Statistics and Its Interface Volume 13 (2020) 193–208

Zero-one-inflated simplex regression models for the analysis of continuous proportion data

Pengyi Liu, Kam Chuen Yuen, Liu-Cang Wu,
Guo-Liang Tian, and Tao Li*

Continuous data restricted in the closed unit interval [0,1] often appear in various fields. Neither the beta distribution nor the simplex distribution provides a satisfactory fitting for such data, since the densities of the two distributions are defined only in the open interval (0,1). To model continuous proportional data with excessive zeros and excessive ones, it is the first time that we propose a *zero-one-inflated simplex* (ZOIS) distribution, which can be viewed as a mixture of the Bernoulli distribution and the simplex distribution. Besides, we introduce a new *minorization–maximization* (MM) algorithm to calculate the *maximum likelihood estimates* (MLEs) of parameters in the simplex distribution without covariates. Likelihood-based inference methods for the ZOIS regression model are also provided. Some simulation studies are performed and the hospital stay data of Barcelona in 1988 and 1990 are analyzed to illustrate the proposed methods. The comparison between the ZOIS model and the *zero-one-inflated beta* (ZOIB) model is also presented.

Keywords and phrases: Continuous proportion data, MM algorithm; Simplex distribution, Zero-one-inflated beta model, Zero-one-inflated simplex model.

1. INTRODUCTION

Many scientific studies in different disciplines yield outcomes in the form of percentages, fractions, rates or proportions that are measured continuously in intervals (0,1), [0,1), (0,1] or [0,1]. Different strategies have been proposed for modeling such continuous proportional data. To fit continuous observations restricted on the open interval (0,1), some authors considered the beta distribution as one of such tools, since its density has various shapes: left-skewed, right-skewed, "U", "J", inverted "J", and uniform depending on the values of the two parameters (see Johnson *et al.*, 1995, §25.1). Beta regression models have been studied by Paolino (2001), Kieschnick and McCullough (2003), Ferrari and Cribari-Neto (2004), Smithson and Verkuilen (2006), Korhonen *et al.* (2007), Espinheira *et al.* (2008a, 2008b), Simas *et al.* (2010), Ferrari and Pinheiro (2011), and so on. Recently, Ospina and Ferrari (2010) proposed mixed continuous–discrete inflated beta distributions to model data observed on [0,1), (0,1] or [0,1]. Ospina and Ferrari (2012) proposes a general class of regression models for continuous proportions when the data contain zeros or ones.

Moreover, as a non-exponential family member, the simplex distribution of Barndorff-Nielsen and Jørgensen (1991) can also be utilized to model continuous proportional data confined in the open interval (0,1). Simulation studies of Zhang and Qiu (2014) showed that the simplex regression model has a better robustness against violation of some distributional assumptions than the beta regression model. Based on these facts, in this paper, we consider the simplex model instead of the beta model.

By employing the simplex distribution, Song and Tan (2000) developed a marginal model for analyzing an eye surgery longitudinal proportional data. Song *et al.* (2004) further modeled heterogeneous dispersion in marginal models. Qiu and Song (2008) proposed a simplex mixed-effects models for longitudinal proportional data. Zhang and Wei (2008) considered maximum likelihood estimation of simplex distribution nonlinear mixed models via the stochastic approximation algorithm. Recently, Zhao *et al.* (2014) considered the Bayesian estimation of simplex distribution nonlinear mixed models for longitudinal data. Quintero and Contreras-Reyes (2018) proposed a mixture simplex model, where the parameters were estimated by an *expectation–maximization* (EM) algorithm.

In practice, usually, proportional data include a non-negligible number of zeros and ones. For these situations, neither the beta distribution nor the simplex distribution provides a satisfactory fitting for such data, since the densities of the two distributions are defined only in the open interval (0,1). To model continuous proportional data with excessive zeros and excessive ones, it is the first time that we propose a so-called *zero-one-inflated simplex* (ZOIS) distribution, which can be viewed as a mixture of the Bernoulli distribution and the simplex distribution. Besides, we provide a new *minorization–maximization* (MM) algorithm to calculate the *maximum likelihood estimate* (MLE) of the mean parameter in the simplex distribution. Two *stochastic representations* (SRs) of the ZOIS random variable are

*Corresponding author.

introduced to facilitate the likelihood-based statistical inferences.

The rest of this paper is organized as follows. In Section 2, we first review some basic properties of the simplex distribution and present a simple procedure to generate i.i.d. random samples from the simplex distribution (see the Appendix), then provide an MM algorithm to calculate MLEs of parameters in the simplex distribution, and introduce a ZOIS distribution via two SRs. In Section 3, likelihood-based inference methods for the ZOIS distribution without covariates and the ZOIS regression model are given. In addition, model selection and goodness-of-fit tests are also provided. Some simulation studies are performed in Section 4. In Section 5, we first fit the hospital stay data of Barcelona in 1988 and 1990 with both ZOIS and ZOIB distributions without/with covariates to illustrate the proposed methods, then we compare the difference between the beta and the simplex distributions for only fitting the continuous part (i.e., the observations in (0,1)). A discussion is presented in Section 6. The algorithm of random variable generation from the simplex distribution is given in the Appendix.

2. ZERO-ONE-INFLATED SIMPLEX MODEL

2.1 The simplex distribution

A continuous random variable X taking values in the open unit interval $(0,1)$ is said to follow the simplex distribution (Barndorff-Nielsen & Jørgensen, 1991), denoted by $X \sim S^{-}(\mu,\sigma^2)$, if its *probability density function* (pdf) is given by

$$f_{\mathrm{S}}(x;\mu,\sigma^2) = [2\pi\sigma^2 x^3(1-x)^3]^{-\frac{1}{2}} \exp\left[-\frac{d(x;\mu)}{2\sigma^2}\right], \tag{2.1}$$

for $x \in (0,1)$, where $\mu \in (0,1)$ is the mean parameter, $\sigma^2\,(>0)$ is the dispersion parameter, and

$$d(x;\mu) \doteq \frac{(x-\mu)^2}{x(1-x)\mu^2(1-\mu)^2} \tag{2.2}$$

is the unit deviance. The mean and variance of X are

$$\begin{aligned} E(X) &= \mu \quad \text{and} \\ \mathrm{Var}(X) &= \mu(1-\mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left[\frac{1}{2\sigma^2\mu^2(1-\mu^2)}\right] \\ &\quad \times\Gamma\left(\frac{1}{2},\ \frac{1}{2\sigma^2\mu^2(1-\sigma)^2}\right), \end{aligned} \tag{2.3}$$

where $\Gamma(a,b) = \int_b^\infty t^{a-1}\mathrm{e}^{-t}\,\mathrm{d}t$ denotes the upper incomplete gamma function.

To generate i.i.d. random samples from the simplex distribution (2.1), in Appendix A.3, we introduce a simple simulation procedure, which is closely related with the inverse Gaussian distribution (Appendix A.1) and the inverse Gaussian mixture distribution (Appendix A.2).

2.2 MLEs of parameters in the simplex distribution via an MM algorithm

Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} S^{-}(\mu,\sigma^2)$, $\{x_i\}_{i=1}^n$ be the corresponding realizations of $\{X_i\}_{i=1}^n$, and $Y_{\text{obs}} = \{x_i\}_{i=1}^n$ denote the observed data. The log-likelihood function of the unknown parameters (μ,σ^2) is given by

$$\ell(\mu,\sigma^2|Y_{\text{obs}}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}D(\mu|Y_{\text{obs}}) + \text{constant},$$

where

$$D(\mu|Y_{\text{obs}}) = \frac{1}{\mu^2(1-\mu)^2}\sum_{i=1}^n \frac{(x_i-\mu)^2}{x_i(1-x_i)}. \tag{2.4}$$

The aim is to calculate the MLEs of the parameters (μ,σ^2). The MLE of μ is

$$\begin{aligned} \hat{\mu} &= \arg\max_{\mu\in(0,1)}\Big[-D(\mu|Y_{\text{obs}})\Big] = \arg\min_{\mu\in(0,1)} D(\mu|Y_{\text{obs}}) \\ &= \arg\min_{\mu\in(0,1)} \log[D(\mu|Y_{\text{obs}})] \\ &= \arg\max_{\mu\in(0,1)}\Big\{-\log[D(\mu|Y_{\text{obs}})]\Big\}, \end{aligned}$$

where

$$\begin{aligned} \log[D(\mu|Y_{\text{obs}})] &= -2[\log(\mu)+\log(1-\mu)] \\ &\quad + \log\left[\sum_{i=1}^n \frac{(x_i-\mu)^2}{x_i(1-x_i)}\right]. \end{aligned}$$

Define

$$z = \sum_{i=1}^n \frac{(x_i-\mu)^2}{x_i(1-x_i)} \quad \text{and} \quad z^{(t)} = \sum_{i=1}^n \frac{(x_i-\mu^{(t)})^2}{x_i(1-x_i)}, \tag{2.5}$$

where $\mu^{(t)}$ denotes the t-th approximate of the MLE $\hat{\mu}$. By using the supporting hyperplane inequality

$$-\log(z) \geqslant 1 - \log(z^{(t)}) - \frac{z}{z^{(t)}},$$

we can construct a Q function as

$$\begin{aligned} Q(\mu|\mu^{(t)}) &= 1 - \log(z^{(t)}) \\ &\quad + 2[\log(\mu)+\log(1-\mu)] \\ &\quad - \frac{1}{z^{(t)}}\sum_{i=1}^n \frac{(x_i-\mu)^2}{x_i(1-x_i)} \end{aligned} \tag{2.6}$$

such that $Q(\mu|\mu^{(t)})$ minorizes $-\log[D(\mu|Y_{\text{obs}})]$ at the point $\mu=\mu^{(t)}$; i.e.,

$$\begin{aligned} Q(\mu|\mu^{(t)}) &\leqslant -\log[D(\mu|Y_{\text{obs}})] \quad \forall\ \mu,\ \mu^{(t)} \in (0,1) \quad \text{and} \\ Q(\mu^{(t)}|\mu^{(t)}) &= -\log[D(\mu^{(t)}|Y_{\text{obs}})]. \end{aligned}$$

According to the MM principle (Lange *et al.*, 2000), the $(t+1)$-th approximate of the MLE $\hat{\mu}$ is given by

$$\mu^{(t+1)} = \arg \max_{\mu \in (0,1)} Q(\mu|\mu^{(t)}).$$

Letting $\mathrm{d}Q(\mu|\mu^{(t)})/\mathrm{d}\mu = 0$, we can obtain $\mu^{(t+1)}$ as the real root of the cubic equation

$$a^{(t)}\mu^3 - (a^{(t)} + b^{(t)})\mu^2 + (b^{(t)} - 2)\mu + 1 = 0, \tag{2.7}$$

where

$$a^{(t)} = \frac{1}{z^{(t)}} \sum_{i=1}^{n} \frac{1}{x_i(1-x_i)} \quad \text{and} \quad b^{(t)} = \frac{1}{z^{(t)}} \sum_{i=1}^{n} \frac{1}{1-x_i},$$

and $z^{(t)}$ is specified by (2.5). In practice, we can take the initial value $\mu^{(0)} = 0.5$.

On the other hand, letting $\partial \ell(\mu, \sigma^2|Y_{\text{obs}})/\partial\sigma^2 = 0$, we can obtain the MLE of σ^2 as

$$\hat{\sigma}^2 = \frac{1}{n\hat{\mu}^2(1-\hat{\mu})^2} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{x_i(1-x_i)}. \tag{2.8}$$

2.3 Zero-one-inflated simplex distribution

Continuous data restricted in the closed unit interval [0,1] often appear in various fields. To model such continuous proportion data with extra zeros and ones, in this paper, we propose a so-called *zero-one-inflated simplex* (ZOIS) distribution, which can be viewed as a mixture of the Bernoulli distribution and the simplex distribution.

2.3.1 The first stochastic representation

Specifically, a continuous random variable Y with support $[0,1]$ is said to follow the ZOIS distribution, denoted by $Y \sim \text{ZOIS}(\lambda, \rho, \mu, \sigma^2)$, if its pdf is

$$\text{zois}(y; \lambda, \rho, \mu, \sigma^2) = \begin{cases} \lambda\rho^y(1-\rho)^{1-y}, & \text{if } y = 0, 1, \\ (1-\lambda) f_{\text{S}}(y; \mu, \sigma^2), & \text{if } y \in (0,1), \end{cases} \tag{2.9}$$

where $\lambda \in [0,1)$ is the mixture parameter, $\rho^y(1-\rho)^{1-y}$ denotes the pmf of the Bernoulli distribution with $\rho \in (0,1)$, and $f_{\text{S}}(\cdot; \mu, \sigma^2)$ denotes the pdf of the simplex distribution $S^-(\mu, \sigma^2)$. In particular, when $\lambda = 0$, the $\text{ZOIS}(\lambda, \rho, \mu, \sigma^2)$ distribution is reduced to the simplex distribution $S^-(\mu, \sigma^2)$.

Let $Z \sim \text{Bernoulli}(\lambda)$, $\eta \sim \text{Bernoulli}(\rho)$, $X \sim S^-(\mu, \sigma^2)$, and (Z, η, X) be mutually independent. Then, the random variable $Y \sim \text{ZOIS}(\lambda, \rho, \mu, \sigma^2)$ has the following *stochastic representation* (SR):

$$Y \overset{\text{d}}{=} Z\eta + (1-Z)X = \begin{cases} \eta, & \text{with probability } \lambda, \\ X, & \text{with probability } 1-\lambda. \end{cases} \tag{2.10}$$

Based on the SR (2.10), we easily obtain

$$\begin{aligned} \Pr(Y=0) &= \Pr(Z=1, \eta=0) = \lambda(1-\rho), \\ \Pr(Y=1) &= \Pr(Z=1, \eta=1) = \lambda\rho, \\ E(Y) &= \lambda\rho + (1-\lambda)E(X) = \lambda\rho + (1-\lambda)\mu, \\ E(Y^2) &= E(Z^2)E(\eta^2) + E[(1-Z)^2]E(X^2) \\ &\quad + E[Z(1-Z)]E(\eta)E(X) \\ &= \lambda\rho + (1-\lambda)E(X^2) \\ &= \lambda\rho + (1-\lambda)[\text{Var}(X) + \mu^2], \\ \text{Var}(Y) &= \lambda\rho(1-\rho) + \lambda(1-\lambda)(\rho-\mu)^2 \\ &\quad + (1-\lambda)\text{Var}(X), \end{aligned}$$

where $\text{Var}(X)$ is given by (2.3).

2.3.2 The second stochastic representation

Alternatively, after the reparameterization of $\lambda = \phi_0 + \phi_1$ and $\rho = \phi_1/(\phi_0 + \phi_1)$, the density (2.9) can be rewritten as

$$\text{zois}(y; \phi_0, \phi_1, \mu, \sigma^2) = \begin{cases} \phi_0, & \text{if } y = 0, \\ \phi_1, & \text{if } y = 1, \\ (1-\phi_0-\phi_1) f_{\text{S}}(y; \mu, \sigma^2), & \text{if } y \in (0,1), \end{cases} \tag{2.11}$$

where $\phi_0, \phi_1, \phi_0 + \phi_1 \in [0,1)$, ϕ_0 denotes the probability of the response being zeros, ϕ_1 denotes the probability of the response being ones, and $f_{\text{S}}(\cdot; \mu, \sigma^2)$ is given by (2.1). We denote the distribution by $Y \sim \text{ZOIS}(\phi_0, \phi_1, \mu, \sigma^2)$. In particular, when $\phi_0 = 0$, the ZOIS distribution is reduced to the *one-inflated simplex* (OIS) distribution (denoted by $\text{OIS}(\phi_1, \mu, \sigma^2)$); when $\phi_1 = 0$, the ZOIS distribution becomes the *zero-inflated simplex* (ZIS) distribution (denoted by $\text{ZIS}(\phi_0, \mu, \sigma^2)$); when $\phi_0 = \phi_1 = 0$, the ZOIS distribution becomes the original simplex distribution $S^-(\mu, \sigma^2)$.

Let $\mathbf{z} = (Z_0, Z_1, Z_2)^\top \sim \text{Multinomial}(1; \phi_0, \phi_1, 1-\phi_0-\phi_1)$, $X \sim S^-(\mu, \sigma^2)$, $\mathbf{z}$ and X be mutually independent (denoted by $\mathbf{z} \perp\!\!\!\perp X$). Then, the random variable $Y \sim \text{ZOIS}(\phi_0, \phi_1, \mu, \sigma^2)$ has the following SR:

$$\begin{aligned} Y &\overset{\text{d}}{=} Z_0 \cdot 0 + Z_1 \cdot 1 + Z_2 \cdot X = Z_1 + Z_2 X \\ &= \begin{cases} 0, & \text{with probability } \phi_0, \\ 1, & \text{with probability } \phi_1, \\ X, & \text{with probability } 1-\phi_0-\phi_1. \end{cases} \end{aligned} \tag{2.12}$$

The SR (2.12) means that $Y \sim \text{ZOIS}(\phi_0, \phi_1, \mu, \sigma^2)$ is a mixture of three distributions: Degenerate(0), Degenerate(1) and $S^-(\mu, \sigma^2)$.

3. LIKELIHOOD-BASED INFERENCE METHODS

3.1 MLEs of parameters via an MM algorithm

Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{ZOIS}(\lambda, \rho, \mu, \sigma^2)$ and $\{y_i\}_{i=1}^n$ be the realizations of $\{Y_i\}_{i=1}^n$. Furthermore, let $Y_{\text{obs}} = \{y_i\}_{i=1}^n$ denote the observed data and $\boldsymbol{\theta} = (\lambda, \rho, \mu, \sigma^2)^\top$ the unknown parameter vector. For the purpose of convenience, we define

$$\mathbb{I}_0 = \{i\colon y_i = 0,\ 1 \leqslant i \leqslant n\},\ \mathbb{I}_1 = \{i\colon y_i = 1,\ 1 \leqslant i \leqslant n\},$$

and $\mathbb{I}_2 = \{i\colon 0 < y_i < 1,\ 1 \leqslant i \leqslant n\}$. In addition, let $n_0 = \#\,\mathbb{I}_0$, $n_1 = \#\,\mathbb{I}_1$, and $m = n_0 + n_1$. In this paper, we assume that both $\mathbb{I}_0$ and $\mathbb{I}_1$ are not empty sets; i.e., n_0 and n_1 cannot be equal to 0. From (2.9), the likelihood function of $\boldsymbol{\theta}$ based on the observed-data is

$$\begin{aligned} L(\boldsymbol{\theta}|Y_{\text{obs}}) &= \left[\prod_{i\in\mathbb{I}_0} \lambda(1-\rho)\right] \times \left[\prod_{i\in\mathbb{I}_1} \lambda\rho\right] \\ &\quad \times \left[\prod_{i\in\mathbb{I}_2} (1-\lambda) f_{\text{S}}(y_i; \mu, \sigma^2)\right] \\ &= \lambda^m (1-\lambda)^{n-m} \cdot \rho^{n_1} (1-\rho)^{m-n_1} \cdot \prod_{i\in\mathbb{I}_2} f_{\text{S}}(y_i; \mu, \sigma^2), \end{aligned}$$

so that the log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= m\log(\lambda) + (n-m)\log(1-\lambda) \\ &\quad + n_1\log(\rho) + (m-n_1)\log(1-\rho) \\ &\quad + \sum_{i\in\mathbb{I}_2} \log[f_{\text{S}}(y_i; \mu, \sigma^2)]. \end{aligned}$$

Therefore, the MLEs of $\boldsymbol{\theta}$ are given by

$$\text{(3.1)} \quad \begin{cases} \hat\lambda &= \dfrac{m}{n}, \quad \hat\rho = \dfrac{n_1}{m}, \\ \hat\mu &= \arg\max\limits_{\mu\in(0,1)} \Big\{ -\log[D_{\mathbb{I}_2}(\mu|Y_{\text{obs}})] \Big\}, \\ \hat\sigma^2 &= \dfrac{1}{(n-m)\hat\mu^2(1-\hat\mu)^2} \sum\limits_{i\in\mathbb{I}_2} \dfrac{(y_i-\hat\mu)^2}{y_i(1-y_i)}, \end{cases}$$

where

$$D_{\mathbb{I}_2}(\mu|Y_{\text{obs}}) = \frac{1}{\mu^2(1-\mu)^2} \sum_{i\in\mathbb{I}_2} \frac{(y_i-\mu)^2}{y_i(1-y_i)}.$$

Let $\mu^{(t)}$ be the t-th approximate of the MLE $\hat\mu$ in the MM algorithm. From (2.6) and (2.7), we know that the $(t+1)$-th approximate $\mu^{(t+1)}$ can be obtained as the real root of the cubic equation

$$\text{(3.2)} \qquad a^{(t)}\mu^3 - (a^{(t)} + b^{(t)})\mu^2 + (b^{(t)} - 2)\mu + 1 = 0,$$

where

$$a^{(t)} = \frac{\sum_{i\in\mathbb{I}_2}[y_i(1-y_i)]^{-1}}{\sum\limits_{i\in\mathbb{I}_2} \dfrac{(y_i-\mu^{(t)})^2}{y_i(1-y_i)}} \quad \text{and} \quad b^{(t)} = \frac{\sum_{i\in\mathbb{I}_2}(1-y_i)^{-1}}{\sum\limits_{i\in\mathbb{I}_2} \dfrac{(y_i-\mu^{(t)})^2}{y_i(1-y_i)}}.$$

Alternatively, if we assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{ZOIS}(\phi_0, \phi_1, \mu, \sigma^2)$, then according to the invariance property of the maximum likelihood estimators, we know that the MLEs of ϕ_0 and ϕ_1 are given by

$$\hat\phi_0 = \hat\lambda(1-\hat\rho) = \frac{n_0}{n} \quad \text{and} \quad \hat\phi_1 = \hat\lambda\hat\rho = \frac{n_1}{n},$$

because the two densities (2.9) and (2.11) are totally same after the reparameterization of $\lambda = \phi_0 + \phi_1$ and $\rho = \phi_1/(\phi_0 + \phi_1)$.

3.2 Bootstrap confidence intervals

For small sample sizes, the bootstrap method is a useful tool to calculate a bootstrap *confidence interval* (CI) for an arbitrary function of $\boldsymbol{\theta} = (\lambda, \rho, \mu, \sigma^2)^\top$, say, $\vartheta = h(\boldsymbol{\theta})$. Let $\hat\vartheta = h(\hat{\boldsymbol{\theta}})$ denote the MLE of ϑ, where $\hat{\boldsymbol{\theta}} = (\hat\lambda, \hat\rho, \hat\mu, \hat\sigma^2)^\top$ are the MLEs of $\boldsymbol{\theta}$ calculated by means of (3.1). Based on the obtained MLEs $\hat{\boldsymbol{\theta}}$, by using the SR (2.10) we can generate $Y_1^* = y_1^*, \ldots, Y_n^* = y_n^* \overset{\text{iid}}{\sim} \text{ZOIS}(\hat\lambda, \hat\rho, \hat\mu, \hat\sigma^2)$. Having obtained $Y_{\text{obs}}^* = \{y_1^*, \ldots, y_n^*\}$, we can calculate the bootstrap replications $\hat{\boldsymbol{\theta}}^*$ and get $\hat\vartheta^* = h(\hat{\boldsymbol{\theta}}^*)$. Independently repeating this process G times, we obtain G bootstrap replications $\{\hat\vartheta_g^*\}_{g=1}^G$. Consequently, the standard error, $\text{se}(\hat\vartheta)$, of $\hat\vartheta$ can be estimated by the sample standard deviation of the G replications, i.e.,

$$\text{(3.3)} \quad \widehat{\text{se}}(\hat\vartheta) = \left\{ \frac{1}{G-1} \sum_{g=1}^{G} [\hat\vartheta_g^* - (\hat\vartheta_1^* + \cdots + \hat\vartheta_g^*)/G]^2 \right\}^{1/2}.$$

The bootstrap CI for ϑ is given by

$$\text{(3.4)} \qquad [\hat\vartheta_L,\ \hat\vartheta_U],$$

where $\hat\vartheta_L$ and $\hat\vartheta_U$ are the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of $\{\hat\vartheta^*\}_{g=1}^G$, respectively.

3.3 Zero-one-inflated simplex regression model

Suppose that we want to investigate the influence of some covariates on the probability (ϕ_0) of the response being zeros, the probability (ϕ_1) of the response being ones and the mean parameter μ. Based on the ZOIS distribution (2.11),

we consider the following ZOIS regression model:

(3.5)
$$\begin{cases} Y_i & \overset{\text{ind}}{\sim} & \text{ZOIS}(\phi_{0i}, \phi_{1i}, \mu_i, \sigma^2), \\ \log\left(\dfrac{\phi_{0i}}{1-\phi_{0i}-\phi_{1i}}\right) & = & \boldsymbol{u}_i^\top \boldsymbol{\alpha}, \\ \log\left(\dfrac{\phi_{1i}}{1-\phi_{0i}-\phi_{1i}}\right) & = & \boldsymbol{v}_i^\top \boldsymbol{\beta}, \\ \log\left(\dfrac{\mu_i}{1-\mu_i}\right) & = & \boldsymbol{x}_i^\top \boldsymbol{\gamma}, \end{cases}$$

where $i = 1, \ldots, n$, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^\top$, $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iq})^\top$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})^\top$ are covariate vectors for subject i and they are not necessarily identical; $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^\top$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)^\top$ are vectors of unknown parameters in the model and $p + q + r < n$. In addition, we assume that σ^2 is the same across all subjects. In practice, it is possible that $u_{i1} = v_{i1} = x_{i1} = 1$ so that $\{\alpha_1, \beta_1, \gamma_1\}$ denote intercepts.

The likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \sigma^2)^\top$ can be factorized into two parts:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \text{zois}(y_i; \phi_{0i}, \phi_{1i}, \mu_i, \sigma^2) = L_1(\boldsymbol{\theta}_1) L_2(\boldsymbol{\theta}_2),$$

where $\boldsymbol{\theta}_1 = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$, $\boldsymbol{\theta}_2 = (\boldsymbol{\gamma}^\top, \sigma^2)^\top$,

$$\begin{aligned} L_1(\boldsymbol{\theta}_1) &= \prod_{i=1}^{n} \phi_{0i}^{I_{\{0\}}(y_i)} \phi_{1i}^{I_{\{1\}}(y_i)} (1-\phi_{0i}-\phi_{1i})^{1-I_{\{0,1\}}(y_i)}, \\ L_2(\boldsymbol{\theta}_2) &= \prod_{i \in \mathbb{I}_2} f_{\text{s}}(y_i; \mu_i, \sigma^2), \end{aligned}$$

$I_{\mathbb{A}}(y)$ is the indicator function,

(3.6)
$$\begin{cases} \phi_{0i} & = & \dfrac{\exp(\boldsymbol{u}_i^\top \boldsymbol{\alpha})}{\Delta}, \\ \phi_{1i} & = & \dfrac{\exp(\boldsymbol{v}_i^\top \boldsymbol{\beta})}{\Delta}, \\ \mu_i & = & \dfrac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\gamma})}, \end{cases}$$

and $\Delta = 1 + \exp(\boldsymbol{u}_i^\top \boldsymbol{\alpha}) + \exp(\boldsymbol{v}_i^\top \boldsymbol{\beta})$. Thus, the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\theta}_1) + \ell_2(\boldsymbol{\theta}_2) = \sum_{i=1}^{n} \ell_1^*(\phi_{0i}, \phi_{1i}) + \sum_{i \in \mathbb{I}_2} \ell_2^*(\mu_i, \sigma^2),$$

where $\ell_2^*(\mu_i, \sigma_i^2) = \log[f_{\text{s}}(y_i; \mu_i, \sigma^2)]$ and

$$\begin{aligned} \ell_1^*(\phi_{0i}, \phi_{1i}) &= I_{\{0\}}(y_i) \log(\phi_{0i}) + I_{\{1\}}(y_i) \log(\phi_{1i}) \\ &\quad + [1 - I_{\{0,1\}}(y_i)] \log(1-\phi_{0i}-\phi_{1i}). \end{aligned}$$

Therefore, the MLEs of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be calculated separately. For the current situation, MM algorithms are not available. Fortunately, Zhang & Qiu (2014) provided an R package named "simplexreg" to calculate the MLEs of parameters in a simplex regression model, and we use this package to compute $\hat{\boldsymbol{\theta}}_2 = (\hat{\boldsymbol{\gamma}}^\top, \hat{\sigma}^2)^\top$.

To calculate the MLEs of $\boldsymbol{\theta}_1$, we first calculate the score function, which is given by

$$\nabla \ell_1(\boldsymbol{\theta}_1) = \frac{\partial \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} = \begin{pmatrix} \dfrac{\partial \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\alpha}} \\ \dfrac{\partial \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\beta}} \end{pmatrix},$$

where

$$\begin{aligned} \frac{\partial \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{n} \left[I_{\{0\}}(y_i) \boldsymbol{u}_i - \frac{\exp(\boldsymbol{u}_i^\top \boldsymbol{\alpha})}{\Delta} \boldsymbol{u}_i \right] \\ &= \sum_{i=1}^{n} \boldsymbol{u}_i [I_{\{0\}}(y_i) - \phi_{0i}] \\ \frac{\partial \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left[I_{\{1\}}(y_i) \boldsymbol{v}_i - \frac{\exp(\boldsymbol{v}_i^\top \boldsymbol{\beta})}{\Delta} \boldsymbol{v}_i \right] \\ &= \sum_{i=1}^{n} \boldsymbol{v}_i [I_{\{1\}}(y_i) - \phi_{1i}]. \end{aligned}$$

The Hessian matrix is

$$\nabla^2 \ell_1(\boldsymbol{\theta}_1) = \frac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^\top} = \begin{pmatrix} \dfrac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top} & \dfrac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}^\top} \\ \dfrac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^\top} & \dfrac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \end{pmatrix},$$

where

$$\begin{aligned} \frac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top} &= -\sum_{i=1}^{n} \frac{\exp(\boldsymbol{u}_i^\top \boldsymbol{\alpha})[1+\exp(\boldsymbol{v}_i^\top \boldsymbol{\beta})]}{\Delta^2} \boldsymbol{u}_i \boldsymbol{u}_i^\top \\ &= -\sum_{i=1}^{n} \phi_{0i}(1-\phi_{0i}) \boldsymbol{u}_i \boldsymbol{u}_i^\top, \\ \frac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= -\sum_{i=1}^{n} \frac{\exp(\boldsymbol{v}_i^\top \boldsymbol{\beta})[1+\exp(\boldsymbol{u}_i^\top \boldsymbol{\alpha})]}{\Delta^2} \boldsymbol{v}_i \boldsymbol{v}_i^\top \\ &= -\sum_{i=1}^{n} \phi_{1i}(1-\phi_{1i}) \boldsymbol{v}_i \boldsymbol{v}_i^\top, \\ \frac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}^\top} &= \sum_{i=1}^{n} \frac{\exp(\boldsymbol{u}_i^\top \boldsymbol{\alpha}) \exp(\boldsymbol{v}_i^\top \boldsymbol{\beta})}{\Delta^2} \boldsymbol{u}_i \boldsymbol{v}_i^\top \\ &= \sum_{i=1}^{n} \phi_{0i} \phi_{1i} \boldsymbol{u}_i \boldsymbol{v}_i^\top. \end{aligned}$$

Therefore, the Newtown–Raphson iteration

$$\boldsymbol{\theta}_1^{(t+1)} = \boldsymbol{\theta}_1^{(t)} - [\nabla^2 \ell_1(\boldsymbol{\theta}_1^{(t)})]^{-1} \nabla \ell_1(\boldsymbol{\theta}_1^{(t)}) \tag{3.7}$$

can be employed to calculate the MLEs of $\boldsymbol{\theta}_1$.

3.4 Model selection and goodness-of-fit tests

The proposed ZOIS model is a mixture of two distributions: The discrete part or zero-and-one inflated part (fitting the Bernoulli data 0's and 1's) and the continuous part (fitting the data in the open unit interval (0,1)). Ospina and Ferrari (2010, 2012) proposed the *zero-one-inflated beta* (ZOIB) distribution (denoted by $Y \sim \text{ZOIB}(\lambda, \rho, p, q)$) with pdf

$$\text{zoib}(y; \lambda, \rho, p, q) = \begin{cases} \lambda \rho^y (1-\rho)^{1-y}, & \text{if } y = 0, 1, \\ (1-\lambda) f_{\text{B}}(y; p, q), & \text{if } y \in (0,1), \end{cases} \tag{3.8}$$

where $\lambda \in [0,1)$ is the mixture parameter, $\rho^y(1-\rho)^{1-y}$ denotes the pmf of the Bernoulli distribution with $\rho \in (0,1)$, and $f_{\text{B}}(y; p, q)$ denotes the pdf of the beta random variable $Y^* \sim \text{Beta}(pq, (1-p)q)$, i.e.,

$$f_{\text{B}}(y; p, q) = \frac{\Gamma(q)}{\Gamma(pq)\Gamma((1-p)q)} y^{pq-1}(1-y)^{(1-p)q-1}, \tag{3.9}$$

where $0 < y < 1$, with $E(Y^*) = p \in (0,1)$ and $\text{Var}(Y^*) = p(1-p)/(1+q)$. Note that (3.8) is a mixture of the Bernoulli distribution and the beta distribution. Alternatively, after the parameterization of $\lambda = \phi_0 + \phi_1$ and $\rho = \phi_1/(\phi_0 + \phi_1)$, the density (3.8) can be rewritten as

$$\text{zoib}(y; \phi_0, \phi_1, p, q) = \begin{cases} \phi_0, & \text{if } y = 0, \\ \phi_1, & \text{if } y = 1, \\ (1-\phi_0-\phi_1) f_{\text{B}}(y; p, q), & \text{if } y \in (0,1), \end{cases} \tag{3.10}$$

where $\phi_0, \phi_1, \phi_0 + \phi_1 \in [0,1)$, ϕ_0 denotes the probability of the response being zeros, ϕ_1 denotes the probability of the response being ones, and $f_{\text{B}}(y; p, q)$ is given by (3.9). We denote the distribution by $Y \sim \text{ZOIB}(\phi_0, \phi_1, p, q)$. For the model selection, we would like to compare the ZOIS and ZOIB models via the *Akaike information criterion* (AIC). In addition, we use the *Kolmogorov–Smirnov* (KS) statistic and *Pearson's chi-squared* statistic for the goodness-of-fit tests of both the simplex and beta distributions for modeling the continuous data in $(0,1)$.

Similar to the ZOIS regression model (3.5), based on the ZOIB distribution (3.10), we consider the following ZOIB regression model:

$$\begin{cases} Y_i \overset{\text{ind}}{\sim} \text{ZOIB}(\phi_{0i}, \phi_{1i}, p_i, q), \\ \log\left(\dfrac{\phi_{0i}}{1-\phi_{0i}-\phi_{1i}}\right) = \boldsymbol{u}_i^\top \boldsymbol{\alpha}, \\ \log\left(\dfrac{\phi_{1i}}{1-\phi_{0i}-\phi_{1i}}\right) = \boldsymbol{v}_i^\top \boldsymbol{\beta}, \\ \log\left(\dfrac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\gamma}, \end{cases} \tag{3.11}$$

where $i = 1, \ldots, n$, $\boldsymbol{u}_i$, $\boldsymbol{v}_i$ and $\boldsymbol{x}_i$ are covariate vectors for subject i and they are not necessarily identical; $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of unknown parameters in the model. In addition, we assume that q is the same across all subjects. For more details on the beta regression and the corresponding parameter estimation, see Ferrari & Cribari-Neto (2004) and Cribari-Neto & Zeileis (2010). Moreover, based on the construction of the ZOIS/ZOIB distributions without covariate and the ZOIS/ZOIB models with covariates, we have observed that the fittings of the discrete data (i.e., zeros and ones) for the two models are identical, indicating that the estimates of λ, ρ, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the ZOIS models are the same as those in the ZOIB models.

4. SIMULATION STUDIES

To evaluate the finite sample performance of the proposed MLEs of $\boldsymbol{\theta}$ for both cases of without and with covariates, we conduct some Monte Carlo simulations. Let $\vartheta = h(\boldsymbol{\theta})$ be an arbitrary function of $\boldsymbol{\theta}$. The performance of the estimator $\hat{\vartheta}$ is assessed by the *mean square error* (MSE), defined by

$$\text{MSE}(\hat{\vartheta}) = E(\hat{\vartheta} - \vartheta)^2 = \text{Var}(\hat{\vartheta}) + [b(\hat{\vartheta}, \vartheta)]^2, \tag{4.1}$$

where $b(\hat{\vartheta}, \vartheta) = E(\hat{\vartheta}) - \vartheta$ denotes the bias of the estimator $\hat{\vartheta}$.

4.1 The case without covariates

To conduct the simulations, we consider the sample size $n = 100, 200, 500, 800, 1000$. The true values of parameters are set as $(\lambda, \rho, \mu, \sigma^2) = (0.2, 0.3, 0.5, 4)$, $(0.5, 0.2, 0.3, 9)$. Based on the SR (2.10), we independently generate

$$Y_1^{(k)}, \ldots, Y_n^{(k)} \overset{\text{iid}}{\sim} \text{ZOIS}(\lambda, \rho, \mu, \sigma^2)$$

for $k = 1, \ldots, K$ $(K = 1000)$.

For the k-th generated sample $Y_{\text{obs}}^{(k)} = \{Y_i^{(k)}\}_{i=1}^n$, the MLEs of $\boldsymbol{\theta} = (\lambda, \rho, \mu, \sigma^2)^\top$ can be calculated according to (3.1) and (3.2), denoted by $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\lambda}^{(k)}, \hat{\rho}^{(k)}, \hat{\mu}^{(k)}, \hat{\sigma}^{2(k)})^\top$. The MSE of each component in $\boldsymbol{\theta}$ is computed in terms of (4.1), denoted by $\text{MSE}(\hat{\lambda}^{(k)})$, $\text{MSE}(\hat{\rho}^{(k)})$, $\text{MSE}(\hat{\mu}^{(k)})$, $\text{MSE}(\hat{\sigma}^{2(k)})$, respectively, where all expectations are replaced by averages. The average MLE for each parameter

Table 1. The average MLE of each parameter and the average MSE of each MLE for the ZOIS distribution

n	Parameter	True value	A-MLE	A-MSE	True value	A-MLE	A-MSE
100			0.1997	0.0017		0.5001	0.0023
200			0.2015	0.0008		0.5012	0.0014
500	λ	0.2	0.2011	0.0003	0.5	0.4995	0.0005
800			0.1996	0.0002		0.5008	0.0003
1000			0.1998	0.0002		0.4991	0.0002
100			0.3049	0.0107		0.2001	0.0032
200			0.3016	0.0054		0.1989	0.0015
500	ρ	0.3	0.3003	0.0021	0.2	0.1989	0.0007
800			0.3019	0.0012		0.1990	0.0004
1000			0.3005	0.0010		0.1997	0.0003
100			0.5012	0.0004		0.3023	0.0012
200			0.5005	0.0002		0.2882	0.0005
500	μ	0.5	0.5003	0.0001	0.3	0.3027	0.0003
800			0.5001	0.0001		0.3023	0.0002
1000			0.4997	0.0000		0.3029	0.0001
100			3.9418	0.4169		8.7896	2.9848
200			3.9618	0.1978		9.0038	1.7063
500	σ^2	4	3.9974	0.0763	9	8.9636	0.6309
800			3.9981	0.0050		8.9925	0.4311
1000			3.9944	0.0043		8.9626	0.3355

A-MLE = Average MLE based on 1000 repetitions.
A-MSE = Average MSE based on 1000 repetitions.

based on the 1000 repetitions and the average MSE for each MLE based on the 1000 repetitions are reported in Table 1.

From Table 1, we have observed the following facts:

(a) For the given values of the four parameters $(\lambda, \rho, \mu, \sigma^2)$, as expected, the differences between the average MLE and its true value become smaller in tendency as the sample size n increases. In addition, the average MSEs of the estimators $\hat{\lambda}$, $\hat{\rho}$, $\hat{\mu}$ and $\hat{\sigma}^2$ also become smaller and smaller as the sample size n increases.

(b) For the given sample size n, the performance of the MLE $\hat{\mu}$ is the best in terms of model error. Furthermore, the performances of both $\hat{\lambda}$ and $\hat{\mu}$ are significantly better than those of $\hat{\rho}$ and $\hat{\sigma}^2$.

4.2 The case with covariates

The sample size n is set to be $100, 200, 500, 800, 1000$, and the ten parameters are set as $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^\top = (1, 0.5, -0.5)^\top, (1.5, 1, -1)^\top$; $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top = (1, 0.5, -0.5)^\top, (1.5, 1, -1)^\top$; $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^\top = (1.5, 0.5, -0.5)^\top, (1, -1, 0.5)^\top$; and $\sigma^2 = 4, 9$. The covariates are distributed as $u_{i1} = 1$, u_{i2}, $u_{i3} \overset{\text{iid}}{\sim} U(-1, 1)$; $v_{i1} = 1$, $v_{i2}, v_{i3} \overset{\text{iid}}{\sim} U(-1, 1)$; $x_{i1} = 1$, $x_{i2} \sim \text{Bernoulli}(0.5)$, $x_{i3} \sim U(0, 5)$. Let $\boldsymbol{u}_i = (u_{i1}, u_{i2}, u_{i3})^\top$, $\boldsymbol{v}_i = (v_{i1}, v_{i2}, v_{i3})^\top$ and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^\top$.

Based on the SR (2.12), we independently (for $k = 1, \ldots, K$ and $K = 1000$) generate

$$Y_i^{(k)} \overset{\text{ind}}{\sim} \text{ZOIS}(\phi_{0i}, \phi_{1i}, \mu_i, \sigma^2) \quad \text{for} \quad i = 1, \ldots, n,$$

where $(\phi_{0i}, \phi_{1i}, \mu_i)$ are determined by (3.6). For the k-th generated sample $Y_{\text{obs}}^{(k)} = \{Y_i^{(k)}\}_{i=1}^n$, the MLEs of $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2\}$ can be calculated according to (3.7) and the R package "simplexreg", denoted by $\hat{\boldsymbol{\theta}}^{(k)} = \{\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)}, \hat{\sigma}^{2(k)}\}$. The MSE of each component in $\boldsymbol{\theta}$ is computed in terms of (4.1), denoted by $\text{MSE}(\hat{\alpha}_j^{(k)})$, $\text{MSE}(\hat{\beta}_j^{(k)})$, $\text{MSE}(\hat{\gamma}_j^{(k)})$, $\text{MSE}(\hat{\sigma}^{2(k)})$, respectively, where $j = 1, 2, 3$. The average MLE for each parameter based on the 1000 repetitions and the average MSE for each MLE based on the 1000 repetitions are displayed in Table 2.

From Table 2, we have observed the following facts:

(a) For the given ten parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and σ^2, as expected, the performances of the MLEs become better and better as the sample size n increases. In addition, the MSEs of estimators $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}^2$ also become smaller and smaller as the sample size n increases.

(b) For the given sample size n, the performance of the MLE $\hat{\boldsymbol{\gamma}}$ is the best in terms of model error. Furthermore, the performances of $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are significantly better than that of $\hat{\sigma}^2$.

5. A REAL EXAMPLE

In this section, we first fit the *hospital stay* (HS) data of Barcelona in 1988 and 1990 with both ZOIS and ZOIB distributions without/with covariates to illustrate the proposed methods, then we compare the difference between the beta and the simplex distributions for only fitting the continuous

Table 2. The average MLE of each parameter and the average MSE of each MLE for the ZOIS regression model

n	Parameter	True value	A-MLE	A-MSE	True value	A-MLE	A-MSE
100			1.0148	0.0920		1.5385	0.1435
200			1.0146	0.0497		1.5150	0.5652
500	α_1	1	0.9965	0.0184	1.5	1.5100	0.0258
800			1.0060	0.0112		1.5042	0.0160
1000			1.0049	0.0094		1.5043	0.0127
100			0.5158	0.1518		1.0353	0.1839
200			0.4946	0.0698		1.0281	0.0865
500	α_2	0.5	0.5122	0.0263	1	1.0071	0.0322
800			0.5004	0.0171		1.0049	0.0218
1000			0.4963	0.0130		1.0047	0.0162
100			−0.5229	0.1582		−1.0548	0.1891
200			−0.5159	0.0701		−1.0284	0.0909
500	α_3	−0.5	−0.5160	0.0281	−1	−1.0133	0.0328
800			−0.4991	0.0167		−1.0033	0.0208
1000			−0.5021	0.0125		−1.0024	0.0171
100			1.0160	0.0898		1.5422	0.1488
200			1.0118	0.0501		1.5147	0.0660
500	β_1	1	0.9985	0.0174	1.5	1.5052	0.0259
800			1.0033	0.0103		1.5048	0.0163
1000			1.0053	0.0090		1.5078	0.0138
100			0.5170	0.1697		1.0344	0.1892
200			0.8098	0.0740		1.0233	0.0844
500	β_2	0.5	0.5067	0.0299	1	1.0123	0.0305
800			0.5081	0.0148		1.0110	0.0207
1000			0.5022	0.0138		1.0045	0.0154
100			−0.5207	0.1605		−1.0601	0.1962
200			−0.5243	0.0729		−1.0361	0.0963
500	β_3	−0.5	−0.5046	0.0270	−1	−1.0154	0.0332
800			−0.5000	0.0168		−1.0034	0.0203
1000			−0.5125	0.0137		−1.0042	0.0160
100			1.5399	0.1025		1.0413	0.1612
200			1.5233	0.0450		1.0001	0.0711
500	γ_1	1.5	1.5044	0.0174	1	1.0019	0.0269
800			1.5100	0.0102		1.0030	0.0160
1000			1.5055	0.0084		0.9994	0.0136
100			0.5048	0.0913		−1.0195	0.1124
200			0.4945	0.0344		−0.9941	0.0477
500	γ_2	0.5	0.4997	0.0146	−1	−0.9958	0.0190
800			0.4994	0.0087		−0.9991	0.0115
1000			0.4955	0.0070		−1.0006	0.0094
100			−0.5112	0.0087		0.4992	0.0120
200			−0.5070	0.0046		0.5040	0.0055
500	γ_3	−0.5	−0.5022	0.0016	0.5	0.5012	0.0028
800			−0.5012	0.0010		0.5007	0.0012
1000			−0.5003	0.0008		0.5021	0.0010
100			3.9744	1.1594		9.0900	6.2986
200			4.0412	0.5569		9.0981	3.0704
500	σ^2	4	3.9785	0.2081	9	9.0098	1.0516
800			4.0059	0.1242		9.0289	0.6670
1000			3.9967	0.0976		8.9609	0.5529

A-MLE = Average MLE based on 1000 repetitions.
A-MSE = Average MSE based on 1000 repetitions.

Table 3. 1988 HS data with 750 patients and some descriptive statistics

Length of stay (days)	Number of patients	Average inappropriate stay (days)	Some descriptive statistics
1	34	0	
2	109	0	
3	41	0.1	
4	42	0.6	Age of patients:
5	30	1	53.4±19.7
6	42	1.4	
7	52	1.4	Gender:
8	36	2	
9	44	2	Male 349 (47%)
10	23	2.7	Female 401 (53%)
11	22	2.7	
12	28	4.3	
13	21	4.2	
14	23	3.3	
15	22	3.5	
[16, 20]	61	5.2	
[21, 30]	68	9	
[31, 40]	24	14.3	
> 40	28	21.6	

Table 4. 1990 HS data with 750 patients and some descriptive statistics

Length of stay (days)	Number of patients	Average inappropriate stay (days)	Some descriptive statistics
1	76	0	
2	74	0.1	
3	45	0.4	
4	39	0.8	Age of patients:
5	34	0.9	55.3±19.5
6	39	1.5	
7	54	2	Gender:
8	40	2	
9	27	2.3	Males 321 (51%)
10	26	3.2	Females 346 (49%)
11	20	4.2	
12	16	4.8	
13	15	3.1	
14	14	1.4	
15	10	1.8	
[16, 20]	30	6.9	
[21, 30]	42	8.9	
[31, 40]	15	10.1	
> 40	17	17.7	

part (i.e., observations in (0,1)) to illustrate the goodness-of-fit tests.

5.1 The hospital stay data of Barcelona

Gange *et al.* (1996) reported a hospital stay data set containing 1383 patients from a study at the Hospital Universitari del Mar (a teaching hospital in Barcelona, Spain) in 1988 with 750 patients and in 1990 with 633 patients, respectively. Each patient was assessed for inappropriate stay on each day through two physicians by using the *appropriateness evaluation protocol* (AEP) method developed by Gertman and Restuccia (1981), see Gange *et al.* (1996) for more detail. The response variable Y is the number of inappropriate days relative to the total number of days that patients spent in the hospital, so Y is the proportion of inappropriate days out of all days spent in the hospital. Tables 3 and 4 list the corresponding HS data in 1988 (with 750 patients) and in 1990 (with 633 patients), and some descriptive statistics. From the two tables, we found out that with the increase of stay days, the average inappropriate stay days may increase too. Figure 1 plots the histograms and box-plots for the proportion of inappropriate stay data (the response Y) in 1988 and in 1990, respectively. From Figure 1, we can see that there are a lot of zeros and ones for the HS data in both 1988 and 1990.

Gange *et al.* (1996) used a logistic regression to model the proportion of inappropriate stay data with binomial and *beta–binomial* (BB) distributions, respectively. They found that the BB distribution provides a better fit to the data by modeling both its mean and dispersion as functions of explanatory variables. In this section, we would like to use the proposed zero-one-inflated simplex distribution $\mathrm{ZOIS}(\lambda, \rho, \mu, \sigma^2)$

(1) to model the proportion of inappropriate stay data in 1988 and 1990, respectively;
(2) to estimate the four parameters $(\lambda, \rho, \mu, \sigma^2)$ without considering covariates;
(3) to investigate the zero-one-inflated simplex regression by considering the effect of some covariates (e.g., sex, age and so on) on the response Y.

5.2 ZOIS and ZOIB distributions without covariates

First we fit the HS data in 1988 with both ZOIS and ZOIB distributions. Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{ZOIS}(\lambda, \rho, \mu, \sigma^2)$ and $\boldsymbol{\theta} = (\lambda, \rho, \mu, \sigma^2)^\top$. By employing the MM algorithm (3.1) and (3.2), we calculate the MLEs of $\boldsymbol{\theta}$ and these results are listed in the third column of Table 5. With $G = 1{,}000$ bootstrap replications, the estimated *standard deviation* (Std) and the 95% bootstrap CIs of each component in $\boldsymbol{\theta}$ are given in the fourth and fifth columns of Table 5. The AIC for the ZOIS distribution is reported in the last co1umn of Table 5. If we only fit the continuous data in $(0, 1)$ with the simplex distribution, the corresponding AIC is -74.868, see the last column of Table 5. Finally, in the goodness-of-fit tests, we use the Kolmogorov–Smirnov and Pearson's χ^2 statistics to model the continuous data in $(0, 1)$ with the simplex dis-

Table 5. MLEs and CIs of parameters without covariates for the HS data in 1988

Model	Parameter	MLE	Std	95% bootstrap CI	AIC
ZOIS	λ	0.6267	0.0177	[0.5933, 0.6600]	1143.326 (ZOIS)
	ρ	0.0638	0.0111	[0.0429, 0.0858]	
	μ	0.4757	0.0127	[0.4517, 0.5006]	−74.868 (Simplex)
	σ^2	6.6739	0.5503	[5.6159, 7.6820]	
ZOIB	p	0.4690	0.0137	[0.4433, 0.4969]	1149.874 (ZOIB)
	q	4.0043	0.3169	[3.4827, 4.7344]	−68.324 (Beta)
	Test	Simplex	Beta		
p-value	KS	0.1216	0.0322		
	Pearson's χ^2	0.2636	0.1618		

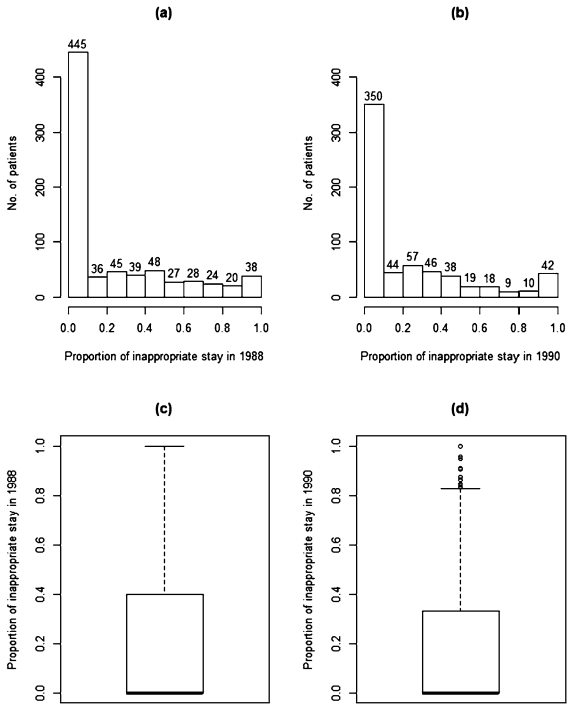

Figure 1. Comparison of histograms and box-plots for the proportion of inappropriate stay in 1988 and in 1990, respectively.

tributions, the resulting p-values are 0.1216 and 0.2636, respectively.

To compare ZOIS with ZOIB distributions, alternatively, we assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{ZOIB}(\lambda, \rho, p, q)$, the corresponding MLEs and CIs of parameters are also reported in Table 5. From the viewpoint of the AIC, the ZOIS (or simplex) distribution fit the data better than the ZOIB (or beta) distribution. On the one hand, at the 0.05 significant level, the p-values of the KS test show that the observations in $(0, 1)$ follow the simplex distribution, but do not follow the beta distribution. On the other hand, at the 0.05 significant level, the p-values of the Pearson's χ^2 test show that we cannot reject H_0: The observations in $(0, 1)$ follow both the simplex and beta distributions. Since the p-value for the simplex distribution is larger than that for the beta distribution, the simplex distribution fits the data in $(0, 1)$ better than the beta distribution.

Next we fit the HS data in 1990 with both ZOIS and ZOIB distributions. Similarly, we display these results in Table 6. Based on the values of AIC, we can see that the ZOIS (or simplex) distribution might not be a good choice for the HS data in 1990. The KS test has the p-value of 8.28×10^{-8} for the simplex distribution, and 1.46×10^{-6} for the beta distribution. Pearson's χ^2 test has the p-value of 0.0009 for the simplex distribution, and 0.0039 for the beta distribution. Therefore, at the 0.05 level of significance, both distributions do not fit the data well.

Figure 2(a) and Figure 2(b) compared three histograms among the observed (black bar), estimated proportion of inappropriate stay with the ZOIS distribution (grey bar) and the ZOIB distribution (white bar) in 1988 (left) and 1990 (right), respectively. The observed proportions are very close to the estimated proportions fitted by both the ZOIS and ZOIB distributions in 1988 and 1990. Figure 2(c) and Figure 2(d) compared the empirical distribution function with the estimated cumulative ZOIS and ZOIB distribution functions based the HS data in 1988 and 1990, respectively. From Figure 2(c), we can see that both the ZOIS and ZOIB distribution are suitable for fitting the HS data in 1988. However, for the HS data in 1990, Figure 2(d) indicates that neither ZOIS nor ZOIB fitted data very well, and these results are consistent with the KS test and Pearson's χ^2 test.

5.3 ZOIS and ZOIB regression models

First we fit the HS data in 1988 with both ZOIS and ZOIB regression models. We consider three covariates: x_1 is the gender of patient (= 0 if male, = 1 if female); x_2 is the age of the patient in years; and x_3 (los, i.e., length of stay) is the total number of days patients spent in hospital. Again, let the response variable Y_i (HS) be the number of inappropriate days of the patient i out of the total number of days that patients spent in hospital, i.e., the proportion of inappropriate days out of all days spent in the hospital. According to (3.5), we consider the following ZOIS regression

Table 6. MLEs and CIs of parameters without covariates for the HS data in 1990

Model	Parameter	MLE	Std	95% bootstrap CI	AIC
ZOIS	λ	0.5703	0.0196	[0.5308, 0.6082]	1010.075 (ZOIS)
	ρ	0.1053	0.0164	[0.0764, 0.1395]	
	μ	0.3988	0.0095	[0.3920, 0.4290]	−101.844 (Simplex)
	σ^2	7.8180	0.6650	[6.5083, 9.1458]	
ZOIB	p	0.3723	0.0127	[0.3477, 0.3961]	1004.759 (ZOIB)
	q	4.1907	0.3330	[3.6220, 4.9473]	−107.154 (Beta)
	Test	Simplex	Beta		
p-value	KS	8.28×10^{-8}	1.46×10^{-6}		
	Pearson's χ^2	0.0009	0.0039		

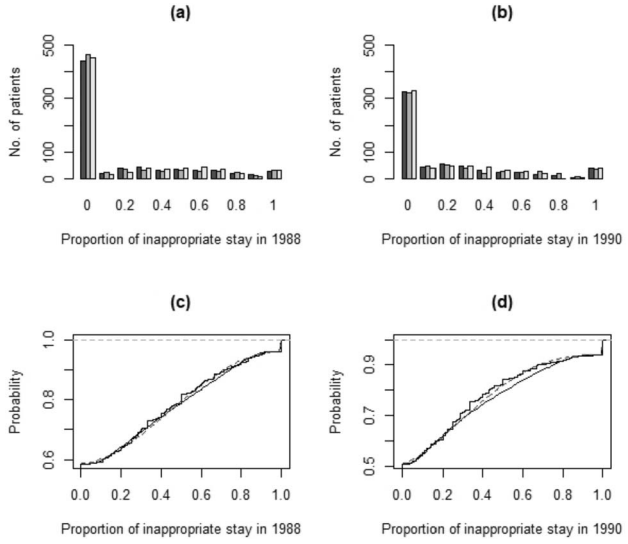

Figure 2. (a)(b) Comparison of histograms among the observed (black bar), estimated proportion of inappropriate stay with the ZOIS distribution (grey bar) and the ZOIB distribution (white bar) in 1988 (left) and 1990 (right), respectively; (c)(d) The horizontal step functions from the empirical distribution functions of hospital stay sample, the curves are estimated cumulative ZOIS distribution functions (black line) and ZOIB distribution functions (red dash line) in 1988 (left) and 1990 (right), respectively.

model:

$$\begin{cases} Y_i \overset{\text{ind}}{\sim} \text{ZOIS}(\phi_{0i}, \phi_{1i}, \mu_i, \sigma^2), \\ \log\left(\dfrac{\phi_{0i}}{1-\phi_{0i}-\phi_{1i}}\right) = \alpha_0 + x_{i1}\alpha_1 + x_{i2}\alpha_2 + x_{i3}\alpha_3, \\ \log\left(\dfrac{\phi_{1i}}{1-\phi_{0i}-\phi_{1i}}\right) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3, \\ \log\left(\dfrac{\mu_i}{1-\mu_i}\right) = \gamma_0 + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3, \end{cases}$$

where $i = 1, \ldots, n$. We employ the Newton–Raphson algorithm (3.7) and the R package "simplxreg" to calculate the MLEs of the regression coefficients $\{\alpha_j, \beta_j, \gamma_j\}_{j=0}^3$ and parameter σ^2, and these results are displayed in the second column of Table 7. With $G = 1{,}000$ bootstrap replications, the estimated Std and the 95% bootstrap CIs of all parameters are given in the third and fourth columns of Table 7. The AIC for the ZOIS regression model is 1029.587, which is also reported in Table 7. If we only fit the continuous data in (0,1) with the simplex regression model, the corresponding AIC is −84.4164.

To compare ZOIS with ZOIB regression models, based on (3.11) we also consider the following ZOIB regression model:

$$\begin{cases} Y_i \overset{\text{ind}}{\sim} \text{ZOIB}(\phi_{0i}, \phi_{1i}, p_i, q), \\ \log\left(\dfrac{\phi_{0i}}{1-\phi_{0i}-\phi_{1i}}\right) = \alpha_0 + x_{i1}\alpha_1 + x_{i2}\alpha_2 + x_{i3}\alpha_3, \\ \log\left(\dfrac{\phi_{1i}}{1-\phi_{0i}-\phi_{1i}}\right) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3, \\ \log\left(\dfrac{p_i}{1-p_i}\right) = \gamma_0 + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3, \end{cases}$$

where $i = 1, \ldots, n$ and the corresponding MLEs, CIs of the regression coefficients $\{\alpha_j, \beta_j, \gamma_j\}_{j=0}^3$ and the parameter q, and the values of AIC are also reported in Table 7.

From the viewpoint of the AIC, the ZOIS (or simplex) model fit the data better than the ZOIB (or beta) distribution. From Table 7, we could see that the x_3 (los) has negative effect on ϕ_{0i} (see, the MLE of α_3). Moreover, with the increase of age, the proportion of inappropriate stay days becomes larger in 1988 (see, the MLE of γ_2), indicating that older patients may spend much time in hospital. In addition, there is no difference for male and female about inappropriate stay days.

Next we fit the HS data in 1990 with both ZOIS and ZOIB regression models. Similarly, we display the corresponding results in Table 8. According to the values of AIC, we can see that the ZOIS (or simplex) model and the ZOIB (or beta) model have no difference for fitting the HS data in 1990. From Table 8, we could see that both the age and

Table 7. MLEs and CIs of regression coefficients for the HS data in 1988

Coefficient	MLE	Std	95% bootstrap CI
α_0	1.5155	0.2787	$[1.0143, 2.0649]^*$
α_1	0.3361	0.1708	$[-0.0093, 0.6722]$
α_2	−0.0057	0.0045	$[-0.0147, 0.0028]$
α_3	−0.0774	0.0094	$[-0.0979, -0.0603]^*$
β_0	−1.5618	0.6826	$[-2.9954, -0.3325]^*$
β_1	0.4716	0.3924	$[-0.2403, 1.2812]$
β_2	−0.0027	0.0106	$[-0.0234, 0.0184]$
β_3	−0.0606	0.0257	$[-0.1243, -0.0241]^*$
Simplex			
γ_0	−0.7223	0.1834	$[-1.0956, -0.3601]^*$
γ_1	−0.1392	0.1035	$[-0.3419, 0.0573]$
γ_2	0.0091	0.0027	$[0.0042, 0.0145]^*$
γ_3	0.0064	0.0036	$[-0.0008, 0.0135]$
σ^2	6.4042	0.5389	$[5.3580, 7.4924]^*$
AIC	1029.587 (ZOIS)	−84.4164 (Simplex)	
Beta			
γ_0	−0.6977	0.1845	$[-1.0457, -0.3298]^*$
γ_1	−0.1776	0.1007	$[-0.3866, 0.0178]$
γ_2	0.0082	0.0026	$[0.0030, 0.0129]^*$
γ_3	0.0083	0.0036	$[0.0013, 0.0154]^*$
q	4.2724	0.3406	$[3.7505, 5.0874]^*$
AIC	1033.796 (ZOIB)	−80.2038 (Beta)	

*Indicating that the CI does not include the zero value.

Table 8. MLEs and CIs of regression coefficients for the HS data in 1990

Coefficient	MLE	Std	95% bootstrap CI
α_0	2.3062	0.3423	$[1.6370, 3.0276]^*$
α_1	0.1017	0.1903	$[-0.2570, 0.4909]$
α_2	−0.0197	0.0050	$[-0.0300, -0.0109]^*$
α_3	−0.1145	0.0152	$[-0.1487, -0.0888]^*$
β_0	−0.9627	0.6403	$[-2.324, 0.2185]$
β_1	−0.0283	0.3469	$[-0.7222, 0.6351]$
β_2	−0.0062	0.0095	$[-0.0240, 0.0127]$
β_3	−0.0562	0.0250	$[-0.1184, -0.0202]^*$
Simplex			
γ_0	−0.8810	0.2107	$[-1.3037, -0.4465]^*$
γ_1	0.1483	0.1166	$[-0.0756, 0.3729]$
γ_2	0.0030	0.0032	$[-0.0033, 0.0091]$
γ_3	0.0078	0.0047	$[-0.0015, 0.0172]$
σ^2	7.6927	0.6715	$[6.3983, 9.0525]^*$
AIC	882.8492 (ZOIS)	−104.2362 (Simplex)	
Beta			
γ_0	−0.7781	0.2124	$[-1.2107, -0.3502]^*$
γ_1	0.1462	0.1099	$[-0.0655, 0.3594]$
γ_2	0.0018	0.0031	$[-0.0042, 0.0079]$
γ_3	0.0047	0.0046	$[-0.0045, 0.0138]$
q	4.2463	0.3347	$[3.7350, 5.0196]^*$
AIC	882.4392 (ZOIB)	−104.6462 (Beta)	

*Indicating that the CI does not include the zero value.

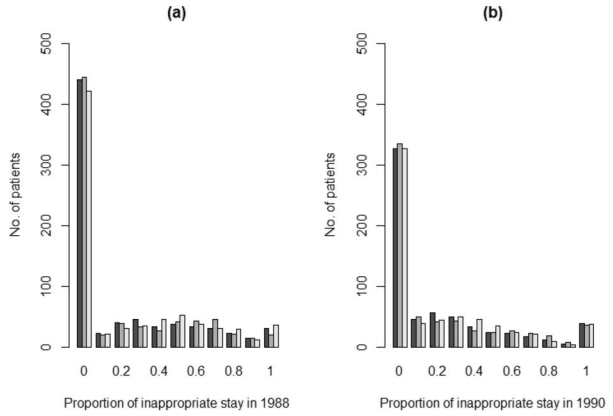

Figure 3. (a)(b) Comparison of histograms among the observed (black bar), estimated proportion of inappropriate stay with the ZOIS regression model (grey bar) and the ZOIB regression model (white bar) based on the HS data in 1988 (left) and 1990 (right), respectively.

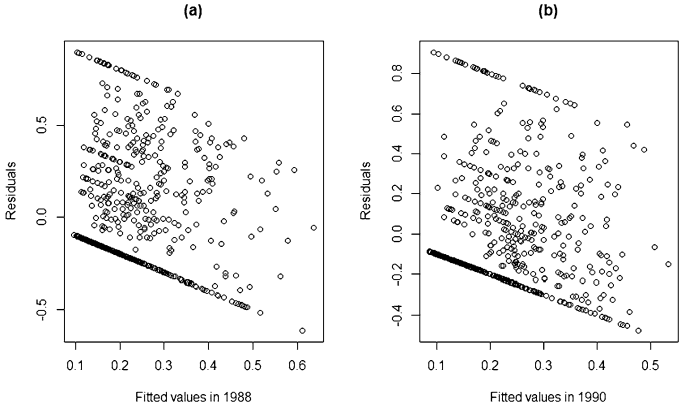

Figure 4. (a)(b)Ordinary residuals against the fitted values for the ZOIS regression model based on the HS data in 1988 and 1990, respectively.

total length of stay have a significant impact on ϕ_{0i} (see, the MLEs of α_2 and α_3), and total length of stay has an impact on ϕ_{1i} (see, the MLE of β_3). However, there is no obvious relation between the continuous part of HS data in 1990 and the three factors, which can be interpreted partially by the fact that neither the ZOIS nor ZOIB models fitted the HS data in 1990 very well as showed in Figure 2(d).

Figure 3(a) and Figure 3(b) compared three histograms among the observed (black bar), estimatd proportion of inappropriate stay with the ZOIS regression model (grey bar) and the ZOIB regression model (white bar) based the HS data in 1988 (left) and 1990 (right), respectively. Obviously, the observed proportions are very close to the estimated proportions fitted by the ZOIS regression model in both 1988 and 1990, indicating that the ZOIS regression model is suitable for fitting the hospital stay data.

Figure 4(a) and Figure 4(b) plot the ordinary residuals against the fitted values for the ZOIS regression model based the HS data in 1988 and 1990, respectively. The two residual plots in Figure 4 do not suggest a lack of fit. Residuals are randomly scatted in the parallelogram, since hospital data are from $[0, 1]$, then $|residuals + fitted\ values| \leqslant 1$. Moreover, for the zero and one inflated data (i.e., the discrete data 0's and 1's), the lower and upper bounds, corresponding to responses equal to zero and one, respectively, are typical of data with only two outcomes. There are similar results in Ospina and Ferrari (2012).

6. DISCUSSION

As a mixture of the Bernoulli distribution (or two degenerate distributions at zero and at one) and the simplex distribution, the proposed ZOIS distribution provides a new tool to analyze continuous proportional data with excessive zeros and excessive ones. We also developed the ZOIS regression models, which allow us to explore the relationship between a set of covariates with the probabilities of observing zero and one values, and the mean of the continuous responses in (0,1). The algorithms for calculating MLEs of parameters and the bootstrapping method for constructing CIs of parameters are given.

Since the observations in the closed unit interval $[0, 1]$ can be decomposed into two parts, the discrete part (the Bernoulli data 0's and 1's) and the continuous part (the data in the open unit interval $(0, 1)$), both the simplex and beta distributions could be used to model the continuous data in $(0, 1)$. Then, both the ZOIS and ZOIB models can be employed to model continuous proportional data with excessive zeros and excessive ones. For the case of without covariates, in Section 5.2, we utilized the KS test, the Pearsion χ^2 test and the AIC to compare the ZOIS and ZOIB distributions. The results showed that the ZOIS is better than the ZOIB distributions in fitting the HS data in 1988, while neither the ZOIS nor the ZOIB are suitable for fitting the HS data in 1990. For the regression models in Section 5.3, AIC indicated that the ZOIS (or simplex) model is more suitable than the ZOIB (or beta) model for fitting the HS data in 1988, while the two models are not much different for fitting the HS data in 1990. Overall, for the HS data, the ZOIS model is a better choice.

In the ZOIS regression model (3.5), we assumed that σ^2 is the same across all subjects. In fact, we could also consider the effect of some covariates on the dispersion parameter σ_i^2 in the future's research. In addition, we did not discuss the problem of variable selection in the ZOIS regression models. Finally, the testing hypotheses under large sample sizes in the ZOIS model for the one-sample and/or the two-sample tests are also our interest in the future.

APPENDIX A. RANDOM VARIABLE GENERATION FROM THE SIMPLEX DISTRIBUTION

A.1 The inverse Gaussian distribution and its generation

A positive random variable X follows the inverse Gaussian (or Wald) distribution with mean parameter $\mu > 0$ and shape parameter $\lambda > 0$, denoted by $X \sim \text{IGaussian}(\mu, \lambda)$, if it has pdf

$$\text{IGaussian}(x|\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left[-\frac{\lambda D(x;\mu)}{2}\right], \tag{A.1}$$

for $x > 0$, where

$$D(x;\mu) \doteq \frac{(x-\mu)^2}{\mu^2 x}. \tag{A.2}$$

An important result (Shuster, 1968) on $X \sim \text{IGaussian}(\mu, \lambda)$ is $\lambda D(X;\mu) \sim \chi^2(1)$, which can be used to generate N i.i.d. samples from the inverse Gaussian distribution. The generation procedure is as follows:

Step 1. Draw $U \sim U(0,1)$ and independently draw $Y \sim \chi^2(1)$;

Step 2. Set $X_1 = \mu + \frac{\mu^2 Y}{2\lambda} - \frac{\mu}{2\lambda}\sqrt{4\mu\lambda Y + \mu^2 Y^2}$ and $X_2 = \frac{\mu^2}{X_1}$;

Step 3. If $U \leqslant \mu/(\mu + X_1)$, return $X = X_1$, else return $X = X_2$.

The corresponding `R` code for generating $X \sim \text{IGaussian}(\mu, \lambda)$ is given by

```
function(N, mu, lambda)
{   # Function name: rigaussian(N, mu, lambda)
# -------------- Aim ------------------------------------
# Generate N i.i.d. samples of x ~ IGaussianDE(mu, lambda)
# with density given by (A.1)
# -------------- Input ----------------------------------
# N      = sample size
# mu     = mean parameter
# lambda = shape parameter
# -------------- Output ---------------------------------
# x_1, ..., x_N ~iid IGaussianDE(mu, lambda)
#################################################################
y <- rchisq(N, 1)
a <- (mu^2/(2 * lambda)) * y
b <- 4 * mu * lambda * y + mu^2 * y^2
x1 <- mu + a - (mu/(2 * lambda)) * sqrt(b)
u <- runif(N)
x <- rep(0, N)
for(i in 1:N) {
if(u[i] < mu/(mu + x1[i])) { x[i] <- x1[i] }
else { x[i] <- mu^2/x1[i] }
}
return(x)
}
```

For the sake of convenience, in this paper, we alternatively denote the inverse Gaussian distribution $X \sim \text{IGaussian}(\mu, 1/\sigma^2)$ by $X \sim \text{IG}(\mu, \sigma^2)$ with density function

$$\text{IG}(x|\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2 x^3}} \exp\left[-\frac{D(x;\mu)}{2\sigma^2}\right], \quad x > 0, \tag{A.3}$$

where $\sigma^2\ (> 0)$ is called scale parameter.

A.2 The inverse Gaussian mixture distribution and its generation

Let $X_1 \sim \text{IG}(\mu, \sigma^2)$, $X_2^{-1} \sim \text{IG}(\mu^{-1}, \sigma^2\mu^2)$, and $X_1 \perp\!\!\!\perp X_2$. The random variable X_2 is called the complementary reciprocal of X_1. Define a new r.v. Y as the mixture of the inverse Gaussian r.v. with its complementary reciprocal; i.e.,

$$Y = \begin{cases} X_1, & \text{with probability } 1-p, \\ X_2, & \text{with probability } p, \end{cases} \tag{A.4}$$

where $p \in [0,1]$. The distribution of Y is called the inverse Gaussian mixture distribution (Jørgensen *et al.*, 1991), denoted by $Y \sim \text{M-IG}(\mu, \sigma^2, p)$, and its pdf is given by

$$\text{M-IG}(y|\mu,\sigma^2,p) = \sqrt{\frac{1}{2\pi\sigma^2 y^3}}\left(1 - p + \frac{py}{\mu}\right) \times \exp\left[-\frac{D(y;\mu)}{2\sigma^2}\right],$$

where $y > 0$.

Note that (A.4) can be rewritten as

$$Y \overset{\text{d}}{=} (1-Z)X_1 + ZX_2, \tag{A.5}$$

where $Z \sim \text{Bernoulli}(p)$ and (Z, X_1, X_2) are mutually independent. Therefore, the SR (A.5) provides a procedure for generating random samples from $Y \sim \text{M-IG}(\mu, \sigma^2, p)$. Furthermore, Jørgensen *et al.* (1991) also obtained the following SR:

$$Y \overset{\text{d}}{=} X_1 + ZX_3, \tag{A.6}$$

where $Z \sim \text{Bernoulli}(p)$, $X_3 \sim \sigma^2\mu^2\chi^2(1)$ and (X_1, Z, X_3) are mutually independent. In this paper, we use the SR (A.6) rather than (A.5) to generate random samples from $Y \sim \text{M-IG}(\mu, \sigma^2, p)$.

A.3 The simplex distribution and its generation

Let $X \sim S^-(\mu, \sigma^2)$ and make a one-to-one transformation $Y = X/(1-X)$. It is easy to show that (see, Zhang & Qiu, 2014)

$$Y \sim \text{M-IG}\left(\frac{\mu}{1-\mu}, \sigma^2(1-\mu)^2, \mu\right). \tag{A.7}$$

Therefore, for a given pair (μ, σ^2) with $\mu \in (0,1)$ and $\sigma^2 > 0$, we first generate $Y = y$ from (A.7), and solve the inverse

transformation $x = y/(1+y)$, then $X = x$ is a random sample from $X \sim S^{-}(\mu, \sigma^2)$.

ACKNOWLEDGEMENTS

G.L. Tian's research was supported by a grant from the National Natural Science Foundation of China (No. 11771199). K.C. Yuen's research was supported by a Seed Fund for Basic Research of the University of Hong Kong (Project Code: 201711159190). L.C. Wu's research was supported by a grant from the National Natural Science Foundation of China (No. 11861041).

Received 9 May 2018

REFERENCES

[1] Barndorff-Nielsen, O.E. and Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis* **39**(1), 106–116. MR1128675

[2] Becker, M.P., Yang, I. and Lange, K. (1997). EM algorithms without missing data. *Statistical Methods in Medical Research* **6**, 38–54.

[3] Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software* **34**(2), 1–24.

[4] Espinheira, P.L., Ferrari, S.L.P. and Cribari-Neto, F. (2008a). Influence diagnostics in beta regression. *Computational Statistics & Data Analysis* **52**(9), 4417–4431. MR2432471

[5] Espinheira, P.L., Ferrari, S.L.P. and Cribari-Neto, F. (2008b). On beta regression residuals. *Journal of Applied Statistics* **35**(4), 407–419. MR2420486

[6] Ferrari, S.L.P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31**(7), 799–815. MR2095753

[7] Ferrari, S.L.P. and Pinheiro, E.C. (2011). Improved likelihood inference in beta regression. *Journal of Statistical Computation and Simulation* **81**(4), 431–443. MR2782138

[8] Gange, S.J., Muñoz, A., Sáez, M. and Alonso, J. (1996). Use of the beta–binomial distribution to model the effect of policy changes on appropriateness of hospital stays. *Applied Statistics* **45**(3), 371–382.

[9] Gertman, P.M. and Restuccia, J.D. (1981). The appropriateness evaluation protocol: A technique for assessing unnecessary days of hospital care. *Medical Care* **19**(8), 855–871.

[10] Hunter, D.R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician* **58**(1), 30–37. MR2055509

[11] Johnson, N., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions* (Second Edition). John Wiley and Sons, New York. MR1299979

[12] Jørgensen, B., Seshadri, V. and Whitmore, G.A. (1991). On the mixture of the inverse Gaussian distribution with its complementary reciprocal. *Scandinavian Journal of Statistics* **18**(1), 77–89. MR1115184

[13] Kieschnick, R. and McCullough, B.D. (2003). Regression analysis of variates observed on $(0,1)$: percentages, proportions and fractions. *Statistical Modelling* **3**(3), 193–213. MR2005473

[14] Korhonen, L., Korhonen, K.T., Stenberg, P.T., Maltamo, M. and Rautiainen, M. (2007). Local models for forest canopy cover with beta regression. *Silva Fennica* **41**(4), 671–685.

[15] Lange, K., Hunter, D.R and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussions). *Journal of Computational and Graphical Statistics* **9**, 1–20. MR1819865

[16] Ospina, R. and Ferrari, S.L.P. (2010). Inflated beta distributions. *Statistical Papers* **51**(1), 111–126. MR2556590

[17] Ospina, R. and Ferrari, S.L.P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis* **56**(6), 1609–1623. MR2892364

[18] Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective*. Advanced Series on Statistical Science & Applied Probability: Volume 4. World Scientific Publishing, Singapore. MR1476674

[19] Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis* **9**(4), 325–346.

[20] Qiu, Z.G., Song, P.X.-K. and Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics* **35**(4), 577–596. MR2468863

[21] Quintero, F.O.L. and Contreras-Reyes, J.E. (2018). Estimation for finite mixture of simplex models: applications to biomedical data. *Statistical Modelling* **18**(2) 129–148. MR3770127

[22] Shuster, J. (1968). On the inverse Gaussian distribution function. *Journal of the American Statistical Association* **63**(324), 1514–1516. MR0235653

[23] Simas, A.B., Barreto-Souza, W. and Rocha, A.V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis* **54**(2), 348–366. MR2756431

[24] Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum likelihood-regression with beta-distributed dependent variables. *Psychological Methods* **11**(1), 54–71.

[25] Song, P.X.-K. and Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics* **56**(2), 496–502.

[26] Song, P.X.-K., Qiu, Z.G. and Tan, M. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal* **46**(5), 540–553. MR2101142

[27] Zhang, P. and Qiu, Z.G. (2014). Regression analysis of proportional data using simplex distribution. *Scientia Sinica Mathematica (in Chinese)* **44**(1), 89–104.

[28] Zhang, W.Z. and Wei, H.J. (2008). Maximum likelihood estimation of simplex distribution nonlinear mixed models via the stochastic approximation algorithm. *Rocky Mountain Journal of Mathematics* **38**(5), 1863–1875. MR2457391

[29] Zhao, Y.Y., Xu, D.K., Duan, X.D. and Dai, L. (2014). Bayesian estimation of simplex distribution nonlinear mixed models for longitudinal data. *International Journal of Applied Mathematics and Statistics* **52**(3), 1–10. MR3224062

Pengyi Liu
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road, Hong Kong
P. R. China
E-mail address: 11750003@mail.sustech.edu.cn

Kam Chuen Yuen
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road, Hong Kong
P. R. China
E-mail address: kcyuen@hku.hk

Liu-Cang Wu
Faculty of Science
Kunming University of Science and Technology
Kunming 650093, Yunnan Province
P. R. China
E-mail address: wuliucang@163.com

Guo-Liang Tian
Department of Statistics and Data Science
Southern University of Science and Technology
Shenzhen 518055, Guangdong Province
P. R. China
E-mail address: tiangl@sustc.edu.cn

Tao Li
Department of Mathematics
Southern University of Science and Technology
Shenzhen 518055, Guangdong Province
P. R. China
E-mail address: lit6@sustc.edu.cn