A multivariate zero-inflated binomial model for the analysis of correlated proportional data

Dianliang Deng\textsuperscript{a}, Yiguang Sun\textsuperscript{a} and Guo-Liang Tian\textsuperscript{b}

\textsuperscript{a}Department of Mathematics and Statistics, University of Regina, Regina, Canada; \textsuperscript{b}Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen City, Guangdong, People’s Republic of China

ABSTRACT
In this paper, a new multivariate zero-inflated binomial (MZIB) distribution is proposed to analyse the correlated proportional data with excessive zeros. The distributional properties of the proposed model are studied. The Fisher scoring algorithm and EM algorithm are given for the computation of estimates of parameters in the proposed MZIB model with/without covariates. The score tests and the likelihood ratio tests are derived for assessing both the zero-inflation and the equality of multiple binomial probabilities in correlated proportional data. A limited simulation study is performed to evaluate the performance of the proposed EM algorithms for the estimation of parameters in the model with/without covariates and to compare the nominal levels and powers of both score tests and likelihood ratio tests. The whitefly data is used to illustrate the proposed methodologies.

1. Introduction
Count and proportional data have been used in a wide variety of fields of study, including education, sociology, psychology, biology, toxicology, epidemiology, insurance, public health, engineering, ecology, econometrics, agriculture, manufacturing and horticulture. When analysing such data, generalized linear models are extensively used. However, these data often present a larger number of zero observations than what would normally arise from the standard count and proportional distributions. When those issues are not properly addressed, the analysis using usual GLMs such as binomial and Poisson models even the over-dispersed GLMs may not provide a good fit and fail to explain the kinds of variation to the actual data. Therefore, statisticians proposed so-called zero-inflated models to fit such data.

The work of zero-inflated models has a long history that could be traced back to at least the 1960s when Cohen [5] and Johnson and Kotz [13] discussed zero-inflated Poisson (ZIP) models without covariates for count data. Later the ZIP models with covariates were studied by Lambert [15] for application to defects in manufacturing. The zero-inflated...
negative binomial (ZINB) models were studied by Deng and Paul [7] for the count data with both zero-inflation and over-dispersion. Hall [9] and Vieira et al. [23] proposed zero-inflated binomial (ZIB) distributions for modelling proportional data with extra zeros. The zero-inflated beta-binomial (ZIBB) models were also applied by Deng and Paul [7]. Moreover, the score tests for zero-inflation in a generalized linear model were studied by Broek [3] and Deng and Paul [6,7]. Hall and Berenhaut [10] developed the score test for heterogeneity and over-dispersion in zero-inflated Poisson and binomial regression models. Jansakul and Hinde [12], Ridout et al. [21], and Min and Gzado [19] also compared the assessing power for testing zero-inflation among the likelihood ratio test, the Wald test and the score test. Most recently, Song [22] established simultaneous statistical modelling of excess zeros, over/underdispersion, and multimodality. Alevizakos and Koukouvinos [1] used the zero-inflated binomial processes with a double exponentially weighted moving average statistic to monitor quality characteristics of high-yield processes. In such processes where a large number of zero observations exists in proportional data, the ZIB models are more appropriate than the ordinary binomial models. Furthermore, Alqawba and Diawara [2] proposed a Markov zero-inflated count time series models based on a joint distribution through copula functions.

As previously stated, most available studies in the area of zero-inflation are concentrated on univariate distributions. As the more complex data frequently arose in many subjects, statisticians have extended univariate distribution to their multivariate analogues (e.g. Fang et al. [8]). Johnson and Kotz [13] introduced multivariate Poisson distribution for modelling several types of defects. Li et al. [17] studied several possible ways to construct multivariate zero-inflated Poisson (MZIP) distribution. Liu and Tian [18] purposed Type I MZIP distribution with comparison to the MZIP distribution in Li et al. [17]. On the other hand, the multivariate binomial distribution was studied by Krishnamoorthy [14]. Chandrasekar and Balakrishnan [4] obtained some properties and a characterization of multivariate binomial distribution. Furthermore, similar to the univariate case, excessive zeros are not unusual to be expected in multivariate correlated proportional data and the univariate ZIB model is typically not sufficient for modelling such data. In order to fix the over-dispersion problem, fit the multivariate proportional data well, as well as have more accurate results, a new distribution called the ‘multivariate zero-inflated binomial (MZIB) distribution’ is proposed in this paper. Such a distribution is developed along the approach of symmetric multivariate distributions by Fang et al. [8] and based on the stochastic representation of the univariate ZIB random variable. The random variable with this new multivariate zero-inflated binomial distribution is assumed to be a $q$-dimensional response vector and is generated by a mixture of a common degenerate distribution with a unit mass point at zero in $\mathbb{R}^q$ and $q$ independent binomial distributions. Further, the correlations among the components of multivariate zero-inflated binomial variable are intuitively addressed although these binomial components are independent. Moreover, different from the random effects ZIB model, our proposed model can give the explicit expression for the correlation coefficients among the components of multivariate zero-inflated binomial variable.

The remainder of this paper is organized as follows. In Section 2, we propose a multivariate zero-inflated binomial distribution, which is inspired by good distributional properties of Type I MZIP distribution of Liu and Tian [18] and driven by the stochastic representation of univariate ZIB random variable. We then obtain joint probability mass function,
joint cumulative distribution function, and mixed moments of the MZIB distribution. The likelihood-based statistical inference about parameters of interest is performed in Section 3. Moreover, the Fisher scoring algorithm and EM algorithm are given for the computation of estimates of parameters on the proposed model with/without covariates. The score tests and the likelihood ratio tests are also developed for assessing the zero-inflation and the equality of all binomial probabilities in this section. In Section 4, simulation studies are performed to evaluate the performance of proposed score tests and the likelihood ratio tests in terms of nominal levels and powers, and of EM algorithm for the computation of estimates of parameters for the proposed MZIB model with/without covariates. The whitefly data is analysed as an application of the proposed methodology in Section 5 with the discussion in Section 6.

2. A multivariate zero-inflated binomial distribution

Let $Z \sim \text{Bernoulli}(1 - \omega)$, $X \sim \text{Binomial}(m, \pi)$, $Z$ and $X$ be independent (denoted by $Z \perp X$). Define a random variable $Y = ZX$, then, $Y$ follows the univariate zero-inflated binomial (ZIB) distribution, denoted by $Y \sim \text{ZIB}(\omega, \pi)$. By virtue of the stochastic representation of the univariate ZIP random variable, it could be naturally extended to a multivariate version. In what follows, we give the definition of the multivariate ZIB distribution, which has the vector form of correlation structure with a common Bernoulli variable $Z$.

**Definition 2.1:** A $q$-dimensional discrete random vector $Y = (Y_1, \ldots, Y_q)^\top$ bounded by a given upper vector $m = (m_1, \ldots, m_q)^\top$ is said to follow a multivariate ZIB distribution with parameters $\omega \in [0, 1)$ and $\pi = (\pi_1, \ldots, \pi_q)^\top \in (0, 1)^q$ if

$$Y \overset{d}{=} ZX = \begin{cases} 0_q & \text{with probability } \omega, \\ X & \text{with probability } 1 - \omega, \end{cases}$$

where $Z \sim \text{Bernoulli}(1 - \omega), X = (X_1, \ldots, X_q)^\top, X_r \sim \text{Binomial}(m_r, \pi_r)$ for $r = 1, \ldots, q$, and $(Z, X_1, \ldots, X_q)$ are mutually independent. The multivariate ZIB distribution is denoted by $Y \sim \text{MZIB}(\omega; m_1, \ldots, m_q; \pi_1, \ldots, \pi_q)$ or $Y \sim \text{MZIB}_q(\omega, m, \pi)$, where $X$ is called the base vector of $Y$.

From the stochastic representation (1), the joint probability mass function of $Y \sim \text{ZIB}_q(\omega, m, \pi)$ can be expressed as

$$f(y | \omega, m, \pi) = \begin{cases} \omega + (1 - \omega) \prod_{r=1}^{q} (1 - \pi_r)^{m_r} & \text{if } y = 0, \\ (1 - \omega) \prod_{r=1}^{q} \left( \frac{m_r}{y_r} \right)^{y_r} (1 - \pi_r)^{m_r - y_r} & \text{if } y \neq 0 \end{cases}$$

$$= \omega \Pr(\xi = y) + (1 - \omega) \Pr(X = y),$$

(2)
where $\xi$ has the degenerate distribution at mass $0$. The corresponding joint cumulative distribution function is given by

$$\Pr(Y \leq y) = \omega + (1 - \omega) \prod_{r=1}^{q} \sum_{s=0}^{\lfloor y_r \rfloor} \binom{m_r}{s} \pi_r^s (1 - \pi_r)^{m_r - s}$$

$$= \omega + (1 - \omega) \prod_{r=1}^{q} B_{1-\pi_r}(m_r - y_r, y_r + 1),$$

where $y = (y_1, \ldots, y_q)^\top$ is a non-negative real vector in $\mathbb{R}^q$, $\lfloor y_r \rfloor$ is the ‘floor’ function of $y_r$, denoting the largest integer less than or equal to $y_r$ and

$$B_{1-\pi_r}(m_r - y_r, y_r + 1) = (m_r - y_r) \left( \begin{array}{c} m_r \\ y_r \end{array} \right) \int_0^{1-\pi_r} t^{m_r-y_r-1} (1-t)^{y_r} \, dt$$

is the regularized incomplete beta function.

Note that although the definition of MZIB distribution based on the stochastic representation has the advantage in the derivation of its properties (see below), the limitation of this definition is that the zero-inflated parameter $\omega$ should be in the interval $[0,1]$. However, it is not necessary to assume $\omega \in [0,1]$. Therefore, we can define the multivariate ZIB distribution based on the probability mass function as follows.

**Definition 2.2:** A $q$-dimensional discrete random vector $Y = (Y_1, \ldots, Y_q)^\top$ bounded by a given upper vector $m = (m_1, \ldots, m_q)^\top$ is said to follow a MZIB distribution with parameters $\omega$ and $\pi = (\pi_1, \ldots, \pi_q)^\top \in (0,1)^q$ if its probability mass function has the form of (2).

From this definition, one can find that it is possible to take $\omega$ less than zero, provided that

$$\omega > -\frac{\prod_{r=1}^{q} (1 - \pi_r)^{m_r}}{1 - \prod_{r=1}^{q} (1 - \pi_r)^{m_r}}$$

with equality for zero-truncation. It is zero-deflation if $\omega$ is negative. Also, the value of $\omega$ should be less than one. (This distribution would degenerate to zero if $\omega = 1$.)

It should be pointed out that when $\omega < 0$, the stochastic representation given in (1) does not hold. Since zero-deflation ($\omega < 0$) seldom happens in proportional data and the current paper concentrates on assessing the zero-inflation in multivariate proportional data, we mainly consider the case $\omega \geq 0$, where (1) can always be used to investigate some properties of MZIB distribution and to make the statistical inference for MZIB model. We now derive the expressions for the moments using this representation. We note that the resulting formulas continue to hold for the case $\omega < 0$, which can be checked using the probabilistic representation (2) for MZIB distribution. Now from (1), the mixed moments
for $Y \sim \text{MZIB}_q(\omega, \boldsymbol{m}, \boldsymbol{\pi})$ can be obtained as follows:

$$E\left(\prod_{r=1}^{q} Y_r^t\right) = (1 - \omega) \prod_{r=1}^{q} E(X_r^t),$$

where $t_1, \ldots, t_q \geq 0$. Also, by setting $\boldsymbol{m\pi} = (m_1\pi_1, \ldots, m_q\pi_q)^\top$, we have

$$\begin{cases}
E(Y) = (1 - \omega)\boldsymbol{m\pi}, \\
E(YY^\top) = (1 - \omega)[\text{diag}(\boldsymbol{m\pi}(1 - \boldsymbol{\pi})) + \boldsymbol{m\pi}(\boldsymbol{m\pi})^\top], \\
\text{cov}(Y) = (1 - \omega)[\text{diag}(\boldsymbol{m\pi}(1 - \boldsymbol{\pi})) + \omega\boldsymbol{m\pi}(\boldsymbol{m\pi})^\top].
\end{cases}$$

Therefore,

$$\text{var}(Y_r) = (1 - \omega)m_r\pi_r[(1 - \pi_r) + \omega m_r\pi_r],$$

$$\text{cov}(Y_r, Y_s) = (1 - \omega)\omega m_r\pi_r m_s\pi_s$$

and the correlation coefficient for $Y_r$ and $Y_s$ is

$$\rho_{rs} \equiv \text{cor}(Y_r, Y_s) = \frac{\omega \sqrt{m_r\pi_r m_s\pi_s}}{\sqrt{(\omega m_r\pi_r + (1 - \pi_r))(\omega m_s\pi_s + (1 - \pi_s))}}, \quad r \neq s. \quad (3)$$

In particular, when $m_r = m_s = m$ and $\pi_r = \pi_s = \pi$, we obtain

$$\rho_{rs} \equiv \text{cor}(Y_r, Y_s) = \frac{\omega m\pi}{\omega m\pi + (1 - \pi)}, \quad r \neq s.$$

It is worthy of note that our proposed model can address the correlations among all components of multivariate zero-inflated binomial variable and give an explicit expressions of correlation coefficients (see (3)). Furthermore, from the expression of the correlation coefficient, one can see that there exist positive (negative) correlations among the components of multivariate zero-inflated binomial variable $Y$ if the parameter $\omega$ is greater (less) than zero although the components of base variable $X$ are independent. The correlation is induced by the imposition of same zero mass probability.

3. Likelihood based inferences for MZIB model

In this section, we consider the statistical inferences for MZIB model. Since the current research focuses on the statistical inference for the zero-inflation of multivariate proportional data with excessive zeros, it is assumed that the value of zero-inflation parameter $\omega$ be in the unit interval $[0, 1)$ and all statistical inferences on MZIB model be based on the stochastic representation (1) in the following sequels.

Let $Y_1, \ldots, Y_n$ be independent random vectors and $Y_i = (Y_{i1}, \ldots, Y_{iq})^\top, \quad i = 1, 2, \ldots, n$, follow the $q$-dimensional ZIB distribution ZIB$_q(\omega_i, \boldsymbol{m_i}, \boldsymbol{\pi_i})$, where for $i = 1, \ldots, n$, $\boldsymbol{m_i} = (m_{i1}, \ldots, m_{iq})^\top$ are the known vectors of binomial denominators, $\boldsymbol{\pi_i} = (\pi_{i1}, \ldots, \pi_{iq})$ are the unknown vectors of binomial probabilities and $\omega_i$ are the unknown zero-inflated parameters. Now suppose $y_i = (y_{i1}, \ldots, y_{iq})^\top$ is the realization of the random vector $Y_i$, then the observed data and associated binomial denominators would be represented by $y_{\text{obs}} = \{y_1, \ldots, y_n\}$ and $\boldsymbol{m_{\text{obs}}} = \{\boldsymbol{m_1}, \ldots, \boldsymbol{m_n}\}$. Furthermore, for convenience, let
\( y_{ir} = \sum_{i=1}^{n} y_{ir}, m_{ir} = \sum_{i=1}^{n} m_{ir} \) for \( r = 1, \ldots, q \) and \( y_{ir} = \sum_{i=1}^{q} y_{ir}, m_{ir} = \sum_{i=1}^{q} m_{ir} \) for \( i = 1, \ldots, n \). Based on the joint probability mass function of \( Y \) given by (2), the likelihood function for the parameters \((\omega, \pi) = (\omega_1, \ldots, \omega_n, \pi_1, \ldots, \pi_n)\) can be obtained as

\[
L(\omega, \pi \mid y_{\text{obs}}) = \prod_{i=1}^{n} \left\{ \left[ \omega_i + (1 - \omega_i) \prod_{r=1}^{q} (1 - \pi_{ir})^{m_{ir}} \right] I(y_{ir} = 0) \right. \\
\times \left. \left( 1 - \omega_i \right) \prod_{r=1}^{q} \left( \frac{m_{ir}}{y_{ir}} \right)^{y_{ir}} (1 - \pi_{ir})^{m_{ir} - y_{ir}} \right] I(y_{ir} \neq 0) \right\}
\]

By reparameterization, let \( \gamma_i = \omega_i/(1 - \omega_i) \). Then the likelihood function for \((\gamma, \pi) = (\gamma_1, \ldots, \gamma_n, \pi_1, \ldots, \pi_n)\) is

\[
L(\gamma, \pi \mid y_{\text{obs}}) = \prod_{i=1}^{n} \left\{ \left[ \frac{\gamma_i}{1 + \gamma_i} + \left( 1 - \frac{\gamma_i}{1 + \gamma_i} \right) \prod_{r=1}^{q} (1 - \pi_{ir})^{m_{ir}} \right] I(y_{ir} = 0) \right. \\
\times \left. \left( 1 - \frac{\gamma_i}{1 + \gamma_i} \right) \prod_{r=1}^{q} \left( \frac{m_{ir}}{y_{ir}} \right)^{y_{ir}} (1 - \pi_{ir})^{m_{ir} - y_{ir}} \right] I(y_{ir} \neq 0) \right\}
\]

so that the log-likelihood function is

\[
\ell(\gamma, \pi \mid y_{\text{obs}}) = -\sum_{i=1}^{n} \ln(1 + \gamma_i) + \sum_{i=1}^{n} \ln \left[ \gamma_i + \prod_{r=1}^{q} (1 - \pi_{ir})^{m_{ir}} \right] I(y_{ir} = 0) \\
+ \sum_{i=1}^{n} \sum_{r=1}^{q} \left[ \ln \left( \frac{m_{ir}}{y_{ir}} \right) + y_{ir} \ln \left( \frac{\pi_{ir}}{1 - \pi_{ir}} \right) + m_{ir} \ln(1 - \pi_{ir}) \right] I(y_{ir} \neq 0).
\]

### 3.1. MLEs of parameter for MZIB model without covariates

Based on the discussion above, we first derive the maximum likelihood estimates of parameters for MZIB model without covariates. In this case, the zero-inflated parameters \( \gamma_i \) and the probabilities \( \pi_i \) are held fixed as \( \gamma \) and \( \pi \). Hence the log-likelihood (5) can be simplified as

\[
\ell(\gamma, \pi \mid y_{\text{obs}}) = -n \ln(1 + \gamma) + \sum_{i=1}^{n} \ln \left[ \gamma + \prod_{r=1}^{q} (1 - \pi_r)^{m_{ir}} \right] I(y_i = 0) \\
+ \sum_{i=1}^{n} \sum_{r=1}^{q} \left[ \ln \left( \frac{m_{ir}}{y_{ir}} \right) + y_{ir} \ln \left( \frac{\pi_r}{1 - \pi_r} \right) + m_{ir} \ln(1 - \pi_r) \right] I(y_i \neq 0). 
\]

### 3.1.1. MLEs of parameters via Fisher scoring algorithm

In this subsection, the Fisher scoring algorithm is derived to calculate the MLEs of the parameters \( \theta = (\gamma, \pi_1, \ldots, \pi_q)^T = (\gamma, \pi^T)^T \), where \( \gamma = \omega/(1 - \omega), \pi = (\pi_1, \ldots, \pi_q)^T \).
The Fisher scoring algorithm is a common method to calculate maximum likelihood estimation. Comparing with EM algorithm, it could have better stability even in multi-parameter cases. In addition, the expected Fisher information matrix should always be positively definite, when the model is not over-parameterized (Lauritzen [16]). However, the Fisher scoring algorithm requires more complex calculation than EM algorithm for deriving the expected Fisher information matrix. Moreover, the expected Fisher information matrix could not be tractable for the complicated models. Since the estimation under multivariate zero-inflated binomial distribution is multi-parameter case, the Fisher scoring algorithm should be studied. Now, based on the equation (6), the score vector is

$$\nabla \ell(\theta | y_{\text{obs}}) = \left( \frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \gamma}, \frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_1}, \ldots, \frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_q} \right)^T,$$

where

$$\frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \gamma} = \frac{-n}{1 + \gamma} + \frac{I(y_i = 0)}{\gamma + \prod_{s=1}^q (1 - \pi_s)^{m_{is}}}.$$

$$\frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_r} = \sum_{i=1}^n \left\{ \frac{m_{ir} I(y_i = 0)}{(1 - \pi_r)(\gamma + \prod_{s=1}^q (1 - \pi_s)^{m_{is}})} + \frac{y_{ir} - m_{ir} \pi_r}{\pi_r (1 - \pi_r)} \right\}$$

for $r = 1, \ldots, q$. In order to apply Fisher scoring algorithm, the Hessian matrix should be obtained first as follows:

$$\nabla^2 \ell(\theta | y_{\text{obs}}) = \nabla^2 \ell(\gamma, \pi | y_{\text{obs}})$$

$$= \begin{pmatrix}
\frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \gamma^2} & \frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \gamma \partial \pi_1} & \cdots & \frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \gamma \partial \pi_q} \\
\frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_1 \partial \gamma} & \frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_1 \partial \pi_1} & \cdots & \frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_1 \partial \pi_q} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_q \partial \gamma} & \frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_q \partial \pi_1} & \cdots & \frac{\partial^2 \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_q \partial \pi_q} 
\end{pmatrix}.$$  

Then, the Fisher information matrix $J(\theta) = -E\nabla^2 \ell(\gamma, \pi | y_{\text{obs}})$ is

$$J(\theta) = J(\gamma, \pi) = \begin{pmatrix} J_{\gamma \gamma} & J_{\gamma \pi} \\ J_{\pi \gamma} & J_{\pi \pi} \end{pmatrix}.$$  

The derivation for the formulas of $J_{\gamma \gamma}, J_{\gamma \pi}, J_{\pi \gamma}$ and $J_{\pi \pi}$ is given in the supplemental file.

Now let $\theta^{(0)}$ be the initial values of the MLEs $\hat{\theta}$ of $\theta = (\gamma, \pi)$. If $\theta^{(t)}$ denote the $t$th approximation of $\hat{\theta}$, then the $(t + 1)$th approximation can be computed by is

$$\theta^{(t+1)} = \theta^{(t)} + J^{-1}(\theta^{(t)}) \nabla \ell(\theta^{(t)} | y_{\text{obs}}).$$

### 3.1.2. MLEs of parameters via the EM algorithm

In this subsection, we will develop the EM algorithm to compute the MLEs of parameters in the proposed MZIB model. Although the Fisher scoring algorithm possesses quadratic
convergence, it may not guarantee the MLEs of \( \omega \) and \( \pi_r, r = 1, 2, \ldots, q \) to be included in the unit interval \([0, 1]\). When the initial value \( (\gamma^{(0)}, \pi^{(0)}) \) of Fisher scoring algorithm is sufficiently near \((\hat{\gamma}, \hat{\pi})\), they converge very fast. However, it is sensitive to initial values under MZIB distribution. When the chosen initial value of \( (\gamma^{(0)}, \pi^{(0)}) \) is far from \((\hat{\gamma}, \hat{\pi})\), they might not converge. Therefore, the expectation-maximization (EM) algorithm is given for the calculation of MLEs in the MZIB model.

The EM algorithm is a popular tool for estimating maximum likelihood estimation in joint statistical models by iterating between E-step and M-step. The E-step represents the expectation of the log-likelihood. The M-step computes parameters maximizing the expected log-likelihood found on the E-step. Then, the unobserved latent variable is determined by these estimated parameters in the next E-step.

For each \( y_i = (y_{i1}, \ldots, y_{iq})^\top \) with \( i = 1, 2, \ldots, n \), based on (1) we introduce independent latent variables

\[
Z_i \sim \text{Bernoulli}(1 - \omega), \quad X_{ir} \sim \text{bin}(m_{ir}, \pi_r)
\]

for \( r = 1, 2, \ldots, q \). We denote the latent/missing data by \( Y_{\text{mis}} = \{z_i, (x_{ir})_{r=1}^q\}_{i=1}^n \) and the complete data by \( Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\} = Y_{\text{mis}} \) where \( z_i, x_{ir} \) are the realizations of \( Z_i \) and \( X_{ir} \), respectively. Thus, the complete-data likelihood function is given by

\[
L(\theta \mid Y_{\text{com}}) = \prod_{i=1}^n \left\{ \omega^{1-z_i}(1 - \omega)^{z_i} \prod_{r=1}^q \left[ \left( \frac{m_{ir}}{x_{ir}} \right)^{x_{ir}} \left( 1 - \pi_r \right)^{m_{ir}-x_{ir}} \right] \right\}
\]

and the complete-data log-likelihood function is proportional to

\[
\ell(\theta \mid Y_{\text{com}}) \propto \sum_{i=1}^n [(1 - z_i) \log \omega + z_i \log(1 - \omega)] + \sum_{i=1}^n \sum_{r=1}^q \left[ x_{ir} \log \frac{\pi_r}{1 - \pi_r} + m_{ir} \log(1 - \pi_r) \right].
\]

The M-step is to calculate the complete-data MLEs, which are given by

\[
\omega = 1 - \frac{1}{n} \sum_{i=1}^n z_i, \quad \pi_r = \frac{\sum_{i=1}^n x_{ir} / \sum_{i=1}^n m_{ir}}{m_r}
\]

for \( r = 1, 2, \ldots, q \). The E-step is to replace \( \sum_{i=1}^n z_i, \sum_{i=1}^n x_{ir} = x_r \) in (12) and (13) with their conditional expectations:

\[
E \left( \sum_{i=1}^n Z_i \mid Y_{\text{obs}}, \theta \right) = n - \omega \sum_{i \in J} \frac{1}{a_i}
\]

and

\[
E \left( \sum_{i=1}^n X_{ir} \mid Y_{\text{obs}}, \theta \right) = y_r + \sum_{i \in J} \frac{\omega}{a_i} m_{ir} \pi_r,
\]

where \( a_i \) is the number of observations in the \( i \)-th group.
where \( J = \{ i : y_i = 0 \} \) and

\[
a_i = \omega + (1 - \omega) \prod_{r=1}^{q} (1 - \pi_r)^{m_{ir}}, \quad i = 1, 2, \ldots, n.
\]

The detail for deriving (14) and (15) is given in A.2 of supplemental file.

Note that the latent variables \( \{ Z_i \}_{i=1}^{n} \) and \( \{ \{ X_{ir}^{q} \}_{r=1}^{q} \}_{i=1}^{n} \) introduced in (11) are independent Bernoulli random variables and binomial random variables, respectively. Thus, the left-hand side of (14) must be less than or equal to \( n \), and the left-hand side in (15) must be between 0 and \( \sum_{i=1}^{n} m_{ir} = m_r \). In other words, the EM algorithm (12)–(15) can guarantee that the MLEs of \( \{ \omega, \{ \pi_r \}_{r=1}^{q} \} \) fall within the unit interval \([0, 1]\), resulting in a clear statistical interpretation for these parameters in the distribution. This is advantage of the EM algorithm over the Fisher scoring algorithm. However, it is worthy of note that the EM algorithm is based on the stochastic representation (1), which intuitively assumes that the zero-inflated parameter \( \omega \) is in the unite interval \([0, 1] \). It does not work for the case of zero-deflation.

Now let \( (\tilde{\omega}, \tilde{\pi}_1, \ldots, \tilde{\pi}_q) \) denote the MLEs of \( (\omega, \pi_1, \ldots, \pi_q) \) obtained via the EM algorithm (12)–(15). Actually, based on the square root of the diagonal elements of the statistical interpretation for these parameters in the distribution. This is advantage of the EM algorithm over the Fisher scoring algorithm. However, it is worthy of note that the EM algorithm is based on the stochastic representation (1), which intuitively assumes that the zero-inflated parameter \( \omega \) is in the unite interval \([0, 1] \). It does not work for the case of zero-deflation.

Now let \( (\tilde{\omega}, \tilde{\pi}_1, \ldots, \tilde{\pi}_q) \) denote the MLEs of \( (\omega, \pi_1, \ldots, \pi_q) \) obtained via the EM algorithm (12)–(15). Actually, based on the square root of the diagonal elements of the estimated inverse Fisher information matrix \( \{ F(\tilde{\omega}, \tilde{\pi}_1, \ldots, \tilde{\pi}_q) \}^{-1} \), the Wald-type confidence intervals for the parameters can be obtained. However, the zero-inflated parameter \( \omega \) and the binomial probabilities \( \pi_1, \pi_2, \ldots, \pi_q \) should be restricted within the unit interval and thus some upper (or lower) limits of these confidence intervals may be larger (or less) than 1 (or 0), resulting in useless confidence intervals. Instead of the Wald-type methods, the bootstrap approach can be used to compute the bootstrap confidence interval for any component of \( (\omega, \pi_1, \ldots, \pi_q) \). At first, the independent sample \( y_1^*, y_2^*, \ldots, y_n^* \) from the distribution \( \text{MZIB}_q(\tilde{\omega}, m, \tilde{\pi}) \) can be generated, where \( \tilde{\omega} \) and \( \tilde{\pi} \) are the MLEs of \( \omega \) and \( \pi \) based on the original sample. Then, based on the generated sample \( y_1^*, y_2^*, \ldots, y_n^* \), the MLE \( (\hat{\omega}^*, \hat{\pi}_1^*, \ldots, \hat{\pi}_p^*) \) of \( (\omega, \pi_1, \ldots, \pi_p) \) can be calculated. Independently repeating this procedure \( G \) times, the \( G \) MLE’s \( \{(\hat{\omega}^*_g, \hat{\pi}_{1g}^*, \ldots, \hat{\pi}_{pg}^*)\}_{g=1}^{G} \) of \( (\omega, \pi_1, \ldots, \pi_p) \) can be obtained and thus the confidence intervals of \( (\omega, \theta_1, \ldots, \theta_p) \) can be constructed by \( [\omega_L, \omega_U], [\pi_{1L}, \pi_{1U}], \ldots, [\pi_{qL}, \pi_{qU}] \), where \( \omega_L, \pi_{1L}, \ldots, \pi_{qL} \) and \( \omega_U, \pi_{1U}, \ldots, \pi_{qU} \) are the 100(\( \alpha / 2 \))% and 100(1 – \( \alpha / 2 \))% percentiles of \( \{(\hat{\omega}^*_g, \hat{\pi}_{1g}^*, \ldots, \hat{\pi}_{pg}^*)\}_{g=1}^{G} \), respectively.

### 3.2. MLEs of parameters for MZIB model with covariates

#### 3.2.1. The formulation of MZIB model with covariates

Again, let \( Y_1, \ldots, Y_n \) be independent random vectors and \( Y_i = (Y_{i1}, \ldots, Y_{iq})^T, \quad i = 1, 2, \ldots, n \), follow the \( q \)-dimensional ZIB distribution \( \text{MZIB}_q(\omega_i, m_i, \pi_i) \), where for \( i = 1, \ldots, n, \quad m_i = (m_{i1}, \ldots, m_{iq})^T \) are the known vectors of binomial denominators, \( \pi_i = (\pi_{i1}, \ldots, \pi_{iq}) \) are the unknown vectors of binomial probabilities and \( \omega_i \) are the unknown zero-inflated parameters. Further, let \( v_i \) and \( u_i \) be the covariates associated with the zero-inflated parameters \( \omega_i \) and binomial probabilities \( \pi_i = (\pi_{i1}, \ldots, \pi_{iq}) \) \( (i = 1, 2, \ldots, n) \), respectively. Now suppose \( Y_i = (y_{i1}, \ldots, y_{iq})^T \) is the realization of the random vector \( Y_i \), then the observed data and associated binomial denominators would be represented by \( y_{obs} = \{ y_1, \ldots, y_n \} \) and \( m_{obs} = \{ m_1, \ldots, m_n \} \). To investigate the relationship between the
parameters \((\omega, \pi)\) and covariates \(v_i\) and \(u_i\), we consider the following regression model:

\[
\begin{align*}
&\begin{cases}
   y_i = (y_{i1}, \ldots, y_{iq})^\top \sim \text{MZIB}(\omega_i, \pi_i, m_i), \quad i = 1, 2, \ldots, n, \\
   \log \left( \frac{\omega_i}{1 - \omega_i} \right) = v_i^\top \alpha, \\
   \log \left( \frac{\pi_{ir}}{1 - \pi_{ir}} \right) = u_i^\top \beta_r, \quad r = 1, 2, \ldots, q,
\end{cases}
\end{align*}
\]

where \(v_i = (1, v_{i1}, \ldots, v_{pi})^\top\) and \(u_i = (1, u_{i1}, \ldots, u_{pi})^\top\) are not necessarily identical covariate vectors associated with the subject \(i\); \(\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)^\top\) and \(\beta_r = (\beta_{r0}, \beta_{r1}, \ldots, \beta_{rp})^\top\) are corresponding regression coefficients. The primary purpose of this section is to estimate the parameter vector \(\theta = (\alpha^\top, \beta_{1r}^\top, \ldots, \beta_{qr}^\top)^\top\).

### 3.2.2. MLEs via the EM algorithm embedded with Fisher scoring algorithms at each M-step

Now, the complete-data likelihood function in Section 3.1.2 now becomes

\[
L(\theta \mid Y_{\text{com}}) = \prod_{i=1}^{n} \left\{ \omega_i^{1-z_i}(1-\omega_i)^{z_i} \prod_{r=1}^{q} \left[ \left( m_{ir} \right)^{x_{ir}} \pi_{ir}^{x_{ir}} (1 - \pi_{ir})^{m_{ir} - x_{ir}} \right] \right\}
\]

and the complete-data log-likelihood function is proportional to

\[
\ell(\theta \mid Y_{\text{com}}) \propto \sum_{i=1}^{n} \left[ (1 - z_i) v_i^\top \alpha - \log(1 + e^{v_i^\top \alpha}) \right] + \sum_{i=1}^{n} \sum_{r=1}^{q} \left[ x_{ir} u_i^\top \beta_r - m_{ir} \log(1 + e^{u_i^\top \beta_r}) \right].
\]

The first and negative second partial derivatives of the complete-data log-likelihood function are given by

\[
\begin{align*}
&\frac{\partial \ell(\theta \mid Y_{\text{com}})}{\partial \alpha} = \sum_{i=1}^{n} (1 - z_i - \omega_i) v_i = v^\top (1 - z - \omega), \\
&\frac{\partial \ell(\theta \mid Y_{\text{com}})}{\partial \beta_r} = \sum_{i=1}^{n} (x_{ir} - m_{ir} \pi_{ir}) u_i = u^\top (x_r - m_r \pi_r), \\
&-\frac{\partial^2 \ell(\theta \mid Y_{\text{com}})}{\partial \alpha \partial \alpha^\top} = \sum_{i=1}^{n} \omega_i (1 - \omega_i) v_i v_i^\top = v^\top \text{diag}[\omega(1 - \omega)] v \triangleq J_{\text{com}}(\alpha), \\
&-\frac{\partial^2 \ell(\theta \mid Y_{\text{com}})}{\partial \beta_r \partial \beta_r^\top} = \sum_{i=1}^{n} m_{ir} \pi_{ir} (1 - \pi_{ir}) u_i u_i^\top = u^\top \text{diag}[m_r \pi_r (1 - \pi_r)] u \triangleq J_{\text{com}}(\beta_r),
\end{align*}
\]

where \(1 = (1, \ldots, 1)^\top, v = (v_1, \ldots, v_n)^\top, u = (u_1, \ldots, u_n)^\top, z = (z_1, \ldots, z_n)^\top, \omega = (\omega_1, \ldots, \omega_n)^\top, x_r = (x_{r1}, \ldots, x_{rn})^\top, m_r = (m_{r1}, \ldots, m_{rn})^\top, \pi_r = (\pi_{r1}, \ldots, \pi_{rn})^\top, \text{diag}[\omega(1 - \omega)] = \text{diag}[\omega_1 (1 - \omega_1), \ldots, \omega_n (1 - \omega_n)], \text{diag}[m_r \pi_r (1 - \pi_r)] = \text{diag}\).
In what follows, based on the likelihood methods we derive score test statistics.

### 3.3. Hypothesis testing in MZIB model without covariates

Now, the M-step is to separately calculate the MLEs of $\alpha$ and $\beta$ via two Fisher scoring algorithms as follows:

$$
\alpha^{(t+1)} = \alpha^{(t)} + J_{com}^{-1}(\alpha^{(t)})v^T (1 - z - \omega(\alpha^{(t)})),
$$

$$
\beta_r^{(t+1)} = \beta_r^{(t)} + J_{com}^{-1}(\beta_r^{(t)})u^T (x_r - m_r \pi_r(\beta_r^{(t)})), \quad r = 1, 2, \ldots, q.
$$

The E-step is to replace the latent variables $z, \{x_r\}_{r=1}^q$ in (16) and (17) by their conditional expectations:

$$
E(Z | Y_{obs}, \theta) = 1 - \text{diag} \left( \frac{\omega}{a} \right) I_0
$$

and

$$
E(X_r | Y_{obs}, \theta) = y_r + \text{diag} \left( \frac{\omega}{a} m_r \pi_r \right) I_0,
$$

where $y_r = (y_{1r}, \ldots, y_{nr})^\top$, $I_0 = (I[y_i = 0], \ldots, I[y_n = 0])^\top$, $\text{diag}(\omega_a) = \text{diag}(\frac{\omega_{a1}}{a_1}, \ldots, \frac{\omega_{an}}{a_n})$, $\text{diag}(\frac{\omega}{a} m_r \pi_r) = \text{diag}(\frac{\omega_{a1}}{a_1} m_{1r} \pi_{1r}, \ldots, \frac{\omega_{an}}{a_n} m_{nr} \pi_{nr})$ with

$$
a_i = a_i(\omega_i, \pi_i) = \omega_i + (1 - \omega_i) \prod_{r=1}^q (1 - \pi_{ir})^{m_{ir}}, \quad i = 1, 2, \ldots, n.
$$

Note that (18) and (19) can be derived in the same way as (14) and (15). Now, let $\hat{\alpha}$ and $\hat{\beta}_r, r = 1, \ldots, q$ are the estimates of the parameters $\alpha, \beta_r, r = 1, 2, \ldots, q$, respectively. Then the asymptotic covariance matrices for $\hat{\alpha}$ and $\hat{\beta}_r, r = 1, \ldots, q$ can be obtained as $\text{cov}(\hat{\alpha}) = J_{com}^{-1}(\hat{\alpha})$, $\text{cov}(\hat{\beta}_r) = J_{com}^{-1}(\hat{\beta}_r), r = 1, 2, \ldots, q$ and thus the corresponding confidence intervals for the components of $\theta$ can be constructed by using the Wald-type method.

#### 3.3.1. Tests for zero-inflation in MZIB model

We should first test the presence of zero-inflation for the multivariate binomial data before the multivariate zero-inflated binomial model is used to fit such data. Based on the score
test method, the test statistic for testing the hypotheses $H_0 : \gamma = 0$ versus $H_a : \gamma \neq 0$ is given by

$$S_1 = \frac{\left\{ \sum_{i=1}^{n} \left[ \prod_{r=1}^{q} \left( m_r \cdot y_{r,i} - y_{r,i} - 1 \right) \right] \right\}^2}{\sum_{i=1}^{n} \left( \prod_{r=1}^{q} \left( m_r \cdot y_{r,i} - y_{r,i} - 1 \right) \right) - \sum_{r=1}^{q} y_{r,i} m_r (m_r - y_{r,i})^{-1}}.$$  

(20)

The details of the derivation for $S_1$ are given in A.3 of supplemental file. Under the null hypothesis $H_0 : \gamma = 0$, the test statistic $S_1$ has an approximately $\chi^2$ distribution with one degree of freedom. The corresponding $p$-value is given by

$$p_1 = \Pr(S_1 > s_1 | H_0) = \Pr\{\chi^2(1) > s_1\},$$

where $s_1$ is the observed value for $S_1$. When $p_1 < \alpha$, we reject the null hypothesis $H_0$ at the $\alpha$ level of significance. Otherwise, we fail to reject $H_0$.

Now for the purpose of comparison, we also give the likelihood ratio test (LRT) for testing the zero-inflation in MZIB model. The LRT statistic has the following form:

$$S_2 = -2\{\ell(0, \hat{\pi}_0 | y_{obs}) - \ell(\hat{\gamma}, \hat{\pi} | y_{obs})\},$$

where $\hat{\pi}_0 = (y_{1} / m_{1}, \ldots, y_{q} / m_{q})^T$ are the MLE’s of $\pi$ under the null hypothesis and $(\hat{\gamma}, \hat{\pi})$ are the unconstrained MLE of $(\gamma, \pi)$, which can be obtained via the Fisher scoring algorithm or EM algorithm given in Sections 3.1.1 and 3.1.2.

Under $H_0$, the LRT statistic $S_2$ has an approximately chi-squared distribution with one degree of freedom and the corresponding $p$-value can be computed as

$$p_2 = \Pr(S_2 > s_2 | H_0) = \Pr\{\chi^2(1) > s_2\},$$

where $s_2$ is the observed value for $S_2$.

Note that the advantage of the score test is that the parameters are estimated only under the null hypothesis not the alternative hypothesis and thus the score test statistic $S_1$ has a closed form, which results in easy computation and application. However, as one can see from the simulation results in Section 4, the score test exhibits some limitation for the application even if the dimension of binomial random vector is moderate. Therefore, the score test is recommended for testing the zero-inflation in MZIB model only for the lower dimensions.

### 3.3.2. Tests for the equality of probabilities in MZIB model

From Section 2, we know that there exists a correlation between any two components for the multivariate zero-inflated binomial model. Another question of interest for the multivariate binomial model is whether all or part components of multivariate binomial distribution share the same binomial probability. Therefore, we develop an approach to testing the equality of probabilities for all/part components in the multivariate zero-inflated binomial model. The null and alternative hypotheses to be tested are

$$H_0 : \pi_1 = \cdots = \pi_q \text{ versus } H_a : \pi_r \neq \pi_s \text{ for at least one pair } r \neq s.$$  

(22)

Same as in Section 3.3.1, the score test statistic can be developed for testing the equality of probabilities for all components in the multivariate binomial variable. It has the following
form:

\[ S_3 = U^\top(\hat{\gamma}, \hat{\pi})\mathbb{I}^{-1}(\hat{\gamma}, \hat{\pi})U(\hat{\gamma}, \hat{\pi}), \]

where \( \hat{\gamma} \) and \( \hat{\pi} \) are the MLEs of \( \gamma \) and \( \pi \), respectively, under the null hypothesis \( H_0 : \pi_1 = \cdots = \pi_q \). \( \hat{\gamma} \) and \( \hat{\pi} \) can be obtained via the Fisher scoring algorithm from the following maximum likelihood equations under the null hypothesis:

\[
\frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \gamma} = \frac{-n}{1 + \gamma} + \sum_{i=1}^{n} \frac{I(y_i = 0)}{\gamma + (1 - \pi)^{m_i}} = 0, \tag{23}
\]

\[
\frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi} = \sum_{i=1}^{n} \left\{ \frac{m_i \gamma I(y_i = 0)}{(1 - \pi)(\gamma + (1 - \pi)^{m_i})} + \frac{y_i - m_i \pi}{\pi(1 - \pi)} \right\} = 0 \tag{24}
\]

or via the EM algorithm which can be derived in the similar way given in Section 3.1.2.

Furthermore, the score function \( U(\gamma, \pi) \) of \( (\gamma, \pi) \) has the form

\[
U(\gamma, \pi) = \left( \frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \gamma}, \frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_1}, \ldots, \frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_q} \right)^\top,
\]

with

\[
\frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \gamma} \bigg|_{\pi = \pi} = \frac{-n}{1 + \gamma} + \sum_{i=1}^{n} \frac{I(y_i = 0)}{\gamma + (1 - \pi)^{m_i}},
\]

\[
\frac{\partial \ell(\gamma, \pi | y_{\text{obs}})}{\partial \pi_r} \bigg|_{\pi = \pi} = \sum_{i=1}^{n} \left\{ \frac{m_i \gamma I(y_i = 0)}{(1 - \pi)(\gamma + (1 - \pi)^{m_i})} + \frac{y_i - m_i \pi}{\pi(1 - \pi)} \right\}
\]

for \( r = 1, \ldots, q \). The expected information matrix \( \mathbb{I}(\gamma, \pi) \) with parameters \( \gamma \) and \( \pi = (\pi_1, \ldots, \pi_q)^\top \) is

\[
\mathbb{I}(\gamma, \pi) = -E(\nabla^2 \ell(\theta | y_{\text{obs}}))
\]

\[
= \begin{pmatrix}
-\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \gamma^2} \right) & -\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \gamma \partial \pi_1} \right) & \cdots & -\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \gamma \partial \pi_q} \right) \\
-\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \pi_1 \partial \gamma} \right) & -\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \pi_1^2} \right) & \cdots & -\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \pi_1 \partial \pi_q} \right) \\
\vdots & \vdots & \ddots & \vdots \\
-\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \pi_q \partial \gamma} \right) & -\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \pi_q \partial \pi_1} \right) & \cdots & -\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \pi_q^2} \right)
\end{pmatrix},
\]
where under the null hypothesis \( \pi_1 = \pi_2 = \cdots = \pi_q = \pi \),

\[
-\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \gamma^2} \right) |_{\pi_s=\pi} = \sum_{i=1}^{n} \left( \frac{1 - (1 - \pi)^{m_i}}{(\gamma + (1 - \pi)^{m_i})(1 + \gamma)^2} \right),
\]

\[
-\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \gamma \partial \pi_r} \right) |_{\pi_s=\pi} = \sum_{i=1}^{n} \frac{m_{ir}(1 - \pi)^{m_i - 1}}{(\gamma + (1 - \pi)^{m_i})(1 + \gamma)},
\]

\[
-\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \pi_r^2} \right) |_{\pi_s=\pi} = \sum_{i=1}^{n} \left( \frac{m_{ir}}{\pi(1 - \pi)(1 + \gamma)} - \frac{\gamma^2 m_{ir}^2(1 - \pi)^{m_i - 2}}{(\gamma + (1 - \pi)^{m_i})(1 + \gamma)} \right)
\]

and

\[
-\mathbb{E} \left( \frac{\partial^2 \ell}{\partial \pi_r \partial \pi_s} \right) |_{\pi_s=\pi} = -\sum_{i=1}^{n} \frac{m_{ir}m_{is}(1 - \pi)^{m_i - 2}\gamma}{(\gamma + (1 - \pi)^{m_i})(1 + \gamma)}
\]

and \( U(\hat{\gamma}, \hat{\pi}) \) and \( \mathbb{I}(\hat{\gamma}, \hat{\pi}) \) are the estimated values of the score function \( U(\gamma, \pi) \) and the expected information matrix \( \mathbb{I}(\gamma, \pi) \) at \( \gamma = \hat{\gamma} \) and \( \pi = \hat{\pi} = \hat{\pi} \cdot 1_q \), respectively.

Under the null hypothesis, the score statistic \( S_3 \) has an approximately \( \chi^2 \) distribution with \( q - 1 \) degrees of freedom and the corresponding \( p \)-value can be computed as

\[
p_3 = \text{Pr}(S_3 > s_3 \mid H_0) = \text{Pr}(\chi^2(q - 1) > s_3),
\]

where \( s_3 \) is the observed value for \( S_3 \). Similar to Section 3.3.1, the likelihood ratio method can also be used to test the equality of parameters in MZIB model. The LRT statistic is

\[
S_4 = -2 \{ \ell(\hat{\gamma}_0, \hat{\pi}_0 \mid y_{\text{obs}}) - \ell(\hat{\gamma}, \hat{\pi} \mid y_{\text{obs}}) \}, \tag{25}
\]

where \( \hat{\gamma}_0, \hat{\pi}_0 \) are the estimators of parameters \( \gamma, \pi \) under the null hypothesis and \( \hat{\gamma}, \hat{\pi} \) are the estimates of parameters \( \gamma, \pi \) under the alternative hypothesis. Although the MLEs for the parameters \( \gamma \) and \( \pi \) under both hypotheses do not have closed forms, \( \hat{\gamma}_0, \hat{\pi}_0 \) can be computed in the same way given in the derivation of \( S_3 \) and \( \hat{\gamma}, \hat{\pi} \) can be calculated via the Fisher scoring algorithm or EM algorithms given in Sections 3.1.1 and 3.1.2 under the alternative hypothesis. Further, under the null hypothesis, the LRT statistic \( S_4 \) in (25) follows an approximately \( \chi^2 \) distribution with \( q - 1 \) degrees of freedom. The corresponding \( p \)-value is given by

\[
p_4 = \text{Pr}(S_4 > s_4 \mid H_0) = \text{Pr}(\chi^2(p - 1) > s_4),
\]

where \( s_4 \) is the observed value for \( S_4 \). Moreover, if the null hypothesis \( H_0 \) in (21) is rejected, then the following hypotheses could be tested:

\[
H_0^* : \pi_{k_1} = \pi_{k_2} = \cdots = \pi_{k_q^*} \quad \text{against} \quad H_1^* : \pi_{k_r} \neq \pi_{k_i}, \tag{26}
\]

for the \( k_1 \)th, \( k_2 \)th, \( \ldots \), \( k_q^* \)th components \( (q^* < q) \). The likelihood ratio test statistic for testing the hypotheses in (26) is given by

\[
S_4^* = -2 \{ \ell(\hat{\gamma}^*, \hat{\pi}^* \mid y_{\text{obs}}) - \ell(\hat{\gamma}^*, \hat{\pi}^* \mid y_{\text{obs}}) \}, \tag{27}
\]

where \( (\hat{\gamma}^*, \hat{\pi}^*) \) are the maximum likelihood estimates of \( (\gamma, \pi) \) under \( H_0^* : \pi_{k_1} = \pi_{k_2} = \cdots = \pi_{k_q^*} \) and can be computed via the Fisher scoring algorithm or EM algorithm. The
parameters \((\hat{\gamma}, \hat{\pi})\) are the unconstrained maximum likelihood estimates of \((\gamma, \pi)\) and can be computed via the same algorithms. Moreover, under the null hypothesis \(H_0^*\) the test statistic \(S_4^*\) has an approximately \(\chi^2\) distribution with \(q^* - 1\) degrees of freedom. The corresponding \(p\)-value is

\[
p_4^* = \Pr(S_4^* > s_4^* | H_0^*) = \Pr(\chi^2_{q^* - 1} > s_4^*),
\]

where \(s_4^*\) is the observed value for \(S_4^*\).

4. Simulation study

In this section, a limited simulation study is carried out to evaluate the performance of the proposed statistical methods in Section 3 for the multivariate ZIB distribution. We first examine the accuracy of point estimates and confidence interval estimates for different parameter settings in the proposed multivariate ZIB models with/without covariates via simulation studies. Next, we establish the validity of four proposed test statistics under a finite sample situation. In terms of the nominal levels and powers, the performance of score test statistics and LRT statistics for the presence of zero-inflation and the equality of probabilities for all components in the multivariate ZIB models are investigated. All simulation studies on test methods are based on the multivariate zero-inflated binomial distribution without regressors to keep the model simple and the study more focused.

4.1. Accuracy of point estimates and interval estimates for MZIB model without covariates

Note that the proposed \(q\)-dimensional multivariate ZIB distribution has \(q + 1\) parameters. We expect that the proposed distribution can yield better data fitting without sacrificing statistical accuracy too much. To evaluate the accuracy of point estimates and confidence intervals for zero-inflated parameter \(\omega\) and the probability parameters \(\pi_1, \ldots, \pi_q\) in the multivariate ZIB model without covariates, we consider five cases for the dimension: \(q = 2, 3, 4, 5\) and 6. The sample size is chosen as \(n = 50, 100\). Parameter configurations can be found in Table 1. First, the procedure for generating the random number \(\{y_i\}_{i=1}^n \overset{iid}{\sim} \text{ZIB}_q(\omega, m, \pi)\) is given as follows:

| Scenarios | \(q\) | \(\omega\) | \((\pi_1, m_1)\) | \((\pi_2, m_2)\) | \((\pi_3, m_3)\) | \((\pi_4, m_4)\) | \((\pi_5, m_5)\) | \((\pi_6, m_6)\) |
|-----------|------|----------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1         | 2    | 0.10     | (0.10, 10)      | (0.20, 6)       |                 |                 |                 |                 |
| 2         | 0.20 |          | (0.20, 6)       | (0.10, 10)      |                 |                 |                 |                 |
| 3         | 3    | 0.10     | (0.10, 8)       | (0.20, 6)       | (0.15, 10)      |                 |                 |                 |
| 4         | 0.20 |          | (0.20, 6)       | (0.25, 10)      | (0.30, 8)       |                 |                 |                 |
| 5         | 4    | 0.10     | (0.10, 10)      | (0.15, 6)       | (0.20, 12)      | (0.15, 8)       |                 |                 |
| 6         | 0.20 |          | (0.20, 8)       | (0.25, 10)      | (0.30, 12)      | (0.10, 6)       |                 |                 |
| 7         | 5    | 0.10     | (0.10, 6)       | (0.25, 12)      | (0.30, 8)       | (0.20, 9)       | (0.15, 11)      |                 |
| 8         | 0.20 |          | (0.25, 10)      | (0.35, 12)      | (0.20, 8)       | (0.15, 9)       | (0.30, 11)      |                 |
| 9         | 6    | 0.10     | (0.20, 10)      | (0.35, 6)       | (0.30, 12)      | (0.10, 8)       | (0.15, 9)       | (0.25, 11)      |
| 10        | 0.20 |          | (0.15, 6)       | (0.25, 8)       | (0.20, 12)      | (0.45, 11)      | (0.10, 10)      | (0.35, 9)       |
(a) Generate $z_1, \ldots, z_n \overset{iid}{\sim} \text{Bernoulli}(1 - \omega)$.

(b) Generate 

\[
\{x_{11}, \ldots, x_{n1}\} \overset{iid}{\sim} \text{Binomial}(m_1, \pi_1), \ldots, \{x_{1q}, \ldots, x_{nq}\} \overset{iid}{\sim} \text{Binomial}(m_q, \pi_q).
\]

(c) Let $x_i = (x_{i1}, \ldots, x_{iq})^\top$. Then $y_i = z_i x_i \sim \text{ZIB}_q(\omega, m, \pi)$ for $i = 1, \ldots, n$.

Calculate the MLEs from the generated sample via EM algorithm (12)–(15) and the 95% bootstrap confidence intervals with repeating times $G = 1000$ for the parameters $\omega, \pi_1, \pi_2, \ldots, \pi_q$. Next, the 1000 samples are independently generated and the corresponding 1000 EM MLEs and 1000 bootstrap confidence intervals of $\omega, \pi_1, \pi_2, \ldots, \pi_q$ are obtained. Further, in Table 2, MLE is the average of the 1000 estimates via the EM algorithm (12)–(15); width and CP of the confidence intervals are the average width and coverage proportion of 1000 bootstrap confidence intervals. As seen in Table 2, the bias are small and the MLE’s are very close to the corresponding true values of parameters and the coverage probabilities are all around 0.95, although the coverage probabilities of zero-inflated parameter $\omega$ is a little less than the nominal level for $q = 5$ and $6$ with $n = 50$. We also conducted the simulation study for moderate number of dimension (e.g. $q = 10$) (the results do not be reported here). The obtained estimates of parameters, width and CP of confidence intervals are consistently close to the true values of parameters and nominal confidence coefficient, which demonstrates the proposed EM algorithm has very good performance even for a moderate dimensional number of multivariate binomial data.

4.2. Accuracy of point estimates and interval estimates for MZIB model with covariates

In this subsection, we perform the limited simulation study to investigate the performance of proposed algorithm for the estimation of regression parameters $\alpha$ and $\beta_1, \ldots, \beta_q$ in the multivariate ZIB model with covariates. The dimension is selected as $q = 2, 3, 4$ and 5. The covariates for the regression model are selected as $u_1 \sim U[1, 3]$ and $u_2 \sim \text{Bin}(1, 0.5)$. The regression parameters are selected as $\alpha = (-2.2, 0.3, -1.0)$, $\beta_1 = (0.3, -0.7, 1.5)$, $\beta_2 = (0.5, -0.8, 2.0)$, $\beta_3 = (1.0, -0.6, 1.8)$, $\beta_4 = (1.2, -1.0, 1.3)$ and $\beta_5 = (1.4, -1.2, 0.7)$. Thus, the logistic models for the parameters $(\omega, \pi)$ and covariates $u_1, u_2$ with regression parameters $\alpha$ and $\beta_r (r = 1, 2, \ldots, q)$ with $q = 2, 3, 4, 5$ are as follows:

\[
\begin{align*}
\mathbf{y} &= (y_1, \ldots, y_q)^\top \sim \text{MZIB}(\omega, \pi, m), \\
\log \left( \frac{\omega}{1 - \omega} \right) &= \alpha_0 + \alpha_1 u_1 + \alpha_2 u_2, \\
\log \left( \frac{\pi_r}{1 - \pi_r} \right) &= \beta_{0r} + \beta_{1r} u_1 + \beta_{2r} u_2 \quad r = 1, 2, \ldots, q.
\end{align*}
\]

Now, based on the above model, the random sample $\mathbf{y}_i, i = 1, 2, \ldots, n$ can be generated as follow:

(a) Generate the covariates $u_{1i} \sim U[1, 3]$ and $u_{2i} \sim \text{Bin}(1, 0.5)$ for $i = 1, 2, \ldots, n$. 
Table 2. Biases of MLEs, widths and coverage probabilities of bootstrap confidence intervals for parameters with the number of dimension \( q = 2, 3, 4, 5 \) and 6.

| \( q \) | \( n \) | Parameter | Bias | Width | CP | Bias | Width | CP |
|---|---|---|---|---|---|---|---|---|
| 2 | 50 | \( \omega \) | -0.0004 | 0.1979 | 0.941 | -0.0027 | 0.2817 | 0.946 |
| 100 | \( \omega \) | -0.0012 | 0.1509 | 0.943 | 0.0004 | 0.2055 | 0.950 |
| | \( \pi_1 \) | -0.0006 | 0.0520 | 0.945 | -0.0003 | 0.0790 | 0.949 |
| | \( \pi_2 \) | -0.0007 | 0.0570 | 0.959 | 0.0003 | 0.0446 | 0.954 |
| 3 | 50 | \( \omega \) | 0.0023 | 0.1735 | 0.925 | -0.0005 | 0.2177 | 0.933 |
| | \( \pi_1 \) | 0.0001 | 0.0723 | 0.951 | 0.0001 | 0.0791 | 0.960 |
| | \( \pi_2 \) | 0.0014 | 0.0846 | 0.944 | -0.0016 | 0.1103 | 0.953 |
| | \( \pi_3 \) | -0.0001 | 0.0672 | 0.946 | 0.0002 | 0.1012 | 0.961 |
| 100 | \( \omega \) | -0.0003 | 0.1258 | 0.940 | -0.0025 | 0.1557 | 0.942 |
| | \( \pi_1 \) | -0.0002 | 0.0508 | 0.949 | -0.0011 | 0.0556 | 0.958 |
| | \( \pi_2 \) | -0.0002 | 0.0597 | 0.947 | 0.0007 | 0.0779 | 0.945 |
| | \( \pi_3 \) | 0.0001 | 0.0476 | 0.937 | 0.0003 | 0.0716 | 0.955 |
| 4 | 50 | \( \omega \) | -0.0030 | 0.1580 | 0.894 | 0.0023 | 0.2172 | 0.941 |
| | \( \pi_1 \) | -0.0003 | 0.0716 | 0.963 | -0.0009 | 0.0791 | 0.952 |
| | \( \pi_2 \) | -0.0009 | 0.0738 | 0.951 | -0.0010 | 0.1102 | 0.941 |
| | \( \pi_3 \) | 0.0006 | 0.0743 | 0.949 | -0.0001 | 0.0825 | 0.956 |
| | \( \pi_4 \) | 0.0008 | 0.0603 | 0.950 | 0.0000 | 0.0659 | 0.936 |
| 100 | \( \omega \) | -0.0008 | 0.1170 | 0.940 | -0.0012 | 0.1548 | 0.944 |
| | \( \pi_1 \) | 0.0005 | 0.0507 | 0.952 | 0.0003 | 0.0558 | 0.948 |
| | \( \pi_2 \) | -0.0009 | 0.0523 | 0.946 | 0.0001 | 0.0779 | 0.948 |
| | \( \pi_3 \) | 0.0007 | 0.0526 | 0.955 | -0.0010 | 0.0583 | 0.944 |
| | \( \pi_4 \) | 0.0005 | 0.0429 | 0.952 | -0.0001 | 0.0467 | 0.953 |
| 5 | 50 | \( \omega \) | 0.0003 | 0.1574 | 0.865 | 0.0014 | 0.2176 | 0.944 |
| | \( \pi_1 \) | 0.0000 | 0.0714 | 0.945 | 0.0002 | 0.0856 | 0.952 |
| | \( \pi_2 \) | 0.0004 | 0.0734 | 0.953 | -0.0005 | 0.0861 | 0.961 |
| | \( \pi_3 \) | 0.0000 | 0.0951 | 0.945 | -0.0008 | 0.0883 | 0.951 |
| | \( \pi_4 \) | 0.0003 | 0.0781 | 0.950 | 0.0005 | 0.0745 | 0.956 |
| | \( \pi_5 \) | -0.0005 | 0.0632 | 0.956 | -0.0009 | 0.0864 | 0.949 |
| 100 | \( \omega \) | -0.0007 | 0.1151 | 0.926 | -0.0022 | 0.1560 | 0.941 |
| | \( \pi_1 \) | -0.0002 | 0.0507 | 0.954 | 0.0011 | 0.0604 | 0.951 |
| | \( \pi_2 \) | -0.0005 | 0.0519 | 0.929 | 0.0004 | 0.0606 | 0.955 |
| | \( \pi_3 \) | 0.0001 | 0.0672 | 0.952 | 0.0004 | 0.0624 | 0.961 |
| | \( \pi_4 \) | 0.0003 | 0.0553 | 0.958 | -0.0001 | 0.0523 | 0.961 |
| | \( \pi_5 \) | -0.0006 | 0.0447 | 0.953 | 0.0009 | 0.0609 | 0.941 |
| 6 | 50 | \( \omega \) | -0.0009 | 0.1573 | 0.877 | 0.0036 | 0.2163 | 0.935 |
| | \( \pi_1 \) | 0.0007 | 0.0740 | 0.942 | -0.0004 | 0.0704 | 0.942 |
| | \( \pi_2 \) | -0.0006 | 0.1142 | 0.953 | 0.0009 | 0.1106 | 0.945 |
| | \( \pi_3 \) | 0.0010 | 0.0776 | 0.947 | 0.0000 | 0.0721 | 0.944 |
| | \( \pi_4 \) | 0.0005 | 0.0621 | 0.946 | 0.0002 | 0.1098 | 0.939 |
| | \( \pi_5 \) | 0.0012 | 0.0698 | 0.954 | -0.0002 | 0.0622 | 0.954 |
| | \( \pi_6 \) | 0.0006 | 0.0766 | 0.950 | -0.0001 | 0.0898 | 0.943 |
| 100 | \( \omega \) | -0.0002 | 0.1157 | 0.931 | -0.0006 | 0.1552 | 0.936 |
| | \( \pi_1 \) | 0.0001 | 0.0525 | 0.947 | -0.0005 | 0.0497 | 0.939 |
| | \( \pi_2 \) | 0.0002 | 0.0809 | 0.968 | 0.0000 | 0.0777 | 0.952 |
| | \( \pi_3 \) | 0.0000 | 0.0549 | 0.951 | -0.0004 | 0.0509 | 0.951 |
| | \( \pi_4 \) | 0.0001 | 0.0439 | 0.943 | -0.0013 | 0.0774 | 0.954 |
| | \( \pi_5 \) | 0.0001 | 0.0494 | 0.945 | -0.0003 | 0.0440 | 0.941 |
| | \( \pi_6 \) | -0.0008 | 0.0542 | 0.953 | -0.0003 | 0.0634 | 0.952 |
(b) Generate the zero-inflated parameters $\omega_i$ and probability parameters $\pi_i = (\pi_{1i}, \ldots, \pi_{qi})$ for $i = 1, 2, \ldots, n$ from the logistic models.

(c) Generate the binomial denominators $m_i = (m_{1i}, \ldots, m_{qi})$ from 5 to 15.

(d) Generate the response $y_i$ from multivariate ZIB distribution $MZIB(m_i, \omega_i, \pi_i)$, $i = 1, 2, \ldots, n$ using the procedure given in Section 4.1.

The sample size is selected as $n = 100, 300$ and 500. Then via EM algorithm (16)–(19) and Wald-type methods the MLEs for the parameters $\alpha$ and $\beta_1, \ldots, \beta_q$ and its MSEs, the 95% confidence intervals and its widths can be calculated from the generated samples. Further, similar to the case in Table 2, with repeating times $G = 1000$, the average values of biases and MSEs of MLEs, widths and coverage probabilities of confidence intervals for the parameters $\alpha$ and $\beta_1, \ldots, \beta_q$ are given in Table 3 with $q = 2, 3$. The simulation results with $q = 4, 5$ can be seen in Table A1 of supplemental file.

From the results given in Table 3 and in Table A1 of the supplemental file, one can see that the proposed EM algorithm for the estimation of regression parameters has a good performance. The biases are small and the MLEs are very close to the corresponding true values of parameters and the coverage probabilities are all around 0.95, although the algorithm shows a little liberty for the estimated parameter $\alpha$ with $q = 2$ and $n = 100$.

### 4.3. Tests for zero-inflation in MZIB model

In this subsection, the performance of both the score test and the likelihood ratio test for testing zero inflation in the multivariate ZIB distribution is conducted by a simulation study. All the simulations in this subsection are performed with $G = 10,000$ replications and covariates are not considered for the purpose of simplicity. The dimensions $q = 2, 3, 4$ are considered for multivariate zero-inflated binomial distribution with sample sizes of $n = 50, 100, 200, 300, 400, 500$. For assessing the powers of proposed test statistics, the zero-inflation parameter is designed to be $\omega = 0.0, 0.01, 0.05, 0.1, 0.2$. Since we are not interested in the inference of binomial probabilities $\pi$ and binomial denominators $m$, we first generate two vectors $\pi$ and $m$ randomly. For a given set of parameters $(n, \omega, m, \pi)$, the samples with multivariate ZIB distribution $ZIB_q(\omega, m, \pi)$ could be generated in the same procedure as that in Section 4.1 and the estimates of parameters $(\omega, \pi)$ for multivariate zero-inflated binomial distribution under LRT alternative hypothesis are computed by EM algorithm.

The simulation results are summarized in Table 4 and in Table A2 of the supplemental file. The results in Table 4 are computed by using the binomial probabilities $\pi$ and $m$, which are randomly generated from 0.05 to 0.20, and 5 to 15, respectively. The results in Table A2 of the supplemental file are computed by using the binomial probabilities $\pi$ and $m$, which are randomly generated from 0.01 to 0.15, and 5 to 20, respectively. These results show the comparison between score test statistic $S_1$ and LRT statistic $S_2$ side by side for both controlling nominal level and powers. Both test statistics hold the nominal level at $\alpha = 0.05$ well. The empirical powers of both tests for detecting zero-inflation increase as the dimensions in multivariate ZIB distribution and the zero-inflated parameter increase. We also perform the simulation study for multivariate zero-inflated binomial distribution with $q = 5+$ dimensions. However, not like the EM algorithm for the estimates of parameters, even for a very small zero-inflation parameter $\omega$ ($\omega = 0.01$) both tests show great
Table 3. Biases and MSEs of MLEs, widths and coverage probabilities of confidence intervals for regression parameters with the dimension $q = 2$ and 3.

| $n$ | $\alpha_{10}$ | $\alpha_{11}$ | $\alpha_{12}$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{20}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{30}$ | $\beta_{31}$ | $\beta_{32}$ |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| **q = 2** | | | | | | | | | | | | |
| 100 | Bias | -0.2948 | 0.1716 | -0.1231 | -0.0021 | 0.0106 | -0.0130 | -0.0174 | 0.0224 | -0.0089 |
| | CP | 0.9600 | 0.9560 | 0.9740 | 0.9360 | 0.9360 | 0.9280 | 0.9130 | 0.9190 | 0.9501 |
| | Width | 7.5060 | 4.7936 | 3.0683 | 1.7440 | 1.1360 | 0.6528 | 1.8381 | 1.2030 | 0.6989 |
| | MSE | 1.9148 | 1.2229 | 0.7828 | 0.4449 | 0.2898 | 0.1665 | 0.4689 | 0.3067 | 0.1783 |
| 300 | Bias | -0.1571 | 0.1396 | -0.1526 | -0.0006 | 0.0105 | -0.0126 | -0.0071 | 0.0128 | -0.0092 |
| | CP | 0.9380 | 0.9420 | 0.9480 | 0.9400 | 0.9370 | 0.9280 | 0.9480 | 0.9410 | 0.9440 |
| | Width | 4.1000 | 2.6239 | 1.6894 | 0.9939 | 0.6475 | 0.3727 | 1.0482 | 0.6866 | 0.3986 |
| | MSE | 1.0459 | 0.6694 | 0.4310 | 0.2536 | 0.1652 | 0.0951 | 0.2674 | 0.1752 | 0.1017 |
| 500 | Bias | -0.0760 | 0.0972 | -0.1253 | -0.0020 | 0.0090 | -0.0082 | -0.0147 | 0.0190 | -0.0120 |
| | CP | 0.9510 | 0.9530 | 0.9440 | 0.9350 | 0.9400 | 0.9290 | 0.9320 | 0.9320 | 0.9290 |
| | Width | 3.1301 | 2.0067 | 1.2769 | 0.7697 | 0.5017 | 0.2883 | 0.8105 | 0.5310 | 0.3081 |
| | MSE | 0.7985 | 0.5119 | 0.3257 | 0.1963 | 0.1280 | 0.0735 | 0.2068 | 0.1355 | 0.0786 |
| **q = 3** | | | | | | | | | | | | |
| 100 | Bias | -0.0921 | 0.0093 | -0.0801 | -0.0090 | 0.0001 | 0.0094 | 0.0170 | -0.0075 | 0.0048 | 0.0132 | -0.0015 | -0.0033 |
| | CP | 0.9660 | 0.9710 | 0.9690 | 0.9170 | 0.9340 | 0.9390 | 0.9440 | 0.9390 | 0.9301 | 0.9410 | 0.9350 | 0.9490 |
| | Width | 4.9825 | 2.2885 | 2.8878 | 1.2186 | 0.5883 | 0.6738 | 1.2612 | 0.6180 | 0.7050 | 1.2949 | 0.6150 | 0.7246 |
| | MSE | 1.2711 | 0.5838 | 0.7367 | 0.3109 | 0.1501 | 0.1719 | 0.3217 | 0.1577 | 0.1798 | 0.3303 | 0.1569 | 0.1848 |
| 300 | Bias | -0.1095 | 0.0497 | -0.0442 | -0.0022 | 0.0035 | 0.0085 | -0.0026 | -0.0023 | 0.0142 | -0.0045 | -0.0047 |
| | CP | 0.9540 | 0.9530 | 0.9560 | 0.9330 | 0.9300 | 0.9260 | 0.9240 | 0.9300 | 0.9400 | 0.9310 | 0.9340 | 0.9350 |
| | Width | 2.7472 | 1.2556 | 1.5657 | 0.6942 | 0.3353 | 0.3853 | 0.7190 | 0.3521 | 0.4035 | 0.7353 | 0.3493 | 0.4147 |
| | MSE | 0.3509 | 0.1605 | 0.1993 | 0.0887 | 0.0428 | 0.0492 | 0.0919 | 0.0450 | 0.0515 | 0.0939 | 0.0446 | 0.0530 |
| 500 | Bias | -0.0439 | 0.0244 | -0.0327 | 0.0028 | 0.0003 | -0.0017 | -0.0037 | 0.0021 | 0.0029 | -0.0063 | 0.0042 | -0.0012 |
| | CP | 0.9580 | 0.9560 | 0.9520 | 0.9430 | 0.9520 | 0.9470 | 0.9400 | 0.9370 | 0.9250 | 0.9340 | 0.9350 | 0.9260 |
| | Width | 2.0987 | 0.9619 | 1.1957 | 0.5362 | 0.2590 | 0.2976 | 0.5560 | 0.2724 | 0.3121 | 0.5679 | 0.2698 | 0.3202 |
| | MSE | 0.5354 | 0.2454 | 0.3050 | 0.1368 | 0.0661 | 0.0759 | 0.1418 | 0.0695 | 0.0796 | 0.1449 | 0.0688 | 0.0817 |
Table 4. Empirical Powers for both the score test and the likelihood ratio test for testing zero-inflation parameter $\omega$ with nominal level $\alpha = 0.05$.

| $q$ | $n$ | $\omega = 0$    | Score | LRT | $\omega = 0.01$ | Score | LRT | $\omega = 0.05$ | Score | LRT | $\omega = 0.10$ | Score | LRT | $\omega = 0.20$ | Score | LRT |
|-----|-----|-----------------|-------|-----|-----------------|-------|-----|-----------------|-------|-----|-----------------|-------|-----|-----------------|-------|-----|
| 2   | 50  | 0.0528          | 0.0467 |     | 0.0872          | 0.1163 |     | 0.1346          | 0.1819 |     | 0.3313          | 0.4204 |     | 0.7578          | 0.8310                             |
|     | 100 | 0.0493          | 0.0428 |     | 0.0919          | 0.1667 |     | 0.2209          | 0.2962 |     | 0.5780          | 0.6705 |     | 0.9664          | 0.9818                             |
|     | 200 | 0.0471          | 0.0436 |     | 0.1846          | 0.2576 |     | 0.3832          | 0.4844 |     | 0.8567          | 0.9095 |     | 0.9977          | 1.000                             |
|     | 300 | 0.0516          | 0.0474 |     | 0.2430          | 0.3365 |     | 0.5185          | 0.6316 |     | 0.9571          | 0.9764 |     | 1.000           | 1.000                             |
|     | 400 | 0.0484          | 0.0449 |     | 0.3053          | 0.4109 |     | 0.6370          | 0.7346 |     | 0.9892          | 0.9948 |     | 1.000           | 1.000                             |
|     | 500 | 0.0497          | 0.0477 |     | 0.3627          | 0.4708 |     | 0.7311          | 0.8144 |     | 0.9975          | 0.9988 |     | 1.000           | 1.000                             |
| 3   | 50  | 0.0419          | 0.0400 |     | 0.3173          | 0.3167 |     | 0.5387          | 0.5410 |     | 0.8851          | 0.8890 |     | 0.9966          | 0.9966                             |
|     | 100 | 0.0368          | 0.0407 |     | 0.4846          | 0.5068 |     | 0.7688          | 0.7847 |     | 0.9896          | 0.9890 |     | 1.000           | 1.000                             |
|     | 200 | 0.0434          | 0.0435 |     | 0.7154          | 0.7500 |     | 0.9516          | 0.9620 |     | 1.000           | 1.000  |     | 1.000           | 1.000                             |
|     | 300 | 0.0460          | 0.0452 |     | 0.8554          | 0.8801 |     | 0.9901          | 0.9927 |     | 1.000           | 1.000  |     | 1.000           | 1.000                             |
|     | 400 | 0.0503          | 0.0475 |     | 0.9258          | 0.9441 |     | 0.9985          | 0.9991 |     | 1.000           | 1.000  |     | 1.000           | 1.000                             |
|     | 500 | 0.0493          | 0.0447 |     | 0.9667          | 0.9758 |     | 0.9996          | 0.9997 |     | 1.000           | 1.000  |     | 1.000           | 1.000                             |
| 4   | 50  | 0.0635          | 0.0214 |     | 0.6021          | 0.5143 |     | 0.8152          | 0.7636 |     | 0.9760          | 0.9684 |     | 1.000           | 1.000                             |
|     | 100 | 0.0512          | 0.0442 |     | 0.8393          | 0.8146 |     | 0.9665          | 0.9577 |     | 0.9999          | 0.9999 |     | 1.000           | 1.000                             |
|     | 200 | 0.0412          | 0.0376 |     | 0.9642          | 0.9618 |     | 0.9986          | 0.9984 |     | 1.000           | 1.000  |     | 1.000           | 1.000                             |
|     | 300 | 0.0421          | 0.0385 |     | 0.9921          | 0.9916 |     | 0.9998          | 0.9998 |     | 1.000           | 1.000  |     | 1.000           | 1.000                             |
|     | 400 | 0.0469          | 0.0471 |     | 0.9984          | 0.9986 |     | 1.000           | 1.000  |     | 1.000           | 1.000  |     | 1.000           | 1.000                             |
|     | 500 | 0.0423          | 0.0442 |     | 0.9999          | 0.9999 |     | 1.000           | 1.000  |     | 1.000           | 1.000  |     | 1.000           | 1.000                             |

powers for testing zero-inflation, but both tests do not hold nominal level well for 5+ dimensional multivariate zero-inflated proportional data. Sometimes both the score test and the likelihood ratio test require a fair size of sample for calculating the inverse of the expected information matrix. Overall, there is not much difference in power between the score test and likelihood ratio test.

4.4. Test for equality of probabilities in MZIB model

In this subsection, the performance of both score test and likelihood ratio test for testing equality of multivariate zero-inflated binomial distribution parameters $\pi_1 = \cdots = \pi_q$ is conducted by a simulation study. All the simulations in this subsection are performed with $G = 10,000$ replications and covariates are not considered for the sake of simplicity. Only one zero-inflation parameter ($\omega = 0.3$) and two sets of dimensions ($q = 2, 3$) are considered for multivariate zero-inflated binomial distribution with sample sizes of $n = 50, 100, 150, 200, 250, 300, 350, 400, 450, 500$. The vector of binomial denominators $m$ is randomly generated. The vector of binomial probabilities is preassigned. For the two-dimension case, $(\pi_1, \pi_2)$ are $(0.3, 0.3), (0.5, 0.5), (0.8, 0.8)$ and $(0.4, 0.5)$. For the three-dimension case, values of $(\pi_1, \pi_2, \pi_3)$ are $(0.3, 0.3, 0.3), (0.5, 0.5, 0.5), (0.8, 0.8, 0.8)$ and $(0.45, 0.5, 0.5)$. For a given set of parameters $(n, \omega, m, \pi)$, the multivariate zero-inflated binomial data could be generated by a similar procedure given in Section 4.1.

The simulation results are summarized in Figure 1 and in Table A3 of the supplemental file. Figure 1 displays the comparisons of performance for empirical levels and empirical powers at the nominal level $\alpha = 0.05$ between the score test (solid line) and the likelihood ratio test (dotted line) for testing the equality of parameters $\pi$ based on two-dimensional
and three-dimensional zero-inflated binomial models. Similar conclusions can also be obtained from Table A3 in the supplemental file.

From the simulation results, both the score test and the likelihood ratio test maintain the nominal level $\alpha = 0.05$ well. The powers of both tests for detecting equality of multivariate zero-inflated binomial parameter $\pi$ are very close in both two-dimensional and three-dimensional simulations with no influence of sample size. Also, both tests show great detecting power even with a very small difference in $\pi$. Overall, there is no difference between the score test and likelihood ratio test in testing the equality of $\pi$ in the multivariate zero-inflated binomial distribution.

5. A real example

In this section, we illustrate the application of the proposed multivariate zero-inflated binomial model to whitefly data. van Iersel et al. [11] studied the purpose of controlling silver leaf whiteflies by using a subirrigation system. The study was designed to determine the effectiveness of controlling silver leaf whiteflies on poinsettia with imidacloprid, which was delivered by a subirrigation system. Imidacloprid is a resilient and powerful chemical (e.g. Natwick et al. [20]), that has low toxicity to mammals, and is used to control silver leaf whiteflies on poinsettia. At the first week of this experiment, researchers placed $m$ adult whiteflies (here, $m$ is considered as the binomial denominators with range 6–15, mean = 9.5 and SD = 1.7) in clip-on leaf cages attached to one leaf per plant and then recorded the number of surviving whiteflies 2 days later, which is considered as the response variable. To measure reproductive inhibition, they removed the fly cages after obtaining the survival count but marked the position of each cage. In the coming week, they placed $m$ adult whiteflies in clip-on leaf cages attached to one leaf on the same plant and recorded the number of surviving whiteflies. This study lasted for consecutive 12 weeks on 54 plants. Therefore, the data can be considered to consist of the 12-dimensional binomial variables, that is, the observed data can be expressed as $\{(y_{i}, m_{i}), i = 1, 2, \ldots, 54\}$ with $y_{i} = (y_{i1}, \ldots, y_{i12})$ and $m_{i} = (m_{i1}, \ldots, m_{i12})$. However, although the design originally called for 648 observations in a balanced design, one observation was lost on each of plant 4 at week 2, plant 7 at week 4, plant 15 at week 11, plant 17 at week 3, plant 20 at week 11 and plant 27 at week 2.
Table 5. Estimates of correlation coefficients among 12 binomial variables.

| Week | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1    | 1.0000 |     |     |     |     |     |     |     |     |     |     |     |
| 2    | 0.1439 | 1.0000 |     |     |     |     |     |     |     |     |     |     |
| 3    | 0.1848 | 0.1295 | 1.0000 |     |     |     |     |     |     |     |     |     |
| 4    | 0.2931 | 0.2232 | 1.0000 |     |     |     |     |     |     |     |     |     |
| 5    | 0.2922 | 0.2585 | 0.3906 | 1.0000 |     |     |     |     |     |     |     |     |
| 6    | 0.2935 | 0.2246 | 0.3250 | 0.4132 | 0.3606 | 1.0000 |     |     |     |     |     |     |
| 7    | 0.2317 | 0.1932 | 0.2323 | 0.3250 | 0.4132 | 0.3606 |     | 1.0000 |     |     |     |     |
| 8    | 0.2488 | 0.1776 | 0.2340 | 0.3445 | 0.3591 | 0.3622 | 0.3505 | 1.0000 |     |     |     |     |
| 9    | 0.2717 | 0.2386 | 0.2252 | 0.3703 | 0.4995 | 0.4645 | 0.3376 | 0.3396 | 1.0000 |     |     |     |
| 10   | 0.2900 | 0.2534 | 0.2219 | 0.3519 | 0.3503 | 0.3537 | 0.3400 | 0.4235 | 0.3484 | 1.0000 |     |     |
| 11   | 0.2523 | 0.2532 | 0.2046 | 0.3774 | 0.4383 | 0.4201 | 0.3076 | 0.3266 | 0.4128 | 0.3673 | 1.0000 |     |
| 12   | 0.1739 | 0.0816 | 0.1080 | 0.2298 | 0.2640 | 0.2519 | 0.1875 | 0.2459 | 0.1414 | 0.2265 | 1.0000 |     |

plants and two observations were lost on plant 18 at weeks 10 and 11, to yield a final data set with \( N = 640 \) observations. Further, by examining the data set, the existence of excessive number of zeros for each week is discovered. The detailed information is shown in Table A4 of the supplemental file. It can be seen that the percentage of zeros in this data set is greater than 50%. Also, Figure A1 in the supplemental file shows the frequency \( y_{ij} \) of alive whiteflies in 3D image for whitefly data.

Now the proposed multivariate zero-inflated binomial model can be used to analyse the whitefly data. At first, the bivariate proportional dataset can be generated from week \( r \) and week \( s \) (\( r, s = 1, 2, \ldots, 12 \) and \( r \neq s \)), denoted by \( D_{rs} = \{ (y_{ir}, y_{is}; m_{ir}, m_{is}), i = 1, 2, \ldots, n \} \). Next the generated data set \( D_{rs} \) can be analysed using bivariate zero-inflated binomial model and the MLEs \( \hat{\omega}, \hat{\pi}_r \) and \( \hat{\pi}_s \) of parameters \( \omega, \pi_r \) and \( \pi_s \) can be computed via EM algorithm. Then, the estimated correlation coefficient \( \hat{\rho}_{rs}^{(i)} \) for each observation \( (y_{ir}, y_{is}; m_{ir}, m_{is}) \) in bivariate binomial data can be computed from (3) with \( \omega = \hat{\omega}, \pi_r = \hat{\pi}_r \) and \( \pi_s = \hat{\pi}_s \):

\[
\hat{\rho}_{rs}^{(i)} = \frac{\hat{\omega} \sqrt{m_{ir} \hat{\pi}_r m_{is} \hat{\pi}_s}}{\sqrt{(\hat{\omega} m_{ir} \hat{\pi}_r + (1 - \hat{\pi}_r))(\hat{\omega} m_{is} \hat{\pi}_s + (1 - \hat{\pi}_s))}}, \quad r \neq s, \; i = 1, 2, \ldots, n.
\]

At last, the estimated correlation coefficient for bivariate zero-inflated binomial data set \( D_{rs} \) is

\[
\hat{\rho}_{rs} = \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}_{rs}^{(i)}, \quad r, s = 1, 2, \ldots, 12
\]

which is the mean of estimated correlation coefficients for all observations in bivariate zero-inflated binomial data. The results for the correlation coefficients among the components of 12-dimensional binomial variables are presented in Table 5. From the results in Table 5, there certainly exist the positive correlations between any two components of the 12 binomial variables, which are induced from the existence of zero-inflation for the bivariate binomial variables.

Next, since the proposed test statistics work well only for the dimension \( q \leq 4 \), for the purpose of illustrating the proposed model, the vector of 12 variables is partitioned into four vectors of three variables and thus obtain four small multivariate proportional data sets denoted weeks 1–3, weeks 4–6, weeks 7–9 and weeks 10–12 by \( D_1, D_2, D_3 \) and \( D_4 \),
respectively. Due to the unbalanced design, the observations on the plants with losing are omitted and the data sets $D_1 D_2, D_3$ and $D_4$ can be expressed as

$$D_1 = \{(y_{i1}, y_{i2}, y_{i3}; m_{i1}, m_{i2}, m_{i3}), i = 1, 2, \ldots, 51\},$$

$$D_2 = \{(y_{i4}, y_{i5}, y_{i6}; m_{i4}, m_{i5}, m_{i6}), i = 1, 2, \ldots, 53\},$$

$$D_3 = \{(y_{i7}, y_{i8}, y_{i9}; m_{i7}, m_{i8}, m_{i9}), i = 1, 2, \ldots, 54\},$$

$$D_4 = \{(y_{i10}, y_{i11}, y_{i12}; m_{i10}, m_{i11}, m_{i12}), i = 1, 2, \ldots, 51\}.$$

The following three-dimensional zero-inflated binomial model is used to analyse the data:

$$f(y; \omega, \pi) = \left[\omega + (1 - \omega) (1 - \pi_1)^{m_1} (1 - \pi_2)^{m_2} (1 - \pi_3)^{m_3}\right]^{y=0} \times \left[(1 - \omega) \frac{m_1}{y_1} \frac{m_2}{y_2} \frac{m_3}{y_3} \pi_1^{y_1} \pi_2^{y_2} \pi_3^{y_3} (1 - \pi_1)^{m_1 - y_1} (1 - \pi_2)^{m_2 - y_2} (1 - \pi_3)^{m_3 - y_3}\right]^{y \neq 0}.$$

Both the score test and the likelihood ratio test are applied to test the existence of zero-inflation and the equality of the binomial probabilities for the three components in each dataset. If there exists the zero-inflation, the MLEs of zero-inflated parameters and binomial probabilities are also computed using the EM algorithm for four data sets $D_1, D_2, D_3$ and $D_4$. The results of data analysis for these data sets are given in Table 6.

From Table 6, one can see that the values of score test statistics and LRT statistics for testing the presence of zero-inflation are very large and thus the $p$-values for these test are very small (near to zero), which show that there exist the zero-inflation in four three-dimensional proportional data sets and thus a positive correlation among the components of these multivariate proportional variables in each data set. Now, by using the bootstrap method given in Section 3.1.2, the 95% confidence intervals of zero-inflation parameter $\omega$ for four data sets are computed as (0.0196, 0.1765), (0.2264, 0.4906), (0.1666, 0.4259) and (0.0588, 0.2545), respectively. These confidence intervals also confirm the existence of zero-inflation and positive correlation among the components in four induced data sets. Further, there exists enough evidence to show that binomial probabilities of three components are not equal for all data sets $D_1, D_2, D_3$ and $D_4$ at the level of significance $\alpha = 0.05$ since all $p$-values of 8 statistics for testing the equality of binomial probabilities are less than 0.05. However, the $p$-values of score tests and LRTs test in weeks 7–9 are 0.03372 and 0.03377, respectively, and thus the proportional probabilities in the components of weeks 7–9 could be equal at the significant level $\alpha = 0.025$. Also, the values of score test and LRT

| Test | Score($p$-value) | LRT($p$-value) | MLE of Parameter $\omega$ | MLE of Parameter $\pi$ |
|------|-----------------|----------------|--------------------------|--------------------------|
| Test $\omega = 0$ | 7639.8219(0.0000) | 14453.2000(0.0000) | 16838.4730(0.0000) | $\omega = 0.025$ |
| $\pi_1 = \pi_2 = \pi_3$ | 67.2800(0.0000) | 288.1016(0.0000) | 203.2252(0.0000) | $\pi_1 = \pi_2 = \pi_3 = 0.1666$ |
| MLE of Parameter | Score($p$-value) | LRT($p$-value) | $\omega$ | $\pi_1$ |
| $\omega_1$ | (0.0980, 0.2643) | (0.3585, 0.4452) | $\pi_1 = 0.0980$ |
| $\omega_2$ | (0.3694, 0.3000) | (0.3046, 0.3249) | $\pi_2 = 0.3694$ |
| $\omega_3$ | (0.3694, 0.3000) | (0.3046, 0.3249) | $\pi_3 = 0.3694$ |
Table 7. The results for analysing three four-dimensional data sets of whitefly data by using MZIB model.

| Data set | $D_5$ | $D_6$ | $D_7$ |
|----------|-------|-------|-------|
| Test $\omega = 0$ Score($p$-value) | 139981.7580(0.0000) | 34758.6047(0.0000) | 26939.4613(0.0000) |
| | LRT($p$-value) | 100.3533(0.0000) | 261.0577(0.0000) | 100.1171(0.0000) |
| Test $\pi_1 = \pi_2 = \pi_3 = \pi_4$ Score($p$-value) | 12.6343(0.0055) | 5.9178(0.1157) | 53.6393(0.0000) |
| | LRT($p$-value) | 12.5544(0.0057) | 5.8597(0.1186) | 51.3813(0.0000) |
| MLE of Parameter | ($\omega, \pi_1, \pi_2$) | (0.1000, 0.2678, 0.3762) | (0.2778, 0.2668, 0.2778) | (0.1176, 0.2233, 0.2828) |
| | ($\pi_3, \pi_4$) | (0.3035, 0.3186) | (0.3168, 0.3378) | (0.2329, 0.4240) |

For testing the equality of probabilities are almost same in each of the four data sets. However, the score test is more sensitive to the zero-inflation than the likelihood ratio test for testing the existence of zero-inflation since the values of score test are much larger than that of the likelihood ratio test for testing zero-inflation.

For the purpose of illustration, the original data are again partitioned into three data sets $D_5$ for weeks 1–4, $D_6$ for weeks 5–8, $D_7$ for weeks 9–12. Hence, each data set consists of four-dimensional binomial variables with four-dimensional binomial denominators. The four-dimensional ZIB model with the corresponding test statistics and EM algorithm is applied to conduct the same analysis for these three four-dimensional binomial data sets as that for four three-dimensional binomial data sets. The results are given in Table 7. The results in Table 7 are similar to that in Table 6. Since all values of statistics for testing the existence of zero-inflation are very large and thus the corresponding $p$-values are near to zero, the data sets $D_5$, $D_6$ and $D_7$ present the zero-inflation and there exist positive correlations among the components in each of three induced four-dimensional binomial data sets. Also, the $p$-values of score test $S_2$ and LRT $S_4$ for data sets $D_5$ and $D_7$ are less than 0.01, which indicates that binomial probabilities of four components could be not equal even at the significant level $\alpha = 0.01$ in $D_5$ and $D_7$. However, the binomial probabilities for weeks 5–8 are probably equal since the $p$-values of score test $S_2$ and LRT $S_4$ are 0.1157 and 0.1186, respectively. Further the 95% bootstrap confidence intervals for zero-inflation parameter $\omega$ in three induced data sets $D_5$, $D_6$ and $D_7$ are (0.0200, 0.2000), (0.1667, 0.3889) and (0.0392, 0.2157), respectively. This also shows that there exist the zero-inflation and positive correlation among the components in three induced four-dimensional proportional data sets $D_5$, $D_6$ and $D_7$.

We also computed the values of score tests and likelihood tests for all possible data sets induced from the original data, such as two-dimensional data sets 1–2 weeks, 2–3 weeks, …, 11–12 weeks, three-dimensional data sets 1–3 weeks, 2–4 weeks, …, 10–12 weeks, …, 11-dimensional data sets 1–11 weeks and 2–12 weeks, and 12-dimensional data set 1–12 weeks (whole data). Although all tests show the strong existence of zero-inflation for these data sets, the test results may not be applicable because the proposed test statistics are not reliable for the multivariate proportional data with 5+ dimensions. However, we may use the bootstrap confidence interval to test the zero-inflation in multivariate binomial data with the moderate dimension number. Now we partition the original data into two six-dimensional data sets $D_8$ and $D_9$ and then analyse them by using the proposed MZIB model. The results are presented in Table 8.

From the results in Table 8, the values of score test statistic for testing the presence of zero-inflation are very large but unreliable. Note that since the EM algorithm is based on the stochastic representation (1) of Definition 2.1 and thus the value of zero-inflation
Table 8. The results for analysing two six-dimensional datasets of whitefly data by using MZIB model.

|                  | Dataset $D_8$ | Dataset $D_9$ |
|------------------|--------------|--------------|
| Test $\omega = 0$ | Score($p$-value) 1092598.3867(0.0000) | Score($p$-value) 368992.8598(0.0000) |
|                  | LRT($p$-value) 76.0592(0.0000) | LRT($p$-value) 49.6176(0.0000) |
| Test $\pi_1 = \cdots = \pi_6$ | Score($p$-value) 32.3973(0.0000) | Score($p$-value) 56.7644(0.0000) |
|                  | LRT($p$-value) 31.9574(0.0000) | LRT($p$-value) 54.0472(0.0000) |
| MLE of Parameter | $(\omega, \pi_1, \pi_2, \pi_3)$ $(0.0600, 0.2575, 0.3602, 0.2891)$ | $(0.0392, 0.2479, 0.2538, 0.2043)$ |
|                  | $(\pi_4, \pi_5, \pi_6)$ $(0.3052, 0.2185, 0.2254)$ | $(0.2593, 0.2160, 0.3974)$ |

Parameter is implicitly assumed to be in the unit interval $[0, 1]$, the smallest values of lower limits in the confidence intervals are at least zero. Further based on Definition 2.1, the hypotheses for testing the presence of zero-inflation should be upper-tailed hypothesis $H_0 : \omega = 0$ versus $H_1 : \omega > 0$. From EM algorithm, the 95% bootstrap confidence intervals for zero-inflation parameter $\omega$ in two induced six-dimensional data sets $D_8$ and $D_9$ are $(0.0000, 0.1200)$ and $(0.0000, 0.0980)$ and the 90% bootstrap confidence intervals are $(0.0200, 0.1200)$ and $(0.0000, 0.0784)$, respectively. By comparing zero with the lower limit of confidence intervals, the null hypothesis is rejected in data set $D_8$ at the level $\alpha = 0.05$ but is not rejected at the level $\alpha = 0.025$. However, the null hypothesis is not rejected at both levels $\alpha = 0.05$ and $0.025$ in data set $D_9$. This means that there is no evidence for the presence of zero-inflation in data set $D_9$ at $\alpha = 0.05$. Furthermore, by using the bootstrap method, the $p$-values are approximately 0.045 in $D_8$ and 0.126 in $D_9$. Based on these $p$-values, we can also get the same conclusions as above.

6. Concluding remarks

A new model for multivariate proportional data, called ‘multivariate zero-inflated binomial model’ has been proposed. The model introduced a common zero-inflated parameter for all components of multivariate binomial variable, which automatically address the correlation among the components. This model can also be regarded as an extension of the widely discussed univariate zero-inflated binomial model in proportional data. The Fisher Scoring algorithm and EM algorithm are derived for the computation of the estimates of parameters in the proposed multivariate model with/without covariates. Score tests and likelihood ratio tests are also proposed to detect the existence of zero-inflation and the equality of the binomial probabilities in the multivariate binomial model. The simulation results demonstrate that the proposed EM algorithm has excellent performance for the computation of MLEs of parameters even for the moderately large dimension number in the MZIB model with/without covariates, and four test statistics maintain the nominal level well for the small dimensional numbers. However, the proposed test statistics do not work well if the dimension numbers of multivariate binomial variables are greater than 5, which is the limitation of these tests. The whitefly data is used to demonstrate the proposed model and inferential methods for the multivariate binomial data. The results show the existence of correlation and zero inflation and the equality of the binomial probabilities in the subsets of whitefly data.

However, it is very unlikely that all components of the binomial random vector are zero-inflated in the same way and/or by the same amount as measured by zero-inflated parameter, even for moderate number of dimension. Therefore, the proposed model has
the limitation for the application. The solution to this question is to introduce more zero-inflated parameters for the components of multivariate zero-inflated binomial variable. Such model has been proposed for multivariate count data with excessive zeros. This model can obviously be extended to analyse the multivariate proportional data with excessive zeros and the corresponding research on such model can be done in the future. Moreover, the proposed score test method had a shortcoming with a large number of dimensions for multivariate proportional data. The main reason is that the denominator of score test statistics is very small, which results in the large variability of score test statistics. In fact, due to the same reason, the score test does not work well even for univariate binomial variable if the binomial denominator is much large. We are considering a modification of score test. For example, the modified score test method may work well for the moderate dimension of multivariate binomial variables. Furthermore, in addition to zero-inflation, there often exists the over-dispersion or under-dispersion in the count/proportional data. Such dispersion should be investigated and the multivariate zero-inflated beta-binomial (MZIBB) model could be used to fit such data with over-dispersion or under-dispersion. We will be doing such research in the future.

Acknowledgments

The authors are very grateful to the editor, associate editor and two referees for their careful reading and valuable comments, which have greatly improved this paper. The research of the first author is partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

Disclosure statement

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

Funding

The research of first author is partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC). Guo-Liang Tian’s research was fully supported by National Natural Science Foundation of China (Grant No. 11771199).

References

[1] V. Alevizakos and C. Koukouvinos, Monitoring of zero-inflated binomial processes with a DEWMA control chart, J. Appl. Stat. 2020. doi:10.1080/02664763.2020.1761950
[2] M. Alqawba and N. Diawara, Copula-based Markov zero-inflated count time series models with application, J. Appl. Stat. 48 (2021), pp. 786–803. doi:10.1080/02664763.2020.1748581.
[3] V.J. Broek, A score test for zero-inflation in a Poisson distribution, Biometrics 51 (1995), pp. 738–743.
[4] B. Chandrasekarn and N. Balakrishnan, Some properties and a characterization of trivariate and multivariate binomial distributions, Statistics 36 (2002), pp. 211–218.
[5] A.C. Cohen, Estimation in Mixtures of Discrete Distributions, Statistical Pub. Society, Calcutta, 1963.
[6] D. Deng and S.R. Paul, Score tests for zero-inflation in generalized linear models, Can. J. Stat. 28 (2000), pp. 563–570.
[7] D. Deng and S.R. Paul, Score tests for zero-inflation and over-dispersion in generalized linear models, Stat. Sin. 15 (2005), pp. 257–276.
[8] K.T. Fang, S. Kotz and K.W. Ng, Symmetric Multivariate and Related Distributions, Chapman and Hall, New York & London, 1990.
[9] D.B. Hall, Zero-inflated Poisson and binomial regression with random effects: A case study, Biometrics 56 (2000), pp. 1030–1039.
[10] D.B. Hall and K.S. Berenhaut, Score tests for heterogeneity and over-dispersion in zero-inflated Poisson and binomial regression models, Can. J. Stat. 30 (2002), pp. 415–430.
[11] M.W. van Iersel, R.D. Oetting and D.B. Hall, Imidacloprid applications by subirrigation for control of silverleaf whitefly (Homoptera: Aleyrodidae) on poinsettia, J. Econ. Entomol. 93 (2000), pp. 813–819.
[12] N Jansakul and J.P. Hinde, Score tests for zero-inflated Poisson models, Comput. Stat. Data Anal. 40 (2002), pp. 75–96.
[13] N.L. Johnson and S. Kotz, Distribution in Statistics: Discrete Distribution, John Wiley & Sons, New York, 1969.
[14] A.S. Krishnamoorthy, Multivariate binomial and Poisson distributions, Sankhya Indian J. Stat. 13 (1951), pp. 117–124.
[15] D. Lambert, Zero-inflated Poisson regression, with an application to defects in manufacturing, Technometrics 34 (1992), pp. 1–14.
[16] L. Lauritzen, BS2 Statistical Inference, Lecture 4, University of Oxford, 2009.
[17] C.H. Li, J.C. Lu, J. Park, K. Kim, P.A. Brinkley and J.P. Peterson, Multivariate zero-inflated Poisson models and their applications, Technometrics 41 (1999), pp. 29–38.
[18] Y. Liu and G.L. Tian, Type I multivariate zero-inflated Poisson distribution with applications, Comput. Stat. Data Anal. 83 (2015), pp. 200–222.
[19] A. Min and C. Czado, Testing for zero-modification in count regression models, Stat. Sin. 20 (2010), pp. 323–341.
[20] E.T. Natwick, J.C. Palumbo and C.E. Engle, Effects of imidacloprid on colonization of aphids and silverleaf whitefly and growth, yield and phytotoxicity in cauliflower, Southwest. Entomol. 21 (1996), pp. 283–292.
[21] M. Ridout, J. Hinde and C.G.B. Demétrio, A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives, Biometrics 57 (2001), pp. 219–223.
[22] K.-S. Song, Simultaneous statistical modelling of excess zeros, over/underdispersion, and multimodality with applications in hotel industry, J. Appl. Stat. 2020. doi:10.1080/02664763.2020.1769577
[23] A.M.C. Vieira, J.P. Hinde and C.G.B. Demétrio, Zero-inflated proportion data models applied to a biological control assay, J. Appl. Stat. 27 (2000), pp. 373–389.