Author/s:
Zhang, P;Calderin, E;Li, S;Wu, X

Title:
On the Type I multivariate zero-truncated hurdle model with applications in health insurance

Date:
2020-01-01

Citation:
Zhang, P., Calderin, E., Li, S. & Wu, X. (2020). On the Type I multivariate zero-truncated hurdle model with applications in health insurance. Insurance: Mathematics and Economics, 90, pp.35-45. https://doi.org/10.1016/j.insmatheco.2019.10.010.

Persistent Link:
https://hdl.handle.net/11343/253886

On the Type I multivariate zero-truncated hurdle model with applications in health insurance

Pengcheng Zhang, Enrique Calderin, Shuanming Li, Xueyuan Wu*

Department of Economics, The University of Melbourne, VIC 3010, AUS

Abstract

In the general insurance modeling literature, there has been a lot of work based on univariate zero-truncated models, but little has been done in the multivariate zero-truncation cases, for instance a line of insurance business with various classes of policies. There are three types of zero-truncation in the multivariate setting: only records with all zeros are missing, zero counts for one or some classes are missing, or zeros are completely missing for all classes. In this paper, we focus on the first case, the so-called Type I zero-truncation, and a new multivariate zero-truncated hurdle model is developed to study it. The key idea of developing such a model is to identify a stochastic representation for the underlying random variables, which enables us to use the EM algorithm to simplify the estimation procedure. This model is used to analyze a health insurance claims dataset that contains claim counts from different categories of claims without common zero observations.

Keywords: Type I multivariate zero-truncation; hurdle model; EM algorithm; health insurance modeling

1 Introduction

Counting data without zero observations are common in insurance. An example is the data sets managed by the claims departments in insurance com-

*Corresponding author. Email: xueyuanw@unimelb.edu.au.

panies which could only have records for policyholders who have positive claim numbers. Practitioners use zero-truncated models to address this issue of missing information. The commonly used univariate zero-truncated models include zero-truncated Poisson (ZTP) model and zero-truncated negative binomial (ZTNB) model (see, e.g., Cameron and Trivedi, 2013).

However, univariate zero-truncated models have obvious limitations, one of which arises from policies with multiple claim types. This is commonly seen in the health insurance industry, where policyholders can make claims on the grounds of disease, accidents or other. The corresponding claims data will usually only contain policies with at least one claim recorded. This type of zero truncation is referred to as Type I multivariate zero-truncation. Throughout this paper, Type I multivariate zero-truncation indicates that there are no observations with all zeros for the variables under consideration. One key characteristic of this type of multivariate zero-truncation is that there is no zero-truncation in respect of any individual variable under consideration, instead, we have zero-modified data for each one of them. This is why univariate zero-truncated models are not directly applicable here.

In the literature, there are a few papers discussing the modeling using Type I zero-truncation data and most of them focused on the bivariate cases (Charalambides, 1984; Jung et al., 2007; Piperigou and Papageorgiou, 2003). However, it is not straightforward to extend the results to multivariate cases and incorporate covariates. Recently, Tian et al. (2018b) proposed a Type I multivariate zero-truncated Poisson distribution to fit some count data of this type. This model started with m independent Poisson distributions, and then used a Bernoulli latent variable to truncate common zeros. A drawback of the model is the restricted choices for the marginal behaviours. To address the issue, we propose the Type I multivariate zero-truncated hurdle model, which allows us to assume different zero count patterns for individual variables and offers us flexible options to describe marginal counts. A univariate hurdle model (Mullahy, 1986) is a two-part model that considers zero counts and positive counts separately. A recent review of hurdle models can be found in Boucher et al. (2007) and Frees (2009). The superiority of hurdle models is that they can handle both zero-deflation and zero-inflation phenomena.

Following the idea in Tian et al. (2018b), we identify a stochastic representation of the Type I zero-truncated counting random variables, which facilitates us to make use of the expectation-maximization (EM) algorithm (Dempster et al., 1977) for the parameters' estimation. The EM algorithm is broadly applied in count models. Karlis (2005) implemented EM algorithm for univariate mixed Poisson models and Ghitany et al. (2012) further extended them to the multivariate cases. Recently, Tian et al. (2018a) pro-

posed the use of EM algorithm to deal with univariate truncation for several distributions. In this paper, we use a similar approach.

The rest of the paper is organized as follows. Section 2 reviews the general Type I multivariate zero-truncated model, followed by two commonly used distributions: Type I multivariate zero-truncated Poisson model and Type I multivariate zero-truncated negative binomial model. The EM algorithms for parameters' estimation are provided in the Appendix A. In Section 3, we propose a Type I multivariate zero-truncated hurdle model, followed with its parameter estimation via EM algorithm and some simulation studies to test the performance of the model. In Section 4, the proposed model is applied to a real health insurance dataset and both model fitting and prediction results are provided. Concluding remarks are given in Section 5.

2 Type I multivariate zero-truncated models

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^\top$ denote a discrete random vector where $Y_j, j = 1, \ldots, m$, are independent of each other. Then $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^\top$ is said to follow the Type I multivariate zero-truncated distribution if

$$\boldsymbol{Y} \stackrel{d}{=} U\boldsymbol{Z} = \begin{cases} \boldsymbol{0}, & U = 0 \\ \boldsymbol{Z}, & U = 1 \end{cases} \tag{2.1}$$

where $U \sim Bernoulli(\pi_0)$, $\pi_0 = \Pr(\boldsymbol{Y} \neq \boldsymbol{0}) = 1 - \prod_{j=1}^m \Pr(Y_j = 0)$ and U is independent of $\boldsymbol{Z}$. The probability mass function (pmf) of $\boldsymbol{Z}$ can be derived as

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \frac{\Pr(\boldsymbol{Y} = \boldsymbol{z})}{\Pr(U = 1)} = \frac{\prod_{j=1}^m \Pr(Y_j = z_j)}{1 - \prod_{j=1}^m \Pr(Y_j = 0)}, \quad \|\boldsymbol{z}\|_1 > 0, \tag{2.2}$$

where $\|\cdot\|_1$ represents the ℓ_1 norm of a vector. An alternative representation is to define $\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{Y} \mid \boldsymbol{Y} \neq \boldsymbol{0}$. This stochastic representation helps us to generate random values of $\boldsymbol{Z}$.

In the following, we review two commonly used Type I multivariate zero-truncated models: the Type I multivariate zero-truncated Poisson(MZTP) model and the Type I multivariate zero-truncated negative binomial(MZTNB) model. The corresponding EM algorithms are included in the Appendix A.

2.1 Type I multivariate zero-truncated Poisson model

Let $Y_j \sim Poisson(\lambda_j)$, for $j = 1, \ldots, m$, then $\boldsymbol{Z}$ is said to follow the Type I multivariate zero-truncated Poisson distribution with the parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$, denoted by $\boldsymbol{Z} \sim ZTP_m^{(I)}(\boldsymbol{\lambda})$. Then $\pi_0 = 1 - \exp(-\sum_{j=1}^m \lambda_j)$. The pmf of $\boldsymbol{Z}$ is

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \frac{1}{1 - \exp(-\sum_{j=1}^m \lambda_j)} \prod_{j=1}^m \frac{\lambda_j^{z_j} e^{-\lambda_j}}{z_j!}, \quad \|\boldsymbol{z}\|_1 > 0. \tag{2.3}$$

Now suppose for each independent individual, $\boldsymbol{Z}_i \sim ZTP_m^{(I)}(\boldsymbol{\lambda}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{im})^\top$. The corresponding observed values are $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})^\top$. Now we introduce some explanatory variables, or covariates, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$. These covariates could be incorporated as $\lambda_{ij} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$ with new parameters $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})^\top$. If we denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top)^\top$, the likelihood function becomes to

$$L(\boldsymbol{\beta} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \prod_{i=1}^n \left[1 - \mathrm{e}^{-\sum_{j=1}^m \lambda_{ij}}\right]^{-1} \prod_{i=1}^n \prod_{j=1}^m \frac{\lambda_{ij}^{z_{ij}} e^{-\lambda_{ij}}}{z_{ij}!}. \tag{2.4}$$

Remark 1 *The exposure time e_i for each sample can be incorporated as $\lambda_{ij} = e_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$.*

2.2 Type I multivariate zero-truncated negative binomial model

Let $Y_j \sim NB(\mu_j, \theta_j)$, for $j = 1, \ldots, m$, then $\boldsymbol{Z}$ is said to follow the Type I multivariate zero-truncated negative binomial distribution with the two parameter vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$, denoted by $\boldsymbol{Z} \sim ZTNB_m^{(I)}(\boldsymbol{\mu}, \boldsymbol{\theta})$. As a result, $\pi_0 = 1 - \prod_{j=1}^m [\theta_j/(\mu_j + \theta_j)]^{\theta_j}$. The pmf of $\boldsymbol{Z}$ is

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \frac{1}{1 - \prod_{j=1}^m \left(\frac{\theta_j}{\mu_j + \theta_j}\right)^{\theta_j}} \prod_{j=1}^m \left[\frac{\Gamma(z_j + \theta_j)}{\Gamma(\theta_j) z_j!} \left(\frac{\mu_j}{\mu_j + \theta_j}\right)^{z_j} \left(\frac{\theta_j}{\mu_j + \theta_j}\right)^{\theta_j}\right]. \tag{2.5}$$

Now suppose for each independent individual, $\boldsymbol{Z}_i \sim ZTNB_m^{(I)}(\boldsymbol{\mu}_i, \boldsymbol{\theta})$, $i = 1, \ldots, n$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})$. Similarly, the covariates could be intro-

duced as $\mu_{ij} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta_j})$. The likelihood function becomes to

$$L(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \prod_{i=1}^{n} \left[1 - \prod_{j=1}^{m} \Big(\frac{\theta_j}{\mu_{ij} + \theta_j} \Big)^{\theta_j} \right]^{-1} \times \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\frac{\Gamma(z_{ij} + \theta_j)}{\Gamma(\theta_j) z_{ij}!} \Big(\frac{\mu_{ij}}{\mu_{ij} + \theta_j} \Big)^{z_{ij}} \Big(\frac{\theta_j}{\mu_{ij} + \theta_j} \Big)^{\theta_j} \right].$$

Remark 2 *The exposure time e_i for each sample can be incorporated as $\mu_{ij} = e_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$.*

3 Type I multivariate zero-truncated hurdle model

In this section we shall define the so-called Type I multivariate zero-truncated hurdle model, then we discuss the model inference in details followed by some illustrative simulation examples. Some distribution properties can be found in Appendix B.

3.1 Model characterization

Since we are considering m counting variables, to allow for greater flexibility in modeling zero-inflation or zero-deflation for each counting variable, we shall assume that each underlying random variable Y_j in (2.1), $j = 1, \ldots, m$, follows a zero-modified distribution, which can be characterized as follows:

$$Y_j \overset{d}{=} U_j W_j = \begin{cases} 0, & U_j = 0, \\ W_j, & U_j = 1, \end{cases} \tag{3.1}$$

where W_j follows a univariate zero-truncated distribution, $U_j \sim Bernoulli(\pi_j)$, $0 < \pi_j < 1$, and U_j is independent of W_j. Again, we assume that all Y_j, $j = 1, \ldots, m$, are independent of each other. Then $\boldsymbol{Z}$ constructed by (2.1) is said to follow the Type I multivariate zero-truncated hurdle distribution with parameter vectors $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^\top$ and $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_m)^\top$, where $\boldsymbol{\Theta}_j$ is the set of parameters related to W_j. Again, let $\pi_0 = 1 - \prod_{j=1}^{m}(1 - \pi_j)$, then the pmf of $\boldsymbol{Z}$ can be expressed as

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \frac{1}{\pi_0} \prod_{j:z_j=0} (1 - \pi_j) \prod_{j:z_j \neq 0} \pi_j f_{W_j}(z_j), \tag{3.2}$$

where $f_{W_j}(z_j) = \Pr(W_j = z_j)$.

Remark 3 *Actually W_j may not be obtained by zero-truncation. It could be generated by shifting a counting random variable (i.e. $W_j - 1$ follows a regular counting distribution). This method is further discussed in the real application given in Section 4.*

3.2 Model inference

Now suppose each $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, independently follows a Type I multivariate zero-truncated hurdle distribution. Taking p covariates into account, the parameters π_{ij}, $i = 1, \ldots, n, j = 1, \ldots, m$, can be modeled as

$$\pi_{ij} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}, \tag{3.3}$$

where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ and $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})^\top$. Then $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top)^\top$ is the whole set of coefficients to determine. We denote $\boldsymbol{\Theta}$ as the set of parameters related to W_{ij}, $i = 1, \ldots, n, j = 1, \ldots, m$, the likelihood function then can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Theta} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = & \prod_{i=1}^{n} \left[1 - \prod_{j=1}^{m} (1 - \pi_{ij}) \right]^{-1} \\ & \times \prod_{i=1}^{n} \left[\prod_{j:z_{ij}=0} (1 - \pi_{ij}) \prod_{j:z_{ij}\neq 0} \pi_{ij} f_{W_{ij}}(z_{ij}) \right]. \end{aligned} \tag{3.4}$$

Remark 4 *The exposure time e_i for each sample can be incorporated as $\mu_{W_{ij}} = e_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha}_j)$ where $\mu_{W_{ij}}$ is the location parameter of W_{ij} and $\boldsymbol{\alpha}_j = (\alpha_{j0}, \alpha_{j1}, \ldots, \alpha_{jp})^\top$. If we denote $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_m^\top)^\top$, then $\boldsymbol{\alpha} \subset \boldsymbol{\Theta}$ as $\boldsymbol{\Theta}$ may also include other parameters of W_{ij}, $i = 1, \ldots, n, j = 1, \ldots, m$.*

The observed log-likelihood function can be divided into two parts:

$$\begin{aligned} \ell_1(\boldsymbol{\beta} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = & \sum_{i=1}^{n} \left[\sum_{j:z_{ij}=0} \log(1 - \pi_{ij}) + \sum_{j:z_{ij}\neq 0} \log \pi_{ij} \right] \\ & - \sum_{i=1}^{n} \log \left(1 - \prod_{j=1}^{m} (1 - \pi_{ij}) \right), \\ \ell_2(\boldsymbol{\Theta} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = & \sum_{i=1}^{n} \sum_{j:z_{ij}\neq 0} f_{W_{ij}}(z_{ij}) = \sum_{j=1}^{m} \sum_{i:z_{ij}\neq 0} f_{W_{ij}}(z_{ij}). \end{aligned}$$

Thus, the maximization procedure can be completed for ℓ_1 and ℓ_2 respectively. For ℓ_2, the estimation can be implemented with respect to the zero-truncation part of each marginal separately. For ℓ_1, we implement the EM algorithm illustrated as follows.

Denote $\boldsymbol{Z'} = (Z'_1, \ldots, Z'_m)^\top$ where $Z'_j = \mathbb{I}(Z_j > 0)$, and $\mathbb{I}(\cdot)$ is the indicator function. The corresponding observed values are denoted by $\boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_n$ where $\boldsymbol{z}'_i = (z'_{i1}, \ldots, z'_{im})^\top$. The observed log-likelihood function ℓ_1 can be rewritten as

$$\begin{aligned}\ell_1(\boldsymbol{\beta} \mid \boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_n) &= \sum_{j=1}^{m}\sum_{i=1}^{n}\left[z'_{ij}\log\pi_{ij} + (1 - z'_{ij})\log(1-\pi_{ij})\right] \\ &\quad - \sum_{i=1}^{n}\log\left(1 - \prod_{j=1}^{m}(1-\pi_{ij})\right).\end{aligned}$$

The independence between U and $\boldsymbol{Z}$ indicates the independence between U and $\boldsymbol{Z'}$. Let $\boldsymbol{Y'} \stackrel{d}{=} U\boldsymbol{Z'}$, then the complete data regarding $\boldsymbol{Y'}$ are $(\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_n)$, where $\boldsymbol{y}'_i = (u_i, \boldsymbol{z}'_i)$, $i = 1, \ldots, n$. The complete-data log-likelihood function given $(\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_n)$ is

$$\ell_1(\boldsymbol{\beta} \mid \boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_n) = \sum_{j=1}^{m}\sum_{i=1}^{n}\left[u_i z'_{ij}\log\pi_{ij} + (1 - u_i z'_{ij})\log(1-\pi_{ij})\right].$$

Given initial values of parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, m$, the EM algorithm is as follows:

- E-step: Replace u_i, $i = 1, \ldots, n$, by their conditional expectations

$$t_i = E(U_i \,|\boldsymbol{z}'_i, \boldsymbol{\beta}) = 1 - \prod_{j=1}^{m}(1-\pi_{ij}),$$

  where $\pi_{ij} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, are obtained using the current values of $\boldsymbol{\beta_j}$.

- M-step: Let

$$\bar{\ell}_{1j}(\boldsymbol{\beta}_j \mid \boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_n) = \sum_{i=1}^{n}\left[t_i z'_{ij}\log\pi_{ij} + (1 - t_i z'_{ij})\log(1-\pi_{ij})\right].$$

  Update the regression parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, m$, respectively by maximizing $\bar{\ell}_{1j}$ using Newton-Raphson method. The first and second

order derivatives are given as follows:

$$\begin{aligned}\frac{\partial \bar{\ell}_{1j}}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n}(t_i z'_{ij} - \pi_{ij})\boldsymbol{x}_i, \\ \frac{\partial^2 \bar{\ell}_{1j}}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} &= -\sum_{i=1}^{n}\pi_{ij}(1-\pi_{ij})\boldsymbol{x}_i\boldsymbol{x}_i^\top.\end{aligned}$$

- Iterate between the E-step and the M-step until some convergence criterion is satisfied, i.e. for two consecutive iterations of the algorithm (t) and $(t-1)$, $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2 < 10^{-8}$.

Remark 5 *The initial values of parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, m$, can be obtained by implementing univariate logistic regression with observed values $\boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_n$. The standard errors of the parameters can be obtained from the Hessian matrix of the observed log-likelihood function or bootstrap method.*

For the case without covariates, the EM algorithm is simplified as follows:

- E-step: Calculate for $i = 1, \ldots, n$,

$$t = t_i = E(U_i \,|\boldsymbol{z}'_i, \boldsymbol{\pi}) = 1 - \prod_{j=1}^{m}(1-\pi_j),$$

  where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^\top$. Initial values for π_j can be set as the proportion of non-zero for each type j.

- M-step: Update π_j, $j = 1, \ldots, m$, through the following equations:

$$\pi_j = \frac{t\sum_{i=1}^{n} z'_{ij}}{n}.$$

3.3 Model testing

In this subsection, we shall conduct two simulation examples to test the performance of our proposed Type I multivariate zero-truncated hurdle (MZTH) model. The examples aim to compare MZTH model with MZTP model when data are simulated from these two models respectively. As we only deal with Poisson marginals in a MZTH model, we use the name of MZTHP to be specific. For simplicity, we set $m = 3$. The data generation process is as follows:

- Simulate 3 numbers from 3 independent corresponding distributions. If the three numbers are simultaneously 0, discard them and continue the simulation until at least one of the three numbers is non-zero. Record them as one observation.
- Continue with this procedure until the sample size reaches $n = 10000$.
- Generate $T = 500$ replications of samples.

Using the EM methods given previously, we can estimate the unknown parameters under different model assumption. We then calculate the average parameter estimate and the variance of the estimate as follows. For parameter γ, its average parameter estimate is

$$\hat{\gamma} = \sum_{t=1}^{T} \hat{\gamma}^{(t)}$$

and the variance of $\hat{\gamma}$ can be estimated as

$$\frac{1}{T}\frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{\gamma}^{(t)} - \hat{\gamma}\right)^2$$

where $\hat{\gamma}^{(t)}$ is the estimation result from the t-th iteration.

Example 1. Covariates are not incorporated here. The parameters used for the generation of MZTHP-type of data are set as:

- $\lambda_1 = 0.5$, $\lambda_2 = 1$, $\lambda_3 = 1.5$;
- $\pi_1 = 0.3$, $\pi_2 = 0.4$, $\pi_3 = 0.5$.

The parameters used to generate MZTP-type of data are: $\lambda_1 = 0.5$, $\lambda_2 = 1$, $\lambda_3 = 1.5$.

Example 2. In this example we introduce a single predictor variable x_i which is generated from a folded standard normal distribution. Assume $\lambda_{ij} = \exp(\alpha_{j0} + \alpha_{j1}x_i)$ and $\pi_{ij} = \frac{\exp(\beta_{j0}+\beta_{j1}x_i)}{1+\exp(\beta_{j0}+\beta_{j1}x_i)}$, $i = 1, 2, \ldots, n$, $j = 1, 2, 3$. The parameters used for the generation of MZTHP-type of data are set as:

- $\alpha_{10} = -1$, $\alpha_{11} = 1$, $\beta_{10} = -1$, $\beta_{11} = 1$,
- $\alpha_{20} = 0.5$, $\alpha_{21} = 0.5$, $\beta_{20} = 0.5$, $\beta_{21} = 0.5$,
- $\alpha_{30} = 1$, $\alpha_{31} = -1$, $\beta_{30} = 1$, $\beta_{31} = -1$.

The parameters used to generate MZTP-type of data are

- $\alpha_{10} = -1$, $\alpha_{11} = 1$,
- $\alpha_{20} = 0.5$, $\alpha_{21} = 0.5$,
- $\alpha_{30} = 1$, $\alpha_{31} = -1$.

The estimation results are summarized in Table 1 and Table 2 when data are generated from MZTHP and MZTP models respectively for Example 1. Table 3-4 summarize the estimation results for Example 2. Average parameter estimates are given in the tables as well as the standard error of the estimates. The average AIC values are calculated under the two models for comparison. For Example 1, the average zero-truncation parameter π_0 is provided.

Several observations can be made form these tables. First, the estimates are anticipated when the model is correctly specified, which verifies the efficiency of our proposed algorithm. Second, when data are generated from the MZTHP model, the estimates suffer using the MZTP model. However, when the MZTP is the underlying model, the MZTHP model can still give us proper estimates for the parameters λ_j, $j = 1, 2, 3$ or α_{jk}, $j = 1, 2, 3$, $k = 1, 2$, though with larger standard errors. This makes sense as the MZTHP model only uses positive numbers to estimate the parameters for λ_j or α_{jk}.

To examine the overall goodness-of-fit, we further compare the AIC statistics from the model fitting. Figure 1 and Figure 2 display the scatter-plots of AIC values between the true model and misspecified model based on the 500 replications in Example 1 and Example 2. One can clearly see that the MZTP model does not work very well when the data are generated from the MZTHP model. Reversely, when dealing with MZTP-type of data, the performances of the MZTHP model and the MZTP model have no significant difference.

In short, The above two examples have demonstrated a better robustness of our proposed MZTHP model than the MZTP model. It showcases that when the underlying model is unknown, our proposed MZTH model seems to be a more robust candidate to study the multivariate claim frequency data with Type I zero-truncation feature.

Table 1: Estimation results of MZTHP-type of data

	MZTHP model		MZTP model	
True value	estimate	s.e.($\times 10^{-5}$)	estimate	s.e.($\times 10^{-5}$)
$\lambda_1 = 0.5$	0.50	2.88	0.43	1.26
$\lambda_2 = 1.0$	1.00	3.28	0.72	2.05
$\lambda_3 = 1.5$	1.50	3.58	1.09	2.61
$\pi_1 = 0.3$	0.30	0.93	-	-
$\pi_2 = 0.4$	0.40	0.98	-	-
$\pi_3 = 0.5$	0.50	1.10	-	-
Ave. π_0	0.79		0.89	
Ave. AIC	68,808.18		70,873.96	

Table 2: Estimation results of MZTP-type of data

	MZTHP model		MZTP model	
True value	estimate	s.e.($\times 10^{-5}$)	estimate	s.e.($\times 10^{-5}$)
$\lambda_1 = 0.5$	0.50	2.95	0.50	1.36
$\lambda_2 = 1.0$	1.00	3.06	1.00	1.91
$\lambda_3 = 1.5$	1.50	3.16	1.50	2.49
π_1	0.39	0.95	-	-
π_2	0.63	1.07	-	-
π_3	0.78	0.92	-	-
Ave. π_0	0.95		0.95	
Ave. AIC	75,241.8		75,238.6	

Table 3: Estimation results of MZTHP-type of data

	MZTHP model		MZTP model	
True value	estimate	s.e.($\times 10^{-5}$)	estimate	s.e.($\times 10^{-5}$)
$\alpha_{10} = -1$	-1.00	6.40	-1.15	4.63
$\alpha_{11} = 1$	1.00	3.56	0.98	3.38
$\alpha_{20} = 0.5$	0.50	2.64	0.28	3.18
$\alpha_{21} = 0.5$	0.50	2.04	0.53	2.63
$\alpha_{30} = 1$	1.00	3.63	0.79	3.47
$\alpha_{31} = -1$	-1.00	6.36	-0.96	4.57
$\beta_{10} = -1$	-1.00	7.40		
$\beta_{11} = 1$	1.00	7.81		
$\beta_{20} = 0.5$	0.50	8.08		
$\beta_{21} = 0.5$	0.50	9.06		
$\beta_{30} = 1$	1.00	7.76		
$\beta_{31} = -1$	-1.00	7.69		
Ave. AIC	86,226.0		89,409.5	

Table 4: Estimation results of MZTP-type of data

	MZTHP model		MZTP model	
True value	estimate	s.e.($\times 10^{-5}$)	estimate	s.e.($\times 10^{-5}$)
$\alpha_{10} = -1$	-1.00	6.78	-1.00	3.74
$\alpha_{11} = 1$	1.00	3.57	1.00	2.41
$\alpha_{20} = 0.5$	0.50	2.57	0.50	2.17
$\alpha_{21} = 0.5$	0.50	2.07	0.50	1.82
$\alpha_{30} = 1$	1.00	3.23	1.00	2.59
$\alpha_{31} = -1$	-1.00	5.74	-1.00	3.81
β_{10}	-0.95	7.42		
β_{11}	1.58	8.58		
β_{20}	1.33	10.39		
β_{21}	1.39	16.12		
β_{30}	2.26	9.78		
β_{31}	-1.62	9.06		
Ave. AIC	88,787.8		88,751.02	

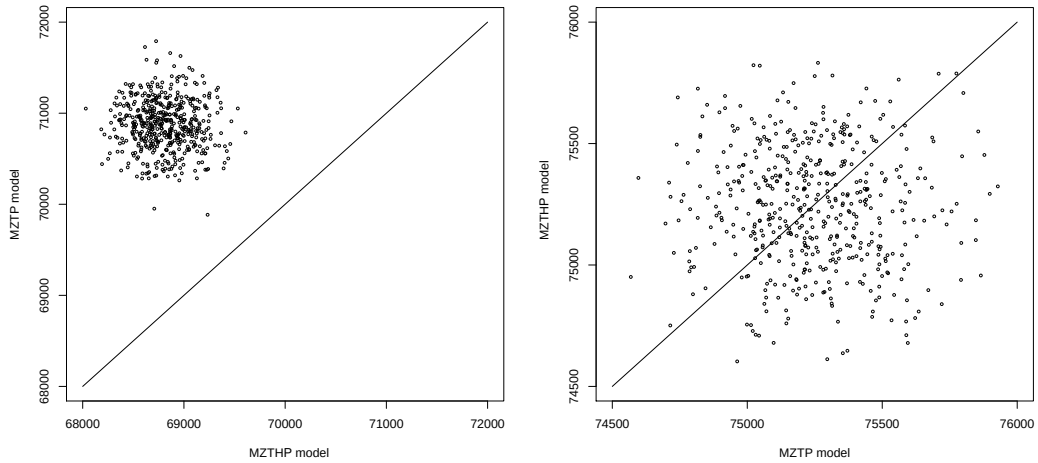

Figure 1: Comparison of AIC between the two models without covariates when data are generated from MZTHP model (left) and MZTP model (right) respectively

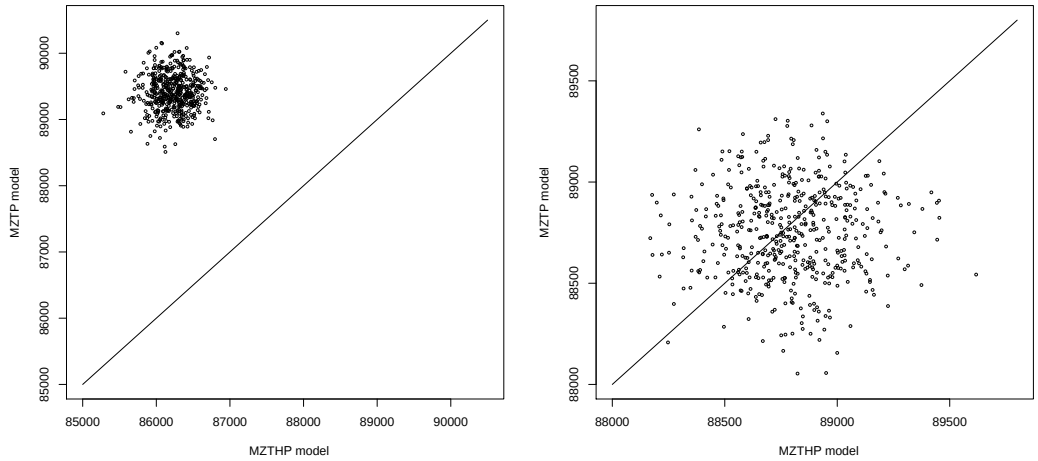

Figure 2: Comparison of AIC between the two models with covariates when data are generated from MZTHP model (left) and MZTP model (right) respectively

4 Application

4.1 Data

In this section we provide a real-life application of our proposed model. The dataset was obtained from a Chinese health fund that contains health insurance claim records in the period of 2015-2017. The underlying health insurance policies provide full insurance covers on payments associated with in-patient and out-patient treatments as well as emergency services. The age of the insured ranges from 28 days to 60 years. There are no out of pocket payments for policyholders, which means that all claim amounts in the dataset are full payment amounts. The dataset consists of 40,030 policyholders. There are three main causes of claims in the dataset: disease, accident and other. As the total number of claims for each policyholder is positive, the number of claims of each type cannot be 0 simultaneously. Apart from information regarding claims, the dataset also contains some explanatory variables such as time exposed to risk, age, gender, region and smoking status. Such information is useful for risk classification and ratemaking purpose.

For this study, we take a random sample of 30,000 as training data to develop the model and the rest is reserved as the holdout sample for validation purposes. The descriptive statistics of claim counts are displayed in Table 5. In this table we present the average total claim counts as well as the average count of each claim type. From the table one can conclude that the youngest insureds (0-7 years of age) have higher total claim counts compared with other age groups. There is no significant difference in average counts among different regions and genders. The smoker group shows lower frequencies than the non-smokers, which somehow contradicts to our common sense. As for the specified claim type, the youngest age group (0-7) are prone to incurring more claims for disease yet less for accidents which varies from the other three age groups. Non-smokers have more frequent claims due to diseases and less frequent claims from accidents which is different from the smokers. We also present the corresponding percentages of various observations in the table. Nearly half of the policyholders are between age 19 and 44. Policyholders aged 8 to 18 only account for 2.1%. It is also noted that majority of the policyholders are non-smokers.

Table 5: Average claim frequency split by categories for each explanatory variable

Category	Count	Disease	Accident	Other	% of obs
age:					
0-7	1.411	1.365	0.017	0.030	18.4
8-18	1.098	0.961	0.067	0.069	2.1
19-44	1.134	1.001	0.066	0.067	46.3
45-60	1.164	1.027	0.053	0.084	33.2
region:					
central	1.215	1.099	0.058	0.058	19.9
north	1.193	1.060	0.058	0.075	6.7
northeast	1.212	1.124	0.041	0.047	22.7
northwest	1.158	1.052	0.051	0.055	2.1
south	1.174	1.022	0.061	0.092	4.7
southwest	1.191	1.063	0.047	0.081	23.9
east	1.168	1.034	0.064	0.070	20.0
gender:					
female	1.180	1.074	0.042	0.065	56.0
male	1.213	1.079	0.066	0.068	44.0
smoke:					
no	1.196	1.078	0.052	0.065	98.0
yes	1.130	0.955	0.084	0.092	2.0

4.2 Analysis

Table 6 gives the observed claim counts of each claim type. The empirical distribution for disease claim frequency obviously has a heavier tail than the ones for accident and other, indicating that Poisson and negative binomial models might not be appropriate for it. Also, our data show negative correlation among different claim types which is summarized in Table 7. This coincides with our proposed model. We start with no covariates for the parameters and use Z_j and W_j, $j = 1, 2, 3$, to represent the claim number and positive count for disease, accident and other. Table 9 gives a comparison among the MZTP model, MZTNB model and our proposed MZTH model. Regarding the marginal distributions of the positive claim counts in the MZTH model, we choose the Poisson inverse Gaussian (PIG) model for the shifted positive counts $W_1 - 1$, the zero-truncated Poisson model (ZTP) for W_2 and the negative binomial (NB) model for the shifted positive counts $W_3 - 1$. The distributional properties for the three models are summarized in Table 8. To further demonstrate the necessity of introducing the hurdle part in our model, we also fit an MZT model with three different marginals, PIG, Poisson and NB without any modifications for Y_1, Y_2 and Y_3 respectively. The result is also included in Table 9. Using AIC and BIC as our main criterion for model selection, one can see that because of the extra flexibility on choices of marginals, the MZT model performs better than the MZTP and MZTNB model, but still not as good as our proposed MZTH model when fitting the training dataset. This is because the MZTH model separates all zero counts from positive counts in the marginals so that the zero counts are fitted more accurately.

Next we add covariates as well as exposure into our MZTH model to see the significance of each explanatory variable. The main results regarding π_j and W_j, $j = 1, 2, 3$, are summarized in Table 10 and Table 11 respectively. It is noticed that some groups classified by certain covariates always have a constant number 1 for W_2 and W_3, causing the estimation of the parameters drifting to infinity. To address this issue, we discard the involved covariates. The standard errors of covariates for π_j are calculated using the Hessian matrix from observed log-likelihood function. Take $\hat{\boldsymbol{\beta}}_1$ for example, its Hessian matrix can be calculated as

$$\mathcal{H}(\hat{\boldsymbol{\beta}}_1) = -\sum_{i=1}^{n} \hat{\pi}_{i1}(1-\hat{\pi}_{i1}) \frac{1-(1-\hat{\pi}_{i2})(1-\hat{\pi}_{i3})}{\left[1-(1-\hat{\pi}_{i1})(1-\hat{\pi}_{i2})(1-\hat{\pi}_{i3})\right]^2} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

where $\hat{\pi}_{ij} = \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j)}{1+\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j)}$, $i = 1, \ldots, n$, $j = 1, 2, 3$.

Table 6: The marginal empirical distribution of three claim types

	Disease		Accident		Other	
Count	Frequency	Percent	Frequency	Percent	Frequency	Percent
0	2,887	9.62	28,448	94.83	28,063	93.54
1	23,511	78.37	1,525	5.08	1,899	6.33
2	2,700	9.00	26	0.09	35	0.12
3	580	1.93	1	0.00	3	0.01
4	164	0.55				
5	76	0.25				
6	37	0.12				
7	18	0.06				
8	13	0.04				
9	7	0.02				
≥ 10	7	0.02				

Table 7: Pearson's correlations between the three claim types

Type	Disease	Accident	Other
Disease	-	-0.299	-0.276
Accident	-0.299	-	-0.032
Other	-0.276	-0.032	-

Table 8: The distributional properties of ZTP, NB and PIG models

Distribution	Parameters	Mean	Variance
ZTP	λ	$\frac{\lambda}{1-e^{-\lambda}}$	$\frac{\lambda+\lambda^2}{1-e^{-\lambda}} - \frac{\lambda^2}{(1-e^{-\lambda})^2}$
NB[a]	μ, σ	μ	$\mu + \sigma\mu^2$
PIG	μ, σ	μ	$\mu + \sigma\mu^2$

[a] Here we use negative binomial of type II.
[b] The covariates can be incorporated with λ or μ via log-link function.

Table 9: Loglikelihood, AIC and BIC of the four models considered

Model	No. of parameters	Loglik	AIC	BIC
MZTP	3	-30,143.14	60,292.28	60,317.21
MZTNB	6	-29,273.42	58,558.84	58,608.69
MZTH	8	-28,841.18	57,698.36	57,764.83
MZT	5	-29,120.88	58,251.76	58,293.30

Table 10: Estimates and t-ratio associated with the covariates of π_j

	π_1		π_2		π_3	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
Intercept	-1.529	-77.260***	-3.942	-150.041***	-3.910	-165.553***
age1(0-7)	1.138	14.350***	-0.446	-4.338***	-0.370	-4.647***
age2(8-18)	-0.900	-7.376***	-0.586	-3.609***	-0.967	-6.119***
age3(19-44)	-0.235	-8.680***	-0.004	-0.127	-0.441	-12.883***
central	0.463	10.309***	0.294	5.265***	0.246	4.449***
north	0.322	4.405***	0.087	0.888	0.302	3.518***
northeast	0.568	11.691***	-0.062	-0.983	0.025	0.440
northwest	0.200	1.395	-0.187	-0.992	-0.140	-0.788
south	0.227	2.762**	0.088	0.764	0.455	4.793***
southwest	0.071	1.833	-0.146	-2.601**	0.310	7.027***
female	0.347	12.525***	-0.207	-5.440***	0.080	2.497*
non-smoker	-0.524	-26.119***	-0.578	-21.661***	-0.448	-18.714***
Loglikelihood	-14,387.16					

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

Table 11: Estimates and t-ratio associated with the covariates of W_j

	W_1		W_2		W_3	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
Intercept	-2.536	-15.400***	-3.729	-14.003***	-3.660	-3.250**
age1(0-7)	1.114	22.798***				
age2(8-18)	-0.395	-2.424*				
age3(19-44)	-0.169	-3.738***				
central	0.231	3.933***			-0.272	-0.533
north	0.128	1.526			0.425	0.771
northeast	0.216	3.800***			-0.128	-0.259
northwest	0.050	0.354			0.098	0.087
south	0.050	0.496			-0.527	-0.656
southwest	-0.040	-0.692			-0.971	-1.810
female	-0.017	-0.439	0.224	0.596	-0.403	-1.199
non-smoker	0.216	1.342			-0.064	-0.058
$\log(\sigma)$	1.212	22.650***			1.506	2.012*
Loglikelihood	-13,021.07		-138.79		-190.50	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

As it can be seen from Table 10, age, region, gender and smoke are all significant covariates on the occurrence of each type of claims. As being empirically observed, the youngest age group is most likely to incur claims by disease compared with other age groups. Smoking status has a positive effect on the occurrence of all three types of claims. For the positive counts, only region and age are significant covariates for modeling μ_{W_1}, yet no covariates play a significant role in modeling μ_{W_2} and μ_{W_3}. This makes sense as the number of claims by accident or other reasons should be totally random in nature so that the impact of the covariates on the number of claims should be negligible.

Overall, we preserve all the covariates when modeling π_j, $j = 1, 2, 3$. As for the positive counts of disease claims, we use region and age factors as well as exposure in PIG model for $W_1 - 1$. For claims due to accident and other causes, only the covariate of exposure is adopted in the ZTP model for W_2 and the NB model for $W_3 - 1$ respectively.

4.3 Predictive performance

Making use of the testing data, we calculate the predicted claim frequencies for the three claim types using the best models obtained above and then

compare with the observed numbers in the data. The formulas for calculating the marginal probabilities can be found in Appendix B. Because of the heterogeneity caused by the covariates, the predicted frequencies are calculated by summing up marginal probabilities of the individual policyholders in the portfolio. Some cells are grouped to comply with the rule of 5.

The goodness-of-fit results for three univariate marginal models are exhibited in Table 12. The small out-of-sample χ^2 statistics suggest a good fit on the marginal counts.The prediction results for three bivariate cases are displayed in Table 13. In this table, the predicted numbers are put in parentheses next to the observed numbers. One can see that the overall performance is acceptable in each case except once cell $(Z_1 = 2, Z_3 = 1)$, where the predicted frequency seriously underestimates the real figure. It indicates that there could be further dependence between Z_1 and Z_3.

In addition, we also generate a joint prediction on the multivariate claim frequency and summarise the results in Table 14, where the predicted numbers are put in parentheses below the observed numbers. It can be seen that the overall prediction suffers due to two cells: $(Z_1 \geq 2, Z_2 = 0, Z_3 = 1)$ and $(Z_1 = 1, Z_2 = 1, Z_3 = 0)$. After excluding the two problematic cells, the overall χ^2 value drops to 20.22 and the corresponding p-value becomes to 0.12.

Table 12: Goodness-of-fit of three univariate marginal models

	Disease		Accident		Other	
Count	Observed	Predicted	Observed	Predicted	Observed	Predicted
0	945	960.01	9,515	9,514.85	9,376	9,386.05
1	7,931	7,861.11	503	505.84	641	631.41
2	879	905.80	12	9.19	13	11.73
3	172	189.87				
4	58	60.94				
5	24	24.86				
6	9	11.77				
7	3	6.17				
≥ 8	9	9.46				
χ^2 (p-value)	5.81 (0.67)		0.79 (0.67)		0.17 (0.92)	

Values are corresponding to $Z_2 \geq 2$ and $Z_3 \geq 2$.

Table 13: Goodness-of-fit of three bivariate marginal models

		Z_2		
		0	1	≥ 2
Z_1	0	497(528.59)	437(423.61)	12(9.32)[a]
	1	7,878(7,787.94)	52(71.86)	
	2	868(897.74)	14(10.37)[a]	
	3	170(188.26)		
	4	57(60.44)		
	5	24(24.66)		
	6	9(11.68)		
	7	3(6.13)		
	≥ 8	9(9.39)		
χ^2 (p-value)		16.08 (0.19)		
		Z_3		
		0	1	≥ 2
Z_1	0	441(426.21)	500(523.40)	13(12.54)[a]
	1	7,834(7,765.47)	94(93.79)	
	2	851(894.87)	47(14.22)[a]	
	3	161(187.61)		
	4	54(60.22)		
	5	20(24.57)		
	6	8(11.64)		
	7	2(6.10)		
	≥ 8	5(9.35)		
χ^2 (p-value)		91.11 (0.00)		
		Z_3		
		0	1	≥ 2
Z_2	0	8,871(8,877.14)	631(625.28)	13(12.54)[b]
	1	493(499.708)	10(6.12)[b]	
	≥ 2	12(9.20)		
χ^2(p-value)		3.47 (0.63)		

[a] Cells are grouped in terms of the count of disease.
[b] Cells are grouped in terms of the count of accident.

Table 14: Goodness-of-fit of multivariate model

			$Z_2 = 0$			$Z_2 = 1$		$Z_2 \geq 2$
		Z_3	0	1	≥ 2	0	≥ 1	≥ 0
Z_1	0			493	13[a]	430	10[a]	12[a]
				(518.29)	(12.42)	(418.50)	(6.13)	(9.32)
	1		7,783	92		50		
			(7,693.22)	(92.90)		(70.96)		
	2		841	46[a]		13[a]		
			(886.91)	(14.09)		(10.24)		
	3		159					
			(186.01)					
	4		53					
			(59.73)					
	5		20					
			(24.38)					
	6		8					
			(11.55)					
	7		2					
			(6.06)					
	≥ 8		5					
			(9.29)					
χ^2 (p-value)					98.68 (0.00)			

[a] Cells are grouped in terms of the count of disease.

5 Concluding remarks

To our best knowledge, few papers have discussed multivariate hurdle models with type I zero truncation. The model proposed in this paper, i.e. the MZTH model, is very flexible when handling multivariate count data with features of zero inflation and/or zero deflation in each dimension as well as various marginal count distributions for the positive observations. We implemented two simulation studies to compare the performance of the MZTHP model with the MZTP model. The results confirmed the better robustness of the former one. The MZTH model was then applied to a real-life health insurance dataset. The in-sample comparison among three candidate models demonstrated the superiority of the MZTH model. The out-of-sample prediction results also demonstrated a satisfactory predictive performance of our MZTH model at single dimensions as well as at the multidimensional level.

It is noted that our proposed model only allows for negative associations between individual components without covariates incorporated. This is caused by our model assumptions. One interesting topic for future research is to employ copulas to deal with a wider range of dependence cases.

Acknowledgements

The authors are grateful to the anonymous reviewer's valuable comments and suggestions on the first draft of this paper which helped them to produce a significantly improved final version. Mr. Pengcheng Zhang is supported by an Australian Government Research Training Program (RTP) Scholarship.

References

[1] Boucher, J. P., Denuit, M. and Guillén, M. (2007). Risk classification for claim counts: a comparative analysis of various zeroinflated mixed Poisson and hurdle models. *North American Actuarial Journal*, **11**(4), 110-131.

[2] Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge University Press.

[3] Charalambides, C. A. (1984). Minimum variance unbiased estimation for the zero class truncated bivariate poisson and logarithmic series distributions. *Metrika*, **31**(1), 115-123.

[4] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1-22.

[5] Frees, E. W. (2009). *Regression modeling with actuarial and financial applications*. Cambridge University Press.

[6] Ghitany, M. E., Karlis, D., Al-Mutairi, D. K. and Al-Awadhi, F. A. (2012). An EM algorithm for multivariate mixed Poisson regression models and its application. *Applied Mathematical Sciences*, **6**(137), 6843-6856.

[7] Jung, B. C., Han, S. M. and Lee, J. (2007). Score tests for testing independence in the zero-truncated bivariate Poisson models. *Communications in Statistics–Theory and Methods*, **36**(3), 599-611.

[8] Karlis, D. (2005). EM algorithm for mixed Poisson and other discrete distributions. *ASTIN Bulletin: The Journal of the IAA*, **35**(1), 3-24.

[9] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, **33**(3), 341-365.

[10] Piperigou, V. E. and Papageorgiou, H. (2003). On truncated bivariate discrete distributions: A unified treatment. *Metrika*, **58**(3), 221-233.

[11] Tian, G. L., Ding, X., Liu, Y. and Tang, M. L. (2018a). Some new statistical methods for a class of zero-truncated discrete distributions with applications. *Computational Statistics*, 1-34.

[12] Tian, G. L., Liu, Y., Tang, M. L. and Jiang, X. (2018b). Type I multivariate zero-truncated/adjusted Poisson distributions with applications. *Journal of Computational and Applied Mathematics*, 344, 132-153.

Appendix

A EM algorithms for two models

A.1 Type I multivariate zero-truncated Poisson model

We can augment the observed data by introducing mutually independent latent variables $U_1, \ldots, U_n$ where $U_i \sim Bernoulli(\pi_{0i})$ to form the complete

data denoted by $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, where $\boldsymbol{y}_i' = (u_i, \boldsymbol{z}_i')$, $i = 1, \ldots, n$, and $u_1, \ldots, u_n$ are observed values of $U_1, \ldots, U_n$. Thus, The complete-data likelihood function is given by

$$L(\boldsymbol{\beta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\lambda_{ij}^{u_i z_{ij}} e^{-\lambda_{ij}}}{(u_i z_{ij})!}$$

and the complete-data log-likelihood function is

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{j=1}^{m} \sum_{i=1}^{n} (u_i z_{ij} \log \lambda_{ij} - \lambda_{ij}) + C_0,$$

where C_0 is a constant not related to $\boldsymbol{\beta}$ and $\lambda_{ij} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

The associated EM algorithm is as follows:

- E-step: Replace u_i, $i = 1, \ldots, n$, by their conditional expectations

  $$t_i = E(U_i \mid \boldsymbol{z}_i, \boldsymbol{\beta}) = 1 - e^{-\sum_{j=1}^{m} \lambda_{ij}},$$

  where $\lambda_{ij} = \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta_j})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, are obtained using the current values of $\boldsymbol{\beta_j}$.

- M-step: Let

  $$\bar{\ell}_j(\boldsymbol{\beta}_j \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \sum_{i=1}^{n} (t_i z_{ij} \log \lambda_{ij} - \lambda_{ij}).$$

  Update the regression parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, m$, respectively by maximizing $\bar{\ell}_j$ using the Newton-Raphson method. The first and second order derivatives are given as follows:

  $$\frac{\partial \bar{\ell}_j}{\partial \boldsymbol{\beta_j}} = \sum_{i=1}^{n} (t_i z_{ij} - \lambda_{ij}) \boldsymbol{x_i}, \qquad \frac{\partial^2 \bar{\ell}_j}{\partial \boldsymbol{\beta_j} \partial \boldsymbol{\beta_j}^\top} = -\sum_{i=1}^{n} \lambda_{ij} \boldsymbol{x_i} \boldsymbol{x_i}^\top.$$

- Iterate between the E-step and the M-step until some convergence criterion is satisfied, i.e. for two consecutive iterations of the algorithm (t) and $(t-1)$, $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2 < 10^{-8}$ where $\|\cdot\|_2$ represents the ℓ_2 norm of a vector.

Remark 6 *Initial values for parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, m$, can be obtained by fitting univariate Poisson models. The standard errors of the parameters can be obtained directly from the Hessian matrix of the observed log-likelihood function or bootstrap method.*

For the case without the covariates, the corresponding E-step and M-step can be simplified as follows:

- E-step: Calculate for $i = 1, \ldots, n$,

$$t = t_i = E(U_i \,|\boldsymbol{z}_i, \boldsymbol{\lambda}) = 1 - e^{-\sum_{j=1}^{m} \lambda_j},$$

  where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$.

- M-step: Update λ_j, $j = 1, \ldots, m$, through the following equations:

$$\lambda_j = \frac{t \sum_{i=1}^{n} z_{ij}}{n}.$$

A.2 Type I multivariate zero-truncated negative binomial model

Using the same augment method as we do in MZTP model, we obtain the complete-data likelihood function given by

$$L(\boldsymbol{\beta}, \boldsymbol{\theta} \,|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left\{ \frac{\Gamma(u_i z_{ij} + \theta_j)}{\Gamma(\theta_j)(u_i z_{ij})!} \Big(\frac{\mu_{ij}}{\mu_{ij} + \theta_j} \Big)^{u_i z_{ij}} \Big(\frac{\theta_j}{\mu_{ij} + \theta_j} \Big)^{\theta_j} \right\},$$

and the complete-data log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta} \,|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{j=1}^{m} \sum_{i=1}^{n} \Big\{ \log\left(\Gamma(u_i z_{ij} + \theta_j)\right) - \log\left(\Gamma(\theta_j)\right) + u_i z_{ij} \log \mu_{ij} + \theta_j \log \theta_j - (u_i z_{ij} + \theta_j) \log(\mu_{ij} + \theta_j) \Big\} + C_1,$$

where C_1 is a constant independent of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\mu_{ij} = \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta_j})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

The EM algorithm can be described as follows:

- E-step: Calculate, for $i = 1, \ldots, n$,

$$t_i = E(U_i \mid \boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) = 1 - \prod_{j=1}^{m} \left(\frac{\theta_j}{\mu_{ij} + \theta_j} \right)^{\theta_j},$$

  where $\mu_{ij} = \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta_j})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, are obtained using the current values of $\boldsymbol{\beta_j}$.

- M-step: Let

$$\bar{\ell}_j(\boldsymbol{\beta}_j, \theta_j \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \sum_{i=1}^{n} \Big\{ t_i \log\left(\Gamma(z_{ij} + \theta_j)\right) - t_i \log\left(\Gamma(\theta_j)\right) + t_i z_{ij} \log \mu_{ij} + \theta_j \log \theta_j - (t_i z_{ij} + \theta_j) \log(\mu_{ij} + \theta_j) \Big\}.$$

  Update the regression parameters $\boldsymbol{\beta}_j$ and θ_j, $j = 1, \ldots, m$, respectively by maximizing $\bar{\ell}_j$ using the Newton-Raphson method. The first and second order derivatives are given as follows:

$$\begin{aligned}
\frac{\partial \bar{\ell}_j}{\partial \boldsymbol{\beta_j}} &= \sum_{i=1}^{n} \frac{(t_i z_{ij} - \mu_{ij})\theta_j}{\mu_{ij} + \theta_j} \boldsymbol{x_i}, \\
\frac{\partial \bar{\ell}_j}{\partial \theta_j} &= \sum_{i=1}^{n} \left[t_i \psi(z_{ij} + \theta_j) - t_i \psi(\theta_j) + \log \frac{\theta_j}{\mu_{ij} + \theta_j} + \frac{\mu_{ij} - t_i z_{ij}}{\mu_{ij} + \theta_j} \right], \\
\frac{\partial^2 \bar{\ell}_j}{\partial \boldsymbol{\beta_j} \partial \boldsymbol{\beta_j}^\top} &= -\sum_{i=1}^{n} \frac{(t_i z_{ij} + \theta_j)\mu_{ij}\theta_j}{(\mu_{ij} + \theta_j)^2} \boldsymbol{x_i} \boldsymbol{x_i}^\top, \\
\frac{\partial^2 \bar{\ell}_j}{\partial \theta_j^2} &= \sum_{i=1}^{n} \left[t_i \psi'(z_{ij} + \theta_j) - t_i \psi'(\theta_j) + \frac{\mu_{ij}^2 + t_i z_{ij} \theta_j}{\theta_j (\mu_{ij} + \theta_j)^2} \right], \\
\frac{\partial^2 \bar{\ell}_j}{\partial \boldsymbol{\beta_j} \partial \theta_j} &= -\sum_{i=1}^{n} \frac{(t_i z_{ij} - \mu_{ij})\mu_{ij}}{(\mu_{ij} + \theta_j)^2} \boldsymbol{x_i}.
\end{aligned}$$

- Iterate between the E-step and the M-step until some convergence criterion is satisfied, i.e. for two consecutive iterations of the algorithm (t) and $(t-1)$, $\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^{(t-1)}\|_2 < 10^{-8}$ where $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$.

Remark 7 *Initial values for parameters $\boldsymbol{\beta}_j$ and θ_j, $j = 1, \ldots, m$, can be obtained by fitting univariate negative binomial models. The standard errors of the parameters can be obtained from the Hessian matrix of the observed log-likelihood function or bootstrap method.*

For the case without covariates, the corresponding E-step and M-step can be simplified as follows:

- E-step: Calculate, for $i = 1, \ldots, n$,

$$t = t_i = E(U_i \,|\boldsymbol{z}_i, \boldsymbol{\mu}, \boldsymbol{\theta}) = 1 - \prod_{j=1}^{m} \left(\frac{\theta_j}{\mu_j + \theta_j} \right)^{\theta_j},$$

  where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$.

- M-step: Let

$$\bar{\ell}_j(\mu_j, \theta_j \,|\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \sum_{i=1}^{n} \Big\{ t \log\left(\Gamma(z_{ij} + \theta_j)\right) - t \log\left(\Gamma(\theta_j)\right) + t z_{ij} \log \mu_j + \theta_j \log \theta_j - (\theta_j + t z_{ij}) \log(\mu_j + \theta_j) \Big\}.$$

  - Update μ_j, $j = 1, \ldots, m$, through the following derivatives:

$$\frac{\partial \bar{\ell}_j}{\partial \mu_j} = \sum_{i=1}^{n} \left[\frac{t z_{ij}}{\mu_j} - \frac{\theta_j + t z_{ij}}{\mu_j + \theta_j} \right] = 0 \Rightarrow \mu_j = \frac{t \sum_{i=1}^{n} z_{ij}}{n}.$$

  - Substitute the value of μ_j obtained by last step into ℓ_j, Update θ_j, $j = 1, \ldots, m$, through the following derivatives:

$$\frac{\partial \bar{\ell}_j}{\partial \theta_j} = \sum_{i=1}^{n} \left[t\psi(z_{ij} + \theta_j) - t\psi(\theta_j) + \log \frac{\theta_j}{\mu_j + \theta_j} + \frac{\mu_j - t z_{ij}}{\mu_j + \theta_j} \right] = 0,$$

    where $\psi(\cdot)$ denotes the digamma function. There is no closed-form solution for the value of θ_j, but it can be obtained numerically through one variable Newton-Raphson method.

B Other properties for multivariate zero truncated hurdle model

B.1 Marginal distributions

We first derive the marginal distribution of single variable. The marginal distribution of Z_j, $j = 1, \ldots, m$, is

$$\Pr(Z_j = z_j) = \begin{cases} \dfrac{f_{Y_j}(z_j)}{\pi_0}, z_j > 0 \\ 1 - \dfrac{\pi_j}{\pi_0}, \; z_j = 0 \end{cases}$$

where $\pi_0 = 1 - \prod_{j=1}^{m} (1 - \pi_j)$.

Proof. If $z_j > 0$,

$$\begin{aligned}
\Pr(Z_j = z_j) &= \sum_{z_1=0}^{\infty} \cdots \sum_{z_{j-1}=0}^{\infty} \sum_{z_{j+1}=0}^{\infty} \cdots \sum_{z_m=0}^{\infty} \Pr(\boldsymbol{Z} = \boldsymbol{z}) \\
&= \frac{f_{Y_j}(z_j)}{\pi_0} \prod_{k=1,k\neq j}^{m} \sum_{z_k=0}^{\infty} f_{Y_k}(z_k) \\
&= \frac{f_{Y_j}(z_j)}{\pi_0}.
\end{aligned}$$

Thus,

$$\Pr(Z_j = 0) = 1 - \sum_{z_j=1}^{\infty} \Pr(Z_j = z_j) = 1 - \sum_{z_j=1}^{\infty} \frac{f_{Y_j}(z_j)}{\pi_0} = 1 - \frac{\pi_j}{\pi_0}.$$

■

Next we discuss the marginal distribution of an arbitrary random subvector of $\boldsymbol{Z}$. Denote $J = (j_1, j_2, \ldots, j_r) \subset (1, 2, \ldots, m)$ where $1 < r < m$ and $J^C = (j_{r+1}, j_{r+2}, \ldots, j_m)$ as the complementary set. Let $\boldsymbol{Z}_r = (Z_{j_1}, Z_{j_2}, \ldots, Z_{j_r})^\top$ and $\boldsymbol{z}_r = (z_{j_1}, z_{j_2}, \ldots, z_{j_r})^\top$, the distribution of $\boldsymbol{Z}_r$ is

$$\Pr(\boldsymbol{Z}_r = \boldsymbol{z}_r) = \begin{cases} \dfrac{1}{\pi_0} \displaystyle\prod_{j\in J} f_{Y_j}(z_j), & \boldsymbol{z}_r \neq \boldsymbol{0}_r \\ \dfrac{\prod_{j\in J}(1-\pi_j) - \prod_{j=1}^{m}(1-\pi_j)}{\pi_0}, & \boldsymbol{z}_r = \boldsymbol{0}_r. \end{cases}$$

Proof. If $\boldsymbol{z}_r \neq \boldsymbol{0}_r$,

$$\begin{aligned}
\Pr(\boldsymbol{Z}_r = \boldsymbol{z}_r) &= \sum_{z_{j_{r+1}}=0}^{\infty} \cdots \sum_{z_{j_m}=0}^{\infty} \Pr(\boldsymbol{Z} = \boldsymbol{z}) \\
&= \frac{1}{\pi_0} \prod_{j\in J} f_{Y_j}(z_j) \prod_{j\in J^C} \sum_{z_j=0}^{\infty} f_{Y_j}(z_j) \\
&= \frac{1}{\pi_0} \prod_{j\in J} f_{Y_j}(z_j).
\end{aligned}$$

Thus,

$$\begin{aligned}
\Pr(\boldsymbol{Z}_r = \boldsymbol{0}_r) &= 1 - \sum_{\|\boldsymbol{z}_r\|_1 > 0} \Pr(\boldsymbol{Z}_r = \boldsymbol{z}_r) \\
&= 1 - \frac{1}{\pi_0} \sum_{\|\boldsymbol{z}_r\|_1 > 0} \prod_{j \in J} f_{Y_j}(z_j) \\
&= 1 - \frac{1}{\pi_0} \left[1 - \sum_{\|\boldsymbol{z}_r\|_1 = 0} \prod_{j \in J} f_{Y_j}(z_j) \right] \\
&= 1 - \frac{1}{\pi_0} \left[1 - \prod_{j \in J} (1 - \pi_j) \right] \\
&= \frac{\prod_{j \in J} (1 - \pi_j) - \prod_{j=1}^{m} (1 - \pi_j)}{\pi_0}.
\end{aligned}$$

■

B.2 Conditional distributions

Following the same notation for J and J^C, let $\boldsymbol{Z}_{m-r} = (Z_{j_{r+1}}, Z_{j_{r+2}}, \ldots, Z_{j_m})^\top$ and $\boldsymbol{z}_{m-r} = (z_{j_{r+1}}, z_{j_{r+2}}, \ldots, z_{j_m})^\top$. The conditional distribution of $\boldsymbol{Z}_r \mid \boldsymbol{Z}_{m-r}$ is

$$\begin{aligned}
&\Pr(\boldsymbol{Z}_r = \boldsymbol{z}_r \mid \boldsymbol{Z}_{m-r} = \boldsymbol{z}_{m-r}) \\
&\qquad = \begin{cases} \displaystyle\prod_{j \in J} f_{Y_j}(z_j), & \boldsymbol{z}_{m-r} \neq \boldsymbol{0}_{m-r} \\ \displaystyle\frac{1}{1 - \prod_{j \in J} (1 - \pi_j)} \prod_{j \in J} f_{Y_j}(z_j), & \boldsymbol{z}_r \neq \boldsymbol{0}_r, \boldsymbol{z}_{m-r} = \boldsymbol{0}_{m-r}. \end{cases}
\end{aligned}$$

Proof. If $\boldsymbol{z}_{m-r} \neq \boldsymbol{0}_{m-r}$,

$$\begin{aligned}
\Pr(\boldsymbol{Z}_r = \boldsymbol{z}_r \mid \boldsymbol{Z}_{m-r} = \boldsymbol{z}_{m-r}) &= \frac{\Pr(\boldsymbol{Z} = \boldsymbol{z})}{\Pr(\boldsymbol{Z}_{m-r} = \boldsymbol{z}_{m-r})} = \frac{\frac{1}{\pi_0} \prod_{j=1}^{m} f_{Y_j}(z_j)}{\frac{1}{\pi_0} \prod_{j \in J^C} f_{Y_j}(z_j)} \\
&= \prod_{j \in J} f_{Y_j}(z_j).
\end{aligned}$$

If $\boldsymbol{z}_{m-r} = \mathbf{0}_{m-r}$, it is obvious that $\boldsymbol{z}_r \neq \mathbf{0}_r$. When $\boldsymbol{z}_r \neq \mathbf{0}_r$, we have

$$\begin{aligned}\Pr(\boldsymbol{Z}_r = \boldsymbol{z}_r \mid \boldsymbol{Z}_{m-r} = \mathbf{0}_{m-r}) &= \frac{\frac{1}{\pi_0}\prod_{j\in J} f_{Y_j}(z_j)\prod_{j\in J^C}(1-\pi_j)}{\frac{1}{\pi_0}\left[\prod_{j\in J^C}(1-\pi_j) - \prod_{j=1}^{m}(1-\pi_j)\right]} \\ &= \frac{1}{1-\prod_{j\in J}(1-\pi_j)}\prod_{j\in J} f_{Y_j}(z_j).\end{aligned}$$

■

B.3 Expectation, variance and covariance

The expectation and variance of Z_j, $j = 1, \ldots, m$, are

$$E(Z_j) = \frac{\mu_{1j}}{\pi_0}, \quad Var(Z_j) = \frac{\mu_{2j}}{\pi_0} - \frac{\mu_{1j}^2}{\pi_0^2},$$

and the covariance between Z_j and Z_k, $j, k = 1, \ldots, m$, $j \neq k$, are

$$Cov(Z_j, Z_k) = -\frac{\mu_{1j}\mu_{1k}(1-\pi_0)}{\pi_0^2} < 0,$$

where $\mu_{1j} = E(Y_j)$, $\mu_{1k} = E(Y_k)$ and $\mu_{2j} = E(Y_j^2)$.

Proof. It is easy to get from $\boldsymbol{Y} \stackrel{d}{=} U\boldsymbol{Z}$. ■

B.4 The moment generating function

The moment generating function of $\boldsymbol{Z}$ is

$$M_{\boldsymbol{Z}}(\boldsymbol{t}) = \frac{\prod_{j=1}^{m} M_{Y_j}(t_j) - (1-\pi_0)}{\pi_0},$$

where $\boldsymbol{t} = (t_1, \ldots, t_m)^\top$.

Proof. The moment generating function of $\boldsymbol{Y}$ is

$$\begin{aligned}M_{\boldsymbol{Y}}(\boldsymbol{t}) &= E\left[\exp(\boldsymbol{t}^\top \boldsymbol{Y})\right] = E\left[\exp(U\boldsymbol{t}^\top \boldsymbol{Z})\right] = E\left\{E\left[\exp(U\boldsymbol{t}^\top \boldsymbol{Z})\,|U\right]\right\} \\ &= E\left[M_{\boldsymbol{Z}}(U\boldsymbol{t})\right] = (1-\pi_0) + \pi_0 M_{\boldsymbol{Z}}(\boldsymbol{t}).\end{aligned}$$

So the moment generating function of $\boldsymbol{Z}$ is

$$M_{\boldsymbol{Z}}(\boldsymbol{t}) = \frac{M_{\boldsymbol{Y}}(\boldsymbol{t}) - (1-\pi_0)}{\pi_0} = \frac{\prod_{j=1}^{m} M_{Y_j}(t_j) - (1-\pi_0)}{\pi_0}.$$

■