A NEW MULTIVARIATE ZERO-INFLATED HURDLE MODEL WITH APPLICATIONS IN AUTOMOBILE INSURANCE

BY

PENGCHENG ZHANG, DAVID PITT AND XUEYUAN WU

ABSTRACT

The fact that a large proportion of insurance policyholders make no claims during a one-year period highlights the importance of zero-inflated count models when analyzing the frequency of insurance claims. There is a vast literature focused on the univariate case of zero-inflated count models, while work in the area of multivariate models is considerably less advanced. Given that insurance companies write multiple lines of insurance business, where the claim counts on these lines of business are often correlated, there is a strong incentive to analyze multivariate claim count models. Motivated by the idea of Liu and Tian (*Computational Statistics and Data Analysis*, **83**, 200-222; 2015), we develop a multivariate zero-inflated hurdle model to describe multivariate count data with extra zeros. This generalization offers more flexibility in modeling the behavior of individual claim counts while also incorporating a correlation structure between claim counts for different lines of insurance business. We develop an application of the expectation-maximization (EM) algorithm to enable the statistical inference necessary to estimate the parameters associated with our model. Our model is then applied to an automobile insurance portfolio from a major insurance company in Spain. We demonstrate that the model performance for the multivariate zero-inflated hurdle model is superior when compared to several alternatives.

KEYWORDS

Multivariate zero-inflation, Hurdle model, EM algorithm, automobile insurance.

1. INTRODUCTION

Generalized linear models (GLMs) have been widely applied in insurance ratemaking over the last few decades. Within the GLM framework, Poisson and negative binomial distributions have been routinely applied to the analysis

Astin Bulletin 52(2), 393–416. doi:10.1017/asb.2021.39

of insurance claim frequencies. Some detailed illustrations can be found in Cameron and Trivedi (1998) and Frees (2009). In automobile insurance, many attempts have also been made in the actuarial literature to find an appropriate distribution to model the dollar cost associated with individual claims. It is often the case that automobile insurance policyholders do not incur any loss in the period of insurance coverage. Even if the insureds do incur a loss, they may opt not to put forward a claim in order to maintain a high level of no claims discount to their annual insurance premium. In short, insurance claim frequency data are often characterized by a large number of zero claims. This, in turn, leads to the study of zero-inflated versions of typical count distributions (see, e.g., Yip and Yau, 2005 and Boucher *et al.*, 2007).

In practice, an insurer may provide multiple lines of insurance business, where claims from different sources are bundled into one single policy. For example, in automobile insurance, one policy may cover the payment in respect of third-party liability and also losses sustained by the insured party. It is clear from data that claims from these two forms of insurance are often triggered from the same event leading to correlation between observed claim counts. Dependence between costs associated with two (or more) different types of claims must be considered when building a robust ratemaking system. As pointed in Bermúdez (2009), even a minor correlation between the claim counts can lead to major differences in ratemaking. Failure to take into account the positive correlation between the claims will often result in premiums which are too low relative to the underlying risk.

In the literature, there were some papers discussing zero-inflated models in the multivariate version. Most of these models concentrated on the case where the marginal claim count distributions are Poisson. Li *et al.* (1999) proposed a multivariate zero-inflated Poisson model as a mixture of $m+2$ components of m-dimensional discrete distributions. The complexity of this model renders the implementation of maximum likelihood estimation not an easy job for large m. Bermúdez and Karlis (2011) considered zero-inflated versions for the multivariate Poisson model with common and full covariance structures. The inference procedure was completed using a Bayesian framework. Using a similar structure as in Li *et al.* (1999), Dong *et al.* (2014) gave a multivariate zero-inflated negative binomial model.

Recently Liu and Tian (2015) examined a new multivariate zero-inflated Poisson model. This model had the advantage of avoiding the computation issues resulting from an increase of dimension. However, all components in this model share a common zero-inflation parameter which restricts scope for application of this model. In addition, the fact that each marginal distribution is assumed to be Poisson imposes a restriction on the ability of the model to work with many different data sets. In an attempt to alleviate these limitations, we propose a multivariate zero-inflated hurdle model. A univariate hurdle model (Mullahy, 1986) is a two-part model that separates the occurrence of

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

an event from the number of those events actually observed. Constructing a multivariate hurdle model by assuming a hurdle distribution in every margin has two advantages. First, it can easily handle the zero-inflation or zero-deflation feature in each margin. Second, it provides users with greater choice in modeling marginal behaviors. See some examples regarding the implementation of hurdle models in Boucher *et al.* (2007). As a result, our proposed multivariate zero-inflated hurdle model has greater in-built flexibility than the multivariate zero-inflated Poisson model considered in Liu and Tian (2015) in respect of diversifying marginal zero-inflation parameters and employing non-Poisson marginal distributions. The hurdle model has also been considered in a multivariate context by Zhang *et al.* (2020), but to study the Type I multivariate zero-truncated data, which has a very different feature from the type of data studied in this paper.

In our work, the inference process is enabled using the EM algorithm (Dempster *et al.*, 1977). The EM algorithm is a two-step iterative method to find the maximum likelihood estimates (MLEs). It is particularly useful when working with zero-inflated models. Examples illustrating the implementation of the EM algorithm in zero-inflated models can be found in Lambert (1992) and Hall (2000). Unfortunately, when covariates are introduced in our model, there is no closed-form representation in the M-step. We could find the optimal values in the M-step using the Newton-Raphson method however this was shown to be computationally expensive. Rai (1993) provided an alternative approach in which the Newton-Raphson method is carried out for only one step in the M-step. This can reduce the computation time considerably.

Our work contributes to the existing literature in several ways. First, we provide a very efficient way to generalize the zero-inflated model from univariate case to multivariate case. Our proposed framework can easily handle high dimensional data without any computational issues. Second, our model differs from the model of Liu and Tian (2015) in the sense that hurdle margins are assumed here instead of just Poisson margins. This generalization preserves the flexibility of capturing types of features in the insurance claims data. It is also worth stressing that no closed-form exists anymore in the M-step due to the introduction of covariates, enabling us to use a generalized EM algorithm for inference. Third, we emphasize the better performance of our proposed model compared with several existing candidate models when fitted using the same insurance data set.

The rest of the article is organized as follows. Section 2 provides the definition of general multivariate zero-inflated distribution and investigates some of its distributional properties. In Section 3, we propose our multivariate zero-inflated hurdle model, followed by the corresponding EM algorithm for model inference. In Section 4, the proposed model is applied to an automobile insurance data set. The last section concludes the paper.

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

2. Multivariate zero-inflated distribution

2.1. Definition

We now define our new multivariate zero-inflated distribution. 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 and defined on $\mathbb{N}$. Then $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^\top$ is said to follow the multivariate zero-inflated distribution if

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

where $U \sim \textit{Bernoulli}(\pi_0)$, $0 < \pi_0 < 1$, and U is independent of $\boldsymbol{Y}$. The symbol "$\overset{d}{=}$" means that the random variables on both sides of the equality share the same distribution. The probability mass function (pmf) of $\boldsymbol{Z}$ can be derived as

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} \Pr(Y_j = 0)\right]^{v} \left[\pi_0 \prod_{j=1}^{m} \Pr(Y_j = z_j)\right]^{1-v}, \tag{2.2}$$

where $\boldsymbol{z} = (z_1, \ldots, z_m)^\top$ is a vector of observed values, $v = \mathbb{I}(\boldsymbol{z} = \mathbf{0}_m)$ and $\mathbb{I}(\,\cdot\,)$ is an indicator function.

2.2. Properties of distribution

2.2.1. *Marginal distributions*

We first derive the marginal distribution for a single variable. The marginal distribution of Z_j is

$$\Pr(Z_j = z_j) = \begin{cases} \pi_0 f_{Y_j}(z_j), & z_j > 0, \\ 1 - \pi_0 + \pi_0 f_{Y_j}(0), & z_j = 0. \end{cases}$$

Proof. If $z_j > 0$,

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

Thus,

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

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

Next we derive an expression for the marginal distribution of an arbitrary random sub-vector 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

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

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

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

Thus,

$$\begin{aligned}\Pr(\boldsymbol{Z}_r=\boldsymbol{0}_r)&=1-\sum_{\boldsymbol{z}_r\neq\boldsymbol{0}_r}\Pr(\boldsymbol{Z}_r=\boldsymbol{z}_r)\\ &=1-\pi_0\sum_{\boldsymbol{z}_r\neq\boldsymbol{0}_r}\prod_{j\in J}f_{Y_j}(z_j)\\ &=1-\pi_0\left[1-\sum_{\boldsymbol{z}_r=\boldsymbol{0}_r}\prod_{j\in J}f_{Y_j}(z_j)\right]\\ &=1-\pi_0\left[1-\prod_{j\in J}f_{Y_j}(0)\right]\\ &=1-\pi_0+\pi_0\prod_{j\in J}f_{Y_j}(0).\end{aligned}$$

□

The marginal distributions can also be obtained from the definition $\boldsymbol{Z}\overset{d}{=}U\boldsymbol{Y}$.

2.2.2. *Conditional distribution*

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 $\boldsymbol{Z}_r|\boldsymbol{Z}_{m-r}$ is

$$\Pr\left(\boldsymbol{Z}_r = \boldsymbol{z}_r | \boldsymbol{Z}_{m-r} = \boldsymbol{z}_{m-r}\right) = \begin{cases} \prod_{j\in J} f_{Y_j}(z_j), & \boldsymbol{z}_{m-r} \neq \boldsymbol{0}_{m-r}, \\ \pi_0^* \prod_{j\in J} f_{Y_j}(z_j), & \boldsymbol{z}_r \neq \boldsymbol{0}_r, \boldsymbol{z}_{m-r} = \boldsymbol{0}_{m-r}, \\ 1 - \pi_0^* + \pi_0^* \prod_{j\in J} f_{Y_j}(0), & \boldsymbol{z}_r = \boldsymbol{0}_r, \boldsymbol{z}_{m-r} = \boldsymbol{0}_{m-r}, \end{cases}$$

where $\pi_0^* = \frac{\pi_0 \prod_{j\in J^C} f_{Y_j}(0)}{1-\pi_0+\pi_0 \prod_{j\in J^C} f_{Y_j}(0)}$.

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

$$\Pr\left(\boldsymbol{Z}_r = \boldsymbol{z}_r | \boldsymbol{Z}_{m-r} = \boldsymbol{z}_{m-r}\right) = \frac{\pi_0 \prod_{j=1}^m f_{Y_j}(z_j)}{\pi_0 \prod_{j\in J^C} f_{Y_j}(z_j)} = \prod_{j\in J} f_{Y_j}(z_j).$$

If $\boldsymbol{z}_{m-r} = \boldsymbol{0}_{m-r}$ and $\boldsymbol{z}_r \neq \boldsymbol{0}_r$,

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

If $\boldsymbol{z}_{m-r} = \boldsymbol{0}_{m-r}$ and $\boldsymbol{z}_r = \boldsymbol{0}_r$,

$$\begin{aligned} \Pr\left(\boldsymbol{Z}_r = \boldsymbol{0}_r | \boldsymbol{Z}_{m-r} = \boldsymbol{0}_{m-r}\right) &= \frac{1-\pi_0+\pi_0 \prod_{j=1}^m f_{Y_j}(0)}{1-\pi_0+\pi_0 \prod_{j\in J^C} f_{Y_j}(0)} \\ &= 1 - \pi_0^* + \pi_0^* \prod_{j\in J} f_{Y_j}(0). \end{aligned}$$ □

2.2.3. *Expectation, variance and covariance*

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

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

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

$$\mathrm{Cov}(Z_j, Z_k) = \pi_0(1-\pi_0)\mu_{1j}\mu_{1k} > 0,$$

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

Proof. This is easily obtained from $\boldsymbol{Z} \overset{d}{=} U\boldsymbol{Y}$. □

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

2.2.4. *Moment generating function*
The moment generating function of $\boldsymbol{Z}$ is

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

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

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

$$\begin{aligned} M_{\boldsymbol{Z}}(\boldsymbol{t}) &= \mathrm{E}\left[\exp\left(\boldsymbol{t}^\top \boldsymbol{Z}\right)\right] = \mathrm{E}\left[\exp\left(U\boldsymbol{t}^\top \boldsymbol{Y}\right)\right] = \mathrm{E}\left\{\mathrm{E}\left[\exp\left(U\boldsymbol{t}^\top \boldsymbol{Y}\right)|U\right]\right\} \\ &= \mathrm{E}\left[M_{\boldsymbol{Y}}(U\boldsymbol{t})\right] = 1 - \pi_0 + \pi_0 M_{\boldsymbol{Y}}(\boldsymbol{t}) = 1 - \pi_0 + \pi_0 \prod_{j=1}^{m} M_{Y_j}(t_j). \end{aligned}$$ □

2.3. Two special cases

In this subsection, we introduce two multivariate zero-inflated distributions where the margins are assumed to be either all Poisson or all negative binomial distributed. Users do not have the flexibility to vary the marginal distribution types for these two distributions. We note that the multivariate zero-inflated Poisson distribution was first proposed by Liu and Tian (2015). To our best knowledge, the multivariate zero-inflated negative binomial distribution has not been studied in the literature before, but it has the same limitations as the Poisson model, which was discussed above in Section 1. We will use these two models in our model comparison in Section 4. For the purpose of simplification, we have put the EM algorithms of parameter estimation for these two models in Appendix A, where the location parameters are all covariate-dependent.

2.3.1. *Multivariate zero-inflated Poisson distribution*
Let $Y_j \sim Poisson(\lambda_j)$, for $j = 1, \ldots, m$. Then $\boldsymbol{Z}$ is said to follow the multivariate zero-inflated Poisson distribution with the parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ and a zero-inflation parameter π_0, denoted by $\boldsymbol{Z} \sim MZIP(\boldsymbol{\lambda}, \pi_0)$. The pmf of $\boldsymbol{Z}$ is

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \left(1 - \pi_0 + \pi_0 e^{-\sum_{j=1}^{m} \lambda_j}\right)^{v} \left(\pi_0 \prod_{j=1}^{m} \frac{\lambda_j^{z_j} e^{-\lambda_j}}{z_j!}\right)^{1-v}, \tag{2.3}$$

where $v = \mathbb{I}(\boldsymbol{z} = \boldsymbol{0}_m)$.

2.3.2. *Multivariate zero-inflated negative binomial distribution*
Let $Y_j \sim NB(\mu_j, \theta_j)$, for $j = 1, \ldots, m$. Then $\boldsymbol{Z}$ is said to follow the multivariate zero-inflated negative binomial distribution with two parameter vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$ and a zero-inflation parameter π_0, denoted by

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

$\boldsymbol{Z} \sim MZINB(\boldsymbol{\mu}, \boldsymbol{\theta}, \pi_0)$. The pmf of $\boldsymbol{Z}$ is

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} \left(\frac{\theta_j}{\mu_j + \theta_j}\right)^{\theta_j}\right]^{v} \times \left[\pi_0 \prod_{j=1}^{m} \frac{\Gamma(z_j + \theta_j)}{\Gamma(\theta_j) z_j!} \left(\frac{\mu_j}{\mu_j + \theta_j}\right)^{z_j} \left(\frac{\theta_j}{\mu_j + \theta_j}\right)^{\theta_j}\right]^{1-v}, \tag{2.4}$$

where $v = \mathbb{I}(\boldsymbol{z} = \boldsymbol{0}_m)$.

3. Multivariate zero-inflated hurdle model

3.1. Model characterization

We shall assume that each underlying random variable Y_j in (2.1) 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 count distribution defined on $\mathbb{N}_+$, $U_j \sim Bernoulli(\pi_j)$, $0 < \pi_j < 1$, and U_j is independent of W_j. Again, we assume that all Y_j are independent of each other. Then $\boldsymbol{Z}$ constructed by (2.1) is said to follow the multivariate zero-inflated hurdle distribution with parameter vectors $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^\top$, $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_m)^\top$ and a zero-inflation parameter π_0. Here $\boldsymbol{\Theta}_j$ is the set of parameters related to W_j. The pmf of $\boldsymbol{Z}$ can be expressed as

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_j)\right]^{v} \times \left[\pi_0 \prod_{j:z_j=0} (1 - \pi_j) \prod_{j:z_j \neq 0} \pi_j f_{W_j}(z_j)\right]^{1-v}, \tag{3.2}$$

where $v = \mathbb{I}(\boldsymbol{z} = \boldsymbol{0}_m)$.

3.2. Model inference

Suppose each $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, independently follows a multivariate zero-inflated hurdle distribution. The corresponding observed values are $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})^\top$. The latent variables are $v_1, \ldots, v_n$, where $v_i = \mathbb{I}(\boldsymbol{z}_i = \boldsymbol{0}_m)$. Now we introduce some covariates, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

$\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$. The parameter π_{ij} can then 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 $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})^\top$. For the purpose of easy interpretation, we do not inject covariates in π_0. We denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ as the set of parameters related to all π_{ij}, and $\boldsymbol{\Theta}$ as the set of parameters related to all W_j, the likelihood function then can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Theta}, \pi_0) = &\prod_{i=1}^{n} \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_{ij}) \right]^{v_i} \\ &\times \prod_{i=1}^{n} \left[\pi_0 \prod_{j:z_{ij}=0} (1 - \pi_{ij}) \prod_{j:z_{ij}\neq 0} \pi_{ij} f_{W_j}(z_{ij}) \right]^{1-v_i}. \end{aligned} \tag{3.4}$$

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

$$\begin{aligned} \ell_1(\boldsymbol{\beta}, \pi_0) = &\sum_{i=1}^{n} v_i \log \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_{ij}) \right] + \sum_{i=1}^{n} (1 - v_i) \log \pi_0 \\ &+ \sum_{i=1}^{n} (1 - v_i) \left[\sum_{j:z_{ij}=0} \log(1 - \pi_{ij}) + \sum_{j:z_{ij}\neq 0} \log \pi_{ij} \right], \\ \ell_2(\boldsymbol{\Theta}) = &\sum_{i=1}^{n} \sum_{j:z_{ij}\neq 0} (1 - v_i) \log f_{W_j}(z_{ij}) = \sum_{j=1}^{m} \sum_{i:z_{ij}\neq 0} \log f_{W_j}(z_{ij}). \end{aligned}$$

Thus, the maximization procedure can be completed for ℓ_1 and ℓ_2, respectively. For ℓ_2, the estimation can proceed in respect of the zero-truncation part of each margin separately. For ℓ_1, we implement the EM algorithm as described below.

Denote $\boldsymbol{Z}' = (Z'_1, \ldots, Z'_m)^\top$ where $Z'_j = \mathbb{I}(Z_j > 0)$. 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

$$\begin{aligned} \ell_1(\boldsymbol{\beta}, \pi_0) = &\sum_{i=1}^{n} v_i \log \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_{ij}) \right] + \sum_{i=1}^{n} (1 - v_i) \log \pi_0 \\ &+ \sum_{j=1}^{m} \sum_{i=1}^{n} (1 - v_i) \left[z'_{ij} \log \pi_{ij} + (1 - z'_{ij}) \log(1 - \pi_{ij}) \right]. \end{aligned}$$

In addition to the known values $\boldsymbol{z}'_i$, suppose we also know the value u'_i, one for each individual to take the value 1 if the observation of common zeros is inflated and 0 otherwise. The complete data log-likelihood function then becomes

$$\begin{aligned}\ell_1^c(\boldsymbol{\beta}, \pi_0) &= \sum_{i=1}^{n} \left[u_i' v_i \log (1 - \pi_0) + (1 - u_i' v_i) \log \pi_0 \right] \\ &\quad + \sum_{j=1}^{m} \sum_{i=1}^{n} \left[z_{ij}' \log \pi_{ij} + (1 - u_i' v_i - z_{ij}') \log (1 - \pi_{ij}) \right].\end{aligned}$$

Note in our case, $v_i z_{ij}' = 0$. Given initial values of parameters $\boldsymbol{\beta}$ and π_0, the EM algorithm proceeds as follows.

- E-step: Replace u_i' by

  $$\bar{u}_i' = \frac{1 - \pi_0}{1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_{ij})}, \quad i = 1, \ldots, n,$$

  where $\pi_{ij} = \frac{\exp (\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{1 + \exp (\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}$.
- M-step:
  - For π_0, we can get

    $$\pi_0 = 1 - \frac{1}{n} \sum_{i=1}^{n} \bar{u}_i' v_i.$$

  - For $\boldsymbol{\beta}$, let

    $$\bar{\ell}_{1j}^c(\boldsymbol{\beta}_j) = \sum_{i=1}^{n} \left[z_{ij}' \log \pi_{ij} + (1 - \bar{u}_i' v_i - z_{ij}') \log (1 - \pi_{ij}) \right].$$

  There is no closed-form representation for $\boldsymbol{\beta}_j$, so we update the regression parameters, respectively, for each j by implementing the Newton–Raphson method for one step. The first- and second-order derivatives are given as follows:

  $$\begin{aligned}\frac{\partial \bar{\ell}_{1j}^c}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n} \left[z_{ij}' - (1 - \bar{u}_i' v_i) \pi_{ij} \right] \boldsymbol{x}_i, \\ \frac{\partial^2 \bar{\ell}_{1j}^c}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} &= - \sum_{i=1}^{n} (1 - \bar{u}_i' v_i) \pi_{ij} (1 - \pi_{ij}) \boldsymbol{x}_i \boldsymbol{x}_i^\top.\end{aligned}$$

- Iterate through the E-step and M-step until some convergence criterion is met, for example, the relative change of observed log-likelihood between two consecutive iterations is smaller than a tolerance level ε.

Remark 1. *The initial values of parameters $\boldsymbol{\beta}_j$ for EM algorithm can be obtained by implementing univariate logistic regression with observed values $z_{1j}', \ldots, z_{nj}'$.*

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

The initial value of parameter π_0 can be set as 0.5. The standard errors for the estimates can be approximated using the approach in Louis (1982).

For the case without covariates incorporated in π_{ij}, the EM algorithm is simplified as shown below.

- E-step: Replace u'_i by

$$\bar{u}' = \bar{u}'_i = \frac{1-\pi_0}{1-\pi_0+\pi_0\prod_{j=1}^{m}(1-\pi_j)}, \quad i=1,\ldots,n.$$

  Initial value for π_j can be set as the proportion of non-zeros for margin j.
- M-step:
  - Update π_0 through the following equation:

$$\pi_0 = 1 - \frac{\bar{u}'}{n}\sum_{i=1}^{n} v_i.$$

  - Update each π_j through the following equation:

$$\pi_j = \frac{\sum_{i=1}^{n} z'_{ij}}{n-\bar{u}'\sum_{i=1}^{n} v_i}, \quad j=1,\ldots,m.$$

4. Application

4.1. Data description

This application is based on an automobile portfolio from a major insurance company operating in Spain in 1995. The whole data set consists of 80,994 policyholders. We have access to a rich set of information for each individual. The detailed description for each predictor is presented in Table 1. The mean of each covariate is also provided in the table to show the proportion of corresponding group. For example, the mean of $v1$ tells us that 16.0% of the policyholders are female.

The simplest policy only includes third-party liability (denoted as Z_1 type) and a set of basic guarantees such as emergency roadside assistance, legal assistance or insurance covering medical costs (denoted as Z_2 type). The comprehensive coverage (damage to one's vehicle caused by any unknown party, for example, damage resulting from theft, flood or fire) and the collision coverage (damage resulting from a collision with another vehicle or object when the policyholder is at fault) are also denoted as Z_2 type. For our purpose, we only select policyholders with full coverage ($v9=0$, $v10=1$) to implement our analysis. This is consistent with the analysis in Bermúdez and Karlis (2017). This subset only contains information for 28,590 policyholders. The empirical joint

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

TABLE 1

THE DESCRIPTION FOR EXPLANATORY VARIABLES.

Variable	Description	Mean
$v1$	= 1 for women; = 0 for men	0.160
$v2$	= 1 when driving in urban area; = 0 otherwise	0.669
$v3$	= 1 when zone is medium risk (Madrid and Catalonia)	0.239
$v4$	= 1 when zone is high risk (northern Spain)	0.194
$v5$	= 1 if the driving license is between 4 and 14 years old	0.257
$v6$	= 1 if the driving license is 15 or more years old	0.719
$v7$	= 1 if the client is in the company for more than 5 years	0.856
$v8$	= 1 if the insured is 30 years old or younger	0.092
$v9$	= 1 if includes comprehensive coverage (except fire)	0.156
$v10$	= 1 if includes comprehensive and collision coverage	0.353
$v11$	= 1 if horsepower is $\geq$ 5500 cc	0.806

TABLE 2

THE EMPIRICAL JOINT DISTRIBUTION OF Z_1 AND Z_2.

	Z_2						
Z_1	0	1	2	3	4	5	6
0	24,408	1916	296	69	12	6	0
1	1068	317	61	21	6	2	2
2	203	71	18	6	2	1	1
3	49	14	8	3	3	1	0
4	11	6	2	0	1	0	0
5	2	0	0	0	0	0	1
6	1	0	0	1	0	0	0
8	0	0	1	0	0	0	0

distribution for claim numbers Z_1 and Z_2 is displayed in Table 2. The overall Pearson's correlation coefficient between these two types of claim is 0.189. This observed correlation is consistent with our model's positive correlation assumption. For our study, we randomly take 70% of the observations from the subset as training data to develop the models, and the remaining 30% are reserved as hold-out sample for validation purpose.

4.2. Model fitting

Prior to fitting our proposed hurdle model, we need to determine the distributions for the zero-truncated univariate components $W_j, j = 1, 2$. In addition to the commonly used zero-truncated Poisson (ZTP) and zero-truncated negative binomial (ZTNB) distributions, we also try unit-shifted Poisson (USP)

TABLE 3

GOODNESS-OF-FIT OF MARGINAL MODELS.

W_1	Observed	ZTP	ZTNB	USP	USNB
1	1033	993.71	1037.84	981.99	1032.45
2	207	266.50	202.80	286.75	209.20
3	54	47.65	52.32	41.87	53.21
4	17	6.39	15.14	4.08	14.45
≥5	4	0.75	6.90	0.32	5.68
χ^2		47.34	1.61	112.32	0.98
LogLik		−924.59	−906.31	−940.51	−905.92
W_2	Observed	ZTP	ZTNB	USP	USNB
1	1624	1562.13	1612.94	1548.65	1623.88
2	265	358.21	281.18	382.08	265.14
3	66	54.76	64.72	47.13	66.48
4	18	6.28	16.70	3.88	18.61
≥5	9	0.62	6.46	0.25	7.89
χ^2		163.54	2.13	403.05	0.18
LogLik		−1258.84	−1221.06	−1283.18	−1220.22

and unit-shifted negative binomial (USNB) distributions. Implementing a unit-shifted distribution on W_j is equivalent to using a standard count distribution for $W_j - 1$. These four univariate models can be implemented in R using the packages gamlss and gamlss.tr. The results are summarized in Table 3. Both the log-likelihood values and the χ^2 statistics indicate that the USNB distribution outperforms the alternatives for both claim type counts. The small differences between the observed and fitted frequencies demonstrate the adequacy of adopting a USNB distribution for both margins.

We next turn to parameter estimation when covariates are introduced in our multivariate zero-inflated hurdle (MZIH) model. The estimation results in Table 4 correspond to π_j, $j = 0, 1, 2$. The 95% confidence interval of π_0 is $(0.318, 0.372)$, where the upper bound is far below the boundary 1. This reveals the zero-inflation feature reflected in the data set. Focusing on the claims of Z_1 type, parameters $v4$ and $v7$ are statistically significant. It can be concluded that policyholders with more years with the insurance company ($v7$) exhibit a lower probability of claiming. On the other hand, driving in a high-risk region ($v4$) is associated with an increase in the probability of making a claim. If we focus on the claims of Z_2 type, we notice that $v3$, $v5$, $v7$ and $v11$ are all statistically significant. This suggests that driving in a zone of medium risk ($v3$), driving license between 4 and 14 years ($v5$) and greater horsepower ($v11$) are all associated with increased chances of a claim in this category. However, clients with the company for more than 5 years ($v7$) have a lower probability of making a claim. Table 5 is associated with the estimates for W_j, $j = 1, 2$. In this table, we report the coefficient estimates when a linear model with logarithmic link

TABLE 4

ESTIMATES AND t-RATIOS ASSOCIATED WITH THE COVARIATES OF π_j IN MZIH MODEL.

	π_1		π_2	
	Estimate	t-ratio	Estimate	t-ratio
Intercept	−0.953	−3.474***	−1.271	−4.786***
$v1$	0.029	0.327	0.041	0.496
$v2$	−0.044	−0.629	0.084	1.305
$v3$	0.098	1.227	0.168	2.336*
$v4$	0.274	3.230**	−0.032	−0.397
$v5$	−0.188	−0.863	0.430	2.014*
$v6$	−0.336	−1.462	0.081	0.360
$v7$	−0.259	−3.022**	−0.359	−4.517***
$v8$	0.107	0.873	0.060	0.527
$v11$	−0.086	−0.846	0.398	4.051***
	Estimate	95% CI		
π_0	0.345	(0.318, 0.372)		

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

TABLE 5

ESTIMATES AND t-RATIOS ASSOCIATED WITH THE COVARIATES OF W_j IN MZIH MODEL.

	W_1		W_2	
	Estimate	t-ratio	Estimate	t-ratio
Intercept	−1.299	−2.594**	−1.350	−2.769**
$v1$	−0.112	−0.663	0.098	0.675
$v2$	−0.013	−0.096	0.001	0.009
$v3$	−0.079	−0.531	0.231	1.821
$v4$	−0.304	−1.884	−0.141	−0.896
$v5$	0.129	0.319	0.162	0.393
$v6$	0.141	0.333	0.039	0.092
$v7$	0.012	0.075	−0.037	−0.274
$v8$	−0.081	−0.366	−0.081	−0.415
$v11$	0.053	0.280	−0.178	−0.964
	Estimate	95% CI		
θ_1	0.678	(0.428, 0.928)		
θ_2	0.498	(0.351, 0.644)		

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

function is used to describe the location parameter in the USNB marginal distributions of the MZIH model. As can be observed from this table, conditional on the occurrence of claims, no predictor is significant for the expected number of W_1 and W_2.

Table 6

Observed and expected frequencies of MZIH model for the joint distribution of Z_1 and Z_2.

	Z_2						
Z_1	0	1	2	3	4	5	6
0	17,104	1342	199	41	8	4	0
	(17,102.66)	(1306.43)	(213.31)	(53.49)	(14.97)	(4.41)	(1.34)
1	736	228	46	16	5	1	1
	(731.96)	(246.71)	(40.28)	(10.10)	(2.83)	(0.83)	(0.25)
2	145	42	10	6	2	1	1
	(148.32)	(49.99)	(8.16)	(2.05)	(0.57)	(0.17)	(0.05)
3	34	7	8	2	2	1	0
	(37.73)	(12.72)	(2.08)	(0.52)	(0.15)	(0.04)	(0.01)
4	9	5	2	0	1	0	0
	(10.25)	(3.45)	(0.56)	(0.14)	(0.04)	(0.01)	(0.00)
5	2	0	0	0	0	0	0
	(2.87)	(0.97)	(0.16)	(0.04)	(0.01)	(0.00)	(0.00)
6	1	0	0	1	0	0	0
	(0.82)	(0.28)	(0.05)	(0.01)	(0.00)	(0.00)	(0.00)

Furthermore, we present in Table 6 the observed and expected frequencies (numbers in the brackets) based on the MZIH model. It can be observed that the overall fit is acceptable. The result of the chi-squared test indicates that only a few cells contribute to this goodness-of-fit. It is thus our belief that the overall quality of fit is good considering the number of cells.

Finally, we compare different available models. Apart from our proposed MZIH model, fitted with the USNB marginal distributions, our candidate models also include the multivariate zero-inflated Poisson (MZIP) and multivariate zero-inflated negative binomial (MZINB) models. It is worth mentioning that for the zero-truncation parts in our MZIH model, only intercepts are adopted to avoid the over-fitting problem. Furthermore, a model with two independent hurdle margins (Ind) is fitted as a benchmark. The marginal zero-truncated distributions in the independent hurdle model are the same as for the MZIH model. Different information criteria are provided in Table 7. By assuming the existence of a zero-inflation parameter, we are able to improve the model fitting considerably. This confirms the fact that extra common zeros exist in our data set. Also the superior performance of MZIH model over MZIP and MZINB models verifies the benefit of having extra flexibility in our MZIH model when handling marginal distributions.

4.3. Predictive analysis

To evaluate the predictive performance, we calculate the predicted claim frequencies and compare these to the observed frequencies from the test sample for the following scenarios: no claims in any line, claims occur in only one of

TABLE 7

INFORMATION CRITERIA OF FOUR FITTED MODELS.

Model	Parameters	LogLik	AIC	BIC
MZIH	25	−13,162.50	26,375.00	26,581.52
MZIP	21	−13,313.94	26,669.88	26,843.36
MZINB	23	−13,195.20	26,436.40	26,626.39
Ind	24	−13,372.87	26,793.73	26,991.99

TABLE 8

PREDICTED FREQUENCIES OF FOUR MODELS UNDER FOUR DIFFERENT SCENARIOS.

(Z_1, Z_2)	Observed	MZIH	MZIP	MZINB	Ind
(0, 0)	7304	7332.54	7330.50	7330.93	7224.20
(>0, 0)	407	399.46	403.20	411.71	506.12
(0, >0)	705	681.36	621.17	684.42	790.03
(>0, >0)	161	163.63	222.13	149.94	56.65
χ^2		1.12	28.26	1.59	221.68

the lines and claims occur in both lines. The candidate models include MZIH, MZIP, MZINB and Ind models. The goodness-of-fit results are shown in Table 8. Our conclusions are consistent with those made based on Table 7. As anticipated, we observe very poor performance from the Ind model, which can be explained by the failure to model the excess common zeros in the data set as well as the positive correlation between the two lines of claims. The more accurate prediction of MZIH model against MZIP and MZINB models is largely due to the introduction of hurdle margins. The tiny discrepancies between observed and predicted frequencies of our MZIH model suggest a satisfactory quality of fit for this particular data set.

4.4. Model comparison

In this subsection, we apply the four models from the previous section using the whole data set with 80,994 policyholders to facilitate the comparison between these models and the ones in Bermúdez (2009) and Bermúdez and Karlis (2012). All eleven covariates are considered in this case. The first candidate model is the best model studied in Bermúdez (2009) that is the zero-inflated bivariate Poisson (ZIBP) model with regressors on the third Poisson parameter λ_3. The second candidate model is the best model studied in Bermúdez and Karlis (2012), which is the 2-finite mixture of bivariate Poisson (2-FMBP) model with regressors on the mixing proportion p. For readers' convenience, description of the ZIBP and 2-FMBP models is given in Appendix B. Our MZIH model is still fitted with the USNB marginal distributions for zero-truncation parts with only significant predictors reserved. The comparative

TABLE 9

INFORMATION CRITERIA OF SIX CANDIDATE MODELS.

Model	Parameters	LogLik	AIC	BIC
MZIH	30	−44,777.18	89,614.35	89,893.41
2-FMBP	53	−44,842.22	89,790.44	90,283.45
MZINB	27	−45,069.46	90,192.92	90,444.08
ZIBP	27	−45,414.80	90,883.60	91,134.76
MZIP	25	−45,471.08	90,992.17	91,224.72
Ind	29	−45,721.80	91,501.59	91,771.36

analysis is shown in Table 9. As is evident from the table, all information criteria agree that our proposed MZIH model still performs the best.

5. CONCLUDING REMARKS

In this article, we introduced a new type of multivariate model, the so called multivariate zero-inflated hurdle (MZIH) model. We started with the definition for the multivariate zero-inflated distribution and provided some of its properties. Next, two special multivariate zero-inflated models, namely the MZIP and MZINB, were presented with associated EM algorithms necessary for estimating their parameters given in Appendix A. However, these two models often cannot accommodate the observed characteristics in the marginal distributions of claim counts. Our proposed MZIH model could address these limitations by allowing any zero-modified distribution for each margin. The separation of zeros from the positive parts provided us with more freedom to deal with zero features in each dimension and the marginal distributions as well. The usefulness of our model was illustrated with the help of real data from automobile insurance. The results from fitting several relevant models were shown and compared. As expected, the superiority of our proposed MZIH model was supported by different information criteria as well as predictive performance.

Our proposed MZIH model provides significantly enhanced flexibility compared to univariate models for the actuary who is dealing with multiple related insurance coverage. Our model takes into account the correlations between observed claim frequencies for different lines of insurance business, meaning that the models for one line of business are enhanced by information in the data relating to other lines of business. Univariate models are unable to incorporate this given their separate focus on individual lines of business. Second, the MZIH model builds on the analysis from univariate modeling of claim counts by adding a hurdle requirement within a multivariate structure. More importantly, despite the significantly enhanced model flexibility inherent in MZIH model analysis, the additional effort in coding and computation is limited. Starting form appropriate values, our program for the EM algorithm only takes several minutes to obtain the results when working on tens of thousands

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

of observations. The estimation aimed at the positive part of each margin can be easily handled in R using publicly available packages. All of these advantages make the MZIH model an attractive candidate when studying claims with the zero-inflation feature in a multivariate context.

R CODE

The authors used the R package gamlss and gamlss.tr in the model fitting of this paper. The authors are happy to share their R code when needed. Interested readers please contact the corresponding author.

ACKNOWLEDGEMENTS

The authors would like to thank reviewers for valuable comments and suggestions that led to significant improvement from the first version. Mr. Pengcheng Zhang was supported by a University of Melbourne Faculty of Business and Economics Doctoral Program Scholarship. Acknowledgements also go to the Research Group of Risk in Insurance and Finance at the University of Barcelona for providing the data used in this paper.

REFERENCES

BOUCHER, J.P., DENUIT, M. and GUILLÉN, M. (2007) Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*, **11**(4), 110–131.

BERMÚDEZ, L. (2009) A priori ratemaking using bivariate Poisson regression models. *Insurance: Mathematics and Economics*, **44**(1), 135–141.

BERMÚDEZ, L. and KARLIS, D. (2011) Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, **48**(2), 226–236.

BERMÚDEZ, L. and KARLIS, D. (2012) A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking. *Computational Statistics and Data Analysis*, **56**, 3988–3999.

BERMÚDEZ, L. and KARLIS, D. (2017) A posteriori ratemaking using bivariate Poisson models. *Scandinavian Actuarial Journal*, **2017**(2), 148–158.

CAMERON, A.C. and TRIVEDI, P.K. (1998) *Regression Analysis of Count Data*. New York: Cambridge University Press.

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.

DONG, C., CLARKE, D.B., YAN, X., KHATTAK, A. and HUANG, B. (2014) Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. *Accident Analysis and Prevention*, **70**, 320–329.

FREES, E.W. (2009) *Regression Modeling with Actuarial and Financial Applications*. New York: Cambridge University Press.

HALL, D.B. (2000) Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, **56**(4), 1030–1039.

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

LAMBERT, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

LI, C., LU, J., PARK, J., KIM, K., BRINKLEY, P. and PETERSON, J. (1999) Multivariate zero-inflated Poisson models and their applications. *Technometrics*, **41**(1), 29–38.

LIU, Y. and TIAN, G.L. (2015) Type I multivariate zero-inflated Poisson distribution with applications. *Computational Statistics and Data Analysis*, **83**, 200–222.

LOUIS, T.A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**(2), 226–233.

MULLAHY, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, **33**(3), 341–365.

RAI, S.N. and MATTHEWS, D.E. (1993) Improving the EM algorithm. *Biometrics*, **49**, 587–591.

YIP, K.C.H. and YAU, K.K.W. (2005) On modeling claim frequency data in general insurance with extra zeros. *Insurance Mathematics and Economics*, **36**(2), 153–163.

ZHANG, P., CALDERÍN-OJEDA, E., LI, S. and WU, X. (2020) On the type I multivariate zero-truncated hurdle model with applications in health insurance. *Insurance Mathematics and Economics*, **90**, 35–45.

PENGCHENG ZHANG
School of Insurance
Shandong University of Finance and Economics
Jinan 250014, China
E-Mail: qingdaozpc@163.com

DAVID PITT
Centre for Actuarial Studies
Department of Economics
The University of Melbourne
Melbourne, VIC 3010, Australia
E-Mail: david.pitt@unimelb.edu.au

XUEYUAN WU(Corresponding author)
Centre for Actuarial Studies
Department of Economics
The University of Melbourne
Melbourne, VIC 3010, Australia
E-Mail: xueyuanw@unimelb.edu.au

APPENDIX

APPENDIX A. EM ALGORITHMS FOR TWO MODELS

A.1. Multivariate zero-inflated Poisson model

Suppose for each independent individual, $\boldsymbol{Z}_i \sim MZIP(\boldsymbol{\lambda}_i, \pi_0)$, $i = 1, \ldots, n$, where $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{im})^\top$. Taking covariates into account, the parameter λ_{ij} can be modeled as

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press

$\lambda_{ij} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$ with new parameters $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})^\top$. If we denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$, the likelihood function becomes

$$L(\boldsymbol{\beta}, \pi_0) = \prod_{i=1}^{n} \left(1 - \pi_0 + \pi_0 e^{-\sum_{j=1}^m \lambda_{ij}}\right)^{v_i} \prod_{i=1}^{n} \left(\pi_0 \prod_{j=1}^{m} \frac{\lambda_{ij}^{z_{ij}} e^{-\lambda_{ij}}}{z_{ij}!}\right)^{1-v_i}.$$

Following the same data augmentation method as for the multivariate zero-inflated hurdle model, we obtain the complete data likelihood function given as

$$L^c(\boldsymbol{\beta}, \pi_0) = \prod_{i=1}^{n} (1 - \pi_0)^{u_i' v_i} \prod_{i=1}^{n} \left(\pi_0 e^{-\sum_{j=1}^m \lambda_{ij}}\right)^{(1-u_i')v_i}$$
$$\times \prod_{i=1}^{n} \left(\pi_0 \prod_{j=1}^{m} \frac{\lambda_{ij}^{z_{ij}} e^{-\lambda_{ij}}}{z_{ij}!}\right)^{1-v_i}.$$

The complete data log-likelihood function is

$$\ell^c(\boldsymbol{\beta}, \pi_0) = \sum_{i=1}^{n} \left[u_i' v_i \log(1 - \pi_0) + (1 - u_i' v_i) \log \pi_0\right]$$
$$+ \sum_{j=1}^{m} \sum_{i=1}^{n} \left[z_{ij} \log \lambda_{ij} - (1 - u_i' v_i)\lambda_{ij}\right] + C_0,$$

where C_0 is a constant not related to $(\boldsymbol{\beta}, \pi_0)$. Note in our case, $v_i z_{ij} = 0$.

The associated EM algorithm is given below.

- E-step: Replace u_i' by

$$\bar{u}_i' = \frac{1 - \pi_0}{1 - \pi_0 + \pi_0 e^{-\sum_{j=1}^m \lambda_{ij}}}, \quad i = 1, \ldots, n,$$

  where $\lambda_{ij} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$.
- M-step:
  – For π_0, we can get

$$\pi_0 = 1 - \frac{1}{n} \sum_{i=1}^{n} \bar{u}_i' v_i.$$

  – For $\boldsymbol{\beta}$, let

$$\bar{\ell}_j^c(\boldsymbol{\beta}_j) = \sum_{i=1}^{n} \left[z_{ij} \log \lambda_{ij} - (1 - \bar{u}_i' v_i)\lambda_{ij}\right].$$

  There is no closed-form representation for $\boldsymbol{\beta}_j$, so we update the regression parameters, respectively, for each j by implementing the Newton–Raphson method for one step. The first- and second-order derivatives are given as follows:

$$\frac{\partial \bar{\ell}_j^c}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n} \left[z_{ij} - (1 - \bar{u}'_i v_i)\lambda_{ij} \right] \boldsymbol{x_i},$$

$$\frac{\partial^2 \bar{\ell}_j^c}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^{\top}} = -\sum_{i=1}^{n} (1 - \bar{u}'_i v_i)\lambda_{ij} \boldsymbol{x_i} \boldsymbol{x_i}^{\top}.$$

For the case without covariates incorporated in λ_{ij}, the corresponding E-step and M-step can be simplified as follows.

- E-step: Replace u_i by

$$\bar{u}' = \bar{u}'_i = \frac{1 - \pi_0}{1 - \pi_0 + \pi_0 e^{-\sum_{j=1}^{m} \lambda_j}}, \quad i = 1, \ldots, n.$$

- M-step:
  - Update π_0 using the following equation:

$$\pi_0 = 1 - \frac{\bar{u}'}{n} \sum_{i=1}^{n} v_i.$$

  - Update each λ_j using the following equation:

$$\lambda_j = \frac{\sum_{i=1}^{n} z_{ij}}{n - \bar{u}' \sum_{i=1}^{n} v_i}, \quad j = 1, \ldots, m.$$

A.2. Multivariate zero-inflated negative binomial model

Suppose for each independent individual, $\boldsymbol{Z}_i \sim MZINB(\boldsymbol{\mu}_i, \boldsymbol{\theta}, \pi_0)$, $i = 1, \ldots, n$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})^{\top}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\top}$. Similarly, the covariates could be introduced as $\mu_{ij} = \exp(\boldsymbol{x}_i^{\top} \boldsymbol{\beta}_j)$. If we denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$, the likelihood function becomes

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \pi_0) = \prod_{i=1}^{n} \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} \left(\frac{\theta_j}{\mu_{ij} + \theta_j} \right)^{\theta_j} \right]^{v_i} \times \prod_{i=1}^{n} \left[\pi_0 \prod_{j=1}^{m} \frac{\Gamma(z_{ij} + \theta_j)}{\Gamma(\theta_j) z_{ij}!} \left(\frac{\mu_{ij}}{\mu_{ij} + \theta_j} \right)^{z_{ij}} \left(\frac{\theta_j}{\mu_{ij} + \theta_j} \right)^{\theta_j} \right]^{1 - v_i}.$$

Following the same data augmentation method as for the multivariate zero-inflated hurdle model, we obtain the complete data likelihood function given as

$$L^c(\boldsymbol{\beta}, \boldsymbol{\theta}, \pi_0) = \prod_{i=1}^{n} (1 - \pi_0)^{u'_i v_i} \prod_{i=1}^{n} \left[\pi_0 \prod_{j=1}^{m} \left(\frac{\theta_j}{\mu_{ij} + \theta_j} \right)^{\theta_j} \right]^{(1 - u'_i) v_i} \times \prod_{i=1}^{n} \left[\pi_0 \prod_{j=1}^{m} \frac{\Gamma(z_{ij} + \theta_j)}{\Gamma(\theta_j) z_{ij}!} \left(\frac{\mu_{ij}}{\mu_{ij} + \theta_j} \right)^{z_{ij}} \left(\frac{\theta_j}{\mu_{ij} + \theta_j} \right)^{\theta_j} \right]^{1 - v_i}.$$

The complete data log-likelihood function is

$$\begin{aligned}\ell^c(\boldsymbol{\beta},\boldsymbol{\theta},\pi_0) = &\sum_{i=1}^{n}\left[u_i'v_i\log(1-\pi_0)+(1-u_i'v_i)\log\pi_0\right]\\ &+\sum_{j=1}^{m}\sum_{i=1}^{n}\Big\{(1-u_i'v_i)\theta_j\log\theta_j-\left[(1-u_i'v_i)\theta_j+z_{ij}\right]\log(\mu_{ij}+\theta_j)\\ &+z_{ij}\log\mu_{ij}+\log\left(\Gamma(z_{ij}+\theta_j)\right)-\log\left(\Gamma(\theta_j)\right)\Big\}+C_1,\end{aligned}$$

where C_1 is a constant not related to $(\boldsymbol{\beta},\boldsymbol{\theta},\pi_0)$. Note in our case, $v_iz_{ij}=0$.

The EM algorithm can be described as follows.

- E-step: Replace u_i by

$$\bar{u}_i' = \frac{1-\pi_0}{1-\pi_0+\pi_0\prod_{j=1}^{m}\left(\frac{\theta_j}{\mu_{ij}+\theta_j}\right)^{\theta_j}}, \quad i=1,\ldots,n,$$

where $\mu_{ij}=\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}_j)$.

- M-step:
  - For π_0, we can get

$$\pi_0 = 1-\frac{1}{n}\sum_{i=1}^{n}\bar{u}_i'v_i.$$

  - For $(\boldsymbol{\beta},\boldsymbol{\theta})$, let

$$\begin{aligned}\bar{\ell}_j^C(\boldsymbol{\beta}_j,\theta_j) = &\sum_{i=1}^{n}\Big\{(1-\bar{u}_i'v_i)\theta_j\log\theta_j-\left[(1-\bar{u}_i'v_i)\theta_j+z_{ij}\right]\log(\mu_{ij}+\theta_j)\\ &+z_{ij}\log\mu_{ij}+\log\left(\Gamma(z_{ij}+\theta_j)\right)-\log\left(\Gamma(\theta_j)\right)\Big\}.\end{aligned}$$

There is no closed-form representation for $\boldsymbol{\beta}_j$ and θ_j, so we update the regression parameters, respectively, for each j by implementing the Newton–Raphson method for one step. The first- and second-order derivatives are given as follows:

$$\frac{\partial\bar{\ell}_j^C}{\partial\boldsymbol{\beta}_j} = \sum_{i=1}^{n}\frac{\left[z_{ij}-(1-\bar{u}_i'v_i)\mu_{ij}\right]\theta_j}{\mu_{ij}+\theta_j}\boldsymbol{x}_i,$$

$$\begin{aligned}\frac{\partial\bar{\ell}_j^C}{\partial\theta_j} = &\sum_{i=1}^{n}\Bigg[(1-\bar{u}_i'v_i)\log\frac{\theta_j}{\mu_{ij}+\theta_j}+\frac{(1-\bar{u}_i'v_i)\mu_{ij}-z_{ij}}{\mu_{ij}+\theta_j}\\ &+\psi(z_{ij}+\theta_j)-\psi(\theta_j)\Bigg],\end{aligned}$$

$$\frac{\partial^2\bar{\ell}_j^C}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^\top} = -\sum_{i=1}^{n}\frac{\left[z_{ij}+(1-\bar{u}_i'v_i)\theta_j\right]\mu_{ij}\theta_j}{(\mu_{ij}+\theta_j)^2}\boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

$$\frac{\partial^2 \bar{\ell}_j^C}{\partial \theta_j^2} = \sum_{i=1}^{n} \left[\frac{(1-\bar{u}_i' v_i)\mu_{ij}^2 + z_{ij}\theta_j}{\theta_j(\mu_{ij}+\theta_j)^2} + \psi_1(z_{ij}+\theta_j) - \psi_1(\theta_j) \right],$$

$$\frac{\partial^2 \bar{\ell}_j^C}{\partial \boldsymbol{\beta}_j \partial \theta_j} = \sum_{i=1}^{n} \frac{\left[z_{ij} - (1-\bar{u}_i' v_i)\mu_{ij}\right]\mu_{ij}}{(\mu_{ij}+\theta_j)^2} \boldsymbol{x_i}.$$

where $\psi(\cdot)$ and $\psi_1(\cdot)$ denote the digamma and trigamma functions, respectively.

For the case without covariates incorporated in μ_{ij}, the corresponding E-step and M-step can be simplified as follows.

- E-step: Replace u_i by

$$\bar{u}' = \bar{u}_i' = \frac{1-\pi_0}{1-\pi_0+\pi_0\prod_{j=1}^{m}\left(\frac{\theta_j}{\mu_j+\theta_j}\right)^{\theta_j}}, \quad i=1,\ldots,n.$$

- M-step:
  - For π_0, we can get

$$\pi_0 = 1 - \frac{\bar{u}'}{n}\sum_{i=1}^{n} v_i.$$

  - For $(\boldsymbol{\mu}, \boldsymbol{\theta})$, let

$$\bar{\ell}_j^C(\mu_j,\theta_j) = \sum_{i=1}^{n} \Big\{ (1-\bar{u}_i' v_i)\theta_j \log\theta_j - \left[(1-\bar{u}_i' v_i)\theta_j + z_{ij}\right]\log(\mu_j+\theta_j) + z_{ij}\log\mu_j + \log\left(\Gamma(z_{ij}+\theta_j)\right) - \log\left(\Gamma(\theta_j)\right) \Big\}.$$

  - Update each μ_j using the following equation:

$$\mu_j = \frac{\sum_{i=1}^{n} z_{ij}}{n - \bar{u}'\sum_{i=1}^{n} v_i}, \quad j=1,\ldots,m.$$

  - Substitute the value of μ_j obtained from the last step into $\bar{\ell}_j^C$, update θ_j using the one variable Newton–Raphson method.

APPENDIX B. DESCRIPTION FOR TWO PREVIOUS MODELS

B.1. Zero-inflated bivariate Poisson (ZIBP)

The bivariate Poisson (BP) regression model is defined as follows. Let Y_1, Y_2 and Y_3 denote three independent Poisson random variables with parameters λ_1, λ_2 and λ_3, respectively. Then $Z_1 = Y_1 + Y_3$ and $Z_2 = Y_2 + Y_3$ follow jointly a BP distribution, denoted as

$\boldsymbol{Z} = (Z_1, Z_2) \sim BP(\lambda_1, \lambda_2, \lambda_3)$. Covariates can be introduced into the model through the following schemes:

$$\lambda_j = \exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}_j\right), \quad j = 1, 2, 3.$$

where $\boldsymbol{x} = (1, x_1, \ldots, x_p)^\top$ and $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})^\top$.

A zero-inflated bivariate Poisson (ZIBP) model is specified by the following pmf:

$$f_{ZIBP}(z_1, z_2) = \begin{cases} \pi_0 f_{BP}(z_1, z_2), & (z_1, z_2) \neq (0, 0), \\ 1 - \pi_0 + \pi_0 f_{BP}(0, 0), & (z_1, z_2) = (0, 0). \end{cases}$$

B.2. 2-finite mixture of bivariate Poisson (2-FMBP)

Let $\boldsymbol{Z_1} = (Z_{11}, Z_{12})$ and $\boldsymbol{Z}_2 = (Z_{21}, Z_{22})$ denote two independent BP random variables with parameters $(\lambda_{11}, \lambda_{12}, \lambda_{13})$ and $(\lambda_{21}, \lambda_{22}, \lambda_{23})$, respectively. Then the 2-finite mixture of bivariate Poisson (2-FMBP) model is defined by the following pmf:

$$f(z_1, z_2) = p f_{\boldsymbol{Z}_1}(z_1, z_2) + (1 - p) f_{\boldsymbol{Z}_2}(z_1, z_2).$$

Covariates can be introduced into the model through the following schemes:

$$\lambda_{kj} = \exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}_{kj}\right), \quad k = 1, 2, \quad j = 1, 2, 3,$$

where $\boldsymbol{x} = (1, x_1, \ldots, x_p)^\top$ and $\boldsymbol{\beta}_{kj} = (\beta_{kj0}, \beta_{kj1}, \ldots, \beta_{kjp})^\top$. Covariates can also be incorporated in the mixing proportion p through a logit-link function.

https://doi.org/10.1017/asb.2021.39 Published online by Cambridge University Press