, we get the MLEs of λ, δ and π which we denote by $\hat{\lambda}_{ML}$, $\hat{\delta}_{ML}$ and $\hat{\pi}_{ML}$ respectively. The related Fisher Information matrix is provided in Appendix 3.

3. Moment estimators

The maximum likelihood estimators though efficient, are algebraically complex and have no closed form expressions in an undercount-adjusted ZIP model. So, from a practical consideration, we explore the performance of the moment estimators. Let us assume that the validation sample size is r and therefore non-validation sample size is $n—r$. To derive the moment estimators we first define the following quantities.

$$T_1 = r^{-1}\sum_{i\in S_v} Y_i^*, \quad T_2 = r^{-1}\sum_{i\in S_v} Y_i, \quad T_3 = n^{-1}\left(\sum_{i\in S_v} Y_i + \sum_{i\in S_{nv}} Y_i\right) \text{ and}$$

$$T_4 = n^{-1}\left(\sum_{i\in S_v} Y_i^2 + \sum_{i\in S_{nv}} Y_i^2\right).$$

It is to be noted that,

$$E(T_1) = (1-\delta)\,\lambda. \tag{15}$$

$$E(T_2) = (1-\delta)\,\lambda\pi. \tag{16}$$

$$E(T_3) = (1-\delta)\,\lambda\pi. \tag{17}$$

$$E(T_4) = (1-\delta)\,\lambda\pi\,(1+\lambda\pi). \tag{18}$$

The moment estimators of π, δ and λ denoted respectively by $\hat{\pi}_{MM}$, $\hat{\delta}_{MM}$ and $\hat{\lambda}_{MM}$ are then given by,

$$\hat{\pi}_{MM} = \frac{T_2}{T_1}. \tag{19}$$

$$\hat{\lambda}_{MM} = \left(\frac{T_4}{T_3} - 1\right)\cdot\frac{T_1}{T_2}. \tag{20}$$

$$\hat{\delta}_{MM} = 1 - \frac{T_3^2}{T_4 - T_3}. \tag{21}$$

Remark 3.1: The moment estimators can be expressed as explicit functions of $T_1 - T_4$. The variance covariance matrix of $(\hat{\pi}_{MM}, \hat{\lambda}_{MM}\ \hat{\delta}_{MM})$ can be derived using the delta method. The detailed expressions are given in Appendix 4.

4. Simulation study

An extensive simulation study is carried out to compare the performance of the maximum likelihood estimators with the Benchmark estimators with respect to their Bias and Mean square error (MSE). The estimates and the standard error of the estimators of the moment estimators (**MM**) and the maximum likelihood estimators (**ML)** are also compared in the tables. Further, different bootstrap confidence intervals of λ and δ are compared with respect to the length and coverage of the confidence intervals.

4.1. Bias and MSE

In this section, we obtain the simulated bias and mean square error (MSE) of the estimators of (λ, δ) for the Benchmark (**BM**) method and that of (λ, δ, π) for the maximum likelihood (**ML**) method. We shall also provide a bootstrap estimate of the bias and the mean square error of the bootstrap bias-corrected estimator. To this end, we first outline the steps for data generation.

Step 1 : For given values of λ and δ, $Y_i^*, i = 1, 2, \ldots, n$ are generated from $ZIP(\lambda, \delta)$. The study is carried out for different prefixed choices of λ and δ. The values are reported in the tables.

Step 2: Given $Y_i^* = t_i$ (from Step 1), we generate the surrogate count Y_i from $Binomial(t_i, \pi)$ for $i = 1, 2, \ldots, n$ and for some fixed values of π. The choices of π are reported in the tables.

Step 3 : A simple random sample of size r is selected from the n units and captured in S_v. For $i \in S_v$ we store the values of (Y_i, Y_i^*) and for $i \in S_{nv}$ we store the values of Y_i only.

Step 4: Based on the data $(Y_i, Y_i^*, i \in S_v, Y_i, i \in S_{nv})$, the estimators are computed using **BM** and **ML** methods. Steps 1-4 are repeated $K = 1000$ times. Let $\hat{\theta}^k$ denote the estimate of a generic parameter θ around the k^{th} simulation $(k = 1, 2, .., K)$. The accuracy and precision of the estimator are assessed based on the simulated bias and MSE respectively, which are given by, Bias $= K^{-1} \sum_{k=1}^{K} (\hat{\theta}^k - \theta)$ and $MSE = K^{-1} \sum_{k=1}^{K} (\hat{\theta}^k - \theta)^2$.

To compute the bootstrap estimate of bias for the **BM** method, note that, the surrogate count (Y) is distributed as $ZIP(\lambda\pi, \delta)$. With π known, we generate $B = 1000$ bootstrap samples using the distribution of Y with the true values of λ and δ replaced by their estimates at the k^{th} simulation (i.e. by $\hat{\lambda}_{BM}^k$ and $\hat{\delta}_{BM}^k$). Now, based on the b^{th} bootstrap sample at the k^{th} simulation, we compute $\hat{\lambda}_{BM}^k(b)$ and $\hat{\delta}_{BM}^k(b)$. The bootstrap estimate of bias is then given by,

$$Bias_{boot}(\hat{\lambda}_{BM}) = K^{-1} \sum_{k=1}^{K} (\hat{\lambda}_{BM}^k(B) - \hat{\lambda}_{BM}^k), \tag{22}$$

$$Bias_{boot}(\hat{\delta}_{BM}) = K^{-1} \sum_{k=1}^{K} (\hat{\delta}_{BM}^k(B) - \hat{\delta}_{BM}^k), \tag{23}$$

where, $\hat{\lambda}_{BM}^k(B) = B^{-1} \sum_{b=1}^{B} \hat{\lambda}_{BM}^k(b)$ and $\hat{\delta}_{BM}^k(B) = B^{-1} \sum_{b=1}^{B} \hat{\delta}_{BM}^k(b)$.

To compute the bootstrap bias of the maximum likelihood estimators, we generate $B = 1000$ bootstrap samples using the distribution of true and surrogate count as in Steps 1-2 with the true values of λ, δ and π replaced at the k^{th} simulation by $\hat{\lambda}_{ML}^k$, $\hat{\delta}_{ML}^k$ and $\hat{\pi}_{ML}^k$ respectively. Based on the validation and non-validation data (generated using Step 3), we next compute the maximum likelihood estimators. Let $\hat{\lambda}_{ML}^k(b)$, $\hat{\delta}_{ML}^k(b)$ and $\hat{\pi}_{ML}^k(b)$ denote respectively the bootstrap estimators of λ, δ and π around the k^{th} simulation and b^{th} bootstrap sample. Defining, $\hat{\lambda}_{ML}^k(B) = B^{-1} \sum_{b=1}^{B} \hat{\lambda}_{ML}^k(b)$, $\hat{\delta}_{ML}^k(B) = B^{-1} \sum_{b=1}^{B} \hat{\delta}_{ML}^k(b)$ and

Table 1. Bootstrap (Boot) and Simulated (Sim) Bias and MSE (in parenthesis) of the estimator of δ, λ and π for different choices of π and (n, r) for **BM** and **ML** methods.

		δ= 0.1		λ= 5		
(n, r)		BM	ML	BM	ML	π ML
			π = 0.7			
(100, 10)	Boot	−1 (13)	14 (18)	1 (865)	−876 (2340)	150 (40)
	Sim	26 (12)	11 (17)	-y110 (861)	−442 (2142)	100 (39)
(100, 30)	Boot	−1 (13)	21 (16)	1 (865)	−554 (1042)	55 (17)
	Sim	26 (12)	−16 (14)	−110 (861)	−543 (1046)	49 (17)
(300, 30)	Boot	−2 (5)	6 (6)	−5 (341)	−527 (727)	78 (12)
	Sim	23 (5)	38 (5)	−76 (340)	−200 (694)	28 (12)
(300, 90)	Boot	−2 (5)	8 (4)	−5 (341)	−295 (314)	29 (4)
	Sim	23 (5)	−1 (4)	−76 (340)	−164 (311)	30 (4)
			π = 0.8			
(100, 10)	Boot	−2 (10)	5 (13)	−5 (721)	−583 (1654)	102 (34)
	Sim	2 (10)	48 (12)	−48 (720)	−224 (1506)	82 (32)
(100, 30)	Boot	−2 (10)	10 (13)	−5 (721)	−547 (934)	48 (11)
	Sim	2 (10)	37(12)	−48 (720)	−8 (881)	23 (11)
(300, 30)	Boot	−2 (4)	0 (5)	−7 (220)	−370 (526)	59 (11)
	Sim	11 (4)	43 (5)	−152 (220)	194 (477)	2 (10)
(300, 90)	Boot	−2 (4)	2 (5)	−7 (220)	−269 (352)	26 (3)
	Sim	11 (4)	22 (5)	−152 (220)	−120 (347)	13 (3)
			π = 0.9			
(100, 10)	Boot	−2 (10)	1 (11)	−15 (724)	−196 (1402)	26 (22)
	Sim	3 (10)	−20 (10)	−129 (727)	217 (1263)	−5 (18)
(100, 30)	Boot	−2 (10)	5 (10)	−15 (724)	−367 (835)	24 (6)
	Sim	3 (10)	−21 (10)	−129 (727)	−228 (819)	2 (5)
(300, 30)	Boot	−1 (3)	0 (3)	−8 (228)	−171 (361)	27 (6)
	Sim	0 (3)	5 (3)	−121 (228)	93 (345)	13 (5)
(300, 90)	Boot	−1 (3)	1 (4)	−8 (228)	−193 (250)	14 (2)
	Sim	0 (3)	29 (4)	−121 (228)	−5 (242)	6 (2)

Note: True value of $\lambda = 5$ and $\delta = 0.1$.

$\hat{\pi}^k_{ML}(B) = B^{-1} \sum_{b=1}^{B} \hat{\pi}^k_{ML}(b)$, the bootstrap bias of the estimators are given by,

$$Bias_{boot}(\hat{\lambda}_{ML}) = K^{-1} \sum_{k=1}^{K} (\hat{\lambda}^k_{ML}(B) - \hat{\lambda}^k_{ML}). \tag{24}$$

$$Bias_{boot}(\hat{\delta}_{ML}) = K^{-1} \sum_{k=1}^{K} (\hat{\delta}^k_{ML}(B) - \hat{\delta}^k_{ML}). \tag{25}$$

$$Bias_{boot}(\hat{\pi}_{ML}) = K^{-1} \sum_{k=1}^{K} (\hat{\pi}^k_{ML}(B) - \hat{\pi}^k_{ML}). \tag{26}$$

Finally, the bias-corrected estimator of a generic scalar parameter θ is given by,

$$\hat{\theta}^k(corrected) = \hat{\theta}^k - (\hat{\theta}^k(B) - \hat{\theta}^k). \tag{27}$$

The mean square error of the bootstrap bias-corrected estimator is then given by,

$$MSE_{boot}(\hat{\theta}) = K^{-1} \sum_{k=1}^{K} (\hat{\theta}^k(corrected) - \theta)^2. \tag{28}$$

Table 2. Bootstrap (Boot) and Simulated (Sim) Bias and MSE (in parenthesis) of the estimator of δ, λ and π for different choices of π and (n, r) for **BM** and **ML** methods.

		δ= 0.3		λ= 5		
(n, r)		BM	ML	BM	ML	π ML
			$\pi = 0.7$			
(100, 10)	Boot	−7 (20)	−3 (24)	−20 (1073)	−1203 (2803)	188 (43)
	Sim	−6 (20)	−65 (24)	105 (1070)	−1050 (2385)	162 (39)
(100, 30)	Boot	−7 (20)	−2 (24)	−20 (1073)	−953 (1497)	81 (17)
	Sim	−6 (20)	16 (24)	105 (1070)	−774 (1461)	44 (17)
(300, 30)	Boot	−3 (7)	0 (8)	−10 (375)	−849 (885)	133 (13)
	Sim	19 (7)	−31 (8)	42 (372)	−1055 (898)	141 (14)
(300, 90)	Boot	−3 (7)	1 (7)	−10 (375)	−542 (543)	51 (6)
	Sim	19 (7)	14 (7)	42 (372)	0 (496)	3 (6)
			$\pi = 0.8$			
(100, 10)	Boot	−3 (23)	−2 (17)	−22 (993)	−541 (2485)	90 (50)
	Sim	−53 (23)	26 (17)	−180 (990)	−79 (2098)	41 (41)
(100, 30)	Boot	−3 (23)	1 (27)	−22 (993)	−592 (1200)	48 (16)
	Sim	−53 (23)	−16 (27)	−180 (990)	−640 (1184)	57 (16)
(300, 30)	Boot	−2 (8)	−1 (10)	−1 (328)	−338 (1079)	56 (18)
	Sim	−35 (8)	−51 (9)	−233 (328)	−164 (991)	13 (16)
(300, 90)	Boot	−2 (8)	1 (17)	−1 (328)	−253 (741)	26 (6)
	Sim	−35 (8)	−104 (16)	−233 (328)	−353 (735)	−27 (6)
			$\pi = 0.9$			
(100, 10)	Boot	−2 (23)	−6 (28)	−12 (867)	−130 (1549)	9 (30)
	Sim	−6 (23)	−23 (28)	−154 (872)	452 (1311)	4 (22)
(100, 30)	Boot	−2 (23)	1 (20)	−12 (867)	−489 (988)	29 (11)
	Sim	−6 (23)	35 (20)	−154 (872)	−348 (963)	61 (10)
(300, 30)	Boot	−1 (8)	0 (6)	−2 (273)	−202 (442)	34 (8)
	Sim	−20 (8)	23 (6)	−70 (272)	−322 (432)	64 (7)
(300, 90)	Boot	−1 (8)	0 (6)	−2 (273)	−244 (274)	21 (3)
	Sim	−20 (8)	−28 (6)	−70 (272)	64 (260)	7 (3)

Note: True value of $\lambda = 5$ and $\delta = 0.3$.

Tables 1 and 2 report the Bias and MSE of the estimators for different combinations of (n, r) and (λ, δ, π). From the results, some salient features are noted. Bias and MSE decrease with increase in the sample size n. The magnitudes of bias and MSE for the estimate of δ are low compared to that of λ. The **ML** method overestimates λ, which is evident from consistent high bias and MSE values as compared to the **BM** method. The MSEs of the bootstrap bias-corrected estimators are close to the simulated MSEs. Importantly, high δ and low π combinations produce damaging results (See Table 2, $\pi = 0.7$ and $\delta = 0.3$ combination). We compare the Bias (MSE) values of the parameter estimator λ with respect to the best case ($\delta = 0.1, \pi = 0.9$) and the worst case ($\delta = 0.3, \pi = 0.7$) scenario corresponding to $(n, r) = (100, 10)$ combination. Simulated bias (MSE) of the estimator of λ, for the **ML** method comes out as 217(1263) under the best case and -1050(2385) under the worst case.

A further computation is done to compare the performance of the moment estimators with the maximum likelihood estimators. Tables 3 and 4 report the estimate (*Est*) and the standard error of the estimates (*SE*) for **MM** and **ML** methods. For $\lambda = 5.0$, the results are reported for different combinations of (n, r) and (π, δ). The results reveal that the moment estimators of λ, δ and π yield larger standard error compared to those under the likelihood method. This feature holds uniformly for all parametric configurations. Notably the standard error of the estimator of π is substantially higher using the method of moment compared to its likelihood counterpart. So, in the next part of our study, we compute the bootstrap confidence intervals for the likelihood based methods only.

Table 3. Estimate and standard error (in parenthesis) of the estimator of δ, λ and π for different choices of π and (n, r) for **MM** and **ML** methods.

	$\delta = 0.1$		$\lambda = 5$		π	
(n, r)	MM	ML	MM	ML	MM	ML
			$\pi = 0.7$			
(100, 10)	0.0982	0.1044	5.0582	4.9726	0.695	0.7043
	(0.1896)	(0.0345)	(1.3481)	(0.3148)	(3.0987)	(0.0188)
(100, 30)	0.0911	0.0967	4.9903	5.0055	0.6976	0.6973
	(0.1049)	(0.0327)	(0.8044)	(0.2884)	(3.0947)	(0.0191)
(300, 30)	0.096	0.1031	5.0215	5.0027	0.6945	0.6988
	(0.1096)	(0.0199)	(0.7732)	(0.181)	(1.794)	(0.0109)
(300, 90)	0.0955	0.0994	4.9964	5.0096	0.698	0.6983
	(0.0612)	(0.0192)	(0.468)	(0.1666)	(1.7982)	(0.011)
			$\pi = 0.8$			
(100, 10)	0.0972	0.1042	4.9881	0.7924	0.8012	
	(0.2108)	(0.0328)	(1.3349)	(0.2777)	(3.7655)	(0.0113)
(100, 30)	0.0911	0.1016	4.9875	4.9876	0.7973	0.799
	(0.1164)	(0.0319)	(0.8267)	(0.2654)	(3.7536)	(0.0121)
(300, 30)	0.0968	0.1032	5.0111	4.9909	0.7958	0.7999
	(0.1223)	(0.019)	(0.7662)	(0.1596)	(2.1819)	(0.0065)
(300, 90)	0.0954	0.1033	4.989	4.9925	0.7987	0.8
	(0.068)	(0.0187)	(0.4805)	(0.1535)	(2.1814)	(0.007)
			$\pi = 0.9$			
(100, 10)	0.0966	0.1023	5.0063	4.9966	0.8959	0.8978
	(0.2331)	(0.0316)	(1.3204)	(0.2552)	(4.4791)	(0.0054)
(100, 30)	0.0919	0.0992	4.9891	4.9808	0.8977	0.8986
	(0.1283)	(0.031)	(0.8519)	(0.2492)	(4.4782)	(0.0059)
(300, 30)	0.0973	0.0995	4.9923	4.9821	0.8989	0.9004
	(0.1352)	(0.0186)	(0.7621)	(0.148)	(2.5987)	(0.003)
(300, 90)	0.0958	0.1002	4.9878	4.9901	0.8992	0.8995
	(0.0747)	(0.0181)	(0.4942)	(0.1445)	(2.5964)	(0.0034)

Note: True value of $\lambda = 5$ and $\delta = 0.1$.

4.2. Bootstrap confidence intervals

To study the finite sample behaviour of the estimators we compute various bootstrap confidence intervals and compare them with the Wald confidence interval. We illustrate the methodology with reference to a generic scalar parameter θ. Note that, for the **BM** method, θ could be λ or δ and for the **ML** method, θ could be λ, δ or π. Based on the $b^{th}(b = 1, 2, .., B)$ bootstrap sample and around the $k^{th}(k = 1, 2, .., K)$ simulation, we obtain the bootstrap estimate of θ, which we shall denote by $\hat{\theta}^k(b)$. Further, let $\hat{\theta}^k(B) = B^{-1}\sum_{b=1}^{B}\hat{\theta}^k(b)$ and the standard error of the estimator is given by,

$$\hat{\sigma}^k(B) = \sqrt{B^{-1}\sum_{b=1}^{B}(\hat{\theta}^k(b) - \hat{\theta}^k(B))^2}. \tag{29}$$

For a typical (kth) simulation, the various bootstrap confidence intervals are described below. For simplicity of notations we have dropped k.

Bootstrap Wald type (*Boot-Z*): The confidence interval for θ is given by, $\hat{\theta}(B) \pm \tau_{\alpha/2} \times \hat{\sigma}(B)$ where, $\tau_{\alpha/2}$ is the upper $100 \times \alpha/2\%$ point of a standard normal distribution and $\hat{\sigma}(B)$ is given in (29).

Bootstrap Studentized type (*Boot-t*): We first compute $Z(b) = \dfrac{\hat{\theta}(b) - \hat{\theta}}{\hat{\sigma}(b)}$, where $\hat{\sigma}(b)$

Table 4. Estimate and standard error (in parenthesis) of the estimator of δ, λ and π for different choices of π and (n, r).

	δ= 0.3		λ= 5		π	
(n, r)	MM	ML	MM	ML	MM	ML
			$\pi = 0.7$			
(100, 10)	0.2903	0.2908	4.9861	4.8892	0.7097	0.7197
	(0.2759)	(0.0498)	(1.6918)	(0.3506)	(3.1624)	(0.0200)
(100, 30)	0.2922	0.2953	4.9927	4.9408	0.6978	0.7018
	(0.151)	(0.0474)	(0.9868)	(0.3224)	(3.1334)	(0.0213)
(300, 30)	0.2952	0.2946	4.985	4.8872	0.704	0.715
	(0.1585)	(0.0281)	(0.9797)	(0.1994)	(1.830)	(0.0116)
(300, 90)	0.2909	0.2952	4.9904	4.961	0.6987	0.7027
	(0.0868)	(0.0274)	(0.568)	(0.1863)	(1.8203)	(0.0122)
			$\pi = 0.8$			
(100, 10)	0.2906	0.2941	4.9755	4.9871	0.8079	0.8091
	(0.3083)	(0.0468)	(1.6797)	(0.3135)	(3.8814)	(0.0122)
(100, 30)	0.2927	0.2938	4.9685	4.9473	0.8018	0.8021
	(0.1704)	(0.0465)	(1.011)	(0.2985)	(3.8616)	(0.0135)
(300, 30)	0.2946	0.2883	4.9912	4.9803	0.8012	0.7984
	(0.1771)	(0.0279)	(0.9745)	(0.1844)	(2.2442)	(0.0074)
(300, 90)	0.2916	0.2886	4.9879	5.0018	0.7988	0.7934
	(0.0975)	(0.0268)	(0.5835)	(0.1737)	(2.238)	(0.0081)
			$\pi = 0.9$			
(100, 10)	0.292	0.2944	4.9805	5.0125	0.9061	0.904
	(0.3416)	(0.0462)	(1.6792)	(0.2881)	(4.6925)	(0.0057)
(100, 30)	0.293	0.3014	4.9672	4.9552	0.9009	0.9052
	(0.1886)	(0.0465)	(1.038)	(0.2826)	(4.6463)	(0.0063)
(300, 30)	0.2951	0.2966	4.9772	4.9799	0.9038	0.9043
	(0.1972)	(0.0268)	(0.9712)	(0.1658)	(2.7058)	(0.0033)
(300, 90)	0.2921	0.2949	4.9905	5.0023	0.8989	0.8997
	(0.1081)	(0.0267)	(0.5998)	(0.1639)	(2.6974)	(0.0038)

Note: True value of $\lambda = 5$ and $\delta = 0.3$.

is the standard deviation of the estimator obtained from the inverse of Fisher Information, based on the b^{th} bootstrap sample. Suppose the upper and lower α% points based on $Z(b)(b = 1, 2, .., B)$ be denoted by q_α and $q_{1-\alpha}$ respectively. Then the confidence interval for θ is given by $[\hat{\theta}(B) + q_{1-\alpha} \times \hat{\sigma}(B), \quad \hat{\theta}(B) + q_\alpha \times \hat{\sigma}(B)]$.

Bootstrap percentile (*Boot-Per*): The confidence interval in this case is given by $[q_{low,\alpha}, q_{up,\alpha}]$, where $q_{up,\alpha}$ and $q_{low,\alpha}$ denote respectively the upper and lower percentiles of $\hat{\theta}(b)(b = 1, 2, .., B)$.

Bootstrap BCa : BCa intervals are based on a bias correction factor and an acceleration factor. We define the bias correction factor as $\hat{z}_0 = \Phi^{-1}(\frac{\#\hat{\theta}(b) < \hat{\theta}}{B})$, where Φ denotes the cdf of a standard normal distribution. The acceleration factor is defined as, $\hat{a}_0 = \frac{\sum_{i=1}^{n} (\hat{\theta}^{\bar{i}} - \hat{\theta}^{i-})^3}{6\{\sum_{i=1}^{n} (\hat{\theta}^{i-} - \hat{\theta}^{\bar{i}})^2\}^{3/2}}$, where $\hat{\theta}^{i-}$ is the jackknife estimate of θ obtained after deleting the i^{th} data point and $\hat{\theta}^{\bar{i}} = n^{-1} \sum_{i=1}^{n} \hat{\theta}^{i-}$. Note that the i^{th} data point comprise the pair

Table 5. Average length and coverage of confidence intervals (in parenthesis) of δ and λ for different choices of π and (n, r) for **BM** and **ML** methods.

		δ= 0.1		λ= 5	
(n, r)	Interval	BM	ML	BM	ML
		π = 0.7			
(100, 10)	Wald	0.138 (0.950)	0.133 (0.880)	1.168 (0.940)	1.221 (0.835)
	Boot-Z	0.137 (0.945)	0.129 (0.880)	1.170 (0.950)	1.803 (0.935)
	Boot-t	0.142 (0.935)	0.134 (0.880)	1.196 (0.960)	1.854 (0.920)
	Boot-Per	0.136 (0.940)	0.128 (0.875)	1.166 (0.960)	1.785 (0.940)
	Boot-Bca	0.139 (0.960)	0.130 (0.870)	1.168 (0.950)	1.871 (0.950)
(100, 30)	Wald	0.138 (0.950)	0.129 (0.905)	1.168 (0.940)	1.118 (0.920)
	Boot-Z	0.137 (0.945)	0.126 (0.910)	1.170 (0.950)	1.283 (0.935)
	Boot-t	0.142 (0.935)	0.132 (0.900)	1.196 (0.960)	1.320 (0.925)
	Boot-Per	0.136 (0.940)	0.125 (0.915)	1.166 (0.960)	1.281 (0.940)
	Boot-Bca	0.139 (0.960)	0.126 (0.905)	1.168 (0.950)	1.282 (0.960)
(300, 30)	Wald	0.080 (0.925)	0.078 (0.930)	0.676 (0.920)	0.706 (0.815)
	Boot-Z	0.079 (0.920)	0.077 (0.930)	0.676 (0.925)	1.068 (0.930)
	Boot-t	0.081 (0.915)	0.078 (0.935)	0.672 (0.920)	1.086 (0.930)
	Boot-Per	0.079 (0.910)	0.077 (0.930)	0.672 (0.920)	1.062 (0.940)
	Boot-Bca	0.079 (0.915)	0.077 (0.915)	0.673 (0.910)	1.079 (0.945)
(300, 90)	Wald	0.080 (0.925)	0.075 (0.945)	0.676 (0.920)	0.650 (0.925)
	Boot-Z	0.079 (0.920)	0.075 (0.950)	0.676 (0.925)	0.732 (0.955)
	Boot-t	0.081 (0.915)	0.076 (0.930)	0.672 (0.920)	0.738 (0.945)
	Boot-Per	0.079 (0.910)	0.074 (0.950)	0.672 (0.920)	0.728 (0.950)
	Boot-Bca	0.079 (0.915)	0.075 (0.950)	0.673 (0.910)	0.733 (0.970)
		π = 0.8			
(100, 10)	Wald	0.128 (0.955)	0.128 (0.935)	1.071 (0.950)	1.079 (0.835)
	Boot-Z	0.128 (0.940)	0.127 (0.935)	1.073 (0.950)	1.545 (0.945)
	Boot-t	0.137 (0.895)	0.134 (0.935)	1.080 (0.960)	1.550 (0.950)
	Boot-Per	0.127 (0.955)	0.126 (0.935)	1.069 (0.955)	1.534 (0.945)
	Boot-Bca	0.131 (0.980)	0.128 (0.915)	1.071 (0.955)	1.604 (0.935)
(100, 30)	Wald	0.128 (0.955)	0.126 (0.950)	1.071 (0.950)	1.042 (0.930)
	Boot-Z	0.128 (0.940)	0.124 (0.950)	1.073 (0.950)	1.191 (0.945)
	Boot-t	0.137 (0.895)	0.131 (0.925)	1.080 (0.960)	1.215 (0.940)
	Boot-Per	0.127 (0.955)	0.123 (0.955)	1.069 (0.955)	1.189 (0.940)
	Boot-Bca	0.131 (0.980)	0.126 (0.940)	1.071 (0.955)	1.194 (0.960)
(300, 30)	Wald	0.075 (0.950)	0.075 (0.900)	0.618 (0.980)	0.627 (0.840)
	Boot-Z	0.075 (0.950)	0.074 (0.900)	0.619 (0.980)	0.922 (0.965)
	Boot-t	0.076 (0.945)	0.075 (0.910)	0.613 (0.980)	0.927 (0.955)
	Boot-Per	0.074 (0.950)	0.074 (0.895)	0.615 (0.985)	0.919 (0.955)
	Boot-Bca	0.075 (0.945)	0.075 (0.895)	0.616 (0.985)	0.937 (0.955)
(300, 90)	Wald	0.075 (0.950)	0.073 (0.925)	0.618 (0.980)	0.601 (0.885)
	Boot-Z	0.075 (0.950)	0.073 (0.915)	0.619 (0.980)	0.677 (0.925)
	Boot-t	0.076 (0.945)	0.074 (0.900)	0.613 (0.980)	0.678 (0.915)
	Boot-Per	0.074 (0.950)	0.072 (0.935)	0.615 (0.985)	0.674 (0.910)
	Boot-Bca	0.075 (0.945)	0.073 (0.915)	0.616 (0.985)	0.675 (0.940)
		π = 0.9			
(100, 10)	Wald	0.124 (0.948)	0.123 (0.932)	1.000 (0.921)	0.993 (0.859)
	Boot-Z	0.123 (0.948)	0.122 (0.929)	1.001 (0.924)	1.254 (0.935)
	Boot-t	0.134 (0.934)	0.131 (0.915)	1.001 (0.921)	1.240 (0.918)
	Boot-Per	0.122 (0.955)	0.121 (0.941)	0.996 (0.921)	1.242 (0.941)
	Boot-Bca	0.126 (0.952)	0.125 (0.931)	0.997 (0.921)	1.280 (0.923)
(100, 30)	Wald	0.124 (0.948)	0.123 (0.934)	1.000 (0.921)	0.975 (0.879)
	Boot-Z	0.123 (0.948)	0.120 (0.937)	1.001 (0.924)	1.068 (0.896)
	Boot-t	0.134 (0.934)	0.129 (0.919)	1.001 (0.921)	1.077 (0.899)
	Boot-Per	0.122 (0.955)	0.119 (0.951)	0.996 (0.921)	1.062 (0.899)
	Boot-Bca	0.126 (0.952)	0.123 (0.948)	0.997 (0.921)	1.070 (0.922)
(300, 30)	Wald	0.072 (0.941)	0.072 (0.953)	0.576 (0.934)	0.577 (0.901)
	Boot-Z	0.072 (0.934)	0.072 (0.950)	0.577 (0.934)	0.750 (0.964)
	Boot-t	0.073 (0.914)	0.073 (0.945)	0.572 (0.934)	0.747 (0.956)
	Boot-Per	0.071 (0.934)	0.071 (0.956)	0.573 (0.934)	0.745 (0.964)
	Boot-Bca	0.072 (0.934)	0.072 (0.953)	0.574 (0.938)	0.754 (0.970)

(*continued*).

Table 5. Continued.

		δ= 0.1		λ= 5	
(n, r)	Interval	BM	ML	BM	ML
(300, 90)	Wald	0.072 (0.941)	0.071 (0.945)	0.576 (0.934)	0.567 (0.930)
	Boot-Z	0.072 (0.934)	0.071 (0.948)	0.577 (0.934)	0.616 (0.948)
	Boot-t	0.073 (0.914)	0.072 (0.945)	0.572 (0.934)	0.616 (0.953)
	Boot-Per	0.071 (0.934)	0.071 (0.940)	0.573 (0.934)	0.613 (0.948)
	Boot-Bca	0.072 (0.934)	0.071 (0.937)	0.574 (0.938)	0.613 (0.953)

Note: True value of $\lambda = 5$ and $\delta = 0.1$.

(Y_i^*, Y_i) for $i \in S_v$ and only Y_i for $i \in S_{nv}$. Finally define,

$$\alpha_{\text{low}} = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + \tau_{1-\alpha}}{1 - \hat{a}_0\left(\hat{z}_0 + \tau_{1-\alpha}\right)}\right\} \quad \text{and} \quad \alpha_{\text{up}} = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + \tau_{\alpha}}{1 - \hat{a}_0\left(\hat{z}_0 + \tau_{\alpha}\right)}\right\}. \tag{30}$$

The BCa interval is given by $[\tilde{q}_{low,\alpha}\ \tilde{q}_{up,\alpha}]$, where $\tilde{q}_{low,\alpha}$ and $\tilde{q}_{up,\alpha}$ denote respectively the $\alpha_{low}th$ and $\alpha_{up}th$ percentiles of the bootstrap distribution of $\hat{\theta}(b)(b = 1, 2, .., B)$.

Wald Interval : The Wald confidence interval for θ is given by $\hat{\theta} \pm \tau_{\alpha/2} \times \hat{\sigma}$, where $\hat{\sigma}$ is the estimate of the standard error of $\hat{\theta}$ obtained from the inverse of Fisher Information.

The lengths of the confidence intervals are obtained by averaging the difference of the upper and lower confidence limits over the K simulations. The empirical coverage is given by the proportion of times (out of K) the random interval contains the true value of the parameter θ. In our study we have chosen $K = 1000$, $B = 1000$ and $\alpha = 0.05$.

Tables 5 and 6 report the average lengths and coverage (within parentheses) of different types of confidence intervals for δ and λ. Results are provided for **BM** and **ML** estimators. The true choice of λ is taken as 5.0. The study is carried out for several prefixed choices of δ and π. The values chosen are reported in the tables. However for δ as large as 0.5 and π as low as 0.5, the likelihood approach reportedly fails to produce valid estimates of the parameters. A reason for this could be that large δ and low π results in too many zeros in the system and in such a case, the validation sample may fail to contribute to the estimation of π. So, essentially, in this situation, the non-validation component of the likelihood function plays a dominating role in the estimation of the nuisance parameter π as well as the parameters of interest (λ, δ) which causes the identifiability problem as mentioned in Section 2.

The results show that the confidence interval of both λ and δ becomes tighter with increase in the sample size n, or, for fixed n, with increase in the subsample size r. The Wald interval gives suboptimal coverage compared to other bootstrap intervals, especially when (n, r) combination is low. The Boot-Z method often overestimates the coverage and results in wider confidence intervals especially when δ and π are both large. The BCa interval returns coverage value close to the nominal level for almost all parametric configurations.

5. Real data study

In this section we analyse the data on daily consumption of beverages by individuals at their workplace [23]. The data was collected in the 1980 Wave II of the National Survey of Personal Health Practices and Consequences (NSPHPC). The beverage consumption

Table 6. Average length and coverage of confidence intervals (in parenthesis) of δ and λ for different choices of π and (n, r) for **BM** and **ML** methods.

		$\delta = 0.3$		$\lambda = 5$	
(n, r)	Interval	BM	ML	BM	ML
		$\pi = 0.7$			
(100, 10)	Wald	0.191 (0.960)	0.187 (0.950)	1.319 (0.950)	1.358 (0.821)
	Boot-Z	0.189 (0.955)	0.188 (0.950)	1.330 (0.955)	1.955 (0.927)
	Boot-t	0.195 (0.960)	0.191 (0.922)	1.318 (0.930)	2.066 (0.894)
	Boot-Per	0.188 (0.955)	0.187 (0.955)	1.328 (0.960)	1.939 (0.944)
	Boot-Bca	0.189 (0.960)	0.188 (0.961)	1.330 (0.955)	2.061 (0.950)
(100, 30)	Wald	0.191 (0.960)	0.187 (0.941)	1.319 (0.950)	1.263 (0.903)
	Boot-Z	0.189 (0.955)	0.187 (0.941)	1.330 (0.955)	1.487 (0.930)
	Boot-t	0.195 (0.960)	0.191 (0.930)	1.318 (0.930)	1.538 (0.903)
	Boot-Per	0.188 (0.955)	0.187 (0.951)	1.328 (0.960)	1.481 (0.935)
	Boot-Bca	0.189 (0.960)	0.187 (0.962)	1.330 (0.955)	1.494 (0.957)
(300, 30)	Wald	0.110 (0.955)	0.109 (0.944)	0.765 (0.955)	0.780 (0.806)
	Boot-Z	0.110 (0.955)	0.109 (0.949)	0.768 (0.955)	1.140 (0.908)
	Boot-t	0.111 (0.950)	0.109 (0.949)	0.761 (0.950)	1.180 (0.888)
	Boot-Per	0.109 (0.960)	0.108 (0.949)	0.764 (0.955)	1.134 (0.913)
	Boot-Bca	0.109 (0.970)	0.108 (0.949)	0.764 (0.950)	1.175 (0.959)
(300, 90)	Wald	0.110 (0.955)	0.108 (0.935)	0.765 (0.955)	0.739 (0.895)
	Boot-Z	0.110 (0.955)	0.108 (0.940)	0.768 (0.955)	0.869 (0.930)
	Boot-t	0.111 (0.950)	0.108 (0.935)	0.761 (0.950)	0.893 (0.930)
	Boot-Per	0.109 (0.960)	0.108 (0.940)	0.764 (0.955)	0.870 (0.930)
	Boot-Bca	0.109 (0.970)	0.108 (0.945)	0.764 (0.950)	0.859 (0.940)
		$\pi = 0.8$			
(100, 10)	Wald	0.185 (0.925)	0.185 (0.965)	1.205 (0.940)	1.238 (0.809)
	Boot-Z	0.184 (0.920)	0.185 (0.950)	1.213 (0.935)	1.741 (0.957)
	Boot-t	0.189 (0.915)	0.189 (0.950)	1.204 (0.940)	1.770 (0.936)
	Boot-Per	0.183 (0.920)	0.184 (0.950)	1.212 (0.940)	1.730 (0.957)
	Boot-Bca	0.184 (0.945)	0.184 (0.979)	1.214 (0.945)	1.809 (0.922)
(100, 30)	Wald	0.185 (0.925)	0.183 (0.924)	1.205 (0.940)	1.171 (0.938)
	Boot-Z	0.184 (0.920)	0.183 (0.917)	1.213 (0.935)	1.351 (0.945)
	Boot-t	0.189 (0.915)	0.187 (0.917)	1.204 (0.940)	1.388 (0.931)
	Boot-Per	0.183 (0.920)	0.182 (0.924)	1.212 (0.940)	1.348 (0.945)
	Boot-Bca	0.184 (0.945)	0.183 (0.917)	1.214 (0.945)	1.363 (0.972)
(300, 30)	Wald	0.108 (0.950)	0.109 (0.962)	0.694 (0.955)	0.717 (0.756)
	Boot-Z	0.107 (0.950)	0.106 (0.947)	0.701 (0.960)	1.015 (0.931)
	Boot-t	0.108 (0.935)	0.109 (0.962)	0.693 (0.955)	1.041 (0.916)
	Boot-Per	0.106 (0.940)	0.105 (0.954)	0.699 (0.965)	1.010 (0.931)
	Boot-Bca	0.106 (0.945)	0.105 (0.954)	0.698 (0.960)	1.040 (0.901)
(300, 90)	Wald	0.108 (0.950)	0.112 (0.899)	0.694 (0.955)	0.699 (0.916)
	Boot-Z	0.107 (0.950)	0.105 (0.882)	0.701 (0.960)	0.776 (0.933)
	Boot-t	0.108 (0.935)	0.112 (0.891)	0.693 (0.955)	0.808 (0.916)
	Boot-Per	0.106 (0.940)	0.105 (0.891)	0.699 (0.965)	0.774 (0.941)
	Boot-Bca	0.106 (0.945)	0.105 (0.908)	0.698 (0.960)	0.769 (0.941)
		$\pi = 0.9$			
(100, 10)	Wald	0.183 (0.945)	0.183 (0.940)	1.133 (0.959)	1.147 (0.901)
	Boot-Z	0.182 (0.945)	0.182 (0.934)	1.140 (0.962)	1.426 (0.967)
	Boot-t	0.187 (0.945)	0.186 (0.943)	1.129 (0.948)	1.420 (0.961)
	Boot-Per	0.181 (0.945)	0.180 (0.934)	1.131 (0.959)	1.416 (0.970)
	Boot-Bca	0.182 (0.948)	0.181 (0.948)	1.136 (0.962)	1.445 (0.948)
(100, 30)	Wald	0.183 (0.945)	0.183 (0.964)	1.133 (0.959)	1.114 (0.911)
	Boot-Z	0.182 (0.945)	0.183 (0.961)	1.140 (0.962)	1.228 (0.916)
	Boot-t	0.187 (0.945)	0.186 (0.961)	1.129 (0.948)	1.247 (0.908)
	Boot-Per	0.181 (0.945)	0.181 (0.961)	1.131 (0.959)	1.223 (0.928)
	Boot-Bca	0.182 (0.948)	0.182 (0.964)	1.136 (0.962)	1.227 (0.936)
(300, 30)	Wald	0.106 (0.966)	0.105 (0.957)	0.654 (0.934)	0.655 (0.855)
	Boot-Z	0.105 (0.962)	0.105 (0.960)	0.657 (0.938)	0.843 (0.925)
	Boot-t	0.106 (0.955)	0.105 (0.952)	0.651 (0.928)	0.842 (0.925)
	Boot-Per	0.104 (0.966)	0.105 (0.957)	0.652 (0.928)	0.839 (0.933)
	Boot-Bca	0.104 (0.969)	0.105 (0.962)	0.651 (0.931)	0.854 (0.919)

(*continued*).

Table 6. Continued.

		δ= 0.3		λ= 5	
(n,r)	Interval	BM	ML	BM	ML
(300, 90)	Wald	0.106 (0.966)	0.105 (0.959)	0.654 (0.934)	0.644 (0.946)
	Boot-Z	0.105 (0.962)	0.105 (0.966)	0.657 (0.938)	0.701 (0.956)
	Boot-t	0.106 (0.955)	0.105 (0.961)	0.651 (0.928)	0.705 (0.946)
	Boot-Per	0.104 (0.966)	0.105 (0.964)	0.652 (0.928)	0.698 (0.953)
	Boot-Bca	0.104 (0.969)	0.105 (0.964)	0.651 (0.931)	0.699 (0.964)

Note: True value of $\lambda = 5$ and $\delta = 0.3$.

Table 7. AIC values of $P(\lambda)$ and $ZIP(\lambda, \delta)$ fitting.

Fitted Dist	Coffee	Tea	Milk
Poisson	9703.26	5386.47	5575.35
ZIP	8685.05	4719.68	5474.55

is measured in terms of the number of glasses or cups consumed by an individual. Let $Y_1^*, Y_2^*, .., Y_n^*$ denote the consumption counts reported by the n individuals. We are provided with consumption data of the n individuals on three beverages namely coffee, tea and milk. We analyse the three data sets separately.

The panels in Figure 1 display the frequency distribution of consumption of the three beverages. Taking a cue from the graph, we have fitted a $Poisson(\lambda)$ and a $ZIP(\lambda, \delta)$ to each of these data sets separately. The AIC values of the fit are reported in Table 7. The values in the table indicate that a *ZIP* model fits better on the data sets than a *Poisson* model. Thus, assuming a $ZIP(\lambda, \delta)$ model, the MLEs of the model parameters are computed. These are denoted by $(\hat{\lambda}^C, \hat{\delta}^C)$, $(\hat{\lambda}^T, \hat{\delta}^T)$ and $(\hat{\lambda}^M, \hat{\delta}^M)$ for the coffee, tea and milk data respectively.

However, survey studies indicate that individuals often under-report their beverage consumptions at workplace. Using the misclassification model given in Equation (2), we generate the surrogate counts $Y_1, Y_2, .., Y_n$ for different choices of $\pi \in [0.75, 0.99]$. The purpose of this study is to investigate the impact of ignoring undercount on the parameter estimates. To this end, we carry out a naive analysis using the surrogate counts for fitting a $ZIP(\lambda, \delta)$ model. The parameter estimates at $k^{th}(k = 1, 2, .., K)$ simulation are denoted by $(\hat{\lambda}^C_{N,k}, \hat{\delta}^C_{N,k})$, $(\hat{\lambda}^T_{N,k}, \hat{\delta}^T_{N,k})$ and $(\hat{\lambda}^M_{N,k}, \hat{\delta}^M_{N,k})$ for the data sets on coffee, tea and milk respectively. The estimates of λ and δ are finally obtained by averaging over the K simulations. The average of the standard errors are obtained from repeated calculation of the inverse of Fisher Information matrix.

The parameter estimates and the standard errors (multiplied by 10^4) of the estimators are reported in Table 8 for the true model and naive model for different choices of π. Here, by **True** and **Naive** we refer to the $ZIP(\lambda, \delta)$ model which is fitted using true counts $Y_i^*(i = 1, 2, .., n)$ and surrogate counts $Y_i(i = 1, 2, .., n)$ respectively.

The results in Table 8 reveal that, as π decreases, the naive estimate of λ gets attenuated and gives smaller standard error. On the contrary, δ is overestimated under the naive model with larger standard error. This feature becomes more prominent for relatively small choices of π. Although this pattern occurs for both coffee and milk consumption data, the tea consumption data on the other hand does not reflect much change under the naive model. One possible reason for this could be that for tea consumption data, the estimate of zero inflation parameter (δ) is quite high under the true model. For such a high value of δ,

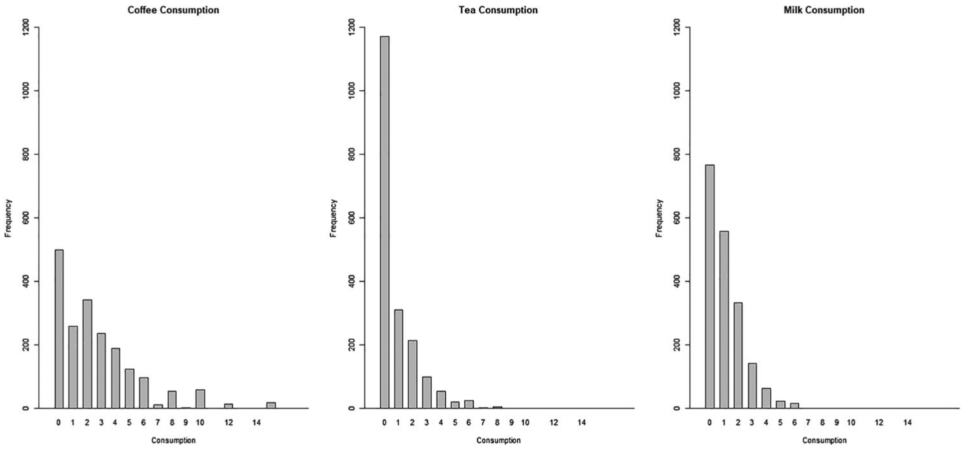

Figure 1. Frequency distribution of the Real Life Data.

Table 8. The parameter estimates and the standard error (in parenthesis) of the estimators corresponding to the true and naive models for the beverage consumption data.

Par	True	Naive ($\pi = 0.70$)	Naive ($\pi = 0.75$)	Naive ($\pi = 0.80$)	Naive ($\pi = 0.85$)	Naive ($\pi = 0.90$)	Naive ($\pi = 0.95$)
				Coffee Consumption			
δ	0.241	0.269	0.264	0.258	0.253	0.249	0.245
	(103)	(117)	(115)	(115)	(111)	(111)	(107)
λ	3.565	2.591	2.755	2.918	3.079	3.241	3.404
	(530)	(495)	(500)	(495)	(504)	(507)	(521)
				Tea Consumption			
δ	0.537	0.548	0.546	0.547	0.547	0.545	0.547
	(159)	(162)	(163)	(162)	(162)	(163)	(162)
λ	1.768	1.447	1.441	1.446	1.444	1.440	1.445
	(550)	(551)	(550)	(551)	(550)	(549)	(550)
				Milk Consumption			
δ	0.209	0.225	0.159	0.219	0.218	0.214	0.212
	(184)	(264)	(74)	(219)	(233)	(227)	(213)
λ	1.403	1.002	0.995	1.137	1.206	1.271	1.337
	(415)	(389)	(139)	(376)	(389)	(395)	(401)

the parameter estimates become unstable. This was earlier noted in the simulation study section also.

6. Concluding remarks

The paper focuses on the joint modelling of excess zeros and under-reporting in count data. The standard Poisson model is modified to accommodate both additional zeros and undercount by introducing two extra parameters; the zero inflation parameter (δ) and the probability of observance (π). In absence of prior knowledge on π, we have made use of validation data. The moment estimators and the likelihood estimators of the model parameters are compared through extensive simulation studies. The finite sample behaviour of the estimators are also looked into using bootstrap resampling technique.

Unlike the equi-dispersion Poisson model, the ZIP distribution incorporates overdispersion through the presence of excess zeros. In real life situations, overdispersion may also occur due to heterogeneity of the data. In such situations, the negative binomial distribution, or COM Poisson model or generalized Poisson distribution become a more plausible assumption. The current research focuses on undercount and some potential examples of under-reporting in survey data have been mentioned at the outset. However, survey data may be inaccurate due to both over-reporting and under-reporting. Extension of the current work in these directions are currently under investigation.

Acknowledgments

The authors thank the referees for careful reading of the manuscript and making constructive suggestions that resulted in a considerably improved present version.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Böhning D, Dietz E, Schlattmann P, et al. The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. J R Stat Soc Ser A (Stat Soc). 1999;162(2):195–209.

[2] Zhou XH, Tu W. Confidence intervals for the mean of diagnostic test charge data containing zeros. Biometrics. 2000;56(4):1118–1125.

[3] Agarwal D, Gelfand A, Citron-Pousty S. Zero-inflated models with application to spatial count data. Environ Ecol Stat. 2002;9(4):341–355.

[4] Hall D. Zero-inflated poisson and binomial regression with random effects: a case study. Biometrics. 2000;56(4):1030–1039.

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

[6] Desjardins CD. Modeling zero-inflated and overdispersed count data: an empirical study of school suspensions. J Exp Educ. 2016;84(3):449–472.

[7] Heilbron DC. Zero-altered and other regression models for count data with added zeros. Biom J. 1994;36(5):531–547.

[8] Sposto R, Preston DL, Shimizu Y, et al. The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in A-bomb survivors. Biometrics. 1992;48(2):605–617.

[9] Anderson C, Bratcher T, Kutran K. Bayesian-estimation of population-density and visibility. Tex J Sci. 1994;46(1):1–12.

[10] Fader PS, Hardie BGS. A note on modelling underreported poisson counts. J Appl Stat. 2000;27(8):953–964.

[11] Bratcher TL, Stamey JD. Estimation of poisson rates with misclassified counts. Biom J. 2002;44(8):946–956.

[12] Stamey JD, Young DM, Cecchini M. A double-sampling approach for maximum likelihood estimation for a Poisson rate parameter with visibility-biased data. Statistica. 2003;63(1):3–11.

[13] Stamey JD, Young DM, et al. Bayesian predictive probability functions for count data that are subject to misclassification. Biom J. 2004;46(5):572–578.

[14] Stamey JD, Young DM, Boese D. A Bayesian hierarchical model for Poisson rate and reporting-probability inference using double sampling. Aust N Z J Stat. 2006;48(2):201–212.

[15] Stamey JD, Young DM. Maximum likelihood estimation for a Poisson rate parameter with misclassified counts. Aust N Z J Stat. 2005;47(2):163–172.

[16] Stamey JD, Young DM, Seaman JW. A Bayesian approach to adjust for diagnostic misclassification between two mortality causes in Poisson regression. Stat Med. 2008;27(13):2440–2452.

[17] Wu W, Stamey J, Kahle D. A Bayesian approach to account for misclassification and overdispersion in count data. Int J Environ Res Public Health. 2015;12(9):10648–10661.

[18] Famoye F, Singh KP. Zero-inflated generalized poisson regression model with an application to domestic violence data. J Data Sci. 2006;4(1):117–130.

[19] Guliani H, Gamtessa S, Çule M. Factors affecting tobacco smoking in ethiopia: evidence from the demographic and health surveys. BMC Public Health. 2019;19(1):938–954.

[20] Handebo S, Birara S, Kassie A, et al. Smoking intensity and associated factors among male smokers in ethiopia: further analysis of 2016 ethiopian demographic and health survey. Biomed Res Int. 2020;2020:1–7.

[21] Johnson NL, Kemp AW, Kotz S. Univariate discrete distributions. 3rd . New York: John Wiley&Sons; 2005.

[22] Sengupta D, Banerjee T, Roy S. Estimation of poisson mean with under-reported counts: a double sampling approach. Aust N Z J Stat. 2020;62(4):508–535.

[23] Mullahy J. Specification and testing of some modified count data models. J Econom. 1986;33(3):341–365. Available from: https://EconPapers.repec.org/RePEc:eee:econom:v:33:y:1986:i:3:p:341-365

Appendices

Appendix 1. Proof of Theorem 2.1

$$\begin{aligned}P(Y=0) &= P(Y=0|Y^*=0)\cdot P(Y^*=0)+\sum_{y^*=1}^{\infty}P(Y=0|Y^*=y^*)\cdot P(Y^*=y^*)\\ &= \delta+(1-\delta)\,\mathrm{e}^{-\lambda}+\sum_{y^*=1}^{\infty}(1-\delta)\,\mathrm{e}^{-\lambda}\frac{\{\lambda(1-\pi)\}^{y^*}}{y^*!}\\ &= \delta+(1-\delta)\,\mathrm{e}^{-\lambda}\left[1+\,\mathrm{e}^{\lambda(1-\pi)}-1\right]=\delta+(1-\delta)\,\mathrm{e}^{-\lambda\pi}.\end{aligned}$$

For $Y>0$,

$$\begin{aligned}P(Y=y) &= \sum_{y^*=y}^{\infty}P(Y^*=y^*)\cdot P(Y=y|Y^*=y^*)=\sum_{y^*=y}^{\infty}(1-\delta)\frac{\mathrm{e}^{-\lambda}\lambda^{y^*}}{y^*!}\frac{y^*!}{y!(y^*-y)!}\pi^{y}(1-\pi)^{y^*-y}\\ &= \frac{(\lambda\pi)^y}{y!}(1-\delta)\,\mathrm{e}^{-\lambda}\sum_{y^*=y}^{\infty}\frac{\{\lambda(1-\pi)\}^{y^*-y}}{(y^*-y)!}=(1-\delta)\,\mathrm{e}^{-\lambda\pi}\frac{\{\lambda\pi\}^y}{y!}.\end{aligned}$$

Thus $Y\sim ZIP(\lambda\pi,\delta)$.

Appendix 2. Fisher information for benchmark estimator

Let $l=l(\lambda,\delta)$. The second order derivatives are given by,

$$\begin{aligned}\frac{\partial^2 l}{\partial\lambda^2} &= -\frac{\delta\pi_0^2\psi(\lambda,\delta,\pi_0)}{\{\delta+\psi(\lambda,\delta,\pi_0)\}^2}\sum_{i=1}^{n}I_i-\frac{1}{\lambda^2}\sum_{i=1}^{n}(1-I_i)y_i.\\ \frac{\partial^2 l}{\partial\lambda\partial\delta} &= \frac{\pi_0\,\mathrm{e}^{-\lambda\pi_0}}{\{\delta+\psi(\lambda,\delta,\pi_0)\}^2}\sum_{i=1}^{n}I_i.\\ \frac{\partial^2 l}{\partial\delta^2} &= -\frac{(1-\,\mathrm{e}^{-\lambda\pi_0})^2}{\{\delta+\psi(\lambda,\delta,\pi_0)\}^2}\sum_{i=1}^{n}I_i-\frac{1}{(1-\delta)^2}\sum_{i=1}^{n}(1-I_i).\end{aligned}\tag{B1}$$

Appendix 3. Fisher information matrix for likelihood estimator

Writing, $l = l(\lambda, \delta, \pi)$, the second order derivatives are given as follows:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \lambda^2} &= -\frac{\delta \psi(\lambda, \delta, 1)}{\{\delta + \psi(\lambda, \delta, 1)\}^2} \sum_{i \in S_v} I_{v,i} - \frac{\pi^2 \delta \psi(\lambda, \delta, \pi)}{\{\delta + \psi(\lambda, \delta, \pi)\}^2} \sum_{i \in S_{nv}} I_{nv,i} \\
&\quad - \frac{1}{\lambda^2} \left\{ \sum_{i \in S_v} (1 - I_{v,i}) y_i^* + \sum_{i \in S_{nv}} (1 - I_{nv,i}) y_i \right\} \\
\frac{\partial^2 l}{\partial \delta \partial \lambda} &= \frac{\mathrm{e}^{-\lambda}}{\{\delta + \psi(\lambda, \delta, 1)\}^2} \sum_{i \in S_v} I_{v,i} + \frac{\pi \, \mathrm{e}^{-\lambda \pi}}{\{\delta + \psi(\lambda, \delta, \pi)\}^2} \sum_{i \in S_{nv}} I_{nv,i} \\
\frac{\partial^2 l}{\partial \pi \partial \lambda} &= -\frac{\{\delta + \psi(\lambda, \delta, \pi) - \lambda \pi\} \, \psi(\lambda, \delta, \pi)}{\{\delta + \psi(\lambda, \delta, \pi)\}^2} \sum_{i \in S_{nv}} I_{nv,i} - \sum_{i \in S_{nv}} (1 - I_{nv,i}) \\
\frac{\partial^2 l}{\partial \delta^2} &= -\frac{(1 - \mathrm{e}^{-\lambda})^2}{\{\delta + \psi(\lambda, \delta, 1)\}^2} \sum_{i \in S_v} I_{v,i} - \frac{(1 - \mathrm{e}^{-\lambda \pi})^2}{\{\delta + \psi(\lambda, \delta, \pi)\}^2} \sum_{i \in S_{nv}} I_{nv,i} \qquad \text{(C1)} \\
&\quad - \frac{1}{(1 - \delta)^2} \cdot \left\{ \sum_{i \in S_v} (1 - I_{v,i}) + \sum_{i \in S_{nv}} (1 - I_{nv,i}) \right\} \\
\frac{\partial^2 l}{\partial \pi \partial \delta} &= \frac{\lambda \, \mathrm{e}^{-\lambda \pi}}{\{\delta + \psi(\lambda, \delta, \pi)\}^2} \sum_{i \in S_{nv}} I_{nv,i} \\
\frac{\partial^2 l}{\partial \pi^2} &= -\frac{\delta \lambda^2 \psi(\lambda, \delta, \pi)}{\{\delta + \psi(\lambda, \delta, \pi)\}^2} \sum_{i \in S_{nv}} I_{nv,i} - \frac{1}{(1 - \pi)^2} \sum_{i \in S_v} y_i^* - \frac{1 - 2\pi}{\pi^2 (1 - \pi)^2} \sum_{i \in S_v} y_i \\
&\quad - \frac{1}{\pi^2} \sum_{i \in S_{nv}} (1 - I_{nv,i}) y_i
\end{aligned}$$

Appendix 4. Variance of moment estimators

Define $T = (T_1, T_2, T_3, T_4)'$ and

$$\begin{aligned}
\hat{\theta} &= \begin{pmatrix} \hat{\pi}_{MM} & \hat{\lambda}_{MM} & \hat{\delta}_{MM} \end{pmatrix}' \\
&= \begin{pmatrix} g_1(T) & g_2(T) & g_3(T) \end{pmatrix}' \\
&= G(T) \quad \text{(say)}. \qquad \text{(D1)}
\end{aligned}$$

We have $V(Y_i^*) = (1 - \delta)\lambda(1 + \lambda\delta)$, $V(Y_i) = (1 - \delta)\lambda\pi(1 + \lambda\delta\pi)$, $V(Y_i^2) = (1 - \delta)\lambda\pi\{4\lambda^2\pi^2 + 6\lambda\pi + 1 + \lambda\pi\delta(1 + \lambda\pi)\}^2$, $Cov(Y_i, Y_i^*) = (1 - \delta)\lambda\pi(1 + \lambda\delta\pi)$.

Now based on a sample of size n and validation sub-sample of size r, the variance-covariance matrix of T is given by,

$$\Sigma^{4x4} = (1 - \delta)\lambda\pi \left((\sigma_{ij}) \right), \qquad \text{(D2)}$$

where,

$$\sigma_{11} = \frac{n(1+\lambda\delta)}{r\pi}, \quad \sigma_{12} = \sigma_{21} = \frac{n(1+\lambda\delta)}{r}, \quad \sigma_{13} = \sigma_{31} = (1+\lambda\delta),$$
$$\sigma_{14} = \sigma_{41} = \lambda\{2\pi + \delta(1+\lambda\pi)\}, \quad \sigma_{22} = \frac{n(1+\lambda\pi\delta)}{r}, \quad \sigma_{23} = \sigma_{32} = (1+\lambda\pi\delta),$$
$$\sigma_{24} = \sigma_{42} = 2\lambda\pi + 1 + \delta\lambda\pi(1+\lambda\pi), \quad \sigma_{33} = 1 + \lambda\pi\delta,$$
$$\sigma_{34} = \sigma_{43} = 2\lambda\pi + 1 + \delta\lambda\pi(1+\lambda\pi), \quad \sigma_{44} = 4\lambda^2\pi^2 + 6\lambda\pi + 1 + \delta\lambda\pi(1+\lambda\pi). \qquad \text{(D3)}$$

Now, the dispersion matrix of of the moment estimators $(\hat{\theta})$ is given by, $G\Sigma G'$, where $G = \frac{\partial\hat{\theta}}{\partial T} = ((g_{ij}))$ and $g_{ij} = \frac{\partial g_i}{\partial T_j}$, computed at the true value of $\theta = (\pi \quad \lambda \quad \delta)'$. The elements of G are given below.

$$g_{11} = \frac{-\pi}{(1-\delta)\lambda}, \quad g_{12} = \frac{1}{(1-\delta)\lambda}, \quad g_{13} = g_{14} = 0, \quad g_{21} = 1,$$
$$g_{22} = \frac{1}{(1-\delta)\pi}, \quad g_{23} = \frac{-(1+\lambda\pi)}{(1-\delta)\lambda\pi^2}, \quad g_{24} = \frac{1}{(1-\delta)\lambda\pi},$$
$$g_{31} = g_{32} = 0 \quad g_{33} = 2\lambda\pi + 1, \quad g_{34} = 1. \qquad \text{(D4)}$$