JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2023.2207020

An R package AZIAD for analysing zero-inflated and zero-altered data

Niloufar Dousti Mousavi[a], Hani Aldirawi[b] and Jie Yang[a]

[a]Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL, USA; [b]Department of Mathematics, California State University, San Bernardino, CA, USA

ABSTRACT

We introduce a newly developed R package AZIAD for analysing zero-inflated or zero-altered data. Compared with existing R packages, AZIAD covers a much larger class of zero-inflated and hurdle models, including both discrete and continuous cases. It provides more accurate parameter estimates, along with the corresponding Fisher information matrix and confidence intervals. It achieves significantly larger power for model identification and selection. To facilitate the potential users, in this paper we provide detailed formulae and theoretical justifications for AZIAD, as well as new theoretical results on zero-inflated and zero-altered models. We use simulation studies to show the advantages of AZIAD functions over existing R packages and provide real data examples and executable R code to illustrate how to use our package for sparse data analysis and model selection.

ARTICLE HISTORY
Received 17 August 2022
Accepted 18 April 2023

KEYWORDS
Kolmogorov-Smirnov (KS) test; zero-inflated model; hurdle model; model selection; Fisher information matrix

MATHEMATICS SUBJECT CLASSIFICATIONS
62-04; 62-08

1. Introduction

Sparse or zero-inflated data arise frequently from a rich variety of scientific disciplines including microbiome [1,2], gene expression [3], health care [4], insurance claim [5], security [6], and more. Modeling sparse data is very challenging due to the high proportion of zero values and the skewness of the distribution. Zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), Poisson hurdle (PH), and negative binomial hurdle (NBH) models have been widely used to model sparse data [2,7,8].

On one hand, more and more zero-inflated models have been proposed under different circumstances. On the other hand, it becomes more and more difficult for researchers and practitioners to choose the most appropriate model for their sparse data. Some partial comparisons among existing models have been done for certain sparse data. For examples, [9] recommended ZINB and NBH models for microbiome data after comparing

CONTACT Jie Yang jyang06@uic.edu Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 S. Morgan Street, Chicago, IL 60607-7045, USA

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2023.2207020

the performance of Poisson, ZIP, PH, NB (negative binomial), ZINB, and NBH models; while [7] indicated that two new models, zero-inflated beta negative binomial (ZIBNB) and beta negative binomial hurdle (BNBH) models, are more appropriate for microbiome data examples.

For analysing zero-inflated data, there are some available `R` packages from the Comprehensive R Archive Network (CRAN, https://cran.r-project.org/) including `bzinb` [10], `hurdlr` [11], `iZID` [12], `gamlss` [13], `pscl` [14], `mazeinda` [15], `mhurdle` [16], `rbtt` [17], `ZIBBSeqDiscovery` [18], `ZIBseq` [19], `zic` [20], `ZIM` [21], `ziphsmm` [22], etc. For example, `iZID` covers 12 discrete distributions including Poisson, NB, BB (beta binomial), BNB (beta negative binomial) and their zero-inflated and hurdle versions. It implemented the bootstrapped Monte Carlo p-value estimates proposed by Aldirawi et al. [7] for identifying a discrete distribution. For packages other than `iZID`, please see [23] for a good review.

The existing literature and packages on analysing zero-inflated data, including the very recent work [7,23], are still limited for three reasons. First, only a small number of zero-inflated models were under consideration at the same time. Discrete and continuous baseline distributions were typically considered separately. Secondly, the accuracy of the parameter estimates and the power of the tests still have room for improvement. Thirdly, the confidence intervals of the parameter values were seldom provided along with their estimates, which is critical for model diagnostics, such as, testing whether the inflation or deflation of zeros exists in the data.

In this paper, we introduce our newly developed `R` package named `AZIAD` for Analyzing Zero-Inflated and Zero-Altered Data, available from the Comprehensive R Archive Network (CRAN, https://cran.r-project.org/package = AZIAD). Compared with the existing `R` packages for similar purposes, our `AZIAD` achieves the following significant improvements: (1) We not only cover discrete baseline distributions including Poisson, geometric, NB, BB, and BNB, but also cover commonly used continuous baseline distributions including normal (or Gaussian), log-normal, half-normal, and exponential distributions along with their zero-inflated and zero-altered (also known as hurdle) models; (2) By more precise specifications on solution forms under different situations and lower/upper bounds needed for numerical optimizations, our `R` functions provide more accurate maximum likelihood estimates (MLE) even under extreme circumstances, which further gains more power when identifying the most appropriate zero-inflated or hurdle models for a given data set; (3) Following [24], we provide not only MLEs for parameters, but also the Fisher information matrix and the corresponding confidence intervals for estimated parameters, which will facilitate the potential users from biological sciences, insurance, health studies, security, ecology, etc, to identify the most appropriate probabilistic model for their dataset and make robust statistical inference based on it.

The rest of this paper is organized as follows. In Section 2, we not only review relevant theoretical results from [24], but also provide new formulae for Fisher information matrices of the zero-inflated (ZI) and zero-altered (ZA or hurdle) models with geometric, BB, BNB, normal, log-normal, half-normal and exponential distributions, as well as ZINB. In Section 3, we summarize and use numerical examples to illustrate the improvements by using our package over existing `R` packages, as well as identifying the most appropriate models for real data. We interpret and discuss indistinguishable pairs of distributions in Section 4.

2. MLE and Fisher information for ZI and ZA models

2.1. MLE and Fisher information for zero-altered or hurdle models

Zero-altered models or hurdle models have been widely used for modelling data with an excess or deficit of zeros (see, for example, [2], for a good review). A general hurdle model consists of a baseline distribution and a component generating the zeros. The baseline distribution could be fairly general with the distribution function $f_{\boldsymbol{\theta}}(y)$ and its zero-truncated version $f_{\rm tr}(y \mid \boldsymbol{\theta}) = [1 - p_0(\boldsymbol{\theta})]^{-1} f_{\boldsymbol{\theta}}(y)$, $y \neq 0$, where $\boldsymbol{\theta}$ is the model parameter(s), and $p_0(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(Y = 0)$ is the probability that $Y = 0$ under the baseline distribution. Following [24], the distribution function of the corresponding hurdle model can written as:

$$f_{\rm ZA}(y \mid \phi, \boldsymbol{\theta}) = \phi \mathbf{1}_{\{y=0\}} + (1 - \phi) f_{\rm tr}(y \mid \boldsymbol{\theta}) \mathbf{1}_{\{y \neq 0\}} \tag{1}$$

where $\phi \in [0, 1]$ is the weight parameter of zeros. Actually, $\phi = P(Y = 0)$ if $Y \sim f_{\rm ZA}$.

If the baseline distribution is discrete with a probability mass function (pmf) $f_{\boldsymbol{\theta}}(y)$, such as Poisson, negative binomial (NB), geometric (Ge), beta binomial (BB) and beta negative binomial (BNB) distributions, then both $f_{\rm tr}(y \mid \boldsymbol{\theta})$ and $f_{\rm ZA}(y \mid \phi, \boldsymbol{\theta})$ are pmfs as well. The corresponding zero-altered or hurdle models may be called as zero-altered Poisson (ZAP) or Poisson hurdle (PH), zero-altered negative binomial (ZANB) or negative binomial hurdle (NBH), zero-altered geometric (ZAGe) or geometric hurdle (GeH), zero-altered beta binomial (ZABB) or beta binomial hurdle (BBH), zero-altered beta negative binomial (ZABNB) or beta negative binomial hurdle (BNBH) models, respectively.

If the baseline distribution is continuous with a probability density function (pdf) $f_{\boldsymbol{\theta}}(y)$, such as Gaussian (or normal), log-normal, half-normal and exponential distributions, then $p_0(\boldsymbol{\theta}) = 0$ and $f_{\rm tr}(y \mid \boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(y)$. In this case, $f_{\rm ZA}(y \mid \phi, \boldsymbol{\theta})$ in (1) represents a mixture distribution consisting of a discrete component at 0 with probability ϕ and a continuous component with density function $(1 - \phi) f_{\boldsymbol{\theta}}(y)$. We will revisit these continuous cases in Section 2.3. In this section, we focus on discrete cases.

The parameters of the hurdle model (1) include both ϕ and $\boldsymbol{\theta}$. In this paper, we adopt the maximum likelihood estimate (MLE) for estimating the parameters (see, for example, Section 1.3.1 in [25], for justifications on adopting MLE). Let $Y_1, \ldots, Y_n$ be a random sample from model (1). Then the likelihood function of $(\phi, \boldsymbol{\theta})$ is

$$L(\phi, \boldsymbol{\theta}) = \phi^{n-m} (1 - \phi)^m \cdot \prod_{i: Y_i \neq 0} f_{\rm tr}(Y_i \mid \boldsymbol{\theta}) \tag{2}$$

where $m = \#\{i : Y_i \neq 0\}$ is the number of nonzero observations. According to Theorem 1 in [24], the maximum likelihood estimate (MLE) of $(\phi, \boldsymbol{\theta})$ that maximizes (2) is

$$\hat{\phi} = 1 - \frac{m}{n}, \quad \hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i: Y_i \neq 0} f_{\rm tr}(Y_i \mid \boldsymbol{\theta}) \tag{3}$$

That is, $\hat{\boldsymbol{\theta}}$ is simply the MLE for the truncated model with distribution function $f_{\rm tr}(y \mid \boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(y)/[1 - p_0(\boldsymbol{\theta})], y \neq 0$.

According to Theorem 3 in [24], under some regularity conditions, the Fisher information matrix of the zero-altered distribution is

$$\mathbf{F}_{\mathrm{ZA}} = \begin{bmatrix} \phi^{-1}(1-\phi)^{-1} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{F}_{\mathrm{ZA}\boldsymbol{\theta}} \end{bmatrix} \tag{4}$$

where

$$\mathbf{F}_{\mathrm{ZA}\boldsymbol{\theta}} = -\frac{1-\phi}{1-p_0(\boldsymbol{\theta})}\left(E\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(Y')}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right] + \frac{p_0(\boldsymbol{\theta})}{1-p_0(\boldsymbol{\theta})}\cdot\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\cdot\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right)$$

and Y' follows the baseline distribution $f_{\boldsymbol{\theta}}(y)$. Note that the sample size n does not show up in (4) since the Fisher information here is of the distribution, not of the sample.

Major advantages for adopting MLE (3) and calculating the Fisher information matrix (4) include: (i) $\hat{\phi}$ and $\hat{\boldsymbol{\theta}}$ are consistent estimators of ϕ and $\boldsymbol{\theta}$, respectively (Theorem 2 in [24]); (ii) $\sqrt{n}(\hat{\phi}-\phi) \xrightarrow{\mathcal{L}} N(0, \phi(1-\phi))$ and $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{F}^{-1}_{\mathrm{ZA}\boldsymbol{\theta}})$, which provides the formulae for building up approximate confidence intervals and relevant hypothesis tests for ϕ and $\boldsymbol{\theta}$ (see, for example, Sections 1.3.3 and 1.3.4 in [25]). For example, an approximate $(1-\alpha)100\%$ confidence interval for ϕ is

$$\phi \in \left(\hat{\phi} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\hat{\phi}(1-\hat{\phi})},\ \hat{\phi} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\hat{\phi}(1-\hat{\phi})}\right) \tag{5}$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ is the upper $\frac{\alpha}{2}$th quantile of the standard normal distribution, and $\alpha \in (0,1)$ is the desired significance level, such as 0.05. If $p_0(\boldsymbol{\theta})$ does not belong to the confidence interval (5), then there is a significant evidence for zero-inflation or deflation.

Therefore, calculating MLE and Fisher information accurately and efficiently is the basis of sound and thorough statistical inference.

Explicit formulae of the gradients ($\partial \log f_{\boldsymbol{\theta}}(y)/\partial\boldsymbol{\theta}$ and $\partial \log p_0(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, need for finding MLEs) and the Fisher information matrices of zero-altered Poisson (ZAP) or Poisson hurdle (PH) model, zero-altered negative binomial (ZANB) or negative binomial hurdle (NBH) model, have been provided in Examples 2 and 3 of [24], respectively. In this section, we will provide explicit formulae for zero-altered geometric, beta binomial, and beta negative binomial models.

Example 2.1 (Zero-altered geometric (ZAGe) or geometric hurdle (GeH) model): has been used, for example, in [26] for modelling count data in econometrics. The pmf of the baseline distribution can be written as $f_p(y) = p(1-p)^y$ with $y \in \{0,1,2,\ldots\}$ and parameter $p \in (0,1)$. In this case, $p_0(p) = p$, $\frac{\partial \log f_p(y)}{\partial p} = \frac{1}{p} - \frac{y}{1-p}$, and $\frac{\partial \log p_0(p)}{\partial p} = \frac{1}{p}$.

According to Theorem 3 in [24], the Fisher information matrix of the ZAGe or GeH distribution can be written as

$$\mathbf{F}_{\mathrm{ZAGe}} = \begin{bmatrix} \dfrac{1}{\phi(1-\phi)} & 0 \\ 0 & \dfrac{1-\phi}{p^2(1-p)} \end{bmatrix}$$

Example 2.2 (Zero-altered beta-binomial (ZABB) or beta-binomial hurdle (BBH) model): has been used by Aldirawi et al. [7] for modelling microbiome data. The pmf of

the baseline distribution with parameters $\boldsymbol{\theta} = (n, \alpha, \beta) \in \mathbb{N} \times (0, \infty) \times (0, \infty)$ is given by $f_{\boldsymbol{\theta}}(y) = \binom{n}{y} \frac{\text{Beta}(y+\alpha, n-y+\beta)}{\text{Beta}(\alpha,\beta)}$, $y \in \{0, 1, 2, \ldots, n\}$. In this case, $p_0(\boldsymbol{\theta}) = \frac{\Gamma(n+\beta)\Gamma(\alpha+\beta)}{\Gamma(n+\alpha+\beta)\Gamma(\beta)}$. Then

$$\begin{aligned}
\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial n} &= \Psi(n+1) - \Psi(n-y+1) + \Psi(n-y+\beta) - \Psi(n+\alpha+\beta) \\
\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial \alpha} &= \Psi(y+\alpha) - \Psi(n+\alpha+\beta) + \Psi(\alpha+\beta) - \Psi(\alpha) \\
\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial \beta} &= \Psi(n-y+\beta) - \Psi(n+\alpha+\beta) + \Psi(\alpha+\beta) - \Psi(\beta) \\
\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial n} &= \Psi(n+\beta) - \Psi(n+\alpha+\beta) \\
\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial \alpha} &= \Psi(\alpha+\beta) - \Psi(n+\alpha+\beta) \\
\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial \beta} &= \Psi(n+\beta) + \Psi(\alpha+\beta) - \Psi(n+\alpha+\beta) - \Psi(\beta)
\end{aligned}$$

where $\Psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is known as the *digamma* function. Note that the range of parameter n can be extended to positive real numbers. According to Theorem 3 in [24], the Fisher information matrix of the ZABB or BBH distribution is

$$\mathbf{F}_{\text{ZABB}} = \begin{bmatrix} \dfrac{1}{\phi(1-\phi)} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{F}_{\text{BB}\boldsymbol{\theta}} \end{bmatrix}$$

where

$$\begin{aligned}
\mathbf{F}_{\text{BB}\boldsymbol{\theta}} = -\frac{(1-\phi)\Gamma(n+\alpha+\beta)\Gamma(\beta)}{\Gamma(n+\alpha+\beta)\Gamma(\beta) - \Gamma(n+\beta)\Gamma(\alpha+\beta)} \left(\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & A_{22} & A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} \right. \\
\left. + \frac{\Gamma(n+\beta)\Gamma(\alpha+\beta)}{\Gamma(n+\alpha+\beta)\Gamma(\beta) - \Gamma(n+\beta)\Gamma(\alpha+\beta)} \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{12} & B_{22} & B_{23} \\ B_{13} & B_{23} & B_{33} \end{bmatrix} \right)
\end{aligned}$$

with

$$\begin{aligned}
A_{11} &= \Psi_1(n+1) - \Psi_1(n+\alpha+\beta) + E\Psi_1(n-Y'+\beta) - E\Psi_1(n-Y'+1) \\
A_{12} &= -\Psi_1(n+\alpha+\beta) \\
A_{13} &= E\Psi_1(n-Y'+\beta) - \Psi_1(n+\alpha+\beta) \\
A_{22} &= \Psi_1(\alpha+\beta) - \Psi_1(n+\alpha+\beta) - \Psi_1(\alpha) + E\Psi_1(Y'+\alpha) \\
A_{23} &= \Psi_1(\alpha+\beta) - \Psi_1(n+\alpha+\beta) \\
A_{33} &= \Psi_1(\alpha+\beta) - \Psi_1(\beta) - \Psi_1(n+\alpha+\beta) + E\Psi_1(n-Y'+\beta) \\
B_{11} &= [\Psi(n+\beta) - \Psi(n+\alpha+\beta)]^2 \\
B_{12} &= [\Psi(n+\beta) - \Psi(n+\alpha+\beta)] \cdot [\Psi(\alpha+\beta) - \Psi(n+\alpha+\beta)] \\
B_{13} &= [\Psi(n+\beta) - \Psi(n+\alpha+\beta)] \cdot
\end{aligned}$$

$$
\begin{aligned}
&[\Psi(n+\beta)+\Psi(\alpha+\beta)-\Psi(n+\alpha+\beta)-\Psi(\beta)]\\
B_{22} &= [\Psi(\alpha+\beta)-\Psi(n+\alpha+\beta)]^2\\
B_{23} &= [\Psi(\alpha+\beta)-\Psi(n+\alpha+\beta)]\cdot\\
&\quad [\Psi(n+\beta)+\Psi(\alpha+\beta)-\Psi(n+\alpha+\beta)-\Psi(\beta)]\\
B_{33} &= [\Psi(n+\beta)+\Psi(\alpha+\beta)-\Psi(n+\alpha+\beta)-\Psi(\beta)]^2
\end{aligned}
$$

where $\Psi_1(\cdot)=\Psi'(\cdot)$ is known as the *trigamma* function, and Y' follows the baseline distribution $f_{\boldsymbol{\theta}}(y)$.

Due to the lack of simple forms, we may use Monte Carlo estimates for calculating $E\Psi_1(\cdot)$ in Example 2.2 and others approximately. For example, $E\Psi_1(n-Y'+\beta)$ can be approximated by $N^{-1}\sum_{i=1}^{N}\Psi_1(n-Y'_i+\beta)$ with simulated $Y'_1,\ldots,Y'_N$ from $f_{\boldsymbol{\theta}}(y)$.

Another computational issue involved in Example 2.2 and others is that $\log[\Gamma(n+\alpha+\beta)\Gamma(\beta)-\Gamma(n+\beta)\Gamma(\alpha+\beta)]$, which is relevant to $\log p_0(\boldsymbol{\theta})$, may be undefined numerically for large n since both $\Gamma(n+\alpha+\beta)$ and $\Gamma(n+\beta)$ are numerical infinity. To overcome this kind of issues, we use the fact $\log(A-B)=\log(1-\exp(\log B-\log A))+\log A$ if $A\geq B$. It can be verified that $\Gamma(n+\alpha+\beta)\Gamma(\beta)>\Gamma(n+\beta)\Gamma(\alpha+\beta)$.

Example 2.3 (Zero-altered beta negative binomial (ZABNB) or beta negative binomial hurdle (BNBH) model): has been recommended by Aldirawi et al. [7] and Aldirawi and Yang [24] for modelling microbiome data. The pmf of the baseline distribution with parameters $\boldsymbol{\theta}=(r,\alpha,\beta)\in\mathbb{N}\times(0,\infty)\times(0,\infty)$ is given by $f_{\boldsymbol{\theta}}(y)=\binom{r+y-1}{y}\frac{\mathrm{Beta}(r+\alpha,y+\beta)}{\mathrm{Beta}(\alpha,\beta)}$, $y\in\{0,1,2,\ldots\}$. Then $p_0(\boldsymbol{\theta})=\frac{\Gamma(r+\alpha)\Gamma(\alpha+\beta)}{\Gamma(r+\alpha+\beta)\Gamma(\alpha)}$. Note that r can also be extended to positive real numbers.

$$
\begin{aligned}
\frac{\partial\log f_{\boldsymbol{\theta}}(y)}{\partial r} &= \Psi(r+y)-\Psi(r)+\Psi(r+\alpha)-\Psi(r+y+\alpha+\beta)\\
\frac{\partial\log f_{\boldsymbol{\theta}}(y)}{\partial \alpha} &= \Psi(r+\alpha)-\Psi(r+y+\alpha+\beta)+\Psi(\alpha+\beta)-\Psi(\alpha)\\
\frac{\partial\log f_{\boldsymbol{\theta}}(y)}{\partial \beta} &= \Psi(y+\beta)-\Psi(r+y+\alpha+\beta)+\Psi(\alpha+\beta)-\Psi(\beta)\\
\frac{\partial\log p_0(\boldsymbol{\theta})}{\partial r} &= \Psi(r+\alpha)-\Psi(r+\alpha+\beta)\\
\frac{\partial\log p_0(\boldsymbol{\theta})}{\partial \alpha} &= \Psi(r+\alpha)+\Psi(\alpha+\beta)-\Psi(r+\alpha+\beta)-\Psi(\alpha)\\
\frac{\partial\log p_0(\boldsymbol{\theta})}{\partial \beta} &= \Psi(\alpha+\beta)-\Psi(r+\alpha+\beta)
\end{aligned}
$$

According to Theorem 3 in [24], the Fisher information matrix of the ZABNB or BNBH distribution is

$$
\mathbf{F}_{\mathrm{ZABNB}}=\begin{bmatrix}\frac{1}{\phi(1-\phi)} & \mathbf{0}^T\\ \mathbf{0} & \mathbf{F}_{\mathrm{BNB}\boldsymbol{\theta}}\end{bmatrix}
$$

where

$$\mathrm{F}_{\mathrm{BNB}\boldsymbol{\theta}} = -\frac{(1-\phi)\Gamma(r+\alpha+\beta)\Gamma(\alpha)}{\Gamma(r+\alpha+\beta)\Gamma(\alpha)-\Gamma(r+\alpha)\Gamma(\alpha+\beta)}\left(\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & A_{22} & A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix}\right.$$
$$\left. + \frac{\Gamma(r+\alpha)\Gamma(\alpha+\beta)}{\Gamma(r+\alpha+\beta)\Gamma(\alpha)-\Gamma(r+\alpha)\Gamma(\alpha+\beta)}\begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{12} & B_{22} & B_{23} \\ B_{13} & B_{23} & B_{33} \end{bmatrix}\right)$$

with

$$A_{11} = E\Psi_1(r+Y') - \Psi_1(r) + \Psi_1(r+\alpha) - E\Psi_1(r+Y'+\alpha+\beta)$$
$$A_{12} = \Psi_1(r+\alpha) - E\Psi_1(r+Y'+\alpha+\beta)$$
$$A_{13} = -E\Psi_1(r+Y'+\alpha+\beta)$$
$$A_{22} = \Psi_1(r+\alpha) - E\Psi_1(r+Y'+\alpha+\beta) + \Psi_1(\alpha+\beta) - \Psi_1(\alpha)$$
$$A_{23} = -E\Psi_1(r+Y'+\alpha+\beta) + \Psi_1(\alpha+\beta)$$
$$A_{33} = E\Psi_1(Y'+\beta) - E\Psi_1(r+Y'+\alpha+\beta) + \Psi_1(\alpha+\beta) - \Psi_1(\beta)$$
$$B_{11} = [\Psi(r+\alpha) - \Psi(r+\alpha+\beta)]^2$$
$$B_{12} = [\Psi(r+\alpha) - \Psi(r+\alpha+\beta)]\cdot[\Psi(r+\alpha) + \Psi(\alpha+\beta) - \Psi(r+\alpha+\beta) - \Psi(\alpha)]$$
$$B_{13} = [\Psi(r+\alpha) - \Psi(r+\alpha+\beta)]\cdot[\Psi(\alpha+\beta) - \Psi(r+\alpha+\beta)]$$
$$B_{22} = [\Psi(r+\alpha) + \Psi(\alpha+\beta) - \Psi(r+\alpha+\beta) - \Psi(\alpha)]^2$$
$$B_{23} = [\Psi(r+\alpha) + \Psi(\alpha+\beta) - \Psi(r+\alpha+\beta) - \Psi(\alpha)]\cdot[\Psi(\alpha+\beta) - \Psi(r+\alpha+\beta)]$$
$$B_{33} = [\Psi(\alpha+\beta) - \Psi(r+\alpha+\beta)]^2$$

As for Ψ_1 and relevant calculations, please see the arguments right after Example 2.2.

2.2. MLE and Fisher information for zero-inflated models

When data is sparse, a zero-inflated (ZI) model is more commonly used in practice, which assumes an excess of zeros (see, for example, [2] for a good review). Similar as the zero-altered (ZA) models in Section 2.1, there is a baseline distribution with distribution function $f_{\boldsymbol{\theta}}(y)$ and parameter(s) $\boldsymbol{\theta}$. We also denote $p_0(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(Y=0)$, if $Y \sim f_{\boldsymbol{\theta}}$. Different from ZA models, a zero-weighting parameter $\phi \in [0,1]$ adds additional probability of zeros to the ZI model. Following [24], we write the distribution function of the corresponding ZI model as

$$f_{\mathrm{ZI}}(y \mid \phi, \boldsymbol{\theta}) = [\phi + (1-\phi)p_0(\boldsymbol{\theta})]\mathbf{1}_{\{y=0\}} + (1-\phi)f_{\boldsymbol{\theta}}(y)\mathbf{1}_{\{y\neq 0\}} \tag{6}$$

Note that if $Y \sim f_{\mathrm{ZI}}$, then $P(Y=0) = \phi + (1-\phi)p_0(\boldsymbol{\theta})$, which is larger than ϕ in general.

When the baseline distribution is either continuous with a pdf $f_{\boldsymbol{\theta}}(y)$ or discrete but with $p_0(\boldsymbol{\theta}) = 0$, the corresponding zero-inflated model (6) is essentially the same as the corresponding zero-altered model (1). Examples include Gaussian or normal, log-normal, half-normal, and exponential distributions. We will revisit them in Section 2.3.

When the baseline distribution is discrete, such as Poisson (P), negative binomial (NB), geometric (Ge), beta binomial (BB) and beta negative binomial (BNB), the corresponding zero-inflated models can be written as ZIP, ZINB, ZIGe, ZIBB, and ZIBNB, respectively.

Given a random sample $Y_1, \ldots, Y_n$ from the zero-inflated model $f_{\mathrm{ZI}}(y|\phi, \boldsymbol{\theta})$, the likelihood function of $(\phi, \boldsymbol{\theta})$ can be written as

$$\begin{aligned} L(\phi, \boldsymbol{\theta}) &= \big[\phi + p_0(\boldsymbol{\theta})(1-\phi)\big]^{n-m} \cdot (1-\phi)^m \prod_{i:Y_i \neq 0} f_{\boldsymbol{\theta}}(Y_i) \\ &= \big[\phi + p_0(\boldsymbol{\theta})(1-\phi)\big]^{n-m} \cdot (1-\phi)^m \big(1-p_0(\boldsymbol{\theta})\big)^m \cdot \prod_{i:Y_i \neq 0} f_{\mathrm{tr}}(Y_i \mid \boldsymbol{\theta}) \end{aligned} \tag{7}$$

where $m = \#\{i : Y_i \neq 0\}$, and $f_{\mathrm{tr}}(y; \boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(y)/[1 - p_0(\boldsymbol{\theta})], y \neq 0$.

According to Theorem 4 in [24], the maximum likelihood estimate $(\hat{\phi}, \hat{\boldsymbol{\theta}})$ maximizing (7) can be obtained as follows:

(0) Determine $\boldsymbol{\theta}_* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L_{\mathrm{tr}}(\boldsymbol{\theta})$, where $L_{\mathrm{tr}}(\boldsymbol{\theta}) = \prod_{i:Y_i \neq 0} f_{\mathrm{tr}}(Y_i; \boldsymbol{\theta})$.

(1) If $m/n \leq 1 - p_0(\boldsymbol{\theta}_*)$, then $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_*$ and $\hat{\phi} = 1 - [1 - p_0(\boldsymbol{\theta}_*)]^{-1} \cdot m/n$.

(2) Otherwise, $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\psi(\boldsymbol{\theta}), \boldsymbol{\theta})$ and $\hat{\phi} = 1 - \psi(\hat{\boldsymbol{\theta}}) \cdot [1 - p_0(\hat{\boldsymbol{\theta}})]^{-1}$, where $\psi(\boldsymbol{\theta}) = \min\{m/n, 1 - p_0(\boldsymbol{\theta})\}$, and $L(\psi, \boldsymbol{\theta}) = (1-\psi)^{n-m}\psi^m \prod_{i:Y_i \neq 0} f_{\mathrm{tr}}(Y_i; \boldsymbol{\theta})$.

Solving for the MLE $(\hat{\phi}, \hat{\boldsymbol{\theta}})$ may involve two maximization problems, $\boldsymbol{\theta}_* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L_{\mathrm{tr}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\psi(\boldsymbol{\theta}), \boldsymbol{\theta})$. The following first order derivatives may be needed by, for example, quasi-Newton algorithms:

$$\frac{\partial \log L_{\mathrm{tr}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i:Y_i \neq 0} \frac{\partial \log f_{\boldsymbol{\theta}}(Y_i)}{\partial \boldsymbol{\theta}} - m \frac{\partial \log[1 - p_0(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial \log L(\psi(\boldsymbol{\theta}), \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{cases} \displaystyle\sum_{i:Y_i \neq 0} \frac{\partial \log f_{\boldsymbol{\theta}}(Y_i)}{\partial \boldsymbol{\theta}} - m \frac{\partial \log[1 - p_0(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} & \text{if } 1 - p_0(\boldsymbol{\theta}) > \dfrac{m}{n} \\ \displaystyle\sum_{i:Y_i \neq 0} \frac{\partial \log f_{\boldsymbol{\theta}}(Y_i)}{\partial \boldsymbol{\theta}} + (n-m) \frac{\partial \log p_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} & \text{if } 1 - p_0(\boldsymbol{\theta}) < \dfrac{m}{n} \end{cases}$$

Note that

$$\frac{\partial \log[1 - p_0(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} = -\frac{p_0(\boldsymbol{\theta})}{1 - p_0(\boldsymbol{\theta})} \cdot \frac{\partial \log p_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Thus for different models, we only need to prepare specific formulae of $\partial \log f_{\boldsymbol{\theta}}(y)/\partial \boldsymbol{\theta}$ and $\partial \log p_0(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ for numerical calculations.

According to Theorem 5 in [24], under some regularity conditions, the Fisher information matrix of a general zero-inflated distribution is

$$\mathbf{F}_{\mathrm{ZI}} = \begin{bmatrix} \dfrac{1 - p_0(\boldsymbol{\theta})}{[\phi + (1-\phi)p_0(\boldsymbol{\theta})](1-\phi)} & \dfrac{p_0(\boldsymbol{\theta})}{\phi + (1-\phi)p_0(\boldsymbol{\theta})} \cdot \dfrac{\partial \log p_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \\ \dfrac{p_0(\boldsymbol{\theta})}{\phi + (1-\phi)p_0(\boldsymbol{\theta})} \cdot \dfrac{\partial \log p_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} & \mathbf{F}_{\mathrm{ZI}\boldsymbol{\theta}} \end{bmatrix} \tag{8}$$

where

$$\mathbf{F}_{\mathrm{ZI}\boldsymbol{\theta}} = -(1-\phi)\left(E\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(Y')}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right] + \frac{\phi p_0(\boldsymbol{\theta})}{\phi+(1-\phi)p_0(\boldsymbol{\theta})}\cdot\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\cdot\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right)$$

and Y' follows the baseline distribution $f_{\boldsymbol{\theta}}(y)$. We denote $\mathbf{F}_{\boldsymbol{\theta}} = -E[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(Y')}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}]$, which is essentially the Fisher information matrix of the baseline distribution.

The Fisher information matrix (8) is more complicated than (4) for zero-altered models. To obtain the asymptotic distributions of MLEs, it is more convenient to obtain the following theorem by applying the formulae (see, for example, Formulae 4.33 and 14.13(b) in [27]) of block matrices and the Sherman-Morrison formula (see, for example, Section 2.1.4 in [28]).

Theorem 2.1: *Suppose* $\phi \in (0,1)$, $|\mathbf{F}_{\boldsymbol{\theta}}| \neq 0$ *and* $|\mathbf{F}_{\mathrm{ZI}}| \neq 0$. *Then*

$$\mathbf{F}_{\mathrm{ZI}}^{-1} = \begin{bmatrix} \phi(1-\phi)\cdot\frac{d_{\boldsymbol{\theta}}\delta_{\boldsymbol{\theta}}}{d_{\boldsymbol{\theta}}\delta_{\boldsymbol{\theta}}-p_0(\boldsymbol{\theta})} & -\frac{\phi p_0(\boldsymbol{\theta})}{d_{\boldsymbol{\theta}}\delta_{\boldsymbol{\theta}}-p_0(\boldsymbol{\theta})}\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\mathbf{F}_{\boldsymbol{\theta}}^{-1} \\ -\frac{\phi p_0(\boldsymbol{\theta})}{d_{\boldsymbol{\theta}}\delta_{\boldsymbol{\theta}}-p_0(\boldsymbol{\theta})}\mathbf{F}_{\boldsymbol{\theta}}^{-1}\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} & \mathbf{D}_{\boldsymbol{\theta}} \end{bmatrix}$$

where $d_{\boldsymbol{\theta}} = \phi + (1-\phi)p_0(\boldsymbol{\theta}) > 0$, $\delta_{\boldsymbol{\theta}} = 1 - \frac{\phi p_0(\boldsymbol{\theta})}{d_{\boldsymbol{\theta}}}\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\mathbf{F}_{\boldsymbol{\theta}}^{-1}\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \neq 0$, *and*

$$\mathbf{D}_{\boldsymbol{\theta}} = \frac{1}{1-\phi}\left[\mathbf{F}_{\boldsymbol{\theta}}^{-1} + \frac{\phi p_0(\boldsymbol{\theta})}{d_{\boldsymbol{\theta}}\delta_{\boldsymbol{\theta}}-p_0(\boldsymbol{\theta})}\mathbf{F}_{\boldsymbol{\theta}}^{-1}\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \log p_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\mathbf{F}_{\boldsymbol{\theta}}^{-1}\right]$$

The proof of Theorem 2.1 is relegated to the Supplementary Materials (Section S.1). With Theorem 2.1, we have $\sqrt{n}(\hat{\phi}-\phi) \xrightarrow{\mathcal{L}} N(0, \phi(1-\phi)\frac{d_{\boldsymbol{\theta}}\delta_{\boldsymbol{\theta}}}{d_{\boldsymbol{\theta}}\delta_{\boldsymbol{\theta}}-p_0(\boldsymbol{\theta})})$ and $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{D}_{\boldsymbol{\theta}})$. The relevant confidence intervals and hypothesis tests can be performed similarly as for zero-altered models.

As mentioned in Section 2.1, we need to calculating MLE and Fisher information accurately and efficiently. Explicit formulae of the Fisher information matrix of zero-inflated Poisson (ZIP) has been provided in Example 5 of [24]. In this section, we provide explicit formulae of the gradients and Fisher information matrices for zero-inflated geometric (ZIGe) and zero-inflated negative binomial (ZINB) models. We relegate the corresponding formulae for zero-inflated beta binomial (ZIBB) and zero-inflated beta negative binomial (ZIBNB) models to the Supplementary Materials (Section S.2).

Example 2.4 (Zero-inflated geometric model (ZIGe)): Same as in Example 2.1, the pmf of the baseline distribution can be written as $f_p(y) = p(1-p)^y$ with $y \in \{0,1,2,\ldots\}$ and parameter $p \in (0,1)$. According to Theorem 5 in [24], the Fisher information matrix of the corresponding ZIGe distribution is

$$\mathbf{F}_{\mathrm{ZIGe}} = \frac{1}{\phi+(1-\phi)p}\begin{bmatrix} \frac{1-p}{1-\phi} & 1 \\ 1 & \frac{(1-\phi)[\phi(1-p)+p(1-\phi)+\phi p^2]}{p^2(1-p)} \end{bmatrix}$$

Example 2.5 (Zero-inflated negative binomial model (ZINB)): Different from Example 3 of [24], we take the form of the pmf of the baseline NB distribution as $f_{\boldsymbol{\theta}}(y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p^r (1-p)^y$ with parameters $\boldsymbol{\theta} = (r, p) \in (0, \infty) \times [0, 1]$, $y \in \{0, 1, 2, \ldots\}$, which is more popular in the statistical literature (see, for example, [29]). In short, the p in Example 3 of [24] is replaced by $1-p$ here. Then $p_0(\boldsymbol{\theta}) = p^r$ and

$$\begin{aligned}
\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial r} &= \Psi(y+r) - \Psi(r) + \log p \\
\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial p} &= \frac{r}{p} - \frac{y}{1-p} \\
\frac{\partial \log p_0(\theta)}{\partial r} &= \log p \\
\frac{\partial \log p_0(\theta)}{\partial p} &= \frac{r}{p}
\end{aligned}$$

where $\Psi(\cdot)$ is the digamma function. According to Theorem 5 in [24], the Fisher information matrix of the ZINB distribution is

$$\mathbf{F}_{\text{ZINB}} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & & \\ A_{13} & & \mathbf{F}_{\text{ZINB}\boldsymbol{\theta}} \end{bmatrix}$$

where $A_{11} = \frac{1-p^r}{[\phi+(1-\phi)p^r](1-\phi)}$, $A_{12} = \frac{p^r \log p}{\phi+(1-\phi)p^r}$, $A_{13} = \frac{rp^{r-1}}{\phi+(1-\phi)p^r}$, and

$$\mathbf{F}_{\text{ZINB}\boldsymbol{\theta}} = -(1-\phi)\left(\begin{bmatrix} B_{11} & B_{12} \\ B_{12} & B_{22} \end{bmatrix} + \frac{\phi p^r}{\phi + (1-\phi)p^r} \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix}\right)$$

with $B_{11} = E\Psi_1(Y'+r) - \Psi_1(r)$, $B_{12} = \frac{1}{p}$, $B_{22} = -\frac{r}{p^2(1-p)}$, $C_{11} = (\log p)^2$, $C_{12} = \frac{r \log p}{p}$, and $C_{22} = \frac{r^2}{p^2}$. Here $\Psi_1(\cdot)$ is the trigamma function (see Example 2.2).

2.3. Zero-altered and zero-inflated models with continuous baseline distributions

As mentioned in Section 2.2, if the baseline distribution $f_{\boldsymbol{\theta}}(y)$ satisfies $p_0(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(Y = 0) = 0$ given $Y \sim f_{\boldsymbol{\theta}}(y)$, then the zero-altered model (1) and the zero-inflated model (6) are the same. We call such kind of models the *zero-altered-zero-inflated* (ZAZI) models. Examples include all continuous baseline distributions, as well as discrete or mixture baseline distributions satisfying $p_0(\boldsymbol{\theta}) = 0$. For ZAZI models with a baseline distribution $f_{\boldsymbol{\theta}}(y)$, its distribution function can be written as

$$f_{\text{ZAZI}}(y \mid \phi, \boldsymbol{\theta}) = \phi \mathbf{1}_{\{y=0\}} + (1-\phi) f_{\boldsymbol{\theta}}(y) \mathbf{1}_{\{y \neq 0\}} \tag{9}$$

Given a random sample $Y_1, \ldots, Y_n \sim f_{\text{ZAZI}}(y \mid \phi, \boldsymbol{\theta})$, the likelihood function of $(\phi, \boldsymbol{\theta})$ is

$$L(\phi, \boldsymbol{\theta}) = \phi^{n-m}(1-\phi)^m \cdot \prod_{i: Y_i \neq 0} f_{\boldsymbol{\theta}}(Y_i) \tag{10}$$

where $m = \#\{i : Y_i \neq 0\}$. Then the MLEs maximizing (10) are $\hat{\phi} = 1 - m/n$ and $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i: Y_i \neq 0} f_{\boldsymbol{\theta}}(Y_i)$, which are similar as (3) for zero-altered models. As a special case

of Theorem 3 in [24], we have the following formulae for the Fisher information matrix of ZAZI distributions.

Theorem 2.2: *Under regularity conditions, the Fisher information matrix of the ZAZI distribution* (9) *is*

$$\mathbf{F}_{\text{ZAZI}} = \begin{bmatrix} \phi^{-1}(1-\phi)^{-1} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{F}_{\text{ZAZI}\boldsymbol{\theta}} \end{bmatrix}$$

where

$$\mathbf{F}_{\text{ZAZI}\boldsymbol{\theta}} = -(1-\phi)\cdot E\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(Y')}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right]$$

and Y' follows the baseline distribution $f_{\boldsymbol{\theta}}(y)$.

In this section, we provide explicit formulae of gradients and Fisher information matrices for commonly used ZAZI models with continuous baseline distributions including Gaussian or normal, log-normal, half-normal, and exponential distributions.

Example 2.6 (Zero-altered-zero-inflated Gaussian model (ZAZIG)): This model has been used by, for example, [30] for analysing longitudinal microbiome data, known as a zero-inflated Gaussian (ZIG) model. The pdf of the baseline distribution with parameters $\theta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty)$ is given by $f_{\boldsymbol{\theta}}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}$. Then

$$\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial \mu} = \frac{y-\mu}{\sigma^2}$$
$$\frac{\partial \log f_{\theta}(y)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(y-\mu)^2}{\sigma^3}$$

According to Theorem 2.2, the Fisher information matrix of the ZAZIG distribution is

$$\mathbf{F}_{\text{ZAZIG}} = \begin{bmatrix} \phi^{-1}(1-\phi)^{-1} & 0 & 0 \\ 0 & \frac{1-\phi}{\sigma^2} & 0 \\ 0 & 0 & \frac{2(1-\phi)}{\sigma^2} \end{bmatrix}$$

Note that in this case, given a random sample $Y_1, \ldots, Y_n$ from ZAZIG, $\hat{\mu} = \frac{1}{m}\sum_{i:Y_i \neq 0} Y_i$ and $\hat{\sigma} = (\frac{1}{m}\sum_{i:Y_i \neq 0}(Y_i - \hat{\mu})^2)^{1/2}$ have explicit formulae, where $m = \#\{i : Y_i \neq 0\}$.

Example 2.7 (Zero-altered-zero-inflated log-normal model (ZAZILN)): This model, also known as zero-inflated log-normal (ZILN) model, has been used by, for example, [31] to study the effects of a prospective DUR intervention programme for randomized clinical trials. The pdf of the baseline distribution with parameters $\boldsymbol{\theta} = (\mu, \sigma) \in \mathbb{R} \times (0, \infty)$ is given by $f_{\boldsymbol{\theta}}(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(\log y - \mu)^2\}$, $y > 0$. Then

$$\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial \mu} = \frac{\log y - \mu}{\sigma^2}$$

$$\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(\log y - \mu)^2}{\sigma^3}$$

According to Theorem 2.2, the Fisher information matrix of the ZAZILN distribution is

$$\mathbf{F}_{\text{ZAZILN}} = \begin{bmatrix} \phi^{-1}(1-\phi)^{-1} & 0 & 0 \\ 0 & \frac{1-\phi}{\sigma^2} & 0 \\ 0 & 0 & \frac{2(1-\phi)}{\sigma^2} \end{bmatrix}$$

which is exactly the same as the one in Example 2.6.

Example 2.8 (Zero-altered-zero-inflated half-normal model (ZAZIHN)): Also known as a zero-inflated half-normal model (ZIHN), this model has been used by, for example, [32] as a candidate distribution for modelling the animal movement distances in the wild. In our notations, the pdf of its baseline distribution with parameter $\boldsymbol{\theta} = \sigma \in (0, \infty)$, known as a standard half-normal distribution, is given by $f_{\boldsymbol{\theta}}(y) = \frac{\sqrt{2}}{\sqrt{\pi\sigma^2}} \exp\{-\frac{y^2}{2\sigma^2}\}, y > 0$. Then $\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{y^2}{\sigma^3}$. According to Theorem 2.2, the Fisher information matrix of the ZAZIHN distribution is

$$\mathbf{F}_{\text{ZAZIHN}} = \begin{bmatrix} \phi^{-1}(1-\phi)^{-1} & 0 \\ 0 & \frac{2(1-\phi)}{\sigma^2} \end{bmatrix}$$

It should be noted that other forms of ZIHN models have also been used in the literature. For example, [33] utilized a more general ZIHN distribution as a prior for a hierarchical Bayesian model. The continuous component of their ZIHN is a general normal distribution $N(\mu, \sigma^2)$ truncated below by zero. In that case, $\boldsymbol{\theta} = (\mu, \sigma)$.

Example 2.9 (Zero-altered-zero-inflated exponential model (ZAZIE)): Also known as zero-inflated exponential model (ZIE), this model has been used by, for example, [34] for modelling casualty rates in ship collision. The pdf of its baseline distribution with parameter $\boldsymbol{\theta} = \lambda \in (0, \infty)$ can be written as $f_{\boldsymbol{\theta}}(y) = \lambda\, \mathrm{e}^{-\lambda y}, y > 0$. Then $\frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial \lambda} = \frac{1}{\lambda} - y$. According to Theorem 2.2, the Fisher information matrix of the ZAZIE distribution is

$$\mathbf{F}_{\text{ZAZIE}} = \begin{bmatrix} \phi^{-1}(1-\phi)^{-1} & 0 \\ 0 & \frac{1-\phi}{\lambda^2} \end{bmatrix}$$

Note that the MLE of λ has an explicit form $\hat{\lambda} = \frac{m}{\sum_{i:Y_i \neq 0} Y_i}$ given a random sample $Y_1, \ldots, Y_n$ from ZAZIE, where $m = \#\{i : Y_i \neq 0\}$.

2.4. Model selection based on KS and likelihood ratio tests

Given so many zero-altered or zero-inflated models listed in Sections 2.1–2.3, a critical question is which model is the most appropriate one for a given dataset.

The Kolmogorov-Smirnov (KS) test has been commonly used for testing whether a random sample $\{Y_1, \ldots, Y_n\}$ comes from a continuous cumulative distribution function $F_{\boldsymbol{\theta}}(y)$

with specified model parameter(s) $\boldsymbol{\theta}$ [35]. It is based on the KS statistic $D_n = \sup_y |F_n(y) - F_{\boldsymbol{\theta}}(y)|$, where $F_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{(-\infty,y]}(Y_i)$ is known as the *empirical distribution function*. Dimitrova et al. [36] extended the KS test for general distributions $F_{\boldsymbol{\theta}}(y)$ with known parameter(s) $\boldsymbol{\theta}$, including discrete and mixed ones. For typical applications, $\boldsymbol{\theta}$ is unknown and an estimate $\hat{\boldsymbol{\theta}}$ is plugged in when calculating D_n, which tends to overestimate the corresponding p-value [7,37,38]. To overcome the biasedness of estimated p-value due to the plugged-in estimated parameters, [7] proposed a bootstrapped Monte Carlo estimate $\frac{\#\{b|D_n^{(b)} > D_n\}+1}{B+1}$ for the p-value in their Algorithm 1, where $D_n^{(b)} = \sup_y |F_n^{(c)}(y) - F_{\hat{\boldsymbol{\theta}}^{(b)}}(y)|$, $F_n^{(c)}(y)$ is the empirical distribution function of a random sample $\mathbf{Y}^{(c)} = \{Y_1^{(c)}, \ldots, Y_n^{(c)}\}$ from $F_{\hat{\boldsymbol{\theta}}^{(b)}}$, $\hat{\boldsymbol{\theta}}^{(b)}$ is the MLE based on a bootstrapped sample $\mathbf{Y}^{(b)} = \{Y_1^{(b)}, \ldots, Y_n^{(b)}\}$ of $\{Y_1, \ldots, Y_n\}$, $b = 1, \ldots, B$, and B is a predetermined large number, typically $B = 1000$. If the estimated p-value is larger than 0.05, we say that the specified distribution passes the KS test for the given dataset.

Since $D_n = \sup_y |F_n(y) - F_{\hat{\boldsymbol{\theta}}}(y)|$ when $\boldsymbol{\theta}$ is unknown, where both F_n and $\hat{\boldsymbol{\theta}}$ are based on the same data $\{Y_1, \ldots, Y_n\}$, a more reasonable bootstrapped version of D_n would be $D_n^{(b)'} = \sup_y |F_n^{(c)}(y) - F_{\hat{\boldsymbol{\theta}}^{(c)}}(y)|$, where $\hat{\boldsymbol{\theta}}^{(c)}$ is the MLE based on $\mathbf{Y}^{(c)} = \{Y_1^{(c)}, \ldots, Y_n^{(c)}\}$ instead of $\mathbf{Y}^{(b)}$. In this paper, we propose the following nested bootstrap estimate for KS test p-value based on the above argument. To make a distinction, we call [7]'s Algorithm 1 as Algorithm 1A and our Algorithm 1 as Algorithm 1B. The corresponding `R` functions are named `kstest.A` and `kstest.B`, respectively. According to our simulation studies in Section 3.4, we recommend `kstest.B` for small sample sizes such as $n = 30, 50$. For larger sample sizes, since the difference in terms of test power is negligible, we recommend `kstest.A` for less computational cost.

Algorithm 1 Nested Bootstrap Algorithm for Estimating p-value of KS Test

1: Given data $\mathbf{Y} = \{Y_1, Y_2, \cdots Y_n\}$, calculate the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and the KS statistic $D_n = \sup_y |F_n(y) - F_{\hat{\boldsymbol{\theta}}}(y)|$.
2: For $b = 1, \ldots, B$, do steps 3∼7.
3: Resample $\mathbf{Y}$ with replacement to get a bootstrapped sample $\mathbf{Y}^{(b)} = \{Y_1^{(b)}, \cdots, Y_n^{(b)}\}$.
4: Calculate the MLE $\hat{\boldsymbol{\theta}}^{(b)}$ of $\boldsymbol{\theta}$ based on $\mathbf{Y}^{(b)}$.
5: Simulate a random sample $\mathbf{Y}^{(c)} = \{Y_1^{(c)}, \ldots, Y_n^{(c)}\}$ from $F_{\hat{\boldsymbol{\theta}}^{(b)}}$.
6: Calculate the MLE $\hat{\boldsymbol{\theta}}^{(c)}$ of $\boldsymbol{\theta}$ based on $\mathbf{Y}^{(c)}$.
7: Calculate the bootstrapped KS statistic $D_n^{(b)'} = \sup_y |F_n^{(c)}(y) - F_{\hat{\theta}^{(c)}}(y)|$, where $F_n^{(c)}(y)$ is the empirical distribution function of $\mathbf{Y}^{(c)}$.
8: Estimate the p-value of the KS test by $\frac{\#\{b|D_n^{(b)'} > D_n\}}{B}$.

In practice, it is not uncommon that two or more distributions pass the KS test for the same dataset, especially when the sample size is moderate or small (see, for example, [7]). In this situation, likelihood ratio tests may be used for pairwise comparisons. More specifically, for testing $H_0 : Y_1, \ldots, Y_n$ iid $\sim f(y; \boldsymbol{\theta})$ with unknown parameter(s) $\boldsymbol{\theta}$

against $H_1 : Y_1, \ldots, Y_n$ iid $\sim g(y; \boldsymbol{\delta})$ with unknown parameter(s) $\boldsymbol{\delta}$, the likelihood ratio test statistic in log scale (see, for example, [7]) can be defined as

$$\Lambda = \log \frac{\prod_{i=1}^{n} f(Y_i; \hat{\boldsymbol{\theta}})}{\prod_{i=1}^{n} g(Y_i; \hat{\boldsymbol{\delta}})}$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\delta}}$ are the corresponding maximum likelihood estimates. Smaller Λ values are in favour of the alternative distribution $g(y; \boldsymbol{\delta})$.

To overcome the possible biasedness due to plugged-in estimated parameters, we adopt the bootstrapped estimate of p-value described by Algorithm 2 of [7]. More specifically, (i) bootstrap samples $\mathbf{Y}^{(b)} = \{Y_1^{(b)}, \ldots, Y_n^{(b)}\}$, $b = 1, \ldots, B$ are obtained from the original data $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$; (ii) MLEs $\hat{\boldsymbol{\theta}}^{(b)}$ and $\hat{\boldsymbol{\delta}}^{(b)}$ are calculated based on $\mathbf{Y}^{(b)}$; (iii) a random sample $\mathbf{Y}^{(c)} = \{Y_1^{(c)}, \ldots, Y_n^{(c)}\}$ is simulated from $f(y; \hat{\boldsymbol{\theta}}^{(b)})$; (iv) bootstrapped test statistic $\Lambda^{(b)}$ is calculated based on $\mathbf{Y}^{(c)}$; and (v) the estimated p-value is $\frac{\#\{b|\Lambda^{(b)} < \Lambda\}}{B}$. If the p-value is less than 0.05, we claim that H_1 is significantly better; otherwise, we stick to H_0.

2.5. Zero-altered model versus zero-inflated model

Clearly a zero-altered model (1) and its corresponding zero-inflated model (6) are connected by sharing the same baseline distribution $f_{\boldsymbol{\theta}}(y)$. Under some conditions, they are actually equivalent due to the following theorem.

Theorem 2.3: *Let $f_{\mathrm{ZA}}(y \mid \phi_{\mathrm{ZA}}, \boldsymbol{\theta})$ be a zero-altered model as in* (1) *and $f_{\mathrm{ZI}}(y \mid \phi_{\mathrm{ZI}}, \boldsymbol{\theta})$ be the corresponding zero-inflated model as in* (6) *with the same baseline distribution $f_{\boldsymbol{\theta}}(y)$.*

(i) *Given $f_{\mathrm{ZI}}(y \mid \phi_{\mathrm{ZI}}, \boldsymbol{\theta})$, we let $\phi_{\mathrm{ZA}} = \phi_{\mathrm{ZI}} + (1 - \phi_{\mathrm{ZI}})p_0(\boldsymbol{\theta})$ and then $f_{\mathrm{ZA}}(y \mid \phi_{\mathrm{ZA}}, \boldsymbol{\theta}) = f_{\mathrm{ZI}}(y \mid \phi_{\mathrm{ZI}}, \boldsymbol{\theta})$ for all y.*

(ii) *Given $f_{\mathrm{ZA}}(y \mid \phi_{\mathrm{ZA}}, \boldsymbol{\theta})$ with $\phi_{\mathrm{ZA}} \geq p_0(\boldsymbol{\theta})$, we let $\phi_{\mathrm{ZI}} = [\phi_{\mathrm{ZA}} - p_0(\boldsymbol{\theta})]/[1 - p_0(\boldsymbol{\theta})]$ and then $f_{\mathrm{ZI}}(y \mid \phi_{\mathrm{ZI}}, \boldsymbol{\theta}) = f_{\mathrm{ZA}}(y \mid \phi_{\mathrm{ZA}}, \boldsymbol{\theta})$ for all y.*

The proof of Theorem 2.3 is relegated to the Supplementary Materials (Section S.1). Theorem 2.3 implies that if $\phi_{\mathrm{ZA}} \geq p_0(\boldsymbol{\theta})$, which indicates that the data is zero-inflated, then a ZA model is equivalent to the corresponding ZI model. Given a random sample from a zero-inflated model $f_{\mathrm{ZI}}(y \mid \phi_{\mathrm{ZI}}, \boldsymbol{\theta})$, a KS test will conclude that some zero-altered model $f_{\mathrm{ZA}}(y \mid \phi_{\mathrm{ZA}}, \boldsymbol{\theta})$ seems fine with the data as well. The other direction is slightly different though. That is, if the true model is a zero-altered one, only when ϕ_{ZA} is at least as large as $p_0(\boldsymbol{\theta})$, a KS test will conclude that some zero-inflated model seems to be true as well. In Section 3.5, we will provide a numerical example such that both ZIBNB and BNBH fit the data well.

3. Software implementation and numerical analysis

In this section, the `R` implementation of the proposed package `AZIZD` is introduced and its numerical analysis is performed. Real data examples are used to illustrate how our package can be used in practice.

Compared with existing R packages each covering only a limited number of baseline distributions, our package covers many more discrete and continuous distributions including Poisson, geometric, negative binomial, beta binomial, beta negative binomial, normal (or Gaussian), log-normal, half-normal, and exponential distributions along with their zero-altered and zero-inflated models, which facilitates the potential users to choose the most appropriate model from a large class of candidates for their dataset. For all the models mentioned above, we provide the corresponding Fisher information matrix and confidence intervals for estimated parameters, which allows users to run hypothesis tests and make further inference.

3.1. Improvements over existing packages on finding MLEs

From Section 2.4, we can see that an accurate MLE is a critical component for both KS test and likelihood ratio test. In this section, we first summarize the improvements of the proposed `AZIAD` over existing R packages on finding MLEs for zero-altered and zero-inflated models.

(1) Compared with other packages, `AZIAD` provides MLEs well even for extreme cases including $\phi = 0$ or $\phi = 1$ (see the comparison analysis in Section 3.3).
(2) Some packages encountered error message `L-BFGS-B needs finite values of "fn"` for some zero-inflated data when using, for example, the function `dis.kstest` in package `iZID`. This issue is solved in `AZIAD` by specifying appropriate lower bound and upper bound for optimizations Section 3.3.
(3) Compared with package `iZID` which also covered MLEs for ZIBNB, BNBH, ZIBB, and BBH models, our results are more reliable (see Example 3.2 for a comparison) by applying Theorem 4 in [24]. More specifically, we separate the two situations (1) $m/n \leq 1 - p_0(\boldsymbol{\theta}_*)$ and (2) $m/n > 1 - p_0(\boldsymbol{\theta}_*)$ and calculate the MLEs in each case.
(4) Some parameters are originally defined as positive integers, such as n in Examples 2.2 and S.1 and r in Examples 2.3 and 2.5, while in practice they could be extended to positive real numbers. In `AZIAD`, we seek for real-valued MLEs by default. In the mean time, we keep the option of integer-valued n or r in response to users' call.

In the rest part of this section, we use examples to illustrate how to use our package to find the MLEs and the improved accuracy by using our package.

Example 3.1: In order to find the MLE for the parameter(s) of zero-inflated or hurdle models, the main function built in `AZIAD` is

```
zih.mle(x, r, p, alpha1, alpha2, n, lambda, mean, sigma,
type = c("zi","h"), dist, lowerbound = 0.01,
upperbound = 10000)
```

where `x` is a sequence of numbers, which could be integers for discrete cases or real numbers for continuous cases; the arguments `r`, `p`, `alpha1`, `alpha2`, `n`, `lambda`, `mean`, and `sigma` are initial values of the corresponding parameters; `dist` could be chosen as `poisson.zihmle`, `geometric.zihmle`, `nb.zihmle`, `nb1.zihmle`,

`bb.zihmle`, `bb1.zihmle`, `bnb.zihmle`, `bnb1.zihmle`, `normal.zihmle`, `halfnorm.zihmle`, `lognorm.zimle`, and `exp.zihmle`, which correspond to zero-inflated or zero-altered Poisson, geometric, negative binomial, negative binomial with integer-valued r, beta binomial, beta binomial with integer-valued n, beta negative binomial, beta negative binomial with integer-valued r, normal, log-normal, half-normal, and exponential distributions, respectively; option `type` indicates the type of distribution is zero-inflated (`zi`) or hurdle/zero-altered (`h`); `lowerbound` and `upperbound` specify the searching range of parameters when maximizing the likelihood function. For instance, in order to calculate the MLE of zero-inflated geometric distribution (see Example 2.4), one may use the following `R` code:

```
R> set.seed(008)
R> x22=sample.h1(2000,phi=0.3,dist='geometric',p=0.3)
R> zih.mle(x22,p=0.2,dist="geometric.zihmle",type="h")
       p  phi  loglik
    0.2942292 0.3015 -4101.05
```

Our estimates $\hat{p} = 0.2942292$ and $\hat{\phi} = 0.3015$ are fairly close to the true parameter values $p = 0.3$ and $\phi = 0.3$.

Example 3.2: To compare the performance of our `AZIAD` package and the existing `iZID` on finding MLEs, we generate random samples from a BNBH model (see Example 2.3) with parameters $(\phi, r, \alpha_1, \alpha_2) = (0.3, 5, 8, 3)$, and a BBH model (see Example 2.2) with parameters $(\phi, n, \alpha_1, \alpha_2) = (0.6, 5, 8, 3)$, each with increasing sample sizes $N = 10^4, 5 \times 10^4, 20 \times 10^4$ and 100×10^4. To see the converging pattern more clearly, we generate nested datasets such that any dataset with a smaller N is a subset of the corresponding datasets with bigger Ns, if they are simulated from the same distribution.

For the each sample, we find the MLEs $(\hat{\phi}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ of the parameters and calculate the aggregated L_1 relative distance (L1RD) $L_1 = |\hat{\phi} - \phi|/|\phi| + \sum_{i=1}^{3} |\hat{\theta}_i - \theta_i|/|\theta_i|$. The implemented `R` codes are as follows:

```
R> set.seed(167)
R> hi1=sample.h1(N=1000000,phi=0.3,dist="bnb",r=5,
  alpha1=8,alpha2=3)
R> hi2=hi1[1:200000]
R> hi3=hi1[1:50000]
R> hi4=hi1[1:10000]
R> mle13=zih.mle(hi4,type="h",r=6,alpha1=9,alpha2=4,
  dist="bnb.zihmle")
R> mle13=zih.mle(hi4,type="h",r=6,alpha1=9,alpha2=4,
  dist="bnb1.zihmle")
R> mle14=bnb.zihmle(hi4,type="h",r=6,alpha1=9,alpha2=4)

R> set.seed(171)
R> hi1=sample.h1(N=1000000,phi=0.6,dist="bb",n=5,
  alpha1 = 8,alpha2=3)
```

Table 1. Comparison of MLE estimates, log-likelihood, and L_1 relative distance (L1RD) for BNB Hurdle distribution with true parameters ($r = 5, \alpha_1 = 8, \alpha_2 = 3, \phi = 0.3$) using `bnb.zihmle` in `iZID` and `zih.mle` in `AZIAD` with `dist=bnb.zihmle` (real-valued r) and `dist=bnb1.zihmle` (integer-valued r) with various sample sizes N.

N	R function	r	α_1	α_2	ϕ	loglike	L1RD
1×10^4	zih.mle(bnb)	3.81	7.69	3.81	0.303	−19232.34	0.555
	zih.mle(bnb1)	4	7.70	3.64	0.303	−19232.35	0.460
	bnb.zihmle	5.83	20.68	3.89	0	777027.7	3.05
5×10^4	zih.mle(bnb)	5.35	7.69	2.67	0.301	−96201.68	0.223
	zih.mle(bnb1)	5	7.59	2.82	0.301	−96201.71	0.113
	bnb.zihmle	30.09	453.25	28.96	0	134618737	70.32
20×10^4	zih.mle(bnb)	5.36	7.88	2.74	0.299	−384542.1	0.175
	zih.mle(bnb1)	5	7.78	2.90	0.299	−384542.2	0.061
	bnb.zihmle	92.90	5492.33	119.79	0	8730091135	743.05
1×10^6	zih.mle(bnb)	5.46	8.01	2.78	0.299	−1921040	0.181
	zih.mle(bnb1)	5	7.97	2.98	0.299	−1921041	0.009
	bnb.zihmle	101.23	6842.71	138.64	0	55719787227	919.80

```
R> hi2=hi1[1:200000]
R> hi3=hi1[1:50000]
R> hi4=hi1[1:10000]
R> mle23=zih.mle(hi1,n=6,alpha1=9,alpha2=4,type="h",
  dist="bb.zihmle")
R> mle23=zih.mle(hi4,n=6,alpha1=9,alpha2=4,type="h",
  dist="bb1.zihmle")
R> mle24=bb.zihmle(hi1,n=6,alpha1=9,alpha2=3,type="h")
```

In Tables 1 and 2, we list and compare the MLEs, log-likelihood, and L1RD obtained by our package (R function `zih.mle` with real-valued or integer-valued MLEs) and the `iZID` package (R functions `bnb.zihmle` and `bb.zihmle`) for BNBH and BBH models. We can see that the estimates based on our functions are more accurate as indicated by smaller L1RD. As the sample size increases, the L1RD based on our MLEs shows an overall decreasing pattern which indicates the convergence of MLEs towards the true parameter values. Similar results are collected for ZIBNB and ZIBB as well but not shown here.

3.2. *Fisher information, confidence interval, and test on zero-inflation*

As mentioned in Sections 2.1 – 2.3, the Fisher information matrix $\mathbf{F}_{\text{ZA}}$, $\mathbf{F}_{\text{ZI}}$ or $\mathbf{F}_{\text{ZAZI}}$ can be calculated by `AZIAD` for zero-altered/hurdle, zero-inflated, or ZAZI models, respectively. Its inverse matrix is approximately the variance-covariance matrix of parameter estimates $(\hat{\phi}, \hat{\boldsymbol{\theta}})$. For example, for zero-altered or hurdle models (see Section 2.1),

$$\sqrt{n}(\hat{\phi} - \phi) \dot{\sim} N(0, \phi(1-\phi)), \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} N(\mathbf{0}, \mathbf{F}^{-1}_{\text{ZA}\boldsymbol{\theta}})$$

for large n. Approximate confidence intervals can be constructed for ϕ and $\boldsymbol{\theta}$ (for example, expression (5) for ϕ).

In our package `AZIAD`, we use the function

```
FI.ZI(x, dist= "poisson", r = NULL, p = NULL, alpha1 =
```

Table 2. Comparison of MLE estimates, log-likelihood, and L_1 relative distance (L1RD) for BB Hurdle distribution with true parameters ($n = 5, \alpha_1 = 8, \alpha_2 = 3, \phi = 0.6$) using `bb.zihmle` in `iZID` and `zih.mle` in `AZIAD` with `dist=bb.zihmle` (real-valued r) and `dist=bb1.zihmle` (integer-valued r) with various sample sizes N.

N	R function	n	α_1	α_2	ϕ	loglike	L1RD
1×10^4	zih.mle(bb)	4.99	7.58	2.84	0.597	−12572.96	0.110
	zih.mle(bb1)	5	7.86	2.96	0.597	−12579.31	0.033
	bb.zihmle	6	9	3	0.597	28775.56	0.33
5×10^4	zih.mle(bb)	4.99	7.57	2.78	0.597	−62688.44	0.129
	zih.mle(bb1)	5	7.85	2.91	0.597	−62721.29	0.051
	bb.zihmle	6	9	3	0.597	143733.9	0.328
20×10^4	zih.mle(bb)	4.99	7.67	2.84	0.600	−249962	0.094
	zih.mle(bb1)	5	7.95	2.97	0.600	−250089.3	0.015
	bb.zihmle	6	9	3	0.600	570424.6	0.325
1×10^6	zih.mle(bb)	4.99	7.66	2.85	0.600	−1249976	0.093
	zih.mle(bb1)	5	7.94	2.97	0.600	−1250610	0.015
	bnb.zihmle	6	9	3	0.600	2850793	0.325

```
NULL, alpha2 = NULL, n = NULL,lambda=NULL, mean=NULL,
sigma=NULL, lowerbound = 0.01, upperbound = 10000)
```

to calculate (the inverse of) the Fisher information matrix at the MLE and 95% approximate confidence intervals for all parameters, where `x` is the data in its vector form; `dist` can be `poisson`, `geometric`, `nb`, `bb`, `bnb`, `normal`, `halfnormal`, `lognormal`, `exponential`, `zip`, `zigeom`, `zinb`, `zibb`, `zibnb`, `zinormal`, `zilognorm`, `zihalfnorm`, `ziexp`, `ph`, `geomh`, `nbh`, `bbh`, and `bnbh` for Poisson, geometric, negative binomial, beta binomial, beta negative binomial, normal/Gaussian, log-normal, half-normal, exponential, their zero-inflated versions and hurdle versions, respectively; `r`, `p`, `alpha1`, `alpha2`, `n`, `lambda`, `mean`, `sigma` provide initial values of distribution parameters; `lowerbound` and `upperbound` are predetermined ranges for MLEs, which could be extended if attained by any of the estimated parameter values.

Example 3.3 (Confidence intervals for ZINB): As an illustration, we apply the `FI.ZI` function to a simulated ZINB dataset with true parameters $\phi = 0.4$, $r = 10$ and $p = 0.2$.

```
R> set.seed(117)
R> N=1000;r=10;p=0.2;phi=0.4;
R> x<-sample.zi1(N,phi=phi,dist="nb",r=r,p=p)
R> FI.ZI(x,r=3,p=0.1, dist="zinb")
$inversefisher
      [,1]     [,2]     [,3]
[1,] 0.32302489 0.0661767 0.02116832
[2,] 0.06617670 35.0781124 -0.53660185
[3,] 0.02116832 -0.5366019 0.03658915
$ConfidenceIntervals
        [,1]     [,2]
CI of phi 0.3937737 0.4642262
CI of r  9.8952803 10.6294496
```

```
CI of p  0.1923346 0.2160458
```

With sample size $N = 1000$, the derived 95% confidence intervals cover the true parameter values fairly well.

Example 3.4 (Test zero-inflation in Poisson data): As an illustration, we simulate a ZIP data with $N = 1000$, $\phi = 0$ and $\lambda = 0.8$, which is actually a regular Poisson data with $\lambda = 0.8$ without zero-inflation. There are 459 zeros which roughly matches the probatility of zero $p_0(\lambda) = 0.449$. Then we use the `FI.ZI` function with distribution Poisson Hurdle (PH) to allow both zero-inflation and deflation.

```
R> set.seed(337)
R> x<-sample.zi1(N=1000,phi=0,dist="poisson",lambda=0.8)
R> table(x)
x
 0  1  2  3  4  5
459 334 153 41 10  3
R> dpois(0,lambda=0.8)
[1] 0.449329
R> FI.ZI(x,lambda=1, dist="ph")
$inversefisher
     [,1]   [,2]
[1,] 0.248319 0.000000
[2,] 0.000000 2.558941
$ConfidenceIntervals
          [,1]   [,2]
CI of Phi  0.4281146 0.4898854
CI of lambda 0.7937409 0.9920343
R> FI.ZI(x,lambda=1, dist="poisson")
$inversefisher
lambda
[1,] 0.818
$ConfidenceIntervals
[1] 0.7619437 0.8740563
```

It turns out that the 95% approximate confidence interval $(0.4281, 0.4899)$ of ϕ based on PH model (`dist="ph"`) covers the baseline zero probability $p_0(\lambda) = 0.4493$, which indicates that neither zero-inflation nor zero-deflation is significant at the 5% level. Actually, the 95% approximate confidence interval $(0.7619, 0.8741)$ of λ based on the regular Poisson model (`dist="poisson"`) covers the true value 0.8 roughly at the centre, which is better than $(0.7937, 0.9920)$ based on the PH model.

3.3. Comparison study in terms of type I error

In this section and Section 3.4, we use simulation studies to compare the performance of our algorithms for model identification with existing `R` functions for the same purposes. In

Table 3. Type I error rates of KS tests on whether the data comes from ZIP, based on $B = 1000$ simulated ZIP datasets with parameters $\phi = 0.3$, $\lambda = 10$ for each sample size N.

Sample size N	30	50	100	200	500
ks.test	0	0	0	0.001	0.005
disc_ks_test	0.074	0.116	0.081	0.112	0.08
dis.kstest	0	0	0	0	0
kstest.A	0	0	0.001	0	0
kstest.B	0	0	0	0	0

Table 4. Type I error rates of KS tests on whether the data comes from ZINB, based on $B = 1000$ simulated ZINB datasets with parameters $\phi = 0.3$, $r = 5$, $p = 0.2$ for each sample size N.

Sample size N	30	50	100	200	500
ks.test	0.001	0.002	0.001	0.004	0.003
disc_ks_test	0.062	0.065	0.099	0.089	0.107
dis.kstest	0.003 (536NA)	0.004 (638NA)	0.003 (903NA)	0.001 (996NA)	All NA
kstest.A	0	0	0	0	0
kstest.B	0	0	0	0	0

Table 5. Type I error rates of KS tests on whether the data comes from ZIBNB, based on $B = 1000$ simulated ZIBNB datasets with parameters $\phi = 0.3$, $r = 3$, $\alpha_1 = 3$, $\alpha_2 = 5$ for each sample size N.

Sample size N	30	50	100	200	500
ks.test	0.058	0	0	0.001	0
disc_ks_test	0.062	0.049	0.067	0.07	0.084
dis.kstest	0.788 (176NA)	0.889 (89NA)	0.921 (87NA)	0.832 (168NA)	0.671 (383NA)
kstest.A	0	0	0	0.002	0.002
kstest.B	0.002	0	0	0.003	0.001

this section we focus on type I errors. That is, given the true distribution, say ZINB, what is the chance that the test erroneously concludes that the distribution is not ZINB due to a p-value less than 0.05. Such an error is known as a *type I error* [39]. Ideally, such a chance is no more than 0.05, known as the *size* of the test.

The `R` functions under comparison in this section include the basic function `ks.test`, function `disc_ks_test` in package `KSgeneral` [36], function `dis.kstest` in package `iZID` [23], and two of our functions in `AZIAD`, `kstest.A` based on [7]'s Algorithm 1 with $(\#b + 1)/(B + 1)$ replaced by $\#b/B$, and `kstest.B` based on our Algorithm 1 described in Section 2.4.

For illustration purposes, we consider four zero-inflated models, ZIP, ZINB, ZIBNB, and ZIBB. For each model, we consider five different samples sizes, $N = 30$, 50, 100, 200, 500, and simulate $B = 1000$ independent datasets for each sample size. For each simulated dataset, we run the five `R` functions under comparison individually targeting the corresponding true model. If the p-value of the test is less than 0.05, we count it as a type I error. The ratios of type I errors (that is, the number of type I errors divided by $B = 1000$) are listed in Tables 3 – 6. For readers' reference, we provide the `R` code for generating Table 4 in the Supplementary Materials (Section S.3).

According to these tables, `disc_ks_test` (package `KSgeneral`) seems to have larger type I error rates than the nominal level 0.05. One concern about function

Table 6. Type I error rates of KS tests on whether the data comes from ZIBB, based on $B = 1000$ simulated ZIBB datasets with parameters $\phi = 0.3, n = 5, \alpha_1 = 8, \alpha_2 = 3$ for each sample size N.

Sample size N	30	50	100	200	500
ks.test	0	0	0.001	0	0
disc_ks_test	0.054	0.06	0.06	0.058	0.066
dis.kstest	0.039 (24NA)	0.05 (34NA)	0.084 (47NA)	0.645 (64NA)	0.947 (53NA)
kstest.A	0.001	0	0	0	0
kstest.B	0.001	0	0	0	0

`dis.kstest` (`iZID`) for ZINB, ZIBNB and ZIBB distributions is its significant portions of `NA`'s due to errors. Our two functions, `kstest.A` and `kstest.B`, and `R` basic function `ks.test` have type I error rates nearly zero, which are satisfactory.

3.4. Comparison study in terms of test power

In this section, we continue the simulation studies in Section 3.3 and compare the power of the tests based on four different `R` functions. More specifically, given the $B = 1000$ datasets simulated from, for example, the ZIP distribution, we run a KS test on whether the data comes from a ZINB distribution, and denote the test as ZIP versus ZINB. If we reject ZINB distribution for 800 times, then the empirical power of the KS test at ZINB is $800/1000 = 0.80$. Given that the type I error rate does not go beyond the nominal level 0.05, the bigger the power is, the better the test performs (see, for example, [39]). We remove function `dis.kstest` of package `iZID` from power analysis since it generates too many `NA`'s for ZINB (see Table 4) or has a type I error rate much higher than 0.05 for ZIBNB (see Table 5) and ZIBB (see Table 6).

Table 7 shows that our functions, `kstest.A` and `kstest.B`, have larger power at ZIP versus ZIBNB. All tests fail at ZIP versus ZINB and ZIP versus ZIBB. It often happens for a test of a simpler model versus a more flexible model, especially when the flexible model could approximate the simpler model well (see Section 4 for a discussion on it). More simulation studies show that our functions have the largest powers at ZIBNB versus ZIP, and at ZIBB versus ZIBNB as well (see Tables S.1 –S.3 in the Supplementary Materials, Section S.4). Note that both `ks.test` and `disc_ks_test` here rely on the MLEs for ZIBB and ZIBNB provided by our package `AZIAD`.

Overall our two functions `kstest.A` and `ks.test.B` are most reliable for KS tests involving zero-inflated models. In practice, we recommend `ks.testB` for small sample sizes such as $N = 30, 50$ (see ZIP versus ZIBNB in Table 7, ZINB versus ZIP in Table S.2, and ZIBNB versus ZIP in Table S.2) and `ks.testA` for large sample size such as $N = 100, 200, 500$, which is faster.

3.5. Real data analysis

Example 3.5 (DedTrivedi Data): In this example, we use a real data `DebTrivedi` from `R` package `MixAll` to illustrate how to use our package `AZIAD` to identify the most appropriate zero-inflated model. The `DebTrivedi` data was obtained from the US National Medial Expenditure Survey [40,41]. It contains 19 variables from 4,406 individuals aged

Table 7. Empirical power of KS tests on whether the data comes from ZINB, ZIBNB or ZIBB, based on $B = 1000$ simulated ZIP datasets with parameters $\phi = 0.3$, $\lambda = 10$ for each sample size N.

Test	Function	$N = 30$	$N = 50$	$N = 100$	$N = 200$	$N = 500$
ZIPvsZINB	ks.test	0.002	0	0.001	0.001	0
	disc_ks_test	0.079	0.087	0.076	0.096	0.093
	kstest.A	0	0	0	0	0
	kstest.B	0	0	0	0	0
ZIPvsZIBNB	ks.test	0.132	0.441	0.844	0.978	0.984
	disc_ks_test	0.115	0.096	0.084	0.099	0.101
	kstest.A	0.749	0.892	0.954	0.973	0.992
	kstest.B	0.73	0.984	0.954	0.973	0.992
ZIPvsZIBB	ks.test	0.002	0.001	0.001	0.005	0.007
	disc_ks_test	0.079	0.084	0.09	0.074	0.091
	kstest.A	0	0	0	0.001	0
	kstest.B	0	0	0	0	0

66 and over. For illustration purpose, we select the variable `ofp`, which is the number of physician office visits.

In order to analyse the data, we first use KS tests to check each distribution covered by our package. Since the sample size 4,406 is fairly large, we use our function `kstest.A` instead of `kstest.B`. Out of 19 distributions under test, we have 7 models with p-value larger than 0.05, including geometric, NB (with real-valued r), NB1 (with integer-valued r), ZIBB, ZIBNB, BBH, and BNBH.

Since there are seven distributions passing the KS test, we further run our likelihood ratio test function `lrt.A` to compare each pair of the candidate distributions. For example, if `d1` represents geometric distribution and `d2` represents NB distribution, we run `lrt.A(d1, d2)`. A p-value less than 0.05 indicates that `d2` is significantly better than `d1`. The relevant `R` codes are relegated to the Supplementary Materials (Section S.3).

Table 8 shows the results of pairwise comparisons of the seven candidate distributions based on `lrt.A(H0,H1)` with row names indicating H_0 and column names indicating H_1. A small p-value implies that the column distribution is significantly better than the row distribution. It should be noted that in general the p-values of `lrt.A(d1,d2)` and `lrt.A(d2,d1)` are not equal. Based on the p-values in Table 8, we conclude that ZIBNB and BNBH are significantly better than geometric, NB, NB1, ZIBB and BBH, while there is no significant difference between ZIBNB and BNBH, which confirms our conclusion in Section 2.5. Note that neither ZIBNB nor BNBH was considered by Deb and Trivedi [40] and Zeileis et al. [41]. Both distributions were recommended by Aldirawi et al. [7] and Aldirawi and Yang [24] for microbiome data analysis.

We run the same procedure based on our function `lrt.B` as well, which is based on `kstest.B`. The results are consistent with `lrt.A`'s. Similarly as in Section 3.4, we recommend `lrt.B` for cases with smaller sample sizes and `lrt.A` for cases with a sample size larger than 100.

Example 3.6 (Omic Data): In this example, we analyse the `Omic` data from [42], which is a list of 229 bacterial and fungal OTUs, for identifying appropriate models. More specifically, for each of the following distributions, Poisson, geometric, negative binomial, beta

Table 8. The *p*-values of pairwise comparisons of the seven candidate models based on `lrt.A` for data `DebTrivedi`, a *p*-value less than 0.05 indicating that the corresponding column model is significantly better than the corresponding row model.

	H_1						
H_0	Geometric	NB	NB1	ZIBB	ZIBNB	BBH	BNBH
Geometric	1	0.88	0.79	0.9	0	0.81	0
NB	0.07	1	0.165	0.835	0	0.49	0
NB1	0.84	0.82	1	0.895	0	0.8	0
ZIBB	0	0.04	0.015	1	0	0.96	0
ZIBNB	0.66	0.725	0.69	0.81	1	0.83	1
BBH	0.005	0.09	0.015	1	0	1	0
BNBH	0.705	0.73	0.67	0.81	0.995	0.735	1

Table 9. Number and percentage of species out of 229 that passed `kstest.A` with *p*-value > 0.05.

Distribution	Number	Percentage
Poisson	0	0%
Geometric	2	0.8%
NB	0	0%
BB	149	65%
BNB	170	74%
ZIP	5	2%
ZIGe	56	24%
ZINB	57	25%
ZIBB	148	65%
ZIBNB	172	75%
PH	4	2%
GeH	55	24%
NBH	56	24%
BBH	181	79%
BNBH	200	87%

binomial, beta negative binomial, and their corresponding zero-inflated and hurdle models, we use our `kstest.A` in `AZIAD` to check how many species out of 229 passed the corresponding KS tests.

Table 9 summarizes the numbers and percentages of species that do not show significant divergence (*p*-value > 0.05). The bigger the number is, the more appropriate the model is for omic data. The relevant `R` codes are displayed below:

```
R> dvect = list("poisson", "zip", "ph", "geometric",
    "zigeom", "geomh",
    "nb", "zinb", "nbh", "bb", "zibb", "bbh", "bnb",
    "zibnb", "bnbh")
R> dmatnew = matrix(, nrow=229, ncol=15)
R> set.seed(473415)
R> for(i in 1:229) for(j in 13:15)
 {dmatnew[i,j] = kstest.A(as.numeric(omic[i,]),
 dist=dvect[j])$pvalue}
R> write.csv(dmatnew, file="kstestBunpolishedfinal.csv")
```

Table 10. Maximum difference between the CDF of ZIP and the CDF of fitted ZINB based on random samples from ZIP($\phi = 0.3, \lambda = 10$) with various sample sizes N.

N	30	50	100	200	500	1000	5000
$\sup_y \lvert F_{\text{ZIP}}(y) - F_{\text{ZINB}}(y)\rvert$	0.166	0.12	0.08	0.03	0.024	0.04	0.0158

Table 11. Maximum difference between the CDF of ZIBNB and the CDF of fitted ZIBB based on random samples from ZIBNB($\phi = 0.3, r = 15, \alpha_1 = 19, \alpha_2 = 10$) with various sample sizes N.

N	30	50	100	200	500	1000	5000
$\sup_y \lvert F_{\text{ZIBNB}}(y) - F_{\text{ZIBB}}(y)\rvert$	0.166	0.1	0.1	0.11	0.04	0.024	0.014

We conclude that Poisson, geometric, negative binomial (NB), zero-inflated Poisson (ZIP), Poisson hurdle (PH) and geometric hurdle (GeH) are not appropriate for sparse microbial features due to their low percentages (no more than 2%), while beta binomial (BB), beta negative binomial (BNB) and their zero-inflated and hurdle versions are much more popular (at least 64%). Among them, BNBH (or ZABNB, see Example 2.3) is the most appropriate model with percentage 87%. Compared with Table 1 in [7] or Table 2 in [24], our results are more reliable although the patterns are similar.

4. Discussion

A major goal targeted in this paper is to identify the underlying distribution F_0 given a random sample $\{Y_1, \ldots, Y_n\}$ from it. Given the data $\{Y_1, \ldots, Y_n\}$, ideally our tests could accomplish two tasks, (i) do not reject F_0 itself; (ii) reject any F_1 which is not F_0.

According to the Glivenko-Cantelli theorem (see, for example, Theorem 19.1 in [43]), the empirical distribution function F_n converges to F_0 uniformly and then Task (i) can be achieved by proper KS tests up to a type I error, which is confirmed by our simulation studies in Section 3.3.

Task (ii) is much more complicated. First of all, in Section 2.5, we show the equivalence of a zero-inflated model and its corresponding hurdle model given that $\phi_{\text{ZA}} \geq p_0(\boldsymbol{\theta})$. Therefore, one could not make a distinction between a ZI model and its corresponding ZA or hurdle model when zero-inflation exists. We have such a real data example in Section 3.5.

Secondly, as shown by simulation studies in Section 3.4, one may not be able to reject F_1 if F_1 is a more flexible model than F_0, especially when F_1 with fitted parameters could approximate F_0 well. In this section, we use ZIP versus ZINB as an example to illustrate why one may not be able to reject F_1 given that F_0 is the true model.

It is known that Poisson(λ) can be approximated by NB(r, p) with a large r and a p close to 1 such that $\lambda = r(1-p)$ (see, for example, [44]). To illustrate how well ZINB could approximate a ZIP distribution, we simulate random samples $\{Y_1, \ldots, Y_N\}$ from ZIP($\phi = 0.3, \lambda = 10$) with various sample sizes N, then fit a ZINB model. In Table 10, we list the maximum difference between the cumulative distribution function (CDF) $F_{\text{ZIP}}(y)$ of ZIP and the CDF $F_{\text{ZINB}}(y)$ of the fitted ZINB. As the sample size N increases, the maximum distance $\sup_y \lvert F_{\text{ZIP}}(y) - F_{\text{ZINB}}(y)\rvert$ decreases, which indicates ZINB approximates ZIP better and better.

Another example is the KS test of ZIBNB versus ZIBB, whose empirical power shows a decreasing pattern as the sample size N increases (see Table S.2 in the Supplementary Materials). We perform a similar numerical study here for ZIBNB versus ZIBB as well. More specifically, we simulate random samples $\{Y_1, \ldots, Y_N\}$ from ZIBNB($\phi = 0.3, r = 15, \alpha_1 = 19, \alpha_2 = 10$) with various sample sizes N, then fit a ZIBB model. Table 11 shows a decreasing pattern of the maximum difference between the two CDFs as the sample size N increases, which indicates ZIBB can approximate ZIBNB better and better.

Disclosure statement

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

Funding

This work was supported in part by the U.S. National Science Foundation (NSF) [grant number DMS-1924859], and CSUSB research fellowship.

References

[1] Xia Y, Sun J, Chen DG. Statistical analysis of microbiome data with R. Singapore: Springer; 2018.
[2] Metwally AA, Aldirawi H, Yang J. A review on probabilistic models used in microbiome studies. Commun Inf Syst. 2018;18(3):173–191.
[3] McDavid A, Gottardo R, Simon N, et al. Graphical models for zero-inflated single cell gene expression. Ann Appl Stat. 2019;13(2):848–873.
[4] Majo MC, van Soest A. The fixed-effects zero-inflated poisson model with an application to health care utilization; 2011. Available from: https://pure.uvt.nl/ws/portalfiles/portal/1341748/2011-083.pdf
[5] Boucher JP, Denuit M, Guillen M. Number of accidents or number of claims? An approach with zero-inflated poisson models for panel data. J Risk Insur. 2009;76(4):821–846.
[6] Chen P, Liu Q, Sun F. Bicycle parking security and built environments. Transp Res D Transp Environ. 2018;62:169–178.
[7] Aldirawi H, Yang J, Metwally AA. Identifying appropriate probabilistic models for sparse discrete omics data. 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); IEEE; 2019. p. 1–4.
[8] Metwally AA, Dai Y, Finn PW, et al. MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies; 2019. R package version 1.1.8. Available from: https://CRAN.R-project.org/package = MetaLonDA
[9] Xu L, Paterson AD, Turpin W, et al. Assessment and selection of competing models for zero-inflated microbiome data. PLoS One. 2015;10(7):Article ID e0129606.
[10] Cho H, Liu C, Park J, et al. bzinb: bvariate zero-inflated negative binomial model estimator; 2018. R package version 1.0.4. Available from: https://CRAN.R-project.org/package = bzinb
[11] Balderama E, Trippe T. hurdlr: zero-inflated and hurdle modelling using bayesian inference; 2017. R package version 0.1. Available from: https://CRAN.R-project.org/package = hurdlr
[12] Wang L, Aldirawi H, Yang J. iZID: identify zero-inflated distributions; 2019. R package version 0.0.1. Available from: https://cran.r-project.org/web/packages/iZID
[13] Stasinopoulos M. gamlss: generalised additive models for location scale and shape; 2022. R package version 0.0.1. Available from: https://CRAN.R-project.org/package = gamlss
[14] Jackman S. pscl: political science computational laboratory; 2020. R package version 0.0.1. Available from: https://CRAN.R-project.org/package = pscl
[15] Albasi A. mazeinda: monotonic association on zero-inflated data; 2018. R package version 0.0.1. Available from: https://CRAN.R-project.org/package = mazeinda

[16] Croissant Y, Carlevaro F, Hoareau S. mhurdle: multiple hurdle tobit models; 2021. R package version 1.3.0. Available from: https://CRAN.R-project.org/package = mhurdle
[17] Waudby-Smith I, Li P. rbtt: alternative bootstrap-based t-test aiming to reduce type-i error for non-negative, zero-inflated data; 2017. R package version 0.1.0. Available from: https://CRAN.R-project.org/package = rbtt
[18] Hu T, Zhou Y. ZIBBSeqDiscovery: zero-inflated beta-binomial modeling of microbiome count data; 2018. R package version 1.0. Available from: https://CRAN.R-project.org/package = ZIBB SeqDiscovery
[19] Peng X, Li G, Liu Z, et al. ZIBseq: differential abundance analysis for metagenomic data via zero-inflated beta regression; 2017. R package version 1.2. Available from: https://CRAN.R-project.org/package = ZIBseq
[20] Jochmann M. zic: Bayesian inference for zero-inflated count models; 2017. R package version 0.9.1. Available from: https://CRAN.R-project.org/package = zic
[21] Yang M, Zamba G, Cavanaugh J. ZIM: zero-inflated models (zim) for count time series with excess zeros; 2018. R package version 1.1.0. Available from: https://CRAN.R-project.org/pack age = ZIM
[22] Xu ZJ, Liu Y. ziphsmm: zero-inflated poisson hidden (semi-)markov models; 2018. R package version 2.0.6. Available from: https://CRAN.R-project.org/package = ziphsmm
[23] Wang L, Aldirawi H, Yang J. Identifying zero-inflated distributions with a new R package iZID. Commun Inf Syst. 2020;20(1):23–44.
[24] Aldirawi H, Yang J. Modeling sparse data using mle with applications to microbiome data. J Stat Theory Pract. 2022;16(1):Article ID 13.
[25] Agresti A. Categorical data analysis. 3rd ed. Hoboken, NJ: Wiley; 2013.
[26] Mullahy J. Specification and testing of some modified count data models. J Econom. 1986;33(3):341–365.
[27] Seber G. A matrix handbook for statisticians. Hoboken, NJ: Wiley; 2008.
[28] Golub G, Loan CV. Matrix computations. 4th ed. Baltimore, MD: Johns Hopkins University Press; 2013.
[29] Hogg RV, McKean JW, Craig AT. Introduction to mathematical statistics. 8th ed. New York, NY: Pearson Education; 2019.
[30] Zhang X, Guo B, Yi N. Zero-inflated Gaussian mixed models for analyzing longitudinal microbiome data. PLoS One. 2020;15(11):Article ID e0242073.
[31] Zhou XH, Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. Biometrics. 1999;55(2):645–651.
[32] Reyna-Hurtado R, Chapman CA, Calme S, et al. Searching in heterogeneous and limiting environments: foraging strategies of white-lipped peccaries (Tayassu pecari). J Mammal. 2012;93(1):124–133.
[33] Chen Z, Dunson DB. Random effects selection in linear mixed models. Biometrics. 2003;59(4):762–769.
[34] Huang D, Hu H, Li Y. Zero-inflated exponential distribution of casualty rate in ship collision. J Shanghai Jiaotong Univ (Sci). 2019;24(6):739–744.
[35] Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951;46(253):68–78.
[36] Dimitrova DS, Kaishev VK, Tan S. Computing the Kolmogorov-Smirnov distribution when the underlying CDF is purely discrete, mixed, or continuous. J Stat Softw. 2020;95(10):1–42.
[37] Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J Am Stat Assoc. 1967;62(318):399–402.
[38] Lilliefors HW. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. J Am Stat Assoc. 1969;64(325):387–389.
[39] Lehmann EL, Romano JP. Testing statistical hypotheses. 3rd ed. New York, NY: Springer; 2005.
[40] Deb P, Trivedi PK. Demand for medical care by the elderly: a finite mixture approach. J Appl Econ. 1997;12(3):313–336.
[41] Zeileis A, Kleiber C, Jackman S. Regression models for count data in r. J Stat Softw. 2008;27(8):1–25.

[42] Tipton L, Müller CL, Kurtz ZD, et al. Fungi stabilize connectivity in the lung and skin microbial ecosystems. Microbiome. 2018;6(1):12.
[43] Vaart AW van der. Asymptotic statistics. New York, NY: Cambridge University Press; 2000.
[44] Teerapabolarn K. An improved poisson to approximate the negative binomial distribution. Int J Pure Appl Math. 2014;91(3):369–373.