Multivariate zero-modified hurdle models in insurance

Pengcheng Zhang[a], Xueyuan Wu[b]

Enrique Calderín-Ojeda, Shuanming Li, David Pitt

[a]School of Insurance, Shandong University of Finance and Economics, China
[b]Centre for Actuarial Studies, Department of Economics
Faculty of Business and Economics, The University of Melbourne, Australia

28th January, 2022

Presentation Outline

1. Research motivation
2. Existing methods
3. Characterisation of the Type I MZT hurdle model
4. Real Applications – Data set One
5. Characterisation of the MZI hurdle model
6. Real Applications – Data set Two
7. Concluding remarks
8. Selected References

Problem: How to build multivariate claim count models for real-life insurance claim data that display one of the following features?

- Claim numbers from different categories are missing common zeros.
- Claim numbers from some/all categories show a zero-inflation feature.

In the general insurance modelling literature, there has been a lot of work based on univariate zero-inflation/deflation models, but little has been done in the multivariate case.

Multivariate zero-truncation

- There are three cases of missing zero count information in the multivariate setting:
  - only records with all zeros are missing (Type I multivariate zero-truncation);
  - zero counts for one or some classes are missing; or
  - zeros are completely missing for all classes.
- Existing methods:
  - Type I multivariate zero-truncated Poisson model (Tian et al., 2018b)
  - Type I multivariate zero-truncated negative binomial model

Data Set One

- The dataset was obtained from a Chinese health fund that contains health insurance claim records in the period of 2015-2017. The dataset consists of 40,030 policyholders (PHs).
- 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 PHs.
- There are three main causes of claims in the dataset: disease, accident and other.
- The total number of claims for each PH is positive.
- Apart from information regarding claims, the dataset also contains some explanatory variables such as time exposed to risk, age, gender, region and smoking status.

Empirical distributions

Table: 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				

Challenges

- In the health insurance dataset, policyholders (PHs) could make multiple claims on grounds of *Disease, Accidents* or *Other* within each year.
- Univariate zero-modified models can't handle the situation properly because of the correlation between different types of claims: negative correlation between Disease and Accidents and between Disease and Other.
- The existing Type I multivariate zero-truncated models, like Type I multivariate zero-truncated Poisson (MZTP) model and Type I multivariate zero-truncated NB (MZTNB) model, can't deal with the inconsistent zero-modification features in the marginal distributions in the dataset.
- To deal with the heterogeneity among the individual policyholders, we need to incorporate covariates into our modelling.

Multivariate zero-inflation

- There are some possibilities of zero-inflation in the multivariate setting:
  - all dimensions show a zero-inflation feature and there are observations with common zeros;
  - some dimensions show a zero-inflation feature and there are observations with common zeros; or
  - no dimensions show a zero-inflation feature but there are observations with common zeros.
- Existing methods:
  - Multivariate zero-inflated Poisson (MZIP) model (Liu and Tian, 2015)
  - Multivariate zero-inflated negative binomial (MZINB) model

Data Set Two

- This dataset is obtained from an automobile portfolio from a major insurance company operating in Spain in 1995. The whole dataset consists of 80,994 PHs.
- 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).
- PHs with comprehensive coverage and PHs with both comprehensive and collision coverage are also denoted as Z_2 type.
- The overall Pearson's correlation coefficient between these two types of claim is 0.189.
- Apart from information regarding claims, the dataset also contains 11 explanatory variables regarding gender, region, driving experience, age, etc.

The empirical distributions of Z_1 and Z_2

	Z_2								
Z_1	0	1	2	3	4	5	6	7	≥ 0
0	71,087	3,722	807	219	51	14	4	0	75,904
1	3,022	686	184	71	26	10	3	1	4,003
2	574	138	55	15	8	4	1	1	796
3	149	42	21	6	6	1	0	1	226
4	29	15	3	2	1	1	0	0	51
5	4	1	0	0	0	0	2	0	7
6	2	1	0	1	0	0	0	0	4
7	1	0	0	1	0	0	0	0	2
8	0	0	1	0	0	0	0	0	1
≥ 0	74,868	4,605	1,071	315	92	30	10	3	80,994

Challenges

- In this dataset, PHs have different types of policies. We choose to use the subset of PHs with both comprehensive and collision coverage as our basis. There are 28,590 PHs in this subset.
- Clearly both Z_1 and Z_2 show zero-inflation features.
- The existing multivariate zero-inflated models, like MZIP model and MZINB model may not deal with the inconsistencies in the marginal distributions in the dataset.
- To deal with the heterogeneity among the individual policyholders, we need to incorporate covariates into our modelling.

Type I multivariate zero-truncated (MZT) models I

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

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

where $U \sim \mathit{Bernoulli}(\pi_0)$, $\pi_0 = \Pr(\boldsymbol{Y} \neq \mathbf{0}) = 1 - \prod_{j=1}^{m} \Pr(Y_j = 0)$ and U is independent of $\boldsymbol{Z}$.

Type I multivariate zero-truncated (MZT) models II

The probability mass function (pmf) of $\boldsymbol{Z}$ can be derived as

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

where $\|\cdot\|_1$ represents the ℓ_1 norm of a vector. An alternative representation is to define $\boldsymbol{Z}\overset{d}{=}\boldsymbol{Y}\mid\boldsymbol{Y}\neq\mathbf{0}$.

- If $Y_j\sim$ Poisson (λ_j), $j=1,\ldots,m$, then $\mathbf{Z}\sim$ MZTP with $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_m)^\top$.
- If $Y_j\sim$ NB (μ_j,θ_j), $j=1,\ldots,m$, then $\mathbf{Z}\sim$ MZTNB with $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_m)^\top$ and $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)^\top$.

Multivariate zero-inflated (MZI) models I

For the previously defined $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^\top$, $\boldsymbol{Z}' = (Z'_1, \ldots, Z'_m)^\top$ is said to follow a MZI distribution if

$$\boldsymbol{Z}' \overset{d}{=} U\boldsymbol{Y} = \begin{cases} \mathbf{0}_m, & U = 0, \\ \boldsymbol{Y}, & U = 1, \end{cases} \tag{2.2}$$

where $U \sim \mathit{Bernoulli}(\pi_0)$, $0 < \pi_0 < 1$, and U is independent of $\boldsymbol{Y}$. The probability mass function (pmf) of $\boldsymbol{Z}'$ can be derived as

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

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.

Multivariate zero-inflated (MZI) models II

Two special cases:

- Let $Y_j \sim Poisson(\lambda_j)$, for $j = 1, \ldots, m$. Then $\boldsymbol{Z}'$ is said to follow the MZIP distribution with the parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ and a zero-inflation parameter π_0, denoted by $\boldsymbol{Z}' \sim MZIP(\boldsymbol{\lambda}, \pi_0)$.
- Let $Y_j \sim NB(\mu_j, \theta_j)$, for $j = 1, \ldots, m$. Then $\boldsymbol{Z}'$ is said to follow the MZINB 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 $\boldsymbol{Z}' \sim MZINB(\boldsymbol{\mu}, \boldsymbol{\theta}, \pi_0)$.

Some properties

Assume that $\pi_0 = \Pr(U = 1) \in (0, 1)$.

- Marginal distributions (note that $\pi_0 \geq \Pr(Y_j > 0)$ in case **Z**):

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

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

- Covariance :

$$Cov(Z_j, Z_k) = -\frac{(1-\pi_0)E[Y_j]E[Y_k]}{\pi_0^2} < 0;$$

$$Cov(Z_j', Z_k') = \pi_0(1-\pi_0)E[Y_j]E[Y_k] > 0.$$

Univariate hurdle models

To allow for greater flexibility in modelling the marginal behaviour of each counting variable, we shall assume that Y_j, $j = 1, \ldots, m$, follows a zero-modified distribution, which can be characterised 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.

Type I MZT hurdle model

Then $\boldsymbol{Z}$ constructed by (2.1, slide 12) is said to follow the Type I multivariate zero-truncated hurdle (MZTH) 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 $\boldsymbol{Z}$ is

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

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

Remark. 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 later.

Model Inference – MZTH model I

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

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

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 can then be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\Theta} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \prod_{i=1}^{n} \frac{\prod_{j:z_{ij}=0} (1 - \pi_{ij}) \prod_{j:z_{ij}\neq 0} \pi_{ij} f_{W_{ij}}(z_{ij})}{1 - \prod_{j=1}^{m} (1 - \pi_{ij})}.$$

Model Inference – MZTH model II

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

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

$$\ell_2(\Theta \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \sum_{i=1}^{n} \sum_{j:z_{ij}\neq 0} f_{W_{ij}}(z_{ij}) = \sum_{j=1}^{m} \sum_{i:z_{ij}\neq 0} f_{W_{ij}}(z_{ij}).$$

- For ℓ_2, the estimation can be implemented in respect of different j independently for positive values.

Model Inference – MZTH model III

- For ℓ_1, we implement an EM algorithm illustrated as follows.

Denote $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_m)^\top$ where $\Delta_j = \mathbb{I}(Z_j > 0)$, and $\mathbb{I}(\cdot)$ is the indicator function. The corresponding observed values are denoted by $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n$ where $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{im})^\top$.

The complete-data log-likelihood function given $(\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')$ is

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

Model Inference – MZTH model IV

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

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

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 $\boldsymbol{\beta_j}$.

Model Inference – MZTH model V

- M-step: Let

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

  Update the regression parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, m$, respectively by maximizing ℓ_{1j} using Newton-Raphson method.

- 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. The initial values of parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, m$, can be obtained by implementing univariate logistic regression with observed values $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n$.

Real Application One

- For this study, we take a random sample of 30,000 as training data to develop the model and the rest is reserved for validation purposes.
- The empirical distribution for Disease claim frequency obviously has a heavier tail than the ones for Accident and Other, indicating that Poisson and NB might not be appropriate for it.
- 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 $W_3 - 1$.
- Using AIC as our main criterion for model selection, our MZTH model performs much better than the other two models when fitting the training dataset.

The model fitting performance of three candidate models without considering coviariates are given below. It shows that the MZTH model performs better than the other two models.

Table: Loglikelihood and AIC of the three models considered

Model	No. of parameters	Loglik	AIC
MZTP	3	-30143.14	60292.28
MZTNB	6	-29273.42	58558.84
MZTH	8	-28841.18	57698.36

Next two tables show the model fitting results of MZTH model with covariates. Note that certain covariates are disgarded when fitting W_2 and W_3 due to non-convergence issue.

Model-fitting results with covariates I

Table: 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			-14387.16			
AIC			28846.32			

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

Model-fitting results with covariates II

Table: 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	-13021.07		-138.79		-190.50	
AIC	26068.14		281.59		401.00	

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

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.
- 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.
- We also generate a prediction on the multivariate claim frequency and summarise the results in a table below, where the predicted numbers are put in parentheses below the observed numbers.

Table: Goodness-of-fit of the marginal models

	Disease (PIG)		Accident (ZTP)		Other (NB)	
Count	Observed	Predicted	Observed	Predicted	Observed	Predicted
0	945	960.01	9515	9514.85	9376	9386.05
1	7931	7861.11	503	505.84	641	631.41
2	879	905.80	12[a]	9.19[a]	13[a]	11.73[a]
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)	

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

Table: 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 (518.29)	13[a] (12.42)	430 (418.50)	10[a] (6.13)	12[a] (9.32)
	1	7783 (7693.22)	92 (92.90)		50 (70.96)		
	2	841 (886.91)	46[a] (14.09)		13[a] (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.

Multivariate zero-inflated hurdle (MZIH) model

When Y_j, $j = 1, \ldots, m$, follows the hurdle model defined in (3.1, slide 17), the $\boldsymbol{Z}'$ constructed by (2.2, slide 14) is said to follow the MZIH 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

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

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

Model Inference – MZIH model I

Suppose each $\boldsymbol{Z}_i'$, $i = 1, \ldots, n$, independently follows an MZIH 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 $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$. The parameter π_{ij} can then be modeled as

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

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.

Model Inference – MZIH model II

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_{ij}, the likelihood function then can be written as

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

Model Inference – MZIH model III

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

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

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.

Model Inference – MZIH model IV

For ℓ_1, we implement the EM algorithm as described below.

Denote $\boldsymbol{\Delta}' = (\Delta'_1, \ldots, \Delta'_m)^\top$ where $\Delta'_j = \mathbb{I}(Z'_j > 0)$. The corresponding observed values are denoted by $\boldsymbol{\delta}'_1, \ldots, \boldsymbol{\delta}'_n$ where $\boldsymbol{\delta}'_i = (\delta'_{i1}, \ldots, \delta'_{im})^\top$.

The observed log-likelihood function ℓ_1 can be rewritten as

$$\ell_1(\boldsymbol{\beta}, \pi_0) = \sum_{i=1}^{n} v_i \log \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_{ij})\right] + \sum_{i=1}^{n} (1 - v_i) \log \pi_0$$

$$+ \sum_{j=1}^{m} \sum_{i=1}^{n} (1 - v_i) \left[\delta'_{ij} \log \pi_{ij} + (1 - \delta'_{ij}) \log(1 - \pi_{ij})\right].$$

Model Inference – MZIH model V

In addition to the known values δ'_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

$$\ell_1^c(\beta, \pi_0) = \sum_{i=1}^{n} \left[u'_i v_i \log(1-\pi_0) + (1-u'_i v_i) \log \pi_0 \right] + \sum_{j=1}^{m} \sum_{i=1}^{n} \left[\delta'_{ij} \log \pi_{ij} + (1-u'_i v_i - \delta'_{ij}) \log(1-\pi_{ij}) \right].$$

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

Model Inference – MZIH model VI

- 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

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

Model Inference – MZIH model VII

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

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

  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.

- Iterate through the E-step and the M-step until some convergence criterion is met, for example, the relative change of observed log-likelihood between two consecutive iterations is $<\varepsilon$.

Remark. The initial values of parameters $\boldsymbol{\beta}_j$ for EM algorithm can be obtained by implementing univariate logistic regression with observed values $z'_{1j}, \ldots, z'_{nj}$. The initial value of parameter π_0 can be set as 0.5. The s.e.'s for the estimates can be approximated using the approach in Louis (1982).

Real Application Two

- To model the bivariate count data properly, we did the same treatment on the data as the one done in Bermúdez and Karlis (2017), i.e. only selecting policyholders with full coverage (both comprehensive and collision coverage, i.e. $v9 = 0, v10 = 1$). This reduced the data to 28,590 policyholders.
- The empirical distributions regarding Z_1 (number of third-party liability claims) and Z_2 (number of other claims) of this subset are given on next page.
- 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.
- Regarding W_j, $j = 1, 2$, in addition to the commonly used ZTP and ZTNB distributions, we also tried unit-shifted Poisson (USP) and unit-shifted negative binomial (USNB) distributions.

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 $\geq$ 5,500 cc	0.806

Empirical distributions of Z_1 and Z_2

	Z_2							
Z_1	0	1	2	3	4	5	6	≥ 0
0	24,408	1,916	296	69	12	6	0	26,707
1	1,068	317	61	21	6	2	2	1,477
2	203	71	18	6	2	1	1	302
3	49	14	8	3	3	1	0	78
4	11	6	2	0	1	0	0	20
5	2	0	0	0	0	0	1	3
6	1	0	0	1	0	0	0	2
8	0	0	1	0	0	0	0	1
≥ 0	25,742	2,324	386	100	24	10	4	28,590

Goodness-of-fit of marginal models

W_1	Observed	ZTP	ZTNB	USP	USNB
1	1,033	993.71	1,037.84	981.99	1,032.45
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	1,624	1,562.13	1,612.94	1,548.65	1,623.88
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		-1,258.84	-1,221.06	-1,283.18	-1,220.22

Model fitting results of π_j in MZIH model with covariates

	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.

Model fitting results of W_j in MZIH model with covariates

	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.

In-sample predictions

	Z_2						
Z_1	0	1	2	3	4	5	6
0	17,104	1,342	199	41	8	4	0
	(17,102.66)	(1,306.43)	(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)

Model comparison

- We also fitted the MZIP and MZINB models. Furthermore, a model with two independent hurdle margins (Ind) is fitted as a benchmark.
- For the zero-truncation parts in our MZIH model, only intercepts are adopted to avoid the over-fitting problem.

Table: Information criteria of four fitted models

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

Predictive Analysis

To evaluate the predictive performance, we calculate the predicted claim frequencies and compare these to the observed ones based on the out-of-sample for the following scenarios:
The candidate models include MZIH, MZIP, MZINB and Ind models.
Our observations are consistent with the in-sample predictions.

(Z_1, Z_2)	Observed	MZIH	MZIP	MZINB	Ind
(0, 0)	7,304	7,332.54	7,330.50	7,330.93	7,224.20
($>$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

Conclusions

- Extra care needs to be taken on zero counts when modelling multivariate insurance count data.
- The proposed two multivariate zero-modified hurdle models are of good use when multivariate count data display certain features in joint zero counts.
- The hurdle model structure on individual dimensions bring great flexibilities that lead to better fitting results.
- General multivariate zero-modified hurdle models can be constructed by zero-inflating the MZTH models (working paper).

Selected References I

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.

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.

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

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.

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.

Selected References II

- Karlis, D. (2005). EM algorithm for mixed Poisson and other discrete distributions. *ASTIN Bulletin: The Journal of the IAA*, **35**(1), 3-24.
- 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.
- Piperigou, V. E. and Papageorgiou, H. (2003). On truncated bivariate discrete distributions: A unified treatment. *Metrika*, **58**(3), 221-233.
- 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.
- 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.
- 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.