Author/s:
Zhang, Pengcheng

Title:
Multivariate Count Regression Models with Applications in Insurance

Date:
2021

Persistent Link:
https://hdl.handle.net/11343/285208

Multivariate Count Regression Models with Applications in Insurance

Pengcheng ZHANG

A Thesis
for
the Faculty
of
Business and Economics

Submitted in Total Fulfilment of the Requirements
for the Degree of Doctor of Philosophy at
The University of Melbourne
Melbourne, Australia

June, 2021

Produced on Archival Quality Paper

Abstract

This thesis proposes several multivariate count regression models in insurance. In today's life, given that insurance companies often 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 real insurance datasets, we construct a number of multivariate count models to deal with various types of characteristics exhibited in the data. Some distributional properties concerning the models are examined. The inferential procedures are completed either via the expectation–maximization (EM) algorithm or the Markov chain Monte Carlo (MCMC) algorithm.

First of all, we investigate multivariate count modeling with dependence. One possible way to construct such model is to implement copula directly on discrete margins. However, likelihood inference under this construction involves the computation of multidimensional rectangle probabilities, which could be computationally expensive especially in the elliptical copula case. Another potential approach is based on the multivariate mixed Poisson model. The crucial work under this method is to find an appropriate multivariate continuous distribution for mixing parameters. By virtue of the copula, this issue could be easily addressed. Under such framework, MCMC method is a feasible strategy for inference. The usefulness of our model is then illustrated through a real-life example. Both the in-sample analysis and out-of-sample prediction demonstrate the superiority of

using copula-based mixture over other types of mixture.

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. In the second part of this thesis, we develop a multivariate zero-inflated hurdle model to describe multivariate count data with extra zeros. Our model offers 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 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 existing multivariate claim count models.

In contrast to the previous section, the third part of this thesis investigates a multivariate zero-truncation problem. In the general insurance modeling literature, there has been a lot of work based on the univariate zero-truncated models, but little has been done in the multivariate zero-truncation case. 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 chapter, 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 without common zero observations.

In the process of insurance underwriting, policyholders tend to report untrue statements especially when they can benefit from this act, like paying lower premium. However, it is a demanding task to fully investigate the unobservable misrepresentation status, and models without accommodating the misrepresentation tend to result in biased estimates. So in the fourth part of this thesis, we address the issue of misrepresentation in multivariate frequency setting. A multivariate Poisson model is developed with a binary misrepresented risk factor incorporated. The fact that each margin depends on this factor complicates the covariance structure. For inference, the EM algorithm is implemented. Two small simulation studies are then carried out to compare the performance of the model adjusted for misrepresentation with the unadjusted model. Then we perform a frequency analysis using real data obtained from the Australian Health Survey 1977-1978, where a binary variable indicating the health score is regarded as the misrepresented factor.

At last, we present some concluding remarks. The limitations of the research conducted in this thesis are discussed, and several potential future research problems are also specified.

Preface

This thesis was completed under the supervision of Assoc. Prof. Xueyuan Wu, Dr. Enrique Calderín-Ojeda and Prof. Shuanming Li in the Centre for Actuarial Studies, The University of Melbourne. Chapter 2 to 5 contain the original research of this thesis, except as otherwise noted.

Chapter 4 is based on the paper "On the Type I multivariate zero-truncated hurdle model with applications in health insurance", was co-authored by Enrique Calderín-Ojeda, Shuanming Li and Xueyuan Wu and is published in *Insurance: Mathematics and Economics* **90**, 35-45.

The research for this paper was carried out by Pengcheng Zhang under the supervision of Xueyuan Wu, Enrique Calderín-Ojeda and Shuanming Li. The paper was written by Pengcheng Zhang, with proofreading and editing by Xueyuan Wu, Enrique Calderín-Ojeda and Shuanming Li.

None of the work appearing here has been submitted for any other qualifications, nor was it carried out prior to Ph.D. candidature enrolment.

Acknowledgments

I would first like to express my sincere gratitude to my supervisors Assoc. Prof. Xueyuan Wu, Dr. Enrique Calderín-Ojeda and Prof. Shuanming Li, for their continuous support and guidance during my Ph.D. study. The diligence and enthusiasm they have for research are contagious and motivational for me, especially during this pandemic period. I am grateful for their unreserved dedication and invaluable advice, which make my Ph.D. experience productive and stimulating.

I would like to acknowledge Prof. David Pitt, Assoc. Prof. Zhuo Jin and Dr. Ping Chen, and all other academic staff in Centre for Actuarial Studies. Their helpful suggestions really inspire me in every aspect of life. I appreciate my colleagues for the inspiring discussions. The time we were working together and the fun we have had in the past periods were truly unforgettable in my whole life.

My sincere thanks also go to the University of Melbourne and the Faculty of Business and Economics. This research is supported by a Faculty of Business and Economics Doctoral Program Scholarship. I gratefully acknowledge the funding source that made my Ph.D. work possible in this tough time.

Lastly, I would like to thank my parents for supporting me spiritually throughout the life. Their wise counsel and sympathetic ear give me much confidence to continue my academic pursuit.

Contents

List of Figures

List of Tables

List of Notation and Symbols

$\mathbb{N}$	$\{0, 1, 2, \ldots\}$
$\mathbb{N}_+$	$\{1, 2, \ldots\}$
$\mathbb{R}_+$	$(0, \infty)$
$\mathrm{E}(X)$	expectation of random variable X
$\mathrm{Var}(X)$	variance of random variable X
$\mathrm{Cov}(X, Y)$	covariance between two random variables X and Y
$\mathrm{Cor}(X, Y)$	correlation coefficient between two random variables X and Y
$K_j(\cdot)$	modified Bessel function of the second kind
$f(\cdot)$	probability density function (pdf) or probability mass function (pmf)
$F(\cdot)$	cumulative distribution function (cdf)
$C(\cdot)$	copula function
$c(\cdot)$	copula density function
$\mathcal{I}$	information matrix
$\lvert\boldsymbol{\Sigma}\rvert$	determinant of matrix $\boldsymbol{\Sigma}$
$\dim(\boldsymbol{\theta})$	dimension of vector $\boldsymbol{\theta}$
$\min(x, y)$	the minimum of x and y
$\phi(\cdot)$	pdf of the standard normal distribution
$\Phi(\cdot)$	cdf of the standard normal distribution
$\stackrel{d}{=}$	random variables on both sides share the same distribution
$\propto$	be proportional to
$\lVert\boldsymbol{x}\rVert_1$	$\sum_{i=1}^{m} \lvert x_i\rvert$
$\mathbb{I}(\cdot)$	indicator function
AIC	Akaike information criterion
BIC	Bayesian information criterion
DIC	deviance information criterion
$\square$	end of the proof

Introduction

Designing a proper tariff structure for an insurance company is a key actuarial task. Such pricing becomes complex when there exists heterogeneity in the portfolio. One way to resolve this problem is referred to as *a priori* ratemaking. It involves classifying the whole portfolio into numerous homogeneous groups so that fair premium is charged based on the risk level of a particular class.

Claim count modeling has always been an important building block in insurance for ratemaking analysis. The number of claims can to a great extent reveal the riskiness of the policyholders. By examining the relation between claim counts and the insureds' characteristics, risk classification can be accomplished. A thorough review of ratemaking procedure when modeling claim count data in automobile insurance can be found in Denuit et al. (2007).

The starting point for modeling the number of claims is the Poisson distribution which was introduced by Siméon-Denis Poisson. This fundamental model handles a number of insurance datasets of interest to actuaries. However, the requirement of mean being equal to variance in Poisson distribution somehow restricts its application. It is often the case that the unobserved heterogeneity in the data will lead to overdispersion (variance greater than mean). This cannot be fully remedied by a simple Poisson model. To account for overdispersion, we can generalize the Poisson distribution to the mixed Poisson distribution by introducing a random heterogeneity term. A detailed illustration of mixed Poisson distributions can be found in Karlis and Xekalaki (2005).

Even the mixed Poisson models fail to capture some important features exhibited in the insurance datasets. For example, the no-claim discount (NCD) system widely adopted by automobile insurers leads to a high proportion of zero claims, as policyholders prefer not to make a claim if the amount is small. Another example is the dataset managed by the claims department in a health insurance company which could only keep records for policyholders who make claims. Counting data without zero category is a common issue in such insurance portfolio. The above facts motivate the need for more complex count models such as zero-inflated model (see Yip and Yau (2005)) and zero-truncated model (see Grogger and Carson (1991)).

A hurdle model (Mullahy (1986)) provides another mechanism to modify standard count distributions. It is motivated by a two-stage decision-making process confronted by individuals. For example, in health care services, the decision to seek the service is an initial process dependent on the individual, while the amount of expenditure or the positive number of visits is a second process dependent on the health care provider. The superiority of hurdle model is that it can handle both zero-deflation and zero-inflation phenomena.

Nowadays it is not uncommon to find the need to model claim counts from different types of coverage in an insurance policy. For example, after a car accident, the driver's comprehensive car insurance policy could face up to three types of claims, e.g., bodily injury claims, damages to policyholder's car and third-party property damages. A private health insurance policy usually covers medical care costs of office-based visits, hospital stays and emergency room visits. When actuaries face the problem of pricing an insurance contract which contains multiple types of coverage, they generally assume independence among each type of claims. However, it remains to be established whether this assumption is realistic. Several studies have shown that there exists dependence among different types of claims

(Frees and Valdez (2008); Frees et al. (2009); Young et al. (2009)). This fact highlights the importance of dependence modeling for insurance datasets.

A natural method to introduce correlations among multivariate counts is to take advantage of common shock variables. This approach proves to be useful in the multivariate Poisson model, since the sum of independent Poisson random variables still follows Poisson distribution. A simple version of the multivariate Poisson model is to assume a common covariance structure, implying that an identical covariance term is shared by each pair of margins. This model has been described by Kocherlakota and Kocherlakota (1992) and Johnson et al. (1997). However, the complicated form of the joint probability function renders the inference a hard job. To address this issue, Tsiona (1999) and Tsiona (2001) proposed a MCMC method under the Bayesian framework. Karlis (2003) implemented an EM algorithm for inference. The application of this multivariate Poisson model in actuarial setting can be found in Bermúdez (2009) and Bermúdez and Karlis (2011).

A more straightforward way to construct a multivariate discrete model with flexible dependence is to employ copula. Copula was introduced by Sklar (1959) in the context of probabilistic metric spaces. However, dependence modeling using copula was introduced to actuarial science until 1998 by Frees and Valdez (1998). That article explored some of practical applications, including estimation of joint life mortality and multiple decrement models. After that, the copula with generalized linear models (GLMs) (see McCullagh and Nelder (1989)) as margins has been a useful tool to describe the dependence in actuarial literature.

A copula is a function that links univariate margins to their full multivariate distribution. The key advantage of the copula approach is the separate treatment of marginal distributions and dependence structure, see Nelsen (2006) for more details on copulas. However, copula models with discrete margins are more

challenging than with continuous margins, both for theoretical and computational reasons (Genest and Nešlehová (2007)).

There are basically two types of copulas: Archimedean copulas and elliptical copulas. With only one dependence parameter embedded in Archimedean copulas, this class of copulas assumes an exchangeable dependence structure and thus could be restrictive in real application. The family of elliptical copulas provides unstructured dependence and allows for both positive and negative relations. So when dealing with more than two dimensions, elliptical copulas are more preferred as they can accommodate flexible pair-wise dependence. The most commonly used one is Gaussian copula. See Song et al. (2009) for the example regarding Gaussian copula on discrete data. Some actuarial applications can also be found in Shi and Valdez (2014) and Frees et al. (2016).

However, Gaussian copula may not be computationally tractable for multivariate discrete outcomes due to the involvement of multidimensional integration. To circumvent this problem, several methods have been developed aiming for the estimation procedure. Zhao and Joe (2005) proposed a composite likelihood estimation to simplify the inference procedure, which is a two-stage method based on both univariate and bivariate margins. The approach enjoys satisfactory approximation when the dependence among each margin is not strong. The jitter approach inspired by the pioneer work of Denuit and Lambert (2005) can also be applied to the estimation process. The key idea is to convert the discrete variables into continuous ones and then to implement copula to the jittered continuous variables. Also, Nikoloulopoulos (2013) introduced a maximum simulated likelihood method, which is based on evaluating the multidimensional integrals with randomized quasi-Monte Carlo method. Furthermore, Bayesian approach can also be used for inference purposes (Pitt et al. (2006)).

A separate line of constructing a multivariate discrete model is based on

mixture of independent Poisson models. One important property for the mixed Poisson models is that they can account for overdispersion. There are very limited candidates for multivariate mixing distributions in the literature. One popular choice is multivariate lognormal (Aitchison and Ho (1989)) due to its flexible covariance structure. Multivariate gamma distribution has also been considered as a potential selection (see Chatelain et al. (2009)), yet it is too complicated to incorporate covariates in this case. Chapter 2 constructs multivariate mixing distributions with the help of copula that offers some advantages. It permits arbitrary distributional choices of margins, while at the same time, the association parameters in copula can capture the dependence among margins. A simplified version is to assume the marginal Poisson distributions share a common mixing parameter and thus the mixing distribution reduces to a univariate one. In this setting, Ghitany et al. (2012) proposed an EM algorithm for the corresponding regression problem.

Similar as in the univariate case, the insurance dataset may also exhibit zero-inflation or zero-truncation feature in the multivariate context. Regression models aiming at addressing these two issues are singled out for special consideration in Chapter 3 and Chapter 4 respectively.

In the literature, there were a few papers discussing multivariate zero-inflated models. 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 within a Bayesian framework. Using a similar structure as in Li et al. (1999), Dong

et al. (2014) presented a multivariate zero-inflated negative binomial model. Liu and Tian (2015) examined a new multivariate zero-inflated Poisson model which could avoid the computation issues caused by large dimensions. In Chapter 3, we further extend this model by assuming each margin follows a hurdle distribution rather than a simple Poisson.

As for the zero-truncation problem in the multivariate setting, there are typically three types: only common zeros for all classes are missing, zeros for one or some classes are missing, or zeros are completely missing for all classes. In this thesis, we study the first type, namely Type I case. To our best knowledge, only a few papers discussed the modeling using Type I zero-truncation data and most of them focused on the bivariate case (Charalambides (1984); Jung et al. (2007); Piperigou and Papageorgiou (2003)). However, those models cannot be readily extended to the multivariate case and to the regression context. Recently, Tian et al. (2018) proposed a Type I multivariate zero-truncated Poisson distribution to fit some count data of this type. This model starts with m independent Poisson distributions, and then uses a Bernoulli latent variable to truncate common zeros. In a similar spirit, we introduce our hurdle model in Chapter 4 which permits more flexibility to handle different marginal behaviors.

Another common issue existing in actuarial modeling is insurance fraud. Policyholders are prone to hide their true status in their best interest when disclosing their information for insurance pricing purposes. For example, in health insurance, smokers purposely conceal the fact of smoking in order to obtain a lower premium level. In automobile insurance, misstatements on annual mileage often occur especially in the process of online underwriting. However, from the insurers' point of view, it is either time consuming or laborious to verify the true status for this particular risk factor. There is thus a strong incentive to build models accounting for potential misrepresentation, which contributes to a more robust

ratemaking system.

There have been a number of papers on the misrepresentation phenomenon in recent literature. Xia and Gustafson (2016) investigated the identifiability of Bayesian regression models where a binary covariate was subject to unidirectional misclassification. The term unidirectional represents the type of measurement errors in categorical variables can only occur in one direction, which is more favorable to respondents. Sun et al. (2017) further established the moment identifiability of regression models for a general misclassified ordinal covariate. Akakpo et al. (2019) adopted the frequentist inference for severity analysis where the underlying loss distribution was assumed to be lognormal. Xia (2018) established the misrepresentation models for some heavy-tailed loss distributions in the Bayesian context. In Chapter 5, we work on the misrepresentation problem in a multivariate setting. Specifically, a misrepresented covariate is introduced in the multivariate Poisson model, on which all margins depend. This renders a more complicated dependence structure.

Chapter 6 presents some concluding remarks and possible areas of future research.

Chapter 1

Preliminaries

In this chapter, we present a brief introduction for count models and some relevant preliminaries. After specifying some commonly used univariate count models, we summarize several methods of generalizing univariate models to multivariate ones. Some numerical methods including Newton-Raphson and EM algorithm are then provided. We conclude the chapter with the Bayesian approach for count models.

1.1 Univariate Count Models

1.1.1 Poisson Model

If we denote the random variable for claim count as Y, then the pmf of the Poisson distribution can be expressed as

$$\Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots, \tag{1.1.1}$$

where $\lambda > 0$ is the mean parameter. One special property the Poisson distribution has is equidispersion, meaning $\mathrm{E}(Y) = \mathrm{Var}(Y)$.

1.1.2 Mixed Poisson Model

The requirement that the mean is equal to the variance in Poisson is not satisfied in many datasets of interest. If the variance exceeds the mean, then we say there

exists overdispersion in the data. To cope with this feature, we can generalize the Poisson distribution to the mixed Poisson distribution by adding a random heterogeneity term:

$$\Pr(Y = y) = \int_0^\infty \frac{(\lambda\alpha)^y e^{-\lambda\alpha}}{y!} g(\alpha) d\alpha, \quad y = 0, 1, 2, \ldots, \tag{1.1.2}$$

where λ is the Poisson parameter, α is the mixing parameter and $g(\alpha)$ is the pdf for the mixing parameter. To ensure identifiability, we assume $\mathrm{E}(\alpha) = 1$. In the following subsections, we introduce some commonly used mixed Poisson distributions.

Negative Binomial Distribution

When the mixing parameter follows a gamma distribution of mean 1 which is expressed as

$$g(\alpha; \phi) = \frac{\phi^\phi}{\Gamma(\phi)} \alpha^{\phi-1} e^{-\phi\alpha}, \quad \alpha, \phi > 0,$$

we obtain the negative binomial distribution:

$$\Pr(Y = y) = \frac{\Gamma(y + \phi)}{\Gamma(\phi) y!} \left(\frac{\lambda}{\lambda + \phi}\right)^y \left(\frac{\phi}{\lambda + \phi}\right)^\phi. \tag{1.1.3}$$

It can be proved that $\mathrm{E}(Y) = \lambda$ and $\mathrm{Var}(Y) = \lambda + \frac{\lambda^2}{\phi}$. This form is often called a negative binomial 2 (NB2).

Poisson-Inverse Gaussian Distribution

When the mixing parameter follows an inverse Gaussian distribution of mean 1 which is expressed as

$$g(\alpha; \phi) = \left(\frac{\phi}{2\pi\alpha^3}\right)^{1/2} \exp\left[-\frac{\phi(\alpha - 1)^2}{2\alpha}\right], \quad \alpha, \phi > 0,$$

we obtain the Poisson-inverse Gaussian distribution:

$$\Pr(Y = y) = \frac{\lambda^y}{y!} \left(\frac{2\phi}{\pi}\right)^{0.5} e^\phi \left(1 + \frac{2\lambda}{\phi}\right)^{-\frac{s}{2}} K_s(z), \tag{1.1.4}$$

where $s = y - 0.5$, $z = (\phi^2 + 2\phi\lambda)^{0.5}$ and $K_j(\cdot)$ denotes the modified Bessel function of the second kind that satisfies the following recursive properties:

$$K_{-0.5}(z) = \left(\frac{\pi}{2z}\right)^{0.5} e^{-z},$$
$$K_{0.5}(z) = K_{-0.5}(z),$$
$$K_{s+1}(z) = K_{s-1}(z) + \frac{2s}{z} K_s(z).$$

It also can be shown that $\mathrm{E}(Y) = \lambda$ and $\mathrm{Var}(Y) = \lambda + \frac{\lambda^2}{\phi}$.

Poisson-Lognormal Distribution

When the mixing parameter follows a lognormal distribution of mean 1 which is expressed as

$$g(\alpha; \phi) = \frac{1}{\sqrt{2\pi}\phi\alpha} \exp\left[-\frac{(\log\alpha + \phi^2/2)^2}{2\phi^2}\right], \quad \alpha, \phi > 0,$$

we obtain the Poisson-lognormal distribution:

$$\Pr(Y = y) = \int_0^\infty \frac{(\lambda\alpha)^y e^{-\lambda\alpha}}{y!} \frac{1}{\sqrt{2\pi}\phi\alpha} \exp\left[-\frac{(\log\alpha + \phi^2/2)^2}{2\phi^2}\right] d\alpha. \tag{1.1.5}$$

Unfortunately, we cannot express this pmf in closed-form. The mean and variance are $\mathrm{E}(Y) = \lambda$ and $\mathrm{Var}(Y) = \lambda + [\exp(\phi^2) - 1]\lambda^2$ respectively.

1.1.3 Other Count Models

We introduce some generalized count models with the help of stochastic representation.

Zero-Inflated Model

Zero-inflated count model provides a way to model count data with excess zeros. Let Y follow a standard count distribution defined on $\mathbb{N}$, then Z is said to follow the zero-inflated distribution if

$$Z \overset{d}{=} UY = \begin{cases} 0, & U = 0, \\ Y, & U = 1, \end{cases} \tag{1.1.6}$$

where $U \sim Bernoulli(\pi)$, $0 < \pi < 1$, and U is independent of Y. The pmf of Z can be derived as

$$\Pr(Z = z) = \begin{cases} 1 - \pi + \pi \Pr(Y = 0), & z = 0, \\ \pi \Pr(Y = z), & z > 0. \end{cases} \tag{1.1.7}$$

Zero-Truncated Model

The model for incompletely observed count data due to zero-truncation is presented. Let Y follow a standard count distribution defined on $\mathbb{N}$, then Z is said to follow the zero-truncated distribution if

$$Y \stackrel{d}{=} UZ = \begin{cases} 0, & U = 0, \\ Z, & U = 1, \end{cases} \tag{1.1.8}$$

where $U \sim Bernoulli(\pi)$, $\pi = 1 - \Pr(Y = 0)$, and U is independent of Z. The pmf of Z can be derived as

$$\Pr(Z = z) = \frac{\Pr(Y = z)}{\Pr(U = 1)} = \frac{\Pr(Y = z)}{\pi}, \quad z > 0. \tag{1.1.9}$$

An alternative representation is to define $Z \stackrel{d}{=} Y|Y > 0$. This stochastic representation helps us to generate random values of Z.

Hurdle Model

Hurdle model is a two-part model that separates the occurrence of an event from the number of those events actually observed. Let W follow a count distribution defined on $\mathbb{N}_+$, then Z is said to follow the hurdle (zero-modified) distribution if

$$Z \stackrel{d}{=} UW = \begin{cases} 0, & U = 0, \\ W, & U = 1, \end{cases} \tag{1.1.10}$$

where $U \sim Bernoulli(\pi)$, $0 < \pi < 1$, and U is independent of W. The pmf of Z can be derived as

$$\Pr(Z = z) = \begin{cases} 1 - \pi, & z = 0, \\ \pi \Pr(W = z), & z > 0. \end{cases} \tag{1.1.11}$$

Remark 1.1.1. Typical choices for W include zero-truncated distribution and unit-shifted distribution.

1.2 Multivariate Count Models

1.2.1 Multivariate Poisson Model

The multivariate Poisson distribution with common covariance is defined as follows. Let

$$Z_j = Y_j + Y_0, \quad j = 1, \ldots, m, \tag{1.2.12}$$

where each Y_j, $j = 0, \ldots, m$, independently follows a simple Poisson distribution with parameter λ_j. Then $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^\top$ is said to follow the multivariate Poisson distribution. The joint pmf of $\boldsymbol{Z}$ is given by

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \exp(-\lambda') \sum_{l=0}^{s} \left[\frac{\lambda_0^l}{l!} \prod_{j=1}^{m} \frac{\lambda_j^{z_j - l}}{(z_j - l)!} \right], \tag{1.2.13}$$

where $\boldsymbol{z} = (z_1, \ldots, z_m)^\top$, $\lambda' = \sum_{j=0}^{m} \lambda_j$ and $s = \min(z_1, \ldots, z_m)$.

1.2.2 Multivariate Mixed Poisson Model

To generalize the mixed Poisson distribution from univariate case to multivariate case, a simplified version is to assume each Poisson margin shares a common mixing parameter. Specifically, the pmf is given as

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = \int_0^\infty \prod_{j=1}^{m} \frac{\exp(-\alpha\lambda_j)(\alpha\lambda_j)^{y_j}}{y_j!} g(\alpha) d\alpha, \tag{1.2.14}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^\top$, $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$ and $g(\alpha)$ denotes the pdf of the univariate mixing parameter. When the mixing parameter α is assumed to follow gamma, inverse Gaussian or lognormal distribution with $\mathrm{E}(\alpha) = 1$, we obtain the multivariate negative binomial, multivariate Poisson-inverse Gaussian or multivariate Poisson-lognormal distribution respectively.

A potential disadvantage of this form is that it only allows for positive correlation between each pair of margins. A further extension to overcome this limitation is to assume a multivariate mixing distribution $g(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$. Specifically, the pmf is given as

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = \int_{\mathbb{R}_+^m} \prod_{j=1}^{m} \frac{\exp(-\alpha_j \lambda_j)(\alpha_j \lambda_j)^{y_j}}{y_j!} g(\boldsymbol{\alpha}) d\boldsymbol{\alpha}. \tag{1.2.15}$$

1.2.3 Copula-Based Model

A copula is a multivariate distribution function with uniform margins. Let $U_1, \cdots, U_m$ be m uniform random variables on $(0, 1)$. Their joint distribution function

$$C(u_1, \ldots u_m) = \Pr(U_1 \leq u_1, \ldots, U_m \leq u_m), \tag{1.2.16}$$

is called a copula.

Of course we seek to use the copula that is based on more than just uniformly distributed variables. Therefore we consider arbitrary marginal distribution functions $F_1(y_1), \cdots, F_m(y_m)$, and define a multivariate distribution function using the copula such that

$$F(y_1, \ldots y_m) = C(F_1(y_1), \ldots, F_m(y_m)). \tag{1.2.17}$$

For continuous outcomes, we can write the joint pdf as

$$f(y_1, \ldots y_m) = \prod_{j=1}^{m} f_j(y_j) \cdot c(F_1(y_1), \ldots, F_m(y_m)), \tag{1.2.18}$$

where $f_j(\cdot)$ is the pdf of the marginal distribution and $c(\cdot)$ is the copula density function. For our interested discrete count data, the joint pmf can be written as

$$f(y_1, \ldots y_m) = \sum_{j_1=1}^{2} \cdots \sum_{j_m=1}^{2} (-1)^{j_1 + \cdots + j_m} C(u_{1,j_1}, \ldots, u_{m,j_m}), \tag{1.2.19}$$

where $u_{j,1} = F_j(y_j - 1)$ and $u_{j,2} = F_j(y_j)$ respectively.

1.3 Numerical Methods

1.3.1 Newton-Raphson Method

The Newton-Raphson method iteratively solves nonlinear equations to determine the values of parameters that maximize the log-likelihood function. Mathematically, we assume the log-likelihood function is $\ell(\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$. Let $\boldsymbol{u}$ denote the score equation with elements

$$u_i = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_i}, \quad i = 1, \ldots, \dim(\boldsymbol{\theta}),$$

and $\boldsymbol{H}$ denote the Hessian matrix with elements

$$h_{ij} = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}, \quad i, j = 1, \ldots, \dim(\boldsymbol{\theta}).$$

Let $\boldsymbol{u}^{(t)}$ and $\boldsymbol{H}^{(t)}$ be $\boldsymbol{u}$ and $\boldsymbol{H}$ evaluated at $\boldsymbol{\theta}^{(t)}$ for step t ($t = 0, 1, 2, \ldots$). Starting from appropriate initial values $\boldsymbol{\theta}^{(0)}$, the Newton-Raphson algorithm is based on the following recurrence relation:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left(\boldsymbol{H}^{(t)}\right)^{-1} \boldsymbol{u}^{(t)}, \tag{1.3.20}$$

assuming that $\boldsymbol{H}^{(t)}$ is nonsingular.

Remark 1.3.1. *(Fisher Scoring)* An alternative iterative method is to replace $\boldsymbol{H}$ by $\mathcal{I} = -\mathrm{E}(\boldsymbol{H})$. The formula for Fisher scoring is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left(\mathcal{I}^{(t)}\right)^{-1} \boldsymbol{u}^{(t)}. \tag{1.3.21}$$

1.3.2 Expectation-Maximization Method

The EM algorithm (Dempster et al. (1977)) is a powerful iterative method for finding the maximum likelihood estimates (MLEs). It provides an alternative to the Newton-Raphson method when directly solving the scoring equation is quite complicated. The EM algorithm has been widely used in mixed Poisson

distribution (see Karlis (2005)), zero-inflated distribution (see Lambert (1992) and Hall (2000)) and zero-truncated distribution (see Tian et al. (2019)).

Formally, let $\boldsymbol{y}$ be the observed data and $\boldsymbol{z}$ be the latent or missing data. Let $(\boldsymbol{y}, \boldsymbol{z})$ be the complete data having the probability function $f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})$. Then the algorithm can be described as follows.

- E-step: At step t, compute

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \mathrm{E}\left[\log f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}\right].$$

- M-step: Maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ as a function of $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(t+1)}$.

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

Remark 1.3.2. *(GEM)* If the M-step is replaced with:

$$\text{M'-step: Find } \boldsymbol{\theta}^{(t+1)} \text{ so that } Q(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) > Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}),$$

the resultant algorithm is called the generalized EM (GEM) algorithm.

1.4 Bayesian Count Models

Bayesian method provides a quite different way to view statistical inference. The traditional approach (often referred to as frequentist) regards parameter values as fixed effects. The Bayesian approach, however, applies probability distributions to parameters. This yields inferences in the form of probability statement regarding the parameters given the data. It can provide relatively simple solutions for models where frequentist method fails, or at best, is difficult to implement.

1.4.1 Prior and Posterior Distributions

The Bayesian approach involves prior and posterior distributions for parameters $\boldsymbol{\theta}$. The prior distribution, denoted as $p(\boldsymbol{\theta})$, describes knowledge about $\boldsymbol{\theta}$ before we see the data $\boldsymbol{y}$. The posterior distribution, denoted as $p(\boldsymbol{\theta}|\boldsymbol{y})$, combines information from the current data $\boldsymbol{y}$ with distribution $f(\boldsymbol{y}|\boldsymbol{\theta})$ and prior information. By Bayes' theorem, we have

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{1.4.22}$$

With the denominator being independent of $\boldsymbol{\theta}$, we can omit this normalizing constant and write

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1.4.23}$$

1.4.2 Markov Chain Monte Carlo (MCMC)

In most cases, the posterior distribution has no closed-form expression. We need to make use of simulation methods to approximate the posterior distribution. The primary method for doing so is MCMC. In the following subsections, two basic Markov chain simulation methods: the Metropolis-Hastings (MH) algorithm and the Gibbs sampler are introduced.

Metropolis-Hastings Algorithm

The MH algorithm was first introduced by Metropolis et al. (1953) and later generalized by Hastings (1970). The algorithm is illustrated as follows. Generate a proposal $\boldsymbol{\theta}^*$ given $\boldsymbol{\theta}^{(t-1)}$ from a proposal distribution $q\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}\right)$ at iteration t. The proposal $\boldsymbol{\theta}^*$ is then accepted with probability $\alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)$ where

$$\alpha\left(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*\right) = \min\left(1, \frac{p\left(\boldsymbol{\theta}^*|\boldsymbol{y}\right) q\left(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y}\right) q\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}\right)}\right) \tag{1.4.24}$$

If the proposal is accepted, then $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$. Otherwise $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

Gibbs Sampler

The Gibbs sampler was introduced by Geman and Geman (1984) in the context of image restoration and later to the statistics literature by Gelfand and Smith (1990). Suppose the parameter vector $\boldsymbol{\theta}$ is divided into d components or subvectors, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$. At each iteration t, $\boldsymbol{\theta}_j^{(t)}$ is sampled from the conditional distribution given all the other components of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_j^{(t)} \sim p\left(\boldsymbol{\theta}_j | \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{j-1}^{(t)}, \boldsymbol{\theta}_{j+1}^{(t-1)}, \ldots, \boldsymbol{\theta}_d^{(t-1)}, \boldsymbol{y}\right), \quad j = 1, \ldots, d. \tag{1.4.25}$$

There are thus d steps at iteration t. Each subvector $\boldsymbol{\theta}_j$ is updated conditional on the latest values of the other components of $\boldsymbol{\theta}$, which are values at iteration t for the components already updated and the values at iteration $t-1$ for the others. The Gibbs sampler can be shown to be a special case of the MH algorithm.

In the following chapters, we present our multivariate count models to deal with different features exhibited in real datasets.

Chapter 2

Bayesian Multivariate Mixed Poisson Model with Copula-Based Mixture

2.1 Introduction

In this chapter, we propose a novel copula implementation based on the multivariate mixed Poisson model. To be specific, a certain copula is implemented on continuous mixing parameters of the Poisson margins. The introduction of copula provides a flexible way to construct the multivariate mixing distribution. It permits high freedom in marginal distribution selections for mixing parameters. The dependence parameters embedded in the copula could also handle the complex relations among margins. In addition, we can get rid of some identifiability and computational issues existing in the discrete copula models (see Genest and Nešlehová (2007)).

For inference, we shall adopt the idea of Bayesian hierarchical model (see, e.g., Ntzoufras (2009)). The main advantage of the Bayesian approach over the frequentist approach is that we can obtain posterior samples for the parameters of interest instead of just point estimates. This would give us some insight for the variability of the quantities. Besides, no effort needs to be made on computing the

multiple integration under Bayesian formulation. However, the over-complexity of our model would make it almost impossible to complete the full Bayesian analysis analytically. Fortunately, a battery of powerful tools have been developed over the past few decades aiming for generating posterior samples. One of the most popular methods is the well-known MCMC method. A detailed illustration can be found in Brooks et al. (2011). Recently, a new type of MCMC implementation has emerged, which is called Stan. Stan uses the No-U-Turn Sampler (NUTS) (see Hoffman and Gelman (2014)) which further extends a type of MCMC algorithm known as Hamiltonian Monte Carlo (HMC) algorithm. It allows us to analyze the models without having to worry about the implementation of the MCMC algorithm itself. In this chapter, the MCMC algorithm is implemented in STAN called from the R package **rstan**.

The rest of the chapter is organized as follows. Section 2.2 reviews the formulation of the multivariate mixed Poisson distribution, followed by our proposed version with copula incorporated. In Section 2.3, we present the inference method based on the MCMC algorithm. In Section 2.4 we conduct some simulation studies. Section 2.5 explores and compares several potential models for a given real dataset. The last section concludes the chapter.

2.2 The Model

2.2.1 Multivariate Mixed Poisson Distribution

Multivariate mixed Poisson distribution, as a generalization of univariate mixed Poisson distribution, can be defined as

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = \int_{\mathbb{R}_+^m} \prod_{j=1}^{m} \frac{\exp(-\lambda_j)\lambda_j^{y_j}}{y_j!} g(\boldsymbol{\lambda}; \cdot) d\boldsymbol{\lambda}, \tag{2.2.1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^\top$ denotes a vector of m counting random variables and $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$, $y_1, \ldots, y_m \in \mathbb{N}$. $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ is a vector of mixing

parameters. $g(\boldsymbol{\lambda};\cdot)$ denotes the pdf of a multivariate mixing distribution.

With the existence of covariates, we shall adopt another formulation that allows us to link certain parameters with given predictors. In this setting, the pmf of a multivariate mixed Poisson distribution can be defined as

$$\Pr(\boldsymbol{Y}=\boldsymbol{y})=\int_{\mathbb{R}_+^m}\prod_{j=1}^{m}\frac{\exp(-\alpha_j\lambda_j)(\alpha_j\lambda_j)^{y_j}}{y_j!}g(\boldsymbol{\alpha};\cdot)d\boldsymbol{\alpha}. \tag{2.2.2}$$

In this case, $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_m)^\top$ is a vector of mixing parameters and λ_j, $j=1,\ldots,m$, are Poisson parameters which can be linked with covariates. $g(\boldsymbol{\alpha};\cdot)$ is the joint pdf for mixing parameters. To ensure the identifiability under this definition, we assume $\mathrm{E}(\alpha_j)=1$. Some properties of the distribution (2.2.2) are given as follows.

- The marginal distribution of Y_j is a univariate mixed Poisson distribution with the marginal mixing distribution denoted by $g_j(\alpha_j;\cdot)$.

- The mean of Y_j is
$$\mathrm{E}(Y_j)=\lambda_j.$$

- The variance of Y_j is
$$\mathrm{Var}(Y_j)=\lambda_j+\lambda_j^2\sigma_j^2,$$
where σ_j^2 is the variance of α_j.

- The covariance between Y_j and $Y_{j'}$, $j,j'=1,\ldots,m$, $j\neq j'$, is
$$\mathrm{Cov}(Y_j,Y_{j'})=\lambda_j\lambda_{j'}\sigma_j\sigma_{j'}\rho_{jj'},$$
where $\rho_{jj'}$ is the correlation coefficient between α_j and $\alpha_{j'}$.

- The correlation coefficient between Y_j and $Y_{j'}$ is
$$\mathrm{Cor}(Y_j,Y_{j'})=\frac{\lambda_j\lambda_{j'}\sigma_j\sigma_{j'}\rho_{jj'}}{\sqrt{\lambda_j+\lambda_j^2\sigma_j^2}\sqrt{\lambda_{j'}+\lambda_{j'}^2\sigma_{j'}^2}},$$
and $|\mathrm{Cor}(Y_j,Y_{j'})|<|\rho_{jj'}|$ if $\rho_{jj'}\neq 0$.

Remark 2.2.1. When $\rho_{jj'} = 0$ for all $j, j' = 1, \ldots, m$ and $j \neq j'$, $(Y_1, \ldots, Y_m)$ follows a mutually independent multivariate mixed Poisson distribution.

Remark 2.2.2. The distribution (2.2.2) has limitation in modeling strong correlation between Y_j and $Y_{j'}$ when $\frac{1}{\lambda_j \sigma_j^2}$ or $\frac{1}{\lambda_{j'} \sigma_{j'}^2}$ is significantly larger than 0, as in this case $|\mathrm{Cor}(Y_j, Y_{j'})| < |\rho_{jj'}|$ significantly.

The most relevant candidate for the mixing distribution is multivariate log-normal (Aitchison and Ho (1989)), given as

$$g(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{m}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \prod_{j=1}^{m} \alpha_j} \exp\left[-\frac{1}{2} (\log \boldsymbol{\alpha} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\alpha} - \boldsymbol{\mu}) \right],$$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ as mean vector and $\boldsymbol{\Sigma}$ as covariance matrix. Using the parameter set $(\boldsymbol{\phi}, \boldsymbol{\theta})$ where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$ and $\boldsymbol{\theta} = (\theta_{12}, \ldots, \theta_{m-1,m})^\top$, the covariance matrix $\boldsymbol{\Sigma}$ takes the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \phi_1^2 & \theta_{12}\phi_1\phi_2 & \cdots & \theta_{1m}\phi_1\phi_m \\ \theta_{12}\phi_1\phi_2 & \phi_2^2 & \cdots & \theta_{2m}\phi_2\phi_m \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1m}\phi_1\phi_m & \theta_{2m}\phi_2\phi_m & \cdots & \phi_m^2 \end{pmatrix}.$$

To ensure identifiability, we let $\mu_j = -\frac{1}{2}\phi_j^2$, $j = 1, \ldots, m$.

A simplified version of the multivariate mixed Poisson distribution given in (2.2.2) can be obtained by assuming each Poisson margin shares a common mixing parameter. Specifically, the pmf is given as

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = \int_0^\infty \prod_{j=1}^{m} \frac{\exp(-\alpha\lambda_j)(\alpha\lambda_j)^{y_j}}{y_j!} g(\alpha; \cdot) d\alpha, \tag{2.2.3}$$

where $g(\alpha; \cdot)$ denotes the pdf of the univariate mixing parameter. For the sake of identifiability, we assume $\mathrm{E}(\alpha) = 1$. It is worth noting that the correlation between each pair of margins is always positive under this formulation.

2.2.2 The Copula Approach

As mentioned in the introduction section, we aim to build the multivariate mixed Poisson model by employing copula in constructing the mixing distribution. In this chapter, we shall consider Gaussian copula in a general case of m.

The cdf of a Gaussian copula is given by

$$C(\boldsymbol{u}; \boldsymbol{\Sigma}) = \Phi_m \left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m); \boldsymbol{\Sigma} \right),$$

where $\boldsymbol{u} = (u_1, \ldots, u_m)^\top \in (0,1)^m$. Here Φ_m is the cdf of a m-variate normal distribution with zero mean vector and covariance matrix equal to the correlation matrix $\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \theta_{12} & \cdots & \theta_{1m} \\ \theta_{12} & 1 & \cdots & \theta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1m} & \theta_{2m} & \cdots & 1 \end{pmatrix}.$$

Φ^{-1} is the inverse cdf of the standard normal distribution. The density of the copula is then given by

$$c(\boldsymbol{u}; \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1/2} \exp\left[\frac{1}{2} \boldsymbol{q}^\top (\boldsymbol{I}_m - \boldsymbol{\Sigma}^{-1}) \boldsymbol{q} \right],$$

with $\boldsymbol{q} = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m))^\top$ and $\boldsymbol{I}_m$ being the m-dimensional identity matrix.

Making use of Gaussian copula to formulate the multivariate mixing distribution $g(\boldsymbol{\alpha}; \cdot)$ in (2.2.2) gives,

$$g(\boldsymbol{\alpha}; \cdot) = \prod_{j=1}^{m} g_j(\alpha_j; \cdot) \cdot c\left(G_1(\alpha_1; \cdot), \ldots, G_m(\alpha_m; \cdot); \boldsymbol{\theta} \right).$$

Here $g_j(\alpha_j; \cdot)$ and $G_j(\alpha_j; \cdot)$ are the pdf and cdf of the mixing parameter α_j respectively. $c(\cdot, \ldots, \cdot; \boldsymbol{\theta})$ is the copula density function with pair-wise dependence parameters $\boldsymbol{\theta} = (\theta_{12}, \ldots, \theta_{m-1,m})^\top$ which are elements of correlation matrix $\boldsymbol{\Sigma}$.

Available choices for $g_j(\alpha_j;\cdot)$ include gamma, inverse Gaussian and lognormal. Given $\mathrm{E}(\alpha_j) = 1$, these three cases all have modified pdf with only one dispersion parameter ϕ_j.

- Gamma:

$$g_j(\alpha_j;\phi_j) = \frac{\phi_j^{\phi_j}}{\Gamma(\phi_j)}\alpha_j^{\phi_j-1}e^{-\phi_j\alpha_j}, \quad \alpha_j,\phi_j > 0,$$

  $\mathrm{Var}(\alpha_j) = 1/\phi_j$.

- Inverse Gaussian:

$$g_j(\alpha_j;\phi_j) = \left(\frac{\phi_j}{2\pi\alpha_j^3}\right)^{1/2}\exp\left[-\frac{\phi_j(\alpha_j-1)^2}{2\alpha_j}\right], \quad \alpha_j,\phi_j > 0,$$

  $\mathrm{Var}(\alpha_j) = 1/\phi_j$.

- Lognormal:

$$g_j(\alpha_j;\phi_j) = \frac{1}{\sqrt{2\pi}\phi_j\alpha_j}\exp\left[-\frac{(\log\alpha_j + \phi_j^2/2)^2}{2\phi_j^2}\right], \quad \alpha_j,\phi_j > 0,$$

  $\mathrm{Var}(\alpha_j) = \exp(\phi_j^2) - 1$.

Within the rest of the chapter, we only consider gamma, inverse Gaussian and lognormal as our potential candidates for marginal mixing distributions $g_j(\alpha_j;\phi_j)$.

When $m = 2$, there is only one dependence parameter θ. Several Archimedean copulas in addition to Gaussian copula are also introduced, as displayed in Table 2.1 and Table 2.2.

2.3 Model Inference

2.3.1 Likelihood Function

We use Gaussian copula to demonstrate the process. Covariates can be incorporated through a log-link function, namely

$$\lambda_{ij} = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}_j), \quad i = 1,\ldots,n, \quad j = 1,\ldots,m,$$

Table 2.1: Characteristics of selected copula families.

Copula	$C_\theta(u,v)$	Range of θ	Kendall's τ
Gaussian	$\Phi_2\left(\Phi^{-1}(u), \Phi^{-1}(v)\middle\vert\theta\right)$	$-1 \leq \theta \leq 1$	$\frac{2}{\pi}\arcsin(\theta)$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta > 0$	$\frac{\theta}{\theta+2}$
Frank	$-\frac{1}{\theta}\log\left[1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right]$	$\theta \neq 0$	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$ [a]
Gumbel	$\exp\left\{-\left[(-\log u)^\theta + (-\log v)^\theta\right]^{1/\theta}\right\}$	$\theta \geq 1$	$\frac{\theta-1}{\theta}$

[a] $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{x}{e^x-1}dx$.

Table 2.2: Copula density of selected copula families.

Copula	$c_\theta(u,v)$
Gaussian	$\frac{1}{\sqrt{1-\theta^2}}\exp\left\{\frac{2\theta\Phi^{-1}(u)\Phi^{-1}(v) - \theta^2\left[\left(\Phi^{-1}(u)\right)^2 + \left(\Phi^{-1}(v)\right)^2\right]}{2(1-\theta^2)}\right\}$
Clayton	$(\theta+1)(uv)^{-\theta-1}(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta-2}$
Frank	$\frac{\theta(1-e^{-\theta})e^{-\theta(u+v)}}{\left[e^{-\theta}-1+(e^{-\theta u}-1)(e^{-\theta v}-1)\right]^2}$
Gumbel	$(uv)^{-1}(\log u \cdot \log v)^{\theta-1}\left[w^{2/\theta-2} + (\theta-1)w^{1/\theta-2}\right]\exp(-w^{1/\theta})$ [a]

[a] $w = (-\log u)^\theta + (-\log v)^\theta$.

where $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})^\top$ is a $(p+1)$-dimensional vector of coefficients for the j-th component, $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ is a $(p+1)$-dimensional vector of covariates for individual i. $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$ is the dispersion parameter set. We denote the observed values as $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ and mixing variables as $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$ where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^\top$ and $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{im})^\top$. Let $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \boldsymbol{\phi}, \boldsymbol{\theta})$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n)$, then the joint likelihood function of $\boldsymbol{\Theta}$ takes the form

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \int_{\mathbb{R}_+^m} \prod_{j=1}^{m} \frac{\exp(-\alpha_{ij}\lambda_{ij})(\alpha_{ij}\lambda_{ij})^{y_{ij}}}{y_{ij}!} g(\boldsymbol{\alpha}_i; \boldsymbol{\phi}, \boldsymbol{\theta}) d\boldsymbol{\alpha}_i,$$

and the joint likelihood function of $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}$ is

$$L(\boldsymbol{\Theta}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\exp(-\alpha_{ij}\lambda_{ij})(\alpha_{ij}\lambda_{ij})^{y_{ij}}}{y_{ij}!}.$$

2.3.2 Priors

For the purpose of illustration, we assume the following prior distributions:

- $\beta_{jk} \sim N(0, \sigma_\beta^2)$, $j = 1, \ldots, m$, $k = 0, \ldots, p$.
- $(\alpha_{i1}, \ldots, \alpha_{im}) \sim C\left(G_1(\alpha_{i1}; \phi_1), \ldots, G_m(\alpha_{im}; \phi_m); \boldsymbol{\theta}\right)$, $i = 1, \ldots, n$.
- $\phi_j \sim Ga(\alpha_\phi, \beta_\phi)$, $j = 1, \ldots, m$.
- $\theta_{j,j'} \sim Uni(-1, 1)$, $j, j' = 1, \ldots, m$, $j < j'$.

Remark 2.3.1. α_ϕ and β_ϕ are shape and rate parameters for gamma distribution. To express prior ignorance, σ_β can be chosen as a relatively large number, and α_ϕ and β_ϕ can be chosen as relatively small numbers. It is not unusual to assume the above common parameters, σ_β, α_ϕ and β_ϕ, unless we have particular prior knowledge on individual cases.

2.3.3 Posterior Estimation

The posterior distribution of the parameters $\boldsymbol{\Theta}$ and mixing variables $\boldsymbol{\alpha}$ is

$$p(\boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \propto L(\boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)\pi(\boldsymbol{\Theta}, \boldsymbol{\alpha}),$$

where

$$\begin{aligned}\pi(\boldsymbol{\Theta}, \boldsymbol{\alpha}) = &\prod_{j=1}^{m}\prod_{k=0}^{p}\phi\left(\frac{\beta_{jk}}{\sigma_\beta}\right) \cdot \prod_{j=1}^{m}\pi_\phi(\phi_j|\alpha_\phi, \beta_\phi) \cdot \prod_{j=1}^{m-1}\prod_{j'>j}\pi_\theta(\theta_{j,j'}) \\ &\times \prod_{i=1}^{n}\prod_{j=1}^{m} g_j(\alpha_{ij}|\phi_j) \cdot \prod_{i=1}^{n} c(G_1(\alpha_{i1}|\phi_1), \ldots, G_m(\alpha_{im}|\phi_m); \boldsymbol{\theta}).\end{aligned}$$

Here $\phi(\cdot)$ denotes the pdf of the standard normal distribution; $\pi_\phi(\cdot)$ denotes the pdf of the prior distribution for ϕ_j; $\pi_\theta(\cdot)$ denotes the pdf of the prior distribution for $\theta_{j,j'}$.

In general, the Bayesian parameter inference is based on the posterior distribution of the parameters under consideration which is proportional to the product of the likelihood function and the priors for all parameters. There is no explicit form for the posterior distribution, yet we could generate posterior samples by virtue of MCMC method. The following computation results are all obtained with the help of **rstan** package in R.

2.4 Simulation Studies

In order to illustrate the performance of our proposed model, two simple simulation studies are carried out in this section.

Study 1. This study aims to test the proposed algorithm in Section 2.3. We simulate data by setting $n = 2,000$, $m = 3$ and $p = 1$ under the framework of (2.2.2), choosing different mixing distribution for each margin, which is gamma, inverse Gaussian and lognormal respectively. For the purpose of simplification,

the single covariate is simulated from $Uni(-1,1)$ distribution. The choices of the priors follow the discussions in Section 2.3.2.

The true parameter values, priors as well as posterior estimation results based on 10,000 iterations (excluding the first 1,000 iterations as burn-in) are summarized in Table 2.3. It is noted that the posterior mean estimates are very close to the true values and they all reside within the 95% credible intervals as well. These results validate our proposed algorithm.

Table 2.3: The parameter values, prior choices and sample posterior estimates.

	Prior			Posterior	
Parameter	Distribution	Mean	Std. Dev.	Mean	95% Cred. Int.
$\beta_{10}=1$	$N(0,100^2)$	0	100	0.993	(0.925, 1.060)
$\beta_{11}=1$	$N(0,100^2)$	0	100	1.029	(0.909, 1.147)
$\beta_{20}=2$	$N(0,100^2)$	0	100	1.964	(1.920, 2.010)
$\beta_{21}=-0.5$	$N(0,100^2)$	0	100	-0.497	(-0.569, -0.426)
$\beta_{30}=1.5$	$N(0,100^2)$	0	100	1.491	(1.312, 1.685)
$\beta_{31}=0.5$	$N(0,100^2)$	0	100	0.545	(0.351, 0.734)
$\phi_1=0.5$	$Ga(0.01,0.01)$	1	10	0.512	(0.471, 0.557)
$\phi_2=1$	$Ga(0.01,0.01)$	1	10	1.095	(0.989, 1.209)
$\phi_3=2$	$Ga(0.01,0.01)$	1	10	2.060	(1.955, 2.170)
$\theta_{12}=-0.5$	$Uni(-1,1)$	0	$1/\sqrt{3}$	-0.477	(-0.526, -0.425)
$\theta_{13}=0$	$Uni(-1,1)$	0	$1/\sqrt{3}$	0.033	(-0.028, 0.096)
$\theta_{23}=0.5$	$Uni(-1,1)$	0	$1/\sqrt{3}$	0.492	(0.441, 0.540)

Study 2. The second study compares the level of dependence between two different copula implementation approaches:

- Constructing copula at the mixing parameter level (continuous);
- Constructing copula at the counting variable level (discrete).

We only consider the bivariate case here as it can effectively demonstrate the difference between the above two options without unnecessary computational challenges. According to Remark 2.2.2, model (2.2.2) has a limitation in modeling

strong correlation at the counting variable level. To illustrate this point, we simulate data by implementing copula on continuous mixing parameters of Y_1 and Y_2, both adopting a gamma mixing distribution. The parameters for data generation are set as follows:

- $\lambda_1 = 1$, $\phi_1 = 1$,
- $\lambda_2 = 2$, $\phi_2 = 2$.

The dependence is measured by Kendall's τ, calculated from the model with copula directly implemented on discrete margins (negative binomial in this case) using the simulated data.

Four different underlying copulas are examined and the relationships are displayed in Figure 2.1, where the horizontal axes give the Kendall's τ of the mixing parameters whilst the vertical axes give the Kendall's τ of the counting variables. In the Gaussian copula case, it is as expected that the two different copula implementations are consistent in the independent case with Kendall's τ both being 0. The curve in the Frank copula case is quite similar to the one in the Gaussian plot, with an exception that τ cannot be 0 in the Frank copula. A consistent observation among the four cases is that the level of correlation between the counting variables is much restricted when data are simulated from model (2.2.2), under the above assumptions of marginal distributions. This confirms the fact that the discrete copula model might capture a broader level of dependence than our proposed continuous copula approach.

2.5 Application

2.5.1 Data Description

This section showcases a real-life application based on the data quantifying the demand for health care in Australia that were collected from the Australian Health

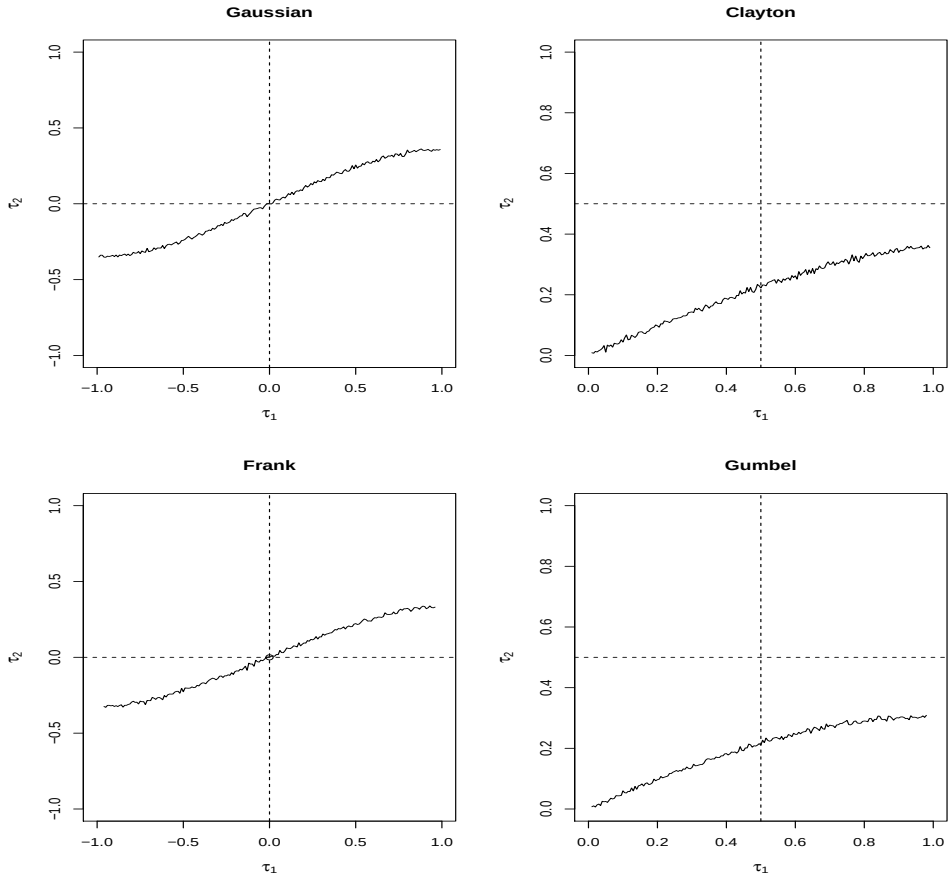

Figure 2.1: Dependence relation plots between two different copula implementation approaches under the measure of Kendall's τ.

Survey 1977-1978. The data were also used in Cameron and Trivedi (1998) to study a variety of count models. The whole dataset consists of 5,190 individuals. For this study, we randomly take 80% of the observations as training data to develop the models and the remaining 20% are reserved as the holdout sample for validation purposes.

We make use of three response variables in the dataset: the number of prescribed medications used in past 2 days (Y_1); the number of non-prescribed medications used in past 2 days (Y_2); and the number of admissions to a hospital, psychiatric hospital, nursing or convalescent home in the past 12 months (Y_3). The descriptive statistics of three marginal counts are displayed in Table 2.4. For each margin, the variance is obviously larger than the mean, indicating an overdispersion associated with our data. Thus our mixed Poisson models would well cope with this feature. In addition to responses, we have access to a set of covariates including respondent's sex, age, income and health status. These selected predictors with description and summary statistics are also displayed in Table 2.4.

Table 2.4: Summary of response variables and selected explanatory variables.

Variables	Description	Mean	Std. Dev.
Responses			
Y_1	The number of prescribed medications used in past 2 days	0.877	1.434
Y_2	The number of non-prescribed medications used in past 2 days	0.358	0.718
Y_3	The number of admissions to a hospital, psychiatric hospital, nursing or convalescent home in the past 12 months	0.175	0.515
Predictors			
Sex	1 if female, 0 if male	0.521	0.500
Age	Age in years divided by 100	0.406	0.205
Income	Annual income in Australian dollars divided by 1000	0.580	0.369
CHCOND1	1 if chronic condition(s) but not limited in activity, 0 otherwise	0.399	0.490
CHCOND2	1 if chronic condition(s) and limited in activity, 0 otherwise	0.118	0.322

2.5.2 Fitted Models

We initially perform a marginal analysis on the claim frequencies of each type. Specifically, three candidate distributions are considered for the counts of each margin: negative binomial (NB), Poisson-inverse Gaussian (PIG) and Poisson-lognormal (PLN). The log-likelihood values for the marginal models are exhibited in Table 2.5. The results indicate the best marginal distributions for Y_1, Y_2 and Y_3 are NB, PLN and PIG respectively. Therefore we choose gamma, lognormal and inverse Gaussian as potential mixing distributions in modeling Y_1, Y_2 and Y_3.

Table 2.5: Log-likelihood values for three margins under NB, PIG and PLN distributions.

Margin	NB	PIG	PLN
Y_1	**-5,328.07**	-5,365.77	-5,373.99
Y_2	-3,251.72	-3,246.42	**-3,245.06**
Y_3	-2,027.90	**-2,025.49**	-2,026.45

Having the best marginal mixing distributions determined, the MCMC algorithm is used to fit our proposed copula model. We run 11,000 iterations in total with the first 1,000 iterations treated as a burn-in period. Within the remaining 10,000 iterations, we reserve every 4th sample to reduce some auto-correlations. Also, four chains are performed in parallel to detect any convergence problems. The choices of the priors still follow the discussions in Section 2.3.2. The posterior estimation results of the model are summarized in Table 2.6. We also display the trace and density plots for 3 dependence parameters in Figure 2.2. For brevity, we omit the plots for other parameters, where similar patterns can be observed.

We make use of the potential scale reduction statistic $\hat{R}$ (see Gelman and Rubin (1992)) to monitor convergence. In our model, $\hat{R}$ for each parameter is close to 1, indicating chains are at equilibrium. There is also no indication of convergence problem as read from the trace plots. The significance of each parameter can be

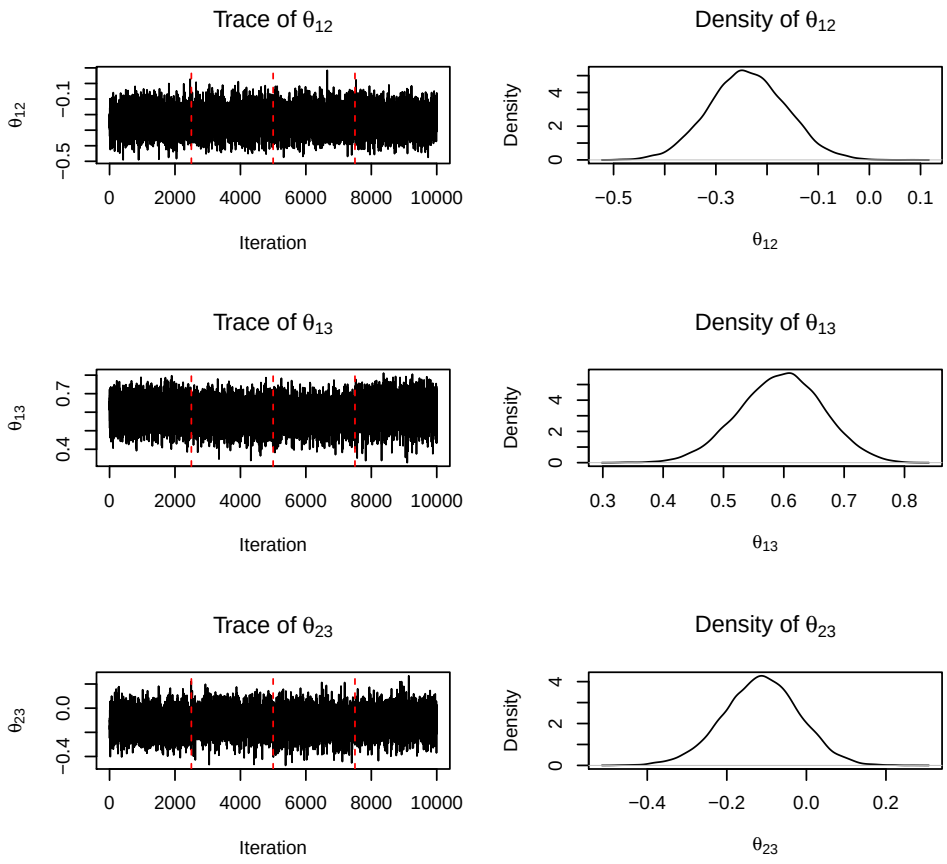

Figure 2.2: Trace and density plots for three dependence parameters.

Table 2.6: Posterior estimates for our proposed copula model.

	Y_1		Y_2		Y_3	
	Mean	95% Cred. Int.	Mean	95% Cred. Int.	Mean	95% Cred. Int.
Intercept	-2.435	(-2.614, -2.259)	-1.142	(-1.338, -0.941)	-2.049	(-2.351, -1.750)
SEX	0.601	(0.507, 0.695)	0.231	(0.105, 0.362)	0.172	(-0.014, 0.359)
AGE	2.233	(1.995, 2.467)	-0.839	(-1.175, -0.517)	0.014	(-0.457, 0.488)
INCOME	-0.027	(-0.156, 0.109)	0.209	(0.035, 0.380)	-0.456	(-0.730, -0.180)
CHCOND1	1.096	(0.987, 1.207)	0.368	(0.235, 0.505)	0.417	(0.203, 0.633)
CHCOND2	1.548	(1.419, 1.679)	0.327	(0.129, 0.520)	1.382	(1.136, 1.626)
ϕ	2.298	(1.932, 2.736)	0.858	(0.775, 0.940)	0.494	(0.357, 0.677)
θ_{12}	-0.238	(-0.382, -0.088)				
θ_{13}	0.595	(0.456, 0.727)				
θ_{23}	-0.114	(-0.305, 0.072)				

deduced from the corresponding 95% credible interval. Several conclusions are drawn from Table 2.6. First of all, females take medications more frequently than males irrespective of the types of medication. Secondly, when people are getting older, they tend to take more prescribed medications but less non-prescribed ones. However, gender and age are not significantly associated with admissions. Further, income is a key factor in modeling the frequency of taking non-prescribed medications as well as admissions, but no evidence showing that it affects the intake of prescribed medications. As expected, chronic conditions are good indicators for all three response variables. In respect of dependence parameters, it can be seen that the frequency of taking prescribed medications and non-prescribed medications are negatively associated, where the former exhibits a positive relation with admissions. Furthermore, there is no strong evidence to suggest significant dependence between the intake of non-prescribed medicines and admissions.

For comparison, we also fit two models with following mixing distributions:

- Multivariate lognormal (MLN) based on (2.2.2),
- Univariate lognormal (ULN) based on (2.2.3).

The posterior estimates are displayed in Table 2.7 for the MLN case and Table 2.8 for the ULN case. In addition, the results for the independent model (named as Ind) with NB, PLN and PIG as margins are given as a benchmark in Table 2.9. The same conclusions can be made regarding the significance of the estimates for covariates. The dependence relations among individual margins obtained in the MLN case agree with the results obtained in the copula model.

Table 2.7: Posterior estimates for the model with MLN as mixing distribution.

	Y_1		Y_2		Y_3	
	Mean	95% Cred. Int.	Mean	95% Cred. Int.	Mean	95% Cred. Int.
Intercept	-2.462	(-2.643, -2.281)	-1.139	(-1.338, -0.943)	-2.055	(-2.362, -1.754)
SEX	0.628	(0.531, 0.726)	0.229	(0.103, 0.356)	0.183	(-0.001, 0.370)
AGE	2.238	(1.999, 2.470)	-0.841	(-1.171, -0.506)	0.012	(-0.472, 0.479)
INCOME	-0.021	(-0.155, 0.113)	0.208	(0.038, 0.377)	-0.461	(-0.742, -0.189)
CHCOND1	1.102	(0.994, 1.213)	0.368	(0.231, 0.509)	0.428	(0.218, 0.645)
CHCOND2	1.566	(1.435, 1.697)	0.326	(0.125, 0.521)	1.401	(1.157, 1.653)
ϕ	0.636	(0.580, 0.692)	0.856	(0.775, 0.938)	1.107	(0.983, 1.239)
θ_{12}	-0.240	(-0.386, -0.084)				
θ_{13}	0.592	(0.453, 0.724)				
θ_{23}	-0.121	(-0.309, 0.062)				

Table 2.8: Posterior estimates for the model with ULN as mixing distribution.

	Y_1		Y_2		Y_3	
	Mean	95% Cred. Int.	Mean	95% Cred. Int.	Mean	95% Cred. Int.
Intercept	-2.439	(-2.617, -2.269)	-1.176	(-1.353, -0.997)	-2.101	(-2.374, -1.835)
SEX	0.593	(0.500, 0.687)	0.229	(0.112, 0.342)	0.200	(0.039, 0.364)
AGE	2.267	(2.043, 2.495)	-0.799	(-1.106, -0.503)	0.090	(-0.314, 0.501)
INCOME	-0.047	(-0.176, 0.081)	0.231	(0.079, 0.380)	-0.447	(-0.686, -0.207)
CHCOND1	1.102	(0.997, 1.210)	0.378	(0.255, 0.501)	0.410	(0.219, 0.604)
CHCOND2	1.558	(1.435, 1.685)	0.336	(0.154, 0.513)	1.364	(1.159, 1.568)
ϕ	0.549	(0.508, 0.590)				

To compare the performance of the above models, we compute a number

Table 2.9: Posterior estimates for the independent model.

	Y_1		Y_2		Y_3	
	Mean	95% Cred. Int.	Mean	95% Cred. Int.	Mean	95% Cred. Int.
Intercept	-2.450	(-2.630, -2.274)	-1.143	(-1.336, -0.945)	-2.028	(-2.334, -1.729)
Sex	0.610	(0.514, 0.708)	0.225	(0.097, 0.355)	0.135	(-0.052, 0.324)
Age	2.253	(2.021, 2.491)	-0.828	(-1.162, -0.495)	0.047	(-0.424, 0.520)
Income	-0.030	(-0.162, 0.102)	0.208	(0.037, 0.377)	-0.481	(-0.758, -0.212)
CHCOND1	1.096	(0.985, 1.207)	0.369	(0.228, 0.505)	0.417	(0.202, 0.632)
CHCOND2	1.554	(1.423, 1.682)	0.324	(0.125, 0.517)	1.368	(1.128, 1.614)
ϕ	2.256	(1.897, 2.688)	0.856	(0.773, 0.941)	0.502	(0.360, 0.690)

of information criteria that are summarized in Table 2.10. The log-likelihood values, AIC and BIC are obtained using posterior means as approximations of MLEs. The DIC statistics in our mixture models are calculated based on the conditional likelihood values. It can be concluded from the table that our copula model outperforms the others in the light of several goodness-of-fit statistics. The improvement of the copula model against MLN model is marginal telling from AIC and BIC, suggesting this trivariate dataset is not very sensitive to the choices of mixing distributions under the premise of pair-wise association being correctly captured in the frequentist context. However, the DIC values indicate that the MLN model performs even worse than the independent model. A possible explanation for that is this Bayesian criterion puts more emphasis on the choices of marginal distributions. The ULN model fits the data worst. It reveals the danger of adopting a uniform mixed distribution when the data actually exhibits various pair-wise dependence features.

We conclude the analysis with the Vuong test (Vuong (1989)) under the frequentist context to examine the closeness of four considered models. The test statistics with corresponding p-values are displayed in Table 2.11. A test statistic with a larger absolute value indicates that the two models being compared are further apart. As the independent model is nested in our copula model, the

Table 2.10: Different information criteria of four candidate models.

Mixture	LogLik	AIC	BIC	DIC
Copula	-9,640.94	19,329.87	19,481.82	18,541.11
MLN	-9,643.57	19,335.14	19,487.09	18,656.23
ULN	-9,808.55	19,655.10	19,775.40	19,340.97
Ind	-9,678.50	19,399.00	19,531.96	18,601.41

comparison between these two models is conducted using the likelihood-ratio test instead. One can observe that the ULN model has the worst performance, which has been foreshadowed by the information criteria in Table 2.10. The better performance of our copula model against the independent model demonstrates the necessity of incorporating three dependence parameters. Although our copula model outperforms the MLN model, the difference between them is not substantial.

Table 2.11: Vuong test statistics and p-values.

		Model 2			
		Copula	MLN	ULN	Ind
Model 1	Copula		1.221	6.451	75.127
			(0.222)	(< 0.001)	(< 0.001)
	MLN			6.427	2.587
				(< 0.001)	(0.010)
	ULN				-5.619
					(< 0.001)

2.5.3 Bivariate Study

To evaluate the impact of copula selection on modeling the dependence between response variables, we conduct a bivariate dependence study between Y_1 and Y_2. Archimedean copulas are taken into account to model the dependence in addition

to Gaussian copula. We also include a case based on bivariate lognormal mixing distribution (BLN) for comparison purpose. Since there exists negative relation between Y_1 and Y_2, only the Frank copula is applicable here. Estimation results for these three models are summarized in Table 2.12.

Table 2.12: Posterior estimates for three bivariate models.

	Gaussian		Frank		BLN	
	Mean	95% Cred. Int.	Mean	95% Cred. Int.	Mean	95% Cred. Int.
Y_1:						
Intercept	-2.449	(-2.627, -2.269)	-2.446	(-2.629, -2.265)	-2.474	(-2.658, -2.293)
SEX	0.610	(0.516, 0.707)	0.609	(0.515, 0.705)	0.637	(0.539, 0.737)
AGE	2.246	(2.014, 2.486)	2.245	(2.011, 2.482)	2.252	(2.020, 2.489)
INCOME	-0.028	(-0.159, 0.104)	-0.030	(-0.164, 0.102)	-0.024	(-0.158, 0.113)
CHCOND1	1.097	(0.987, 1.207)	1.096	(0.985, 1.208)	1.104	(0.995, 1.216)
CHCOND2	1.553	(1.424, 1.684)	1.553	(1.423, 1.682)	1.570	(1.437, 1.701)
ϕ_1	2.263	(1.906, 2.693)	2.273	(1.903, 2.722)	0.638	(0.582, 0.695)
Y_2:						
Intercept	-1.145	(-1.341, -0.953)	-1.145	(-1.341, -0.949)	-1.141	(-1.340, -0.945)
SEX	0.234	(0.103, 0.363)	0.233	(0.104, 0.362)	0.230	(0.101, 0.358)
AGE	-0.838	(-1.176, -0.503)	-0.833	(-1.168, -0.498)	-0.838	(-1.176, -0.502)
INCOME	0.210	(0.041, 0.377)	0.210	(0.043, 0.373)	0.209	(0.042, 0.377)
CHCOND1	0.368	(0.232, 0.505)	0.368	(0.230, 0.504)	0.368	(0.232, 0.504)
CHCOND2	0.328	(0.125, 0.523)	0.324	(0.129, 0.520)	0.324	(0.125, 0.523)
ϕ_2	0.859	(0.777, 0.944)	0.858	(0.776, 0.943)	0.860	(0.775, 0.944)
θ	-0.231	(-0.373, -0.088)	-1.618	(-2.720, -0.598)	-0.229	(-0.379, -0.083)
LogLik		-7,727.80		-7,727.33		-7,729.73
DIC		14,891.43		14,894,43		15,002.70

Several interesting observations can be made from the results. First, estimates for covariates and dispersion parameters coincide with each other under the two copula models. The log-likelihood and DIC results indicate that there is no substantial difference between these two copulas for this particular dataset. Second, we notice the closeness of estimated regression coefficients for Y_1 and Y_2 under this bivariate case and the previous trivariate case, in both the copula model and the MLN model. They are also consistent with the estimates in

marginal analysis. It implies that the two-step composite likelihood estimation method proposed by Zhao and Joe (2005) might provide good approximations for estimates under these two models. Overall, the results suggest superior fits of the copula models to the BLN mixture model.

2.5.4 Predictive Analysis

We conclude the chapter with an illustration for predictive analysis using the out-of-sample. The predicted frequencies are calculated in contrast to observed ones under three different mixture models mentioned in Section 2.5.2. χ^2 statistics and corresponding p-values are also provided for comparison.

Marginal predictions for the three response variables are provided in Table 2.13. The small χ^2 statistic with a p-value greater than 0.05 indicates the sufficiency of our copula model in describing the marginal behaviors. One can see that on individual dimensions, the prediction performance of our copula model is better than MLN and ULN cases in general. This again confirms the benefit of separating marginal distributions from dependence structure through the implementation of the copula model. The model using MLN mixture has almost equal prediction power as the copula model except for Y_1. It implies that the first margin suffers a biased estimate when using lognormal as mixing distribution instead of gamma. Again, the model with ULN mixing distribution has the worst performance.

Next, we would like to evaluate the predictive performance under the multivariate case. In our setting, we compare the predicted numbers with the observed ones in the following situations: no occurrence for each event, occurrence for only one event, occurrence for two events and all three events. The results are given in Table 2.14. The model with ULN mixing distribution results in poor predictions in the multivariate context as it fails to consider different dependence relations

Table 2.13: Goodness-of-fit for three margins under three models.

Y_1	Observed	Copula	MLN	ULN
0	629	597.88	595.75	585.75
1	199	228.69	231.74	237.18
2	101	98.82	99.38	103.42
3	53	49.47	48.89	51.06
4	29	26.77	25.97	26.79
5	12	15.07	14.45	14.52
6	8	8.68	8.33	8.05
7	3	5.07	4.94	4.56
$\geq$8	4	7.55	8.55	6.66
χ^2		9.16	10.82	11.68
p-value		0.33	0.21	0.17
Y_2	Observed	Copula	MLN	ULN
0	764	763.54	763.35	741.66
1	215	205.50	205.73	234.01
2	40	49.51	49.53	50.30
3	15	13.12	13.11	9.70
$\geq$4	4	6.33	6.29	2.33
χ^2		3.39	3.36	8.42
p-value		0.49	0.50	0.08
Y_3	Observed	Copula	MLN	ULN
0	896	899.54	899.44	883.19
1	117	110.42	110.73	133.89
2	19	19.80	19.48	17.62
$\geq$3	6	8.24	8.36	3.31
χ^2		1.05	1.04	4.62
p-value		0.79	0.79	0.20

among the responses. Our copula model outperforms the other two models, although the improvement against the MLN case is only marginal. This is consistent with the observations in our previous in-sample analysis. The χ^2 statistic and p-value again show excellence and sufficiency of adopting Gaussian copula to deal with the dependence structure in this trivariate case.

Table 2.14: Predicted counts of three models under eight different scenarios.

(Y_1, Y_2, Y_3)	Observed	Copula	MLN	ULN
(0, 0, 0)	401	393.68	392.44	394.68
(>0, 0, 0)	255	266.26	267.16	246.97
(0, >0, 0)	155	147.81	146.84	128.85
(0, 0, >0)	53	41.59	41.71	43.11
(>0, >0, 0)	85	91.80	93.02	112.68
(>0, 0, >0)	55	62.03	62.06	56.90
(0, >0, >0)	20	14.78	14.76	19.11
(>0, >0, >0)	14	20.06	20.01	35.69
χ^2		9.07	9.41	28.02
p-value		0.25	0.22	0.00

2.6 Concluding Remarks

To generalize the mixed Poisson model from the univariate case to the multivariate case, several methods are applicable. One is to assume multivariate lognormal as mixing distribution. The other is to let each Poisson margin share a same mixing parameter. However, these two methods are limited in incorporating flexible dependence relations and/or marginal distributions. To address these two limitations, we propose a novel multivariate mixed Poisson model with copula implemented on continuous mixing parameters. This is a different treatment of copula in multivariate count modeling compared with the traditional way of constructing copula on the counts directly. For the multivariate case, Gaussian copula

is a prior choice due to the flexibility of dealing with pair-wise associations. When restricted to the bivariate case, some Archimedean copulas are also applicable. In respect of the marginal mixing distributions, three typical choices, namely gamma, inverse Gaussian and lognormal, are introduced.

As for the parameter estimation, we apply the idea of hierarchical modeling under the Bayesian framework. The implementation of MCMC method makes the whole inference process applicable. One major advantage of the Bayesian approach over frequentist approach is that we could obtain a distribution for the quantity of interest instead of just a point estimate. This provides some insight to capture the uncertainty of our estimates. When the sample size is relatively large and priors are non-informative, the posterior mean estimates are good approximations to MLEs as illustrated in our simulation studies.

A comparison of a number of multivariate mixed Poisson models is conducted with some real data from the Australian Health Survey. The example expresses the superiority of our proposed copula model in comparison with the others. The improvement achieved by the copula model are grounded in the flexibility in both selecting marginal distributions and modeling the dependence relations. By facilitating the out-of-sample observations in the dataset, predictive analysis shows that our copula model is sufficient to predict this dataset both at marginal and multidimensional levels. Failure to properly deal with the dependence among individual response variables would lead to poor predictions, especially in the multivariate case.

Chapter 3

Multivariate Zero-Inflated Hurdle Model

3.1 Introduction

This chapter aims to address the issue of multivariate zero-inflation. It is often the case in automobile insurance that 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-claim discount (NCD) to their annual insurance premium. In short, insurance claim frequency data is often characterized by a large number of zero claims. This, in turn, leads to the study of zero-inflated versions of the 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. Such facts give rise to our study on multivariate zero-inflation with dependence.

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 datasets. In an attempt to alleviate these limitations, we propose our multivariate zero-inflated hurdle model. Assuming each marginal distribution follows a hurdle distribution has several advantages. First, it can easily handle the zero-inflation or zero-deflation feature in each margin. Second, it provides us with greater choices to describe marginal behaviors.

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 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 and Matthews (1993) provides 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.

The rest of the chapter is organized as follows. Section 3.2 provides the definition of general multivariate zero-inflated distribution and investigates some of its distributional properties. In Section 3.3, we propose our multivariate zero-inflated hurdle model, followed by the corresponding EM algorithm for model inference. In Section 3.4, the proposed model is applied to an automobile insurance

dataset. The last section concludes the chapter.

3.2 Multivariate Zero-Inflated Distribution

3.2.1 Definition

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

$$\boldsymbol{Z} \overset{d}{=} U\boldsymbol{Y} = \begin{cases} \boldsymbol{0}_m, & U = 0, \\ \boldsymbol{Y}, & U = 1, \end{cases} \tag{3.2.1}$$

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

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} \Pr(Y_j = 0)\right]^{v} \left[\pi_0 \prod_{j=1}^{m} \Pr(Y_j = z_j)\right]^{1-v}, \tag{3.2.2}$$

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

3.2.2 Properties of Distribution

Marginal Distributions

We first derive the marginal distribution for a single variable.

Proposition 3.2.1. 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,

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

□

Next we derive an expression for the marginal distribution of an arbitrary random sub-vector of $\boldsymbol{Z}$.

Proposition 3.2.2. 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,

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

□

Conditional Distribution

Proposition 3.2.3. 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

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

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

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

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

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

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

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

□

Expectation, Variance and Covariance

Proposition 3.2.4. The expectation and variance of Z_j, $j = 1, \ldots, m$, are

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

and the covariance between Z_j and $Z_{j'}$, $j, j' = 1, \ldots, m$, $j \neq j'$, is

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

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

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

Moment Generating Function

Proposition 3.2.5. 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

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

□

3.2.3 Two Commonly Used Distributions

In this subsection, we introduce two commonly used multivariate zero-inflated models. The corresponding EM algorithms for inference are given in Appendix A.

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

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

Multivariate Zero-Inflated Negative Binomial Distribution

Let $Y_j \sim NB(\lambda_j, \phi_j)$, for $j = 1, \ldots, m$. Then $\boldsymbol{Z}$ is said to follow the multivariate zero-inflated negative binomial distribution with two parameter vectors $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$ and a zero-inflation parameter π_0, denoted by $\boldsymbol{Z} \sim MZINB(\boldsymbol{\lambda}, \boldsymbol{\phi}, \pi_0)$. The pmf of $\boldsymbol{Z}$ is

$$\begin{aligned}\Pr(\boldsymbol{Z} = \boldsymbol{z}) = &\left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} \left(\frac{\phi_j}{\lambda_j + \phi_j}\right)^{\phi_j}\right]^{v} \\ &\times \left[\pi_0 \prod_{j=1}^{m} \frac{\Gamma(z_j + \phi_j)}{\Gamma(\phi_j) z_j!} \left(\frac{\lambda_j}{\lambda_j + \phi_j}\right)^{z_j} \left(\frac{\phi_j}{\lambda_j + \phi_j}\right)^{\phi_j}\right]^{1-v}.\end{aligned}$$

3.3 Multivariate Zero-Inflated Hurdle Model

3.3.1 Model Characterization

We shall assume that each underlying random variable Y_j in (3.2.1) follows a zero-modified distribution, which can be characterized as follows:

$$Y_j \overset{d}{=} U_j W_j = \begin{cases} 0, & U_j = 0, \\ W_j, & U_j = 1, \end{cases} \tag{3.3.3}$$

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 (3.2.1) 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

$$\begin{aligned}\Pr(\boldsymbol{Z} = \boldsymbol{z}) = &\left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_j)\right]^{v} \\ &\times \left[\pi_0 \prod_{j:z_j=0} (1 - \pi_j) \prod_{j:z_j \neq 0} \pi_j f_{W_j}(z_j)\right]^{1-v}.\end{aligned} \tag{3.3.4}$$

3.3.2 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 $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$. The parameter π_{ij} can then be modeled as

$$\pi_{ij} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}, \tag{3.3.5}$$

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

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Theta}, \pi_0) = &\prod_{i=1}^{n} \left[1 - \pi_0 + \pi_0 \prod_{j=1}^{m} (1 - \pi_{ij}) \right]^{v_i} \\ &\times \prod_{i=1}^{n} \left[\pi_0 \prod_{j:z_{ij}=0} (1 - \pi_{ij}) \prod_{j:z_{ij}\neq 0} \pi_{ij} f_{W_j}(z_{ij}) \right]^{1-v_i}. \end{aligned} \tag{3.3.6}$$

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

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

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:

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

- 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 smaller than a tolerance level ε.

Remark 3.3.1. The initial values of parameters $\boldsymbol{\beta}_j$ for EM algorithm can be obtained by implementing univariate logistic regression with observed values $z'_{1j}, \ldots, z'_{nj}$. 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:

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

3.4 Application

3.4.1 Data Description

This application is based on an automobile portfolio from a major insurance company operating in Spain in 1995. The dataset contains information for 80,994 policyholders. Eleven covariates are considered in our analysis. The detailed description for each predictor is presented in Table 3.1. 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 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 excluded from this simplest policy. This simplest type of policy forms the baseline group, while the variable $v9$ denotes the policies which also include comprehensive coverage (except fire) and the variable $v10$ denotes policies which also include comprehensive and collision coverage. This dataset was previously used in Bermúdez (2009) and Bermúdez and Karlis (2012) for analysis of bivariate count models. To achieve a better comparison between our model and other models

including the ones in Bermúdez (2009) and Bermúdez and Karlis (2012), we use the whole dataset for model fitting.

The empirical joint distribution for claim numbers Z_1 and Z_2 is displayed in Table 3.2. The overall Pearson's correlation between these two types of claim is 0.187. This observed correlation is consistent with our model's assumption that there exists a positive correlation between the two margins. We also analyze the correlation when both claim counts are not zero. The Pearson's correlation drops to 0.126, indicating a small right-tailed dependence. Thus our model which assumes independence in the right tail will likely work very well with these data.

Table 3.1: 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

3.4.2 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) and unit-shifted negative binomial (USNB) distributions. Implementing a unit-shifted

Table 3.2: Empirical joint distribution of Z_1 and Z_2.

Z_1	Z_2							
	0	1	2	3	4	5	6	7
0	71,087	3,722	807	219	51	14	4	0
1	3,022	686	184	71	26	10	3	1
2	574	138	55	15	8	4	1	1
3	149	42	21	6	6	1	0	1
4	29	15	3	2	1	1	0	0
5	4	1	0	0	0	0	2	0
6	2	1	0	1	0	0	0	0
7	1	0	0	1	0	0	0	0
8	0	0	1	0	0	0	0	0

distribution on W_j is equivalent to using a standard count distribution for $W_j - 1$. The comparison results are summarized in Table 3.3. 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 distance between observed and fitted frequencies demonstrates the adequacy of adopting a USNB distribution for both margins.

We next turn to parameter estimates when covariates are introduced in our multivariate zero-inflated hurdle (MZIH) model. The estimation results in Table 3.4 correspond to π_j, $j = 0, 1, 2$. The 95% confidence interval of π_0 is $(0.300, 0.322)$, where the upper bound is far away from the boundary 1. This reveals the zero-inflation feature reflected in the dataset. Focusing on the third party liability claims, parameters $v2$ and $v4$-$v7$ are statistically significant. It can be concluded that factors like driving in urban area ($v2$), drivers with more experience ($v5$, $v6$), policyholders with more years with the insurance company ($v7$) reduce the probability of claiming. Whereas driving in a high risk region could increase the chances of making a claim. If we focus on the rest of guarantees, we notice that

Table 3.3: Goodness-of-fit of marginal models.

W_1	Observed	ZTP	ZTNB	USP	USNB
1	4,003	3,859.22	4,026.70	3,814.73	3,999.98
2	796	1,023.18	780.45	1,100.20	813.68
3	226	180.85	199.72	158.65	202.65
4	51	23.97	57.31	15.25	53.55
5	7	2.54	17.51	1.10	14.56
$\geq$6	7	0.24	8.31	0.07	5.59
χ^2		293.32	11.12	958.45	7.48
LogLik		-3,546.53	-3,483.17	-3,604.39	-3,481.01
W_2	Observed	ZTP	ZTNB	USP	USNB
1	4,605	4,375.24	4,649.53	4,302.22	4,603.02
2	1,071	1,398.38	1,026.10	1,520.45	1,079.10
3	315	297.96	298.98	268.67	308.13
4	92	47.62	97.68	31.65	93.24
5	30	6.09	33.99	2.80	29.01
$\geq$6	13	0.71	19.73	0.21	13.50
χ^2		436.79	6.34	1,321,19	0.28
LogLik		-4,864.86	-4,755.15	-4,963.00	-4,751.31

$v2$-$v5$, $v7$, $v9$-$v11$ are all statistically significant. It tells us that driving in an urban area ($v2$), driving in a zone of medium risk ($v3$), driving license between 4 and 14 years ($v5$), policy with comprehensive and collision coverage ($v9$, $v10$), greater horsepower ($v11$) would increase the chances of a claim in this category. However, driving in a high risk zone ($v4$) and clients with the company more than 5 years ($v7$) can decrease the occurrence probability. Table 3.5 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 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, only $v9$ is a significant predictor for the expected number of W_2.

Table 3.4: Estimates and t-ratios associated with the covariates of π_j.

	π_1		π_2	
	Estimate	t-ratio	Estimate	t-ratio
Intercept	-0.932	-6.615***	-3.330	-18.449***
$v1$	0.021	0.448	0.031	0.557
$v2$	-0.077	-2.123*	0.117	2.655**
$v3$	0.005	0.123	0.227	4.614***
$v4$	0.280	6.361***	-0.161	-2.898**
$v5$	-0.296	-2.561*	0.432	2.925**
$v6$	-0.453	-3.707***	0.058	0.377
$v7$	-0.175	-3.692***	-0.326	-5.837***
$v8$	0.108	1.564	0.102	1.247
$v9$	0.051	0.999	3.831	52.118***
$v10$	0.051	1.325	2.207	42.112***
$v11$	0.070	1.552	0.363	5.999***
	Estimate	95% CI		
π_0	0.311	(0.300, 0.322)		

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

Table 3.5: Estimates and t-ratios associated with the covariates of W_j.

	W_1		W_2	
	Estimate	t-ratio	Estimate	t-ratio
Intercept	-1.178	-5.022***	-1.429	-5.588***
$v1$	-0.112	-1.290	0.023	0.338
$v2$	0.008	0.118	-0.025	-0.447
$v3$	-0.105	-1.330	0.067	1.137
$v4$	-0.071	-0.900	-0.034	-0.451
$v5$	-0.067	-0.350	0.071	0.355
$v6$	-0.066	-0.324	0.029	0.138
$v7$	-0.005	-0.065	-0.119	-1.744
$v8$	0.066	0.543	-0.019	-0.193
$v9$	0.109	1.178	0.832	7.460***
$v10$	0.040	0.564	0.170	1.505
$v11$	0.019	0.231	-0.078	-0.892
	Estimate	95% CI		
ϕ_1	0.695	(0.561, 0.829)		
ϕ_2	0.830	(0.699, 0.961)		

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

3.4.3 Model Comparison

In this section, we aim to compare different available models. Apart from our proposed MZIH model, fitted with the USNB marginal distributions, candidates also include the MZIP and MZINB models as well. It is worth mentioning that for zero-truncation parts in MZIH model, only significant predictors are adopted to avoid over-fitting problem. Furthermore, a model with two independent hurdle margins (Ind) is fitted as a benchmark. Regarding the distributions for zero-truncation parts in the independent hurdle model, they are same as in our MZIH model. Different information criteria are listed in Table 3.6. By assuming the existence of zero-inflation parameter, we are able to improve the model fitting considerably. This confirms the fact of extra common zeros in our dataset. Also the superior performance of MZIH model over MZIP and MZINB models verifies the flexibility of our model for handling marginal distributions.

It is interesting to compare our models with previous ones. The best model from Bermúdez (2009) is the zero-inflated bivariate Poisson (ZIBP) model with regressors on the third Poisson parameter λ_0. And the best model from Bermúdez and Karlis (2012) 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 are given in Appendix B. The results of comparison are shown in Table 3.6. As noticed from the table, all information criteria agree that our purposed MZIH model performs best.

3.4.4 Predictive Analysis

To evaluate the predictive performance, we randomly take 70% of the observations from the whole sample as training data to develop the models, and the remaining 30% are reserved as the hold-out sample for validation purposes. We then calculate the predicted claim frequencies and compare these to the observed ones based

Table 3.6: Information criteria of several relevant fitted models.

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

on the out-of-sample for the following scenarios: no claims in any line, claims occur in only one of the lines and claims occur in both lines. Model candidates include MZIH, MZIP, MZINB and Ind models. The goodness-of-fit results are shown in Table 3.7. Conclusions are consistent with those made for Table 3.6. As anticipated, we observe very poor performance of the Ind model, which can be explained by failure to consider the excess of zeros in the dataset and the positive correlation between the two lines. The more accurate prediction of MZIH model against MZIP and MZINB is largely due to the introduction of hurdle margins. The tiny discrepancies between observed and predicted frequencies of our model suggest a satisfactory quality of fit for this particular dataset.

Table 3.7: Predicted frequencies of four models under four different scenarios.

(Z_1, Z_2)	Observed	MZIH	MZIP	MZINB	Ind
(0, 0)	21,319	21,331.53	21,313.06	21,322.39	21,050.73
(>0, 0)	1,151	1,133.47	1,257.60	1,184.15	1,408.17
(0, >0)	1,462	1,438.02	1,280.19	1,451.14	1,719.48
(>0, >0)	367	395.98	448.16	341.32	120.63
χ^2		2.80	49.56	2.94	592.10

3.5 Concluding Remarks

In this chapter we introduce a new type of multivariate model, the so called multivariate zero-inflated hurdle model (MZIH). We start with the definition for the multivariate zero-inflated distribution and provide some of its properties. Next, two commonly used models, namely the MZIP and MZINB, are reviewed with associated EM algorithms necessary for estimating their parameters given in Appendix A. However, these two models often cannot accommodate observed characteristics in the marginal distributions for claim counts. To overcome these restrictions, we present the MZIH model where each margin is assumed to follow a zero-modified distribution. The separation of zeros from the positive parts provides us with more freedom to deal with zero features in each dimension and the marginal distributions as well. The usefulness of our model is illustrated with the help of real data from automobile insurance. The results from fitting several relevant models are shown and compared. As expected, the superiority of our proposed model is supported by different information criteria as well as predictive performance.

Chapter 4

Type I Multivariate Zero-Truncated Hurdle Model

4.1 Introduction

The previous chapter resolves the multivariate zero-inflation problem, whereas in this chapter we deal with another often encountered issue in insurance: multivariate zero-truncation. Counting data without zero observations are common in actuarial literature. Univariate zero-truncated models have been routinely applied by practitioners to the analysis of this type of missing information. The commonly used ones include zero-truncated Poisson (ZTP) model and zero-truncated negative binomial (ZTNB) model (see Grogger and Carson (1991)). However, univariate zero-truncated models have obvious limitations, one of which arises from policies with multiple claim types. This is not uncommon 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 chapter, 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.

Aiming at the Type I multivariate zero-truncation, Tian et al. (2018) proposed a multivariate zero-truncated Poisson distribution. A drawback of this model is the restricted choice for the marginal behaviors. To address the issue, we propose a 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. Following the idea in Tian et al. (2018), we identify a stochastic representation of the Type I zero-truncated counting random variables, which facilitates us to make use of a novel EM algorithm (Dempster et al. (1977)) for the parameters' estimation.

The rest of this chapter is organized as follows. Section 4.2 gives the definition of general Type I multivariate zero-truncated distribution, followed by some distributional properties. Two commonly used distributions: Type I multivariate zero-truncated Poisson distribution and Type I multivariate zero-truncated negative binomial distribution are then presented. The associated EM algorithms for parameters' estimation are provided in Appendix C. In Section 4.3, we propose a new Type I multivariate zero-truncated hurdle model, followed with its parameter estimation via EM algorithm and some simulation studies to test the model performance. In Section 4.4, the proposed model is applied to a real health insurance dataset and both model fitting and prediction results are exhibited. Concluding remarks are given in Section 4.5.

4.2 Type I Multivariate Zero-Truncated Distribution

4.2.1 Definition

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 Type I multivariate zero-truncated distribution if

$$\boldsymbol{Y} \stackrel{d}{=} U\boldsymbol{Z} = \begin{cases} \boldsymbol{0_m}, & U = 0, \\ \boldsymbol{Z}, & U = 1, \end{cases} \tag{4.2.1}$$

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

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \frac{\Pr(\boldsymbol{Y} = \boldsymbol{z})}{\Pr(U = 1)} = \frac{\prod_{j=1}^m \Pr(Y_j = z_j)}{\pi_0}, \quad \|\boldsymbol{z}\|_1 > 0, \tag{4.2.2}$$

An alternative representation is to define $\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{Y} | \boldsymbol{Y} \neq \boldsymbol{0}_m$. This stochastic representation helps us to generate random values of $\boldsymbol{Z}$.

4.2.2 Properties of Distribution

Marginal Distributions

We first derive the marginal distribution of a single variable.

Proposition 4.2.1. The marginal distribution of Z_j is

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

where $\pi_0 = 1 - \prod_{j=1}^m f_{Y_j}(0)$.

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,

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

□

Next we derive an expression for the marginal distribution of an arbitrary random sub-vector of $\boldsymbol{Z}$.

Proposition 4.2.2. 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

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

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

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

Thus,

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

□

Conditional Distribution

Proposition 4.2.3. 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

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

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

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

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

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

□

Expectation, Variance and Covariance

Proposition 4.2.4. The expectation and variance of Z_j, $j = 1, \ldots, m$, are

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

and the covariance between Z_j and $Z_{j'}$, $j, j' = 1, \ldots, m$, $j \neq j'$, is

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

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

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

Moment Generating Function

Proposition 4.2.5. The moment generating function of $\boldsymbol{Z}$ is

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

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

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

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

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

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

□

4.2.3 Two Commonly Used Distributions

In this subsection, we introduce two commonly used Type I multivariate zero-truncated models. The corresponding EM algorithms for inference are given in Appendix C.

Type I Multivariate Zero-Truncated Poisson Distribution

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 $\boldsymbol{Z} \sim MZTP(\boldsymbol{\lambda})$. As a result, $\pi_0 = 1 - e^{-\sum_{j=1}^m \lambda_j}$. The pmf of $\boldsymbol{Z}$ is

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \frac{1}{1 - e^{-\sum_{j=1}^m \lambda_j}} \prod_{j=1}^m \frac{\lambda_j^{z_j} e^{-\lambda_j}}{z_j!}.$$

Type I Multivariate Zero-Truncated Negative Binomial Distribution

Let $Y_j \sim NB(\lambda_j, \phi_j)$, for $j = 1, \ldots, m$. Then $\boldsymbol{Z}$ is said to follow the Type I multivariate zero-truncated negative binomial distribution with two parameter vectors $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top$, denoted by $\boldsymbol{Z} \sim MZTNB(\boldsymbol{\lambda}, \boldsymbol{\phi})$. As a result, $\pi_0 = 1 - \prod_{j=1}^m [\phi_j/(\lambda_j + \phi_j)]^{\phi_j}$. The pmf of $\boldsymbol{Z}$ is

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

4.3 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.

4.3.1 Model Characterization

We shall assume that each underlying random variable Y_j in (4.2.1) follows a zero-modified distribution, which can be characterized as follows:

$$Y_j \stackrel{d}{=} U_j W_j = \begin{cases} 0, & U_j = 0, \\ W_j, & U_j = 1, \end{cases} \tag{4.3.3}$$

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 (4.2.1) 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 $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_m)^\top$. Here $\boldsymbol{\Theta}_j$ is the set of parameters related to W_j. Let $\pi_0 = 1 - \prod_{j=1}^m (1 - \pi_j)$, then the pmf of $\boldsymbol{Z}$ can be expressed as

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \frac{1}{\pi_0} \prod_{j:z_j=0} (1 - \pi_j) \prod_{j:z_j \neq 0} \pi_j f_{W_j}(z_j). \tag{4.3.4}$$

4.3.2 Model Inference

Suppose each $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, independently follows a Type I multivariate zero-truncated hurdle distribution. 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 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

$$\pi_{ij} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}, \tag{4.3.5}$$

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

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

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

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

$$\ell_2(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{j:z_{ij}\neq 0} \log f_{W_j}(z_{ij}) = \sum_{j=1}^{m} \sum_{i:z_{ij}\neq 0} \log f_{W_j}(z_{ij}).$$

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. 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 independence between U and $\boldsymbol{Z}$ indicates the independence between U and $\boldsymbol{Z'}$. 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

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

In addition to the known values $\boldsymbol{z}'_i$, suppose we also know the value u_i from $U_i \sim Bernoulli(\pi_{0i})$. The complete data log-likelihood function then becomes

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

Given initial values of parameters $\boldsymbol{\beta}$, the EM algorithm proceeds as follows.

- E-step: Replace u_i by

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

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

- M-step: Let

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

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

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

- 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 smaller than a tolerance level ε.

Remark 4.3.1. The initial values of parameters $\boldsymbol{\beta}_j$ for EM algorithm can be obtained by implementing univariate logistic regression with observed values $z'_{1j}, \ldots, z'_{nj}$. The standard errors for the estimates can be approximated from the Hessian matrix of the observed log-likelihood function.

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

- E-step: Replace u_i by

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

  Initial value for π_j can be set as the proportion of non-zeros for margin j.

- M-step: Update each π_j through the following equation:

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

4.3.3 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 margins 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 $n = 10,000$.
- Generate $T = 500$ replications of samples.

Using the EM algorithms provided, we can estimate the unknown parameters under different model assumptions. We then calculate the average parameter estimate and the variance of the estimate as follows. For parameter γ, its average parameter estimate is

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

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 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 4.1 and Table 4.2 when data are generated from MZTHP and MZTP models respectively for Example 1. Table 4.3-4.4 summarize the estimation results for Example 2. Average parameter estimates are given in the tables as well as the standard errors 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 also 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 algorithms. 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}, $j = 1, 2, 3$, $k = 0, 1$, though with larger standard errors. This makes sense as the MZTHP model only uses positive numbers to estimate the parameters λ_j or α_{jk}.

To examine the overall goodness-of-fit, we further compare the AIC statistics from the model fitting. Figure 4.1 and Figure 4.2 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 performance of the MZTHP model and the MZTP model has no significant difference.

In short, the above two examples have demonstrated 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 multivariate zero-truncation feature.

4.4 Application

4.4.1 Data Description

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,

Table 4.1: Estimation results of MZTHP-type of data.

	MZTHP model		MZTP model	
True value	estimate	s.e.($\times 10^{-3}$)	estimate	s.e.($\times 10^{-3}$)
$\lambda_1 = 0.5$	0.50	0.64	0.43	0.28
$\lambda_2 = 1.0$	1.00	0.73	0.72	0.46
$\lambda_3 = 1.5$	1.50	0.80	1.09	0.58
$\pi_1 = 0.3$	0.30	0.21		
$\pi_2 = 0.4$	0.40	0.22		
$\pi_3 = 0.5$	0.50	0.25		
Ave. π_0	0.79		0.89	
Ave. AIC	68,808.18		70,873.96	

Table 4.2: Estimation results of MZTP-type of data.

	MZTHP model		MZTP model	
True value	estimate	s.e.($\times 10^{-3}$)	estimate	s.e.($\times 10^{-3}$)
$\lambda_1 = 0.5$	0.50	0.66	0.50	0.30
$\lambda_2 = 1.0$	1.00	0.68	1.00	0.43
$\lambda_3 = 1.5$	1.50	0.71	1.50	0.56
π_1	0.39	0.21		
π_2	0.63	0.24		
π_3	0.78	0.21		
Ave. π_0	0.95		0.95	
Ave. AIC	75,241.75		75,238.55	

Table 4.3: Estimation results of MZTHP-type of data.

	MZTHP model		MZTP model	
True value	estimate	s.e.($\times 10^{-3}$)	estimate	s.e.($\times 10^{-3}$)
$\alpha_{10} = -1$	-1.00	1.43	-1.15	1.04
$\alpha_{11} = 1$	1.00	0.80	0.98	0.75
$\alpha_{20} = 0.5$	0.50	0.59	0.28	0.71
$\alpha_{21} = 0.5$	0.50	0.46	0.53	0.59
$\alpha_{30} = 1$	1.00	0.81	0.79	0.78
$\alpha_{31} = -1$	-1.00	1.42	-0.96	1.02
$\beta_{10} = -1$	-1.00	1.65		
$\beta_{11} = 1$	1.00	1.75		
$\beta_{20} = 0.5$	0.50	1.81		
$\beta_{21} = 0.5$	0.50	2.03		
$\beta_{30} = 1$	1.00	1.74		
$\beta_{31} = -1$	-1.00	1.72		
Ave. AIC	86,225.97		89,409.50	

Table 4.4: Estimation results of MZTP-type of data.

	MZTHP model		MZTP model	
True value	estimate	s.e.($\times 10^{-3}$)	estimate	s.e.($\times 10^{-3}$)
$\alpha_{10} = -1$	-1.00	1.52	-1.00	0.84
$\alpha_{11} = 1$	1.00	0.80	1.00	0.54
$\alpha_{20} = 0.5$	0.50	0.57	0.50	0.48
$\alpha_{21} = 0.5$	0.50	0.46	0.50	0.41
$\alpha_{30} = 1$	1.00	0.72	1.00	0.58
$\alpha_{31} = -1$	-1.00	1.28	-1.00	0.85
β_{10}	-0.95	1.66		
β_{11}	1.58	1.92		
β_{20}	1.33	2.32		
β_{21}	1.39	3.60		
β_{30}	2.26	2.19		
β_{31}	-1.62	2.02		
Ave. AIC	88,787.80		88,751.02	

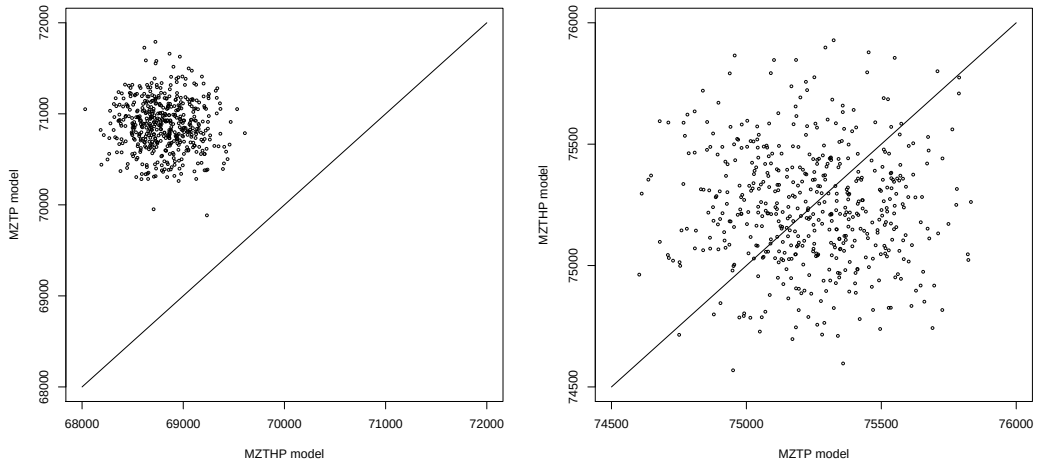

Figure 4.1: Comparison of AIC between the two models without covariates when data are generated from MZTHP model (left) and MZTP model (right) respectively.

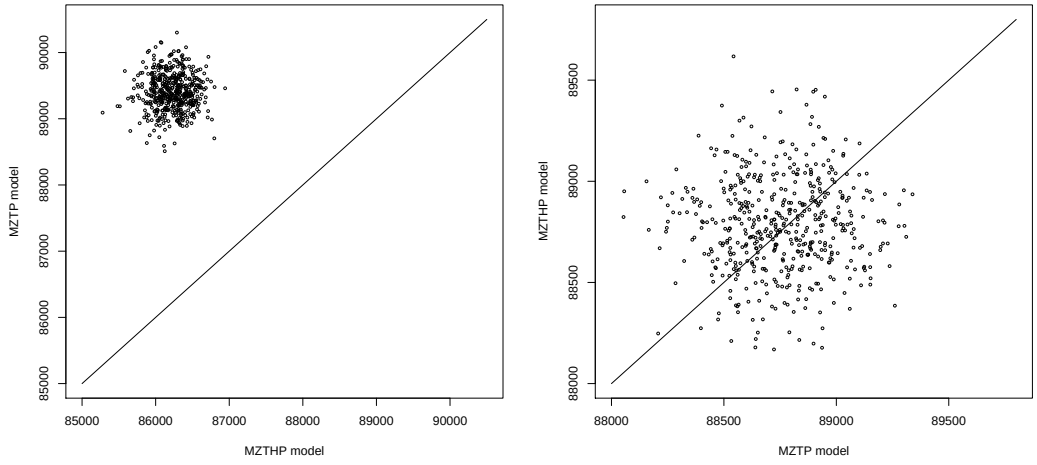

Figure 4.2: Comparison of AIC between the two models with covariates when data are generated from MZTHP model (left) and MZTP model (right) respectively.

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 purposes.

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 4.5. In this table we present the average of total claim counts as well as the average counts 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 frequency than the non-smoker one, which somehow contradicts to our common sense. As for the specified claim type, the youngest age group (0-7) is prone to incurring more claims for disease yet less for accident and other compared with other age groups. Non-smokers have more frequent claims due to disease and less frequent claims from accident compared with 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%. There is an approximately equal number of females and males in the portfolio. It is also noted that the majority of the policyholders are non-smokers.

Table 4.5: Average claim frequencies under each category.

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

4.4.2 Model Analysis

Table 4.6 gives the observed claim counts of each claim type. The empirical distribution for disease claim frequencies 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 correlations among different claim types which is summarized in Table 4.7. This coincides with our model assumption. We start with the no covariate case and use Z_j and W_j, $j = 1, 2, 3$, to represent the claim numbers and positive counts for disease, accident and other. Table 4.8 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 count $W_1 - 1$, the zero-truncated Poisson model (ZTP) for W_2 and the negative binomial (NB) model for the shifted count $W_3 - 1$. These models are selected based on AIC statistics. To further demonstrate the necessity of introducing the hurdle part in our model, we also fit a multivariate zero-truncated (MZT) model with three different margins, namely PIG, Poisson and NB without any modifications for Y_1, Y_2 and Y_3 respectively. The result is also included in Table 4.8. Using AIC and BIC as our main criteria for model selection, one can see that because of the extra flexibility on choices of margins, 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 margins so that both zero and positive parts are fitted more accurately.

Next we add covariates into our MZTH model to see the significance of each explanatory variable. The main results regarding π_j and W_j are summarized in Table 4.9 and Table 4.10 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

Table 4.6: The marginal empirical distributions of three claim types.

	Disease		Accident		Other	
Count	Observed	Percent	Observed	Percent	Observed	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 4.7: Pearson's correlation coefficients among three claim types.

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

Table 4.8: Log-likelihood values, AIC and BIC of four candidate models.

Model	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	57,698.37	57,764.84
MZT	5	-29,093.36	58,196.73	58,238.27

estimation of the parameters drifting to infinity. To address this issue, we discard the involved covariates. The standard errors of $\hat{\boldsymbol{\beta}}_j$ can be approximated using the Hessian matrix from the observed log-likelihood function ℓ_1. The Hessian matrix is given as follows:

$$\frac{\partial^2 \ell_1}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} = \sum_{i=1}^{n} \frac{\pi_{ij}(\pi_{ij} - \pi_{i0})}{\pi_{i0}^2} \boldsymbol{x}_i \boldsymbol{x}_i^\top, \quad \frac{\partial^2 \ell_1}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_{j'}^\top} = \sum_{i=1}^{n} \frac{\pi_{ij}\pi_{ij'}(1 - \pi_{i0})}{\pi_{i0}^2} \boldsymbol{x}_i \boldsymbol{x}_i^\top,$$

for $j, j' = 1, 2, 3$, $j \neq j'$, where $\pi_{i0} = 1 - \prod_{j=1}^{3}(1 - \pi_{ij})$ and $\pi_{ij} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}$.

Table 4.9: Estimates and t-ratios 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	-5.524***	-3.942	-14.218***	-3.910	-14.432***
age1(0-7)	1.138	7.543***	-0.446	-2.905**	-0.370	-2.710**
age2(8-18)	-0.900	-2.171*	-0.586	-1.410	-0.967	-2.342*
age3(19-44)	-0.235	-2.388*	-0.004	-0.044	-0.441	-4.581***
central	0.463	3.272**	0.294	2.083*	0.246	1.767
north	0.322	1.649	0.087	0.442	0.302	1.589
northeast	0.568	3.896***	-0.062	-0.422	0.025	0.179
northwest	0.200	0.553	-0.187	-0.515	-0.140	-0.394
south	0.227	1.048	0.088	0.395	0.455	2.156*
southwest	0.071	0.505	-0.146	-1.016	0.310	2.251*
female	0.347	3.725***	-0.207	-2.201*	0.080	0.880
non-smoker	-0.524	-2.010*	-0.578	-2.215*	-0.448	-1.756
LogLik	-14,387.16					

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

As can be seen from Table 4.9, most covariates are significant 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 claims by disease and accident. For the positive counts, only age and region are significant covariates for modeling the expected number of W_1, yet no covariate plays a significant role

Table 4.10: Estimates and t-ratios 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
ϕ	0.298	18.680***			0.222	1.337
LogLik	-13,021.07		-138.79		-190.50	

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

in modeling the expected count of W_2 and 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 is negligible.

Overall, we preserve all the covariates when modeling π_j. As for the positive counts of disease claims, we use age and region factors as well as exposure in the 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.

4.4.3 Predictive Performance

Making use of the test data, we calculate the predicted claim frequencies for the three claim types using the best model obtained above and then compare with the observed numbers in the data. 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 4.11. 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 4.12. 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 one cell ($Z_1 \geq 2, Z_3 = 1$), 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 frequencies and summarize the results in Table 4.13, 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 result is acceptable considering the number of cells.

Table 4.11: Goodness-of-fit under three univariate cases.

	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[a]	9.32[a]	13[a]	12.54[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 $Z_3 \geq 2$.

4.5 Concluding Remarks

To our best knowledge, few papers have discussed multivariate models with Type I zero-truncation. The model proposed in this chapter, 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 implement two simulation studies to compare the performance of the MZTHP model with the MZTP model. The results confirm the better robustness of the former one. The MZTH model is then applied to a real-life health insurance dataset. The in-sample comparison among four candidate models demonstrates the superiority of the MZTH model. The out-of-sample prediction results also exhibit satisfactory predictive performance of our MZTH model at single dimensions as well as at the multidimensional level.

Table 4.12: Goodness-of-fit under three bivariate cases.

		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 4.13: Goodness-of-fit under the multivariate case.

			$Z_2 = 0$			$Z_2 = 1$		$Z_2 \geq 2$
		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.

Chapter 5

Multivariate Poisson Model with A Misrepresented Covariate Incorporated

5.1 Introduction

In insurance underwriting, misrepresentation represents a scenario where an applicant purposely reports untrue status on particular risk factor(s). Such phenomenon frequently happens in both life and non-life insurance sectors. Typical examples of risk factors with high misrepresentation risks include the smoking status (health insurance) and annual mileage (automobile insurance). However, it is not an easy job to individually investigate the true status of the affected risk factors by the insurance companies due to potentially high costs. This gives rise to the necessity of model adjustments allowing for misrepresentation.

Regarding the misrepresentation problem, the paper of Xia and Gustafson (2016) seems to be the first one to study it from a statistical point of view. Later the misrepresented model was introduced in actuarial setting to study insurance fraud, see Xia (2018) and Akakpo et al. (2019). However, these papers only dealt with the problem of misrepresentation in the univariate case. In practice, an insurance policy can often cover multiple types of claims, highlighting the need of

introducing misrepresentation in multivariate dependence modeling.

One straightforward method of introducing correlations amongst multivariate count variables is through a common additive error. Kocherlakota and Kocherlakota (1992) and Johnson et al. (1997) gave a detailed discussion for the one-factor multivariate Poisson model. However, assuming a common covariance structure in this multivariate Poisson model has several limitations. First, it cannot accommodate flexible pair-wise correlation. Aiming at this restriction, Karlis and Meligkotsidou (2005) proposed an extension which allows for full covariance among Poisson margins. Another drawback is the failure to account for overdispersion. One possible attempt to alleviate this limitation is to construct the model based on negative binomial margins (Winkelmann (2000)). In this chapter, we consider the misrepresentation problem based on the one-factor multivariate Poisson model. Specifically, a misrepresented binary risk factor is incorporated in each margin of the multivariate Poisson. The resultant model then could address the above two issues simultaneously. As for inference, we implement the EM algorithm proposed by Dempster et al. (1977). The lack of no closed-form representation in the M-step motivates us to use a generalized EM algorithm (see Rai and Matthews (1993)), where the Newton-Raphson method in the M-step is carried out for only one step. The EM algorithm is working more efficiently by this technique.

The rest of the chapter is organized as follows. In Section 5.2, we develop our multivariate Poisson model that accounts for misrepresentation. The corresponding EM algorithm is provided for inference. The impact of misrepresentation on model estimates is highlighted in Section 5.3 through two simulation studies. In section 5.4, we present the application of the model to the data from Australian Health Survey. Our conclusions are discussed in Section 5.5.

5.2 The Model

5.2.1 Multivariate Poisson Model

For readers' convenience, we restate the definition of the multivariate Poisson distribution with common covariance. It can be defined as follows. Let

$$Z_j = Y_j + Y_0, \quad j = 1, \ldots, m, \tag{5.2.1}$$

where each Y_j, $j = 0, \ldots, m$, independently follows a simple Poisson distribution with parameter λ_j. Then $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^\top$ is said to follow the multivariate Poisson distribution. The joint pmf of $\boldsymbol{Z}$ is given by

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \exp(-\lambda') \sum_{l=0}^{s} \left[\frac{\lambda_0^l}{l!} \prod_{j=1}^{m} \frac{\lambda_j^{z_j - l}}{(z_j - l)!} \right], \tag{5.2.2}$$

where $\boldsymbol{z} = (z_1, \ldots, z_m)^\top$, $\lambda' = \sum_{j=0}^{m} \lambda_j$ and $s = \min(z_1, \ldots, z_m)$. We may allow for covariates $\boldsymbol{x} = (1, x_1, \ldots, x_p)^\top$ by considering that

$$\lambda_j = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_j), \quad j = 1, \ldots, m, \tag{5.2.3}$$

where $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})^\top$ denotes the corresponding vector of regression coefficients. For the purpose of easy interpretation, we do not inject covariates in λ_0.

5.2.2 Misrepresentation

As motivated at the beginning of this chapter, now we further introduce an unobserved binary risk factor v that is subject to misrepresentation. For this predictor, we assume $v = 0$ results in the outcome that is favorable to the individual policyholder, whereas $v = 1$ leads to an unfavorable result. With a log-link function, the whole covariate set is formulated as

$$\lambda_j = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_j + \beta_{jv} v), \quad j = 1, \ldots, m, \tag{5.2.4}$$

where β_{jv} is the regression coefficient for the risk factor v.

Due to the misstatement on this risk factor, we denote v^* to be the observed and reported status. Then under the assumption of unidirectional misrepresentation, we have

$$\Pr(v = 1|v^* = 1) = 1, \quad \Pr(v = 1|v^* = 0) = q, \quad 0 < q < 1,$$

where q is referred to as the prevalence of misrepresentation (see Akakpo et al. (2019)). It is assumed that once the true status v is known, the response variable Y_j is no longer dependent on the reported status v^*. The unidirectional property of the underlying factor ensures the identifiability of the model. The multivariate Poisson model with misrepresentation considered above becomes

$$f_{\boldsymbol{Z}}(\boldsymbol{z}|\boldsymbol{x}, v^*) = \left[f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z})\right]^{v^*} \left[qf_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}) + (1-q)f_{\boldsymbol{Z}}^{(0)}(\boldsymbol{z})\right]^{1-v^*}, \tag{5.2.5}$$

where

$$f_{\boldsymbol{Z}}^{(r)}(\boldsymbol{z}) = f_{\boldsymbol{Z}}(\boldsymbol{z}|\boldsymbol{x}, v = r), \quad r = 0, 1.$$

The properties of the model are given as follows. When $v^* = 0$,

- The margin Z_j follows a 2-mixture Poisson distribution with parameters $\lambda_j^{(1)} + \lambda_0$, $\lambda_j^{(0)} + \lambda_0$ and mixing proportion q, where $\lambda_j^{(1)} = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_j + \beta_{jv})$ and $\lambda_j^{(0)} = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_j)$.

- The mean of Z_j is

$$\mathrm{E}(Z_j) = q\lambda_j^{(1)} + (1-q)\lambda_j^{(0)} + \lambda_0.$$

- The variance of Z_j is

$$\mathrm{Var}(Z_j) = q\lambda_j^{(1)} + (1-q)\lambda_j^{(0)} + \lambda_0 + q(1-q)\left[\lambda_j^{(1)} - \lambda_j^{(0)}\right]^2.$$

  So it can account for overdispersion.

- The covariance between Z_j and $Z_{j'}$, $j, j' = 1, \ldots, m$, $j \neq j'$, is

$$\mathrm{Cov}(Z_j, Z_{j'}) = \lambda_0 + q(1-q)\left[\lambda_j^{(1)} - \lambda_j^{(0)}\right]\left[\lambda_{j'}^{(1)} - \lambda_{j'}^{(0)}\right].$$

  So

$$\mathrm{Cov}(Z_j, Z_{j'}) \begin{cases} > \lambda_0, & \beta_{jv}\beta_{j'v} > 0, \\ = \lambda_0, & \beta_{jv}\beta_{j'v} = 0, \\ < \lambda_0, & \beta_{jv}\beta_{j'v} < 0. \end{cases}$$

When $v^* = 1$,

- The margin Z_j follows a univariate Poisson distribution with parameter $\lambda_j^{(1)} + \lambda_0$.

- The mean of Z_j is

$$\mathrm{E}(Z_j) = \lambda_j^{(1)} + \lambda_0.$$

- The variance of Z_j is

$$\mathrm{Var}(Z_j) = \lambda_j^{(1)} + \lambda_0.$$

- The covariance between Z_j and $Z_{j'}$, $j, j' = 1, \ldots, m$, $j \neq j'$, is

$$\mathrm{Cov}(Z_j, Z_{j'}) = \lambda_0$$

5.2.3 Model Inference

Suppose we have a random sample of size n. The corresponding observed values are $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})^\top$, $i = 1, \ldots, n$. Now we introduce some covariates, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$, and observed binary risk factors $v_1^* \ldots, v_n^*$. To further explore the potential effects of predictors on the prevalence parameter q_i, we shall allow for the incorporation of covariates in q_i via a logit-link function:

$$q_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha})}, \tag{5.2.6}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_p)^\top$. The likelihood function can be expressed as

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \left[f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}_i) \right]^{v_i^*} \left[q_i f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}_i) + (1-q_i) f_{\boldsymbol{Z}}^{(0)}(\boldsymbol{z}_i) \right]^{1-v_i^*}, \tag{5.2.7}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \beta_{1v}, \ldots, \beta_{mv}, \lambda_0, \boldsymbol{\alpha})$ is the parameter set to estimate. The log-likelihood function is

$$\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{n} v_i^* \log f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}_i) + \sum_{i=1}^{n} (1-v_i^*) \log \left[q_i f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}_i) + (1-q_i) f_{\boldsymbol{Z}}^{(0)}(\boldsymbol{z}_i) \right].$$

The last summation contains a mixture of two components, which complicates the maximization problem. We implement EM algorithm by introducing u_i as the unobserved binary indicator on whether observation i is misrepresented. Furthermore, we assume the value of Y_0 is known. The complete data likelihood function then becomes

$$\begin{aligned} L^c(\boldsymbol{\Theta}) = & \prod_{i=1}^{n} \left[f_{Y_0}(y_{i0}) \prod_{j=1}^{m} f_{Y_j}^{(1)}(y_{ij}) \right]^{v_i^*} \prod_{i=1}^{n} \left[q_i f_{Y_0}(y_{i0}) \prod_{j=1}^{m} f_{Y_j}^{(1)}(y_{ij}) \right]^{u_i(1-v_i^*)} \\ & \times \prod_{i=1}^{n} \left[(1-q_i) f_{Y_0}(y'_{i0}) \prod_{j=1}^{m} f_{Y_j}^{(0)}(y'_{ij}) \right]^{(1-u_i)(1-v_i^*)}. \end{aligned}$$

where

$$f_{Y_j}^{(r)}(y_{ij}) = e^{-\lambda_{ij}^{(r)}} \frac{\left[\lambda_{ij}^{(r)} \right]^{y_{ij}}}{y_{ij}!}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \quad r = 0, 1,$$

with $\lambda_{ij}^{(1)} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \beta_{jv})$, $\lambda_{ij}^{(0)} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$, $y_{ij} = z_{ij} - y_{i0}$ and $y'_{ij} = z_{ij} - y'_{i0}$. The complete data log-likelihood function is then written as

$$\begin{aligned} \ell^c(\boldsymbol{\Theta}) = & \sum_{i=1}^{n} \left[l_i^{(1)} - v_i^* \right] \log q_i + \sum_{i=1}^{n} l_i^{(0)} \log(1-q_i) \\ & + \sum_{i=1}^{n} l_i^{(1)} (-\lambda_0 + y_{i0} \log \lambda_0) + \sum_{i=1}^{n} l_i^{(0)} (-\lambda_0 + y'_{i0} \log \lambda_0) \\ & + \sum_{j=1}^{m} \left\{ \sum_{i=1}^{n} l_i^{(1)} \left[-\lambda_{ij}^{(1)} + y_{ij} \log \lambda_{ij}^{(1)} \right] + \sum_{i=1}^{n} l_i^{(0)} \left[-\lambda_{ij}^{(0)} + y'_{ij} \log \lambda_{ij}^{(0)} \right] \right\} + C, \end{aligned}$$

where $l_i^{(1)} = v_i^* + u_i(1 - v_i^*)$, $l_i^{(0)} = (1 - u_i)(1 - v_i^*)$, C is a constant not related to $\boldsymbol{\Theta}$. The corresponding EM algorithm is illustrated as follows:

- E-step:

- Replace u_i by

$$\bar{u}_i = \frac{q_i f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}_i)}{q_i f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}_i) + (1 - q_i) f_{\boldsymbol{Z}}^{(0)}(\boldsymbol{z}_i)}, \quad i = 1, \ldots, n.$$

- Replace y_{i0} by

$$\bar{y}_{i0} = \begin{cases} \dfrac{\lambda_0 f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}_i - \mathbf{1}_m)}{f_{\boldsymbol{Z}}^{(1)}(\boldsymbol{z}_i)}, & \min(z_{i1}, \ldots, z_{im}) > 0, \\ 0, & \min(z_{i1}, \ldots, z_{im}) = 0, \end{cases}$$

  where $\mathbf{1}_m = (1, \ldots, 1)^\top$ denotes the vector of dimension m with all elements equal to 1.

- Replace y'_{i0} by

$$\bar{y}'_{i0} = \begin{cases} \dfrac{\lambda_0 f_{\boldsymbol{Z}}^{(0)}(\boldsymbol{z}_i - \mathbf{1}_m)}{f_{\boldsymbol{Z}}^{(0)}(\boldsymbol{z}_i)}, & \min(z_{i1}, \ldots, z_{im}) > 0, \\ 0, & \min(z_{i1}, \ldots, z_{im}) = 0. \end{cases}$$

- M-step: The expectation of complete data log-likelihood function $\bar{\ell}^c(\boldsymbol{\Theta})$ can then be decomposed into several parts:

$$\begin{aligned} \bar{\ell}_q^c(\boldsymbol{\alpha}) &= \sum_{i=1}^n \left[\bar{l}_i^{(1)} - v_i^*\right] \log q_i + \sum_{i=1}^n \bar{l}_i^{(0)} \log(1 - q_i), \\ \bar{\ell}_{y_0}^c(\lambda_0) &= \sum_{i=1}^n \bar{l}_i^{(1)}(-\lambda_0 + \bar{y}_{i0} \log \lambda_0) + \sum_{i=1}^n \bar{l}_i^{(0)}(-\lambda_0 + \bar{y}'_{i0} \log \lambda_0), \\ \bar{\ell}_{y_j}^c(\boldsymbol{\beta}_j, \beta_{jv}) &= \sum_{i=1}^n \bar{l}_i^{(1)} \left[-\lambda_{ij}^{(1)} + \bar{y}_{ij} \log \lambda_{ij}^{(1)}\right] + \sum_{i=1}^n \bar{l}_i^{(0)} \left[-\lambda_{ij}^{(0)} + \bar{y}'_{ij} \log \lambda_{ij}^{(0)}\right], \end{aligned}$$

  where $\bar{l}_i^{(1)} = v_i^* + \bar{u}_i(1 - v_i^*)$, $\bar{l}_i^{(0)} = (1 - \bar{u}_i)(1 - v_i^*)$, $\bar{y}_{ij} = z_{ij} - \bar{y}_{i0}$, $\bar{y}'_{ij} = z_{ij} - \bar{y}'_{i0}$. The maximization step can thus be accomplished separately.

- For $\bar{\ell}_q^c(\boldsymbol{\alpha})$, there is no closed-form representation for $\boldsymbol{\alpha}$, so we update the regression parameters by implementing Newton-Raphson method for one step. The first and second order derivatives are given as follows:

$$\frac{\partial \bar{\ell}_q^c}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} (1 - v_i^*)(\bar{u}_i - q_i)\boldsymbol{x}_i,$$
$$\frac{\partial^2 \bar{\ell}_q^c}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top} = -\sum_{i=1}^{n} (1 - v_i^*) q_i (1 - q_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

Remark 5.2.1. If no covariates are incorporated in the form of q, there exists a closed-form for each individual in the M-step:

$$q = \frac{\sum_{i=1}^{n} (1 - v_i^*)\bar{u}_i}{\sum_{i=1}^{n} (1 - v_i^*)}.$$

- For $\bar{\ell}_{y_0}^c(\lambda_0)$,

$$\lambda_0 = \frac{1}{n} \sum_{i=1}^{n} \left[\bar{l}_i^{(1)} \bar{y}_{i0} + \bar{l}_i^{(0)} \bar{y}'_{i0} \right].$$

- For $\bar{\ell}_{y_j}^c(\boldsymbol{\beta}_j, \beta_{jv})$, there is no closed-form representation for $(\boldsymbol{\beta}_j, \beta_{jv})$, 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:

$$\frac{\partial \bar{\ell}_{y_j}^c}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n} \left[\bar{l}_i^{(1)} \bar{y}_{ij} + \bar{l}_i^{(0)} \bar{y}'_{ij} - \bar{l}_i^{(1)} \lambda_{ij}^{(1)} - \bar{l}_i^{(0)} \lambda_{ij}^{(0)} \right] \boldsymbol{x}_i,$$
$$\frac{\partial \bar{\ell}_{y_j}^c}{\partial \beta_{jv}} = \sum_{i=1}^{n} \left[\bar{l}_i^{(1)} \bar{y}_{ij} - \bar{l}_i^{(1)} \lambda_{ij}^{(1)} \right],$$
$$\frac{\partial^2 \bar{\ell}_{y_j}^c}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top} = -\sum_{i=1}^{n} \left[\bar{l}_i^{(1)} \lambda_{ij}^{(1)} + \bar{l}_i^{(0)} \lambda_{ij}^{(0)} \right] \boldsymbol{x}_i \boldsymbol{x}_i^\top,$$
$$\frac{\partial^2 \bar{\ell}_{y_j}^c}{\partial \beta_{jv}^2} = -\sum_{i=1}^{n} \bar{l}_i^{(1)} \lambda_{ij}^{(1)},$$
$$\frac{\partial^2 \bar{\ell}_{y_j}^c}{\partial \boldsymbol{\beta}_j \partial \beta_{jv}} = -\sum_{i=1}^{n} \bar{l}_i^{(1)} \lambda_{ij}^{(1)} \boldsymbol{x}_i.$$

We can 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 smaller than a tolerance level ε.

Remark 5.2.2. The initial values of parameters $\boldsymbol{\beta}_j$ and β_{jv} for EM algorithm can be obtained by implementing univariate Poisson regression with observed values $z_{1j}, \ldots, z_{nj}$. The initial values of parameters $\boldsymbol{\alpha}$ can be set as $\mathbf{0}$. The initial value of parameter λ_0 can be set as 1. The standard errors for the estimates can be approximated via standard bootstrap method.

5.3 Simulation

In this section, two simulation studies are carried out. The first example studies the effect of the prevalence parameter q on other estimates under the multivariate Poisson models with and without considering misrepresentation, that are referred to as adjusted and unadjusted models respectively. The second study aims to evaluate the performance of adjusted and unadjusted models when a single predictor is also embedded in the form of q.

5.3.1 Simulation 1

For illustration, we set the sample size $n = 10,000$. Without loss of generality, we set $m = 2$. A single covariate x_i is generated from a standard normal distribution. The data generation process from the model with misrepresentation is demonstrated as follows.

- v_i^* is generated from $Bernoulli(0.1)$.
- For the individual with $v_i^* = 1$, $v_i = 1$. If $v_i^* = 0$, then v_i is generated from $Bernoulli(q)$.

- y_{i0} is generated from the Poisson model with $\lambda_0 = 1$; y_{i1} is generated from the Poisson model with $\lambda_{i1} = e^{-1+x_i-2v_i}$; y_{i2} is generated from the Poisson model with $\lambda_{i2} = e^{-2-x_i+v_i}$.

- Set $z_{i1} = y_{i1} + y_{i0}$ and $z_{i2} = y_{i2} + y_{i0}$.

A total of 99 equally spaced values of q are considered in the interval $(0, 1)$. Then the adjusted and unadjusted multivariate Poisson models are fitted to the generated data. The corresponding plots are displayed in Figure 5.1 and Figure 5.2.

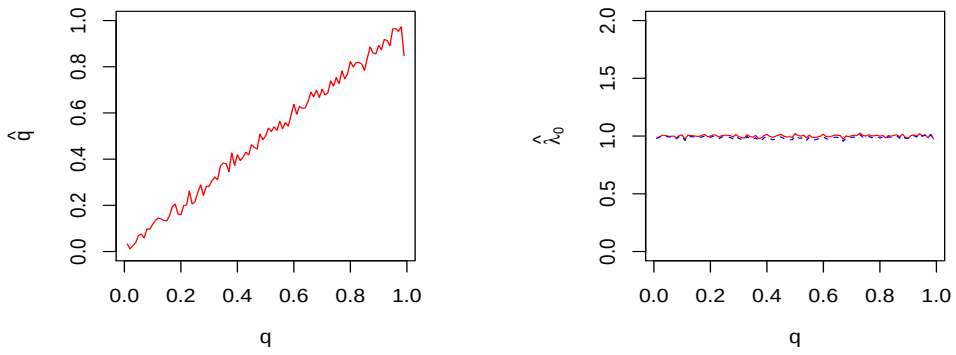

Figure 5.1: Effect of the prevalence parameter q on the estimates of q and λ_0. Red solid curves represent the adjusted model and blue dotted curves represent the unadjusted model.

Several interesting observations can be made from the figures. First, the adjusted and unadjusted models are consistent when $q = 0$. Second, the parameter estimates under the adjusted model are consistent with the true values for majority of the explored values of q. However, the parameter estimates for β_{10}, β_{1v}, β_{20}, β_{2v} become more volatile under the adjusted model when q approaches 1. A possible explanation is when q is getting closer to 1, the implied dummy variable v has a greater proportion of being 1, making this predictor less differentiable from the

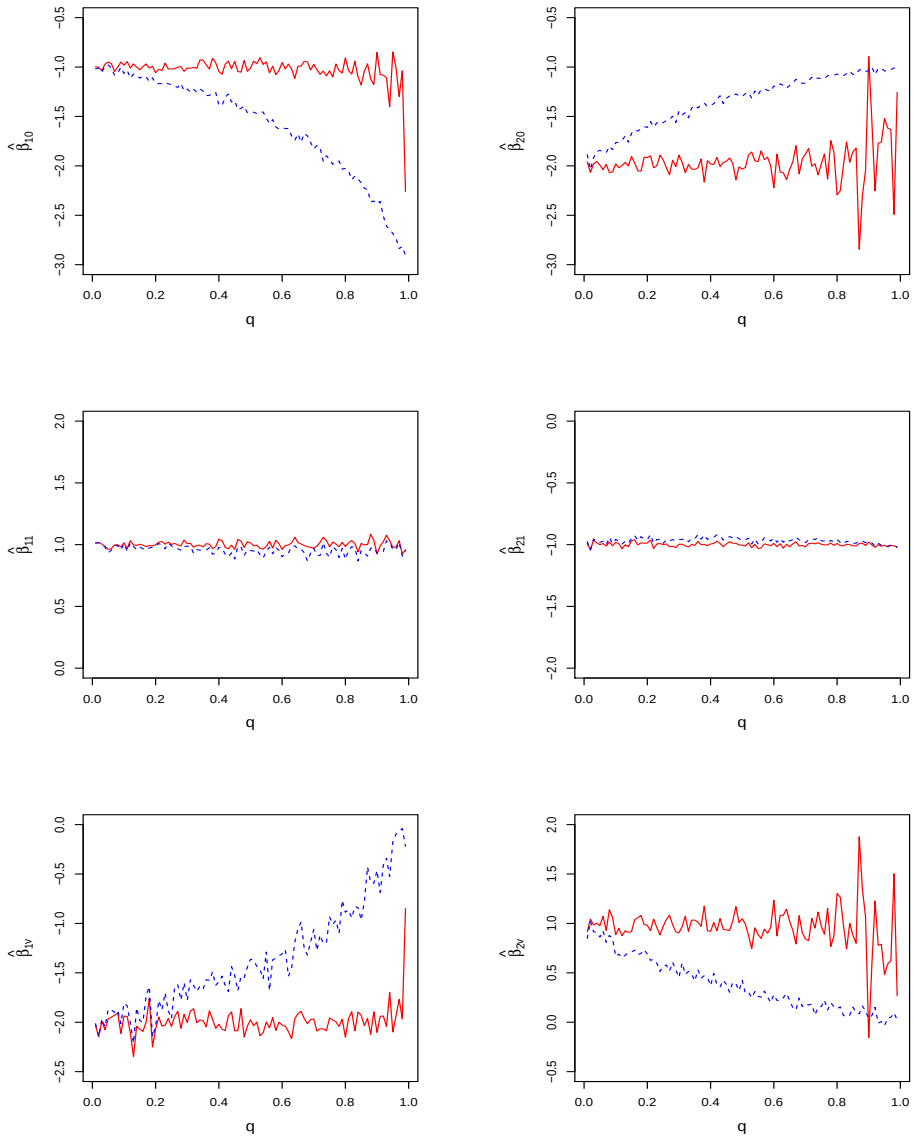

Figure 5.2: Effect of the prevalence parameter q on the estimates of β_{10}, β_{11}, β_{1v} and β_{20}, β_{21}, β_{2v}. Red solid curves represent the adjusted model and blue dotted curves represent the unadjusted model.

intercept. It would therefore result in higher variance in the estimates of β_{10}, β_{1v}, β_{20} and β_{2v}. In the extreme case of $q = 1$, this implied dummy variable v is completely non-differentiable from the intercept, causing an identifiability problem in the adjusted model. Third, estimates of λ_0, β_{11} and β_{21} can be correctly captured regardless of model selection under this setting, whilst other parameter estimates suffer in the unadjusted model. The estimates of those parameters become increasingly biased as q gets larger. When $q = 1$, the estimates of β_{1v} and β_{2v} are both roughly equal to 0 in the unadjusted model, indicating v^* becomes an irrelevant predictor in this model. Interestingly, the estimates of $\beta_{10} + \beta_{1v}$ and $\beta_{20} + \beta_{2v}$ are barely affected by the underlying q, which are both approximately equal to the true values.

5.3.2 Simulation 2

In this simulation, we further introduce a single predictor x_i in the form of q_i, namely $q_i = \frac{\exp(-1+x_{1i})}{1+\exp(-1+x_{1i})}$. The setting for other parameters is same as the one in Simulation 1. The estimation results based on 200 replications are provided in Table 5.1. Mean estimates with their standard errors under the adjusted and unadjusted models are given.

As expected, the estimates for all parameters are consistent with the true values when the model is correctly specified. However, we observe that when the same set of covariates is embedded in the form of q_i, all estimates suffer bias in the unadjusted model, yet the influence on λ_0 is only marginal.

5.4 Application

5.4.1 Data Description

This application showcases a real-life application based on the data quantifying the demand for health care in Australia that were collected from the Australian

Table 5.1: Estimation results of data generated from the adjusted model.

	adjusted model		unadjusted model	
parameter	estimate	s.e.($\times 10^{-3}$)	estimate	s.e.($\times 10^{-3}$)
$\alpha_0 = -1$	-1.00	8.68		
$\alpha_1 = 1$	1.00	6.38		
$\lambda_0 = 1$	1.00	0.76	0.98	0.76
$\beta_{10} = -1$	-1.00	2.46	-1.23	1.90
$\beta_{11} = 1$	1.00	2.01	0.66	1.95
$\beta_{1v} = -2$	-2.01	5.94	-1.37	9.54
$\beta_{20} = -2$	-2.01	3.98	-1.56	2.19
$\beta_{21} = -1$	-1.00	1.48	-0.84	1.52
$\beta_{2v} = 1$	1.01	4.39	0.73	3.28

Health Survey 1977-1978. The data were also used in Cameron and Trivedi (1998) to study a variety of count models. The whole dataset consists of 5,190 individuals. The selected covariates with detailed description are summarized in Table 5.2. A self-assessment health index was also obtained from the questionnaire, where a high score indicates bad health. We construct a binary factor HSCORE to indicate whether the score is higher than 5 or not. In our analysis, we assume the policyholders underestimate the health index for some purposes, e.g., to qualify for lower premiums in health insurance. This rationalizes the assumption of unidirectional misrepresentation. We make use of two response variables in the dataset, where Z_1 represents the number of consultations with a doctor or a non-doctor health professional (chemist, optician, physiotherapist, social worker, district community nurse, chiropodist or chiropractor) in the past two weeks, and Z_2 denotes the total number of prescribed and non-prescribed medications used in past 2 days. The corresponding empirical joint distribution for Z_1 and Z_2 is presented in Table 5.3

Table 5.2: Summary of selected explanatory variables.

Variables	Description
SEX	1 if female, 0 if male
AGE	Age in years divided by 100
INCOME	Annual income in Australian dollars divided by 1000
CHCOND1	1 if chronic condition(s) but not limited in activity, 0 otherwise
CHCOND2	1 if chronic condition(s) and limited in activity, 0 otherwise
HSCORE	1 if general health questionnaire score greater than 5, 0 otherwise

Table 5.3: Empirical joint distribution of Z_1 and Z_2.

Z_1	Z_2								
	0	1	2	3	4	5	6	7	8
0	1,910	1,072	442	229	95	35	19	14	10
1	212	218	177	100	67	41	24	10	17
2	70	56	55	31	25	14	9	5	6
3	11	17	12	11	6	4	2	0	0
4	6	11	12	6	6	2	2	1	0
5	3	2	3	2	2	2	4	0	1
6	4	0	3	3	4	2	0	1	1
7	4	8	6	6	4	2	0	0	3
8	5	2	5	2	3	0	1	0	0
≥ 9	4	3	8	3	6	1	2	1	2

5.4.2 Model Fitting

We fit the multivariate (bivariate in this case) Poisson models with and without considering misrepresentation to the data described previously. In the first adjusted model we shall not incorporate any covariates in modeling q. The main estimation results are given in Table 5.4 and Table 5.5. Standard errors are calculated based on 1,000 bootstrap replications. We examine the model fitting in the light of the log-likelihood value as well as commonly used information criteria including AIC and BIC. All statistics demonstrate the superiority of adjusted model over the unadjusted one. The estimated prevalence parameter q within this bivariate adjusted Poisson model is 0.181, indicating a chance of 18.1% of misreporting for any respondent claiming a health score no more than 5. The dependence parameter λ_0 in the unadjusted model is 0.062 with a 95% confidence interval of (0.046, 0.076). It indicates the significance of the covariance term in this model. However, λ_0 is not significant in our adjusted model, suggesting that we could possibly start from two independent single Poisson models when misrepresentation is incorporated.

We then turn to the effect of each predictor on Z_1 and Z_2. Our findings from Table 5.4 and Table 5.5 are generally consistent. Females are more frequent to visit doctors or non-doctors and take medicines than males. The usage of health care and medications increases with age. Annual income is a good indicator for predicting the frequency of visits, yet it has no association with the intake of medications. As anticipated, chronic conditions are significant predictors for both response variables. Intuitively, the health score can reveal the risk level of respondents as observed from the tables. It is worth noting that the estimates of β_{1v} and β_{2v} get inflated in the adjusted model compared with the unadjusted one.

Next, we would like to explore the potential effect of predictors on the prevalence parameter. The corresponding results are exhibited in Table 5.6. We observe

Table 5.4: Estimation results for the unadjusted bivariate Poisson model.

	λ_1		λ_2	
	Estimate	t-ratio	Estimate	t-ratio
Intercept	-1.930	-21.160***	-1.387	-24.001***
SEX	0.267	5.381***	0.447	14.816***
AGE	1.233	9.972***	1.424	18.595***
INCOME	-0.251	-3.574***	0.016	0.397
CHCOND1	0.530	9.134***	0.780	21.214***
CHCOND2	1.225	18.856***	1.079	25.340***
HSCORE	0.794	13.313***	0.423	10.086***
	Estimate	95% CI		
λ_0	0.062	(0.046, 0.076)		
LogLik		-12,666.51		
AIC		25,363.02		
BIC		25,461.34		

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

some slight differences in estimates compared with the previous adjusted model. Gender is no longer a significant predictor on Z_1 when covariates are incorporated in q. It can be concluded from the table that females and people with chronic conditions are more likely to misstate their health scores. Other explanatory variables seem to be irrelevant on predicting q.

For this case study, both AIC and BIC reveal the evidence of misrepresentation in this particular dataset. This phenomenon more frequently occurs in the groups of females and people with chronic conditions. It is crucial for the insurance companies to recognize the presence of misrepresentation and its potential impact in order to better manage relevant risks.

Table 5.5: Estimation results for the adjusted bivariate Poisson model without covariates incorporated in q.

	λ_1		λ_2	
	Estimate	t-ratio	Estimate	t-ratio
Intercept	-2.762	-23.383***	-1.448	-24.976***
SEX	0.192	3.616***	0.420	13.609***
AGE	1.213	8.912***	1.384	17.851***
INCOME	-0.195	-2.503*	0.015	0.358
CHCOND1	0.392	6.179***	0.726	20.574***
CHCOND2	0.836	11.886***	0.967	23.501***
HSCORE	2.364	35.579***	0.722	21.571***
	Estimate	95% CI		
q	0.181	(0.157, 0.206)		
λ_0	0.012	(0.000, 0.026)		
LogLik		-11,931.72		
AIC		23,895.43		
BIC		24,000.30		

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

Table 5.6: Estimation results for the adjusted bivariate Poisson model with covariates incorporated in q.

	λ_1		λ_2		q	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
Intercept	-2.629	-18.460***	-1.405	-23.498***	-2.189	-6.933***
SEX	0.042	0.646	0.377	11.172***	0.541	3.234**
AGE	1.300	7.676***	1.411	17.606***	-0.270	-0.687
INCOME	-0.245	-2.433*	0.000	0.000	0.216	0.929
CHCOND1	0.242	2.659**	0.691	17.266***	0.422	2.053*
CHCOND2	0.610	6.490***	0.904	19.806***	0.922	4.107***
HSCORE	2.447	34.964***	0.732	22.306***		
	Estimate	95% CI				
λ_0	0.017	(0.003, 0.031)				
LogLik		-11,914.40				
AIC		23,870.80				
BIC		24,008.44				

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

5.5 Concluding Remarks

In this chapter, we introduce a misrepresented covariate in the multivariate Poisson model. Furthermore, predictors are included in the form of prevalence parameter to detect the inclination of being dishonest for specific groups. The incorporation of misrepresentation in the count model helps to deal with more complicated covariance structure. On the other hand, the restriction of equidispersion in the Poisson model could be alleviated to some extent. Through two simulation studies, we conclude that failure to accommodate the existence of misrepresentation often leads to biased estimates. The application part uses a dataset from Australian Health Survey by treating the health score as the misrepresented covariate. All goodness-of-fit statistics agree that the adjusted model outperforms the unadjusted one.

Chapter 6

Concluding Remarks and Further Research

In this thesis, we investigated several multivariate count regression models with dependence. The work was motivated by the characteristics of several real datasets.

From the perspective of dependence modeling, we considered a multivariate mixed Poisson model with copula constructed on continuous mixing parameters in Chapter 2. This formulation allowed for rich dependence structures and different marginal selections. The superiority of our copula model was supported by a real dataset form Australian Health Survey when compared with several alternatives.

When the dimension is greater than two, we only implemented the Gaussian copula on mixture as a typical example in our model. One limitation of Gaussian copula is its inability to handle tail dependence. A possible solution is to use other types of copulas on mixture to accommodate a wider variety of dependence structures.

Analogous to the univariate setting, data might also exhibit zero-inflation or zero-truncation feature in the multivariate case. Chapter 3 and Chapter 4 aimed to address these two issues respectively. Both models started with independent hurdle margins. A Bernoulli variable was then applied to add or truncate common zeros. The assumption on that each margin followed a hurdle distribution per-

mitted high freedom in dealing with various zero features and different marginal selections. The interpretation of the two hurdle models was illustrated in the context of two real datasets from automobile insurance and health insurance respectively. The satisfactory performance of our proposed models was confirmed by several goodness-of-fit statistics.

Assuming independent margins in these two models can easily handle high dimensional data without any computational issues. However, it fails to cope with more complicated association structures. It therefore highlights the need to employ copula to deal with a wider range of dependence cases in future research.

In the process of insurance underwriting, policyholders have a tendency to misreport their status on particular risk factor(s) in their best interest. This misrepresentation problem in insurance was explored in Chapter 5. We incorporated a binary misrepresented risk factor in the multivariate Poisson model. This formulation permitted more flexible dependence relations. As illustrated in the given simulation studies, failure to take misrepresentation into consideration would often lead to biased estimates. Our adjusted model was compared to the unadjusted one using the data from the same source as in Chapter 2.

Chapter 5 leads to several possibilities for future research. First, the unobserved feature of misrepresentation can also be shared by the regression models on the claim frequency and severity. This can relax the independence assumption between the two parts in traditional sense. Second, the binary misrepresented covariate can be generalized to an ordinal one. Third, we must admit that the prevalence parameter is subject to the model selection. It would be interesting to explore the effect of model selection on misrepresentation.

Overall, this thesis demonstrated how to carry out inferential procedures under both the Bayesian and the frequentist framework. In Chapter 2, the inference was completed in the context of Bayesian statistics by virtue of MCMC. We could

easily derive posterior summaries for our interested quantities to account for the uncertainty. In Chapter 3, Chapter 4 and Chapter 5, we adopted the GEM algorithm within the frequentist setting. Different techniques were applied to find the standard errors of estimates in these three chapters.

Throughout this thesis we adopted *a priori* ratemaking mechanism based on cross-sectional data. The underlying assumption is that all the factors influencing a risk could be identified, measured and introduced in the tariff system. The classified groups are then assumed to be homogeneous. However, this is not always realistic as there may exist some risk factors that are not taken into account in the *a priori* structure. For example, in motor vehicle insurance, a driver's aggressiveness is difficult to quantify. In health insurance, it is also hard to rate the impact of environmental factors on policyholder's health. As a result, tariff groups can still be quite heterogeneous. This gives rise to *a posteriori* mechanism, which sets an individual premium based on the past claim history of each policyholder. In that respect, the longitudinal models can be applied to deal with the involved panel data. In addition to the dependence among different types of claims, we should also take serial dependence or time dependence into account in relevant models.

Appendix A

EM Algorithms for Two Multivariate Zero-Inflated 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 $\lambda_{ij} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$ 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

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

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

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

  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. The first and second order derivatives are given as follows:

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

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

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, $\boldsymbol{Z}_i \sim MZINB(\boldsymbol{\lambda}_i, \boldsymbol{\phi}, \pi_0)$, $i = 1, \ldots, n$, where $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{im})^\top$. Similarly, the covariates could be introduced as $\lambda_{ij} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta_j})$. The likelihood function becomes

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\phi}, \pi_0) = &\prod_{i=1}^{n}\left[1-\pi_0+\pi_0\prod_{j=1}^{m}\left(\frac{\phi_j}{\lambda_{ij}+\phi_j}\right)^{\phi_j}\right]^{v_i} \\ &\times\prod_{i=1}^{n}\left[\pi_0\prod_{j=1}^{m}\frac{\Gamma(z_{ij}+\phi_j)}{\Gamma(\phi_j)z_{ij}!}\left(\frac{\lambda_{ij}}{\lambda_{ij}+\phi_j}\right)^{z_{ij}}\left(\frac{\phi_j}{\lambda_{ij}+\phi_j}\right)^{\phi_j}\right]^{1-v_i}. \end{aligned}$$

Following the same data augmentation method as for the multivariate zero-

inflated hurdle model, we obtain the complete data likelihood function given as

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

The complete data log-likelihood function is

$$\ell^c(\boldsymbol{\beta}, \boldsymbol{\phi}, \pi_0) = \sum_{i=1}^{n} \left[u_i' v_i \log(1-\pi_0) + (1 - u_i' v_i) \log \pi_0 \right]$$
$$+ \sum_{j=1}^{m} \sum_{i=1}^{n} \{ (1 - u_i' v_i) \phi_j \log \phi_j - [(1 - u_i' v_i)\phi_j + z_{ij}] \log(\lambda_{ij} + \phi_j)$$
$$+ z_{ij} \log \lambda_{ij} + \log \left(\Gamma(z_{ij} + \phi_j) \right) - \log \left(\Gamma(\phi_j) \right) \} + C_1,$$

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

The EM algorithm can be described as follows.

- E-step: Replace u_i by

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

  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 $(\boldsymbol{\beta}, \boldsymbol{\phi})$, let

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

There is no closed-form representation for $\boldsymbol{\beta}_j$ and ϕ_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:

$$\begin{aligned}
\frac{\partial \bar{\ell}_j^C}{\partial \boldsymbol{\beta_j}} &= \sum_{i=1}^{n} \frac{[z_{ij} - (1 - \bar{u}_i' v_i)\lambda_{ij}]\,\phi_j}{\lambda_{ij} + \phi_j} \boldsymbol{x_i}, \\
\frac{\partial \bar{\ell}_j^C}{\partial \phi_j} &= \sum_{i=1}^{n} \Bigg[(1 - \bar{u}_i' v_i) \log \frac{\phi_j}{\lambda_{ij} + \phi_j} + \frac{(1 - \bar{u}_i' v_i)\lambda_{ij} - z_{ij}}{\lambda_{ij} + \phi_j} \\
&\qquad + \psi(z_{ij} + \phi_j) - \psi(\phi_j) \Bigg], \\
\frac{\partial^2 \bar{\ell}_j^C}{\partial \boldsymbol{\beta_j} \partial \boldsymbol{\beta_j}^\top} &= -\sum_{i=1}^{n} \frac{[z_{ij} + (1 - \bar{u}_i' v_i)\phi_j]\,\lambda_{ij}\phi_j}{(\lambda_{ij} + \phi_j)^2} \boldsymbol{x_i} \boldsymbol{x_i}^\top, \\
\frac{\partial^2 \bar{\ell}_j^C}{\partial \phi_j^2} &= \sum_{i=1}^{n} \left[\frac{(1 - \bar{u}_i' v_i)\lambda_{ij}^2 + z_{ij}\phi_j}{\phi_j(\lambda_{ij} + \phi_j)^2} + \psi_1(z_{ij} + \phi_j) - \psi_1(\phi_j) \right], \\
\frac{\partial^2 \bar{\ell}_j^C}{\partial \boldsymbol{\beta_j} \partial \phi_j} &= \sum_{i=1}^{n} \frac{[z_{ij} - (1 - \bar{u}_i' v_i)\lambda_{ij}]\,\lambda_{ij}}{(\lambda_{ij} + \phi_j)^2} \boldsymbol{x_i},
\end{aligned}$$

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

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

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

- M-step:

- For π_0, we can get

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

- For $(\boldsymbol{\lambda}, \boldsymbol{\phi})$, let

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

  - Update each λ_j using the following equation:

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

  - Substitute the value of λ_j obtained from the last step into $\bar{\ell}_j^C$, update ϕ_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_0, Y_1 and Y_2 denote three independent Poisson random variables with parameters λ_0, λ_1 and λ_2 respectively. Then $Z_1 = Y_1 + Y_0$ and $Z_2 = Y_2 + Y_0$ follow jointly a BP distribution, denoted as $\boldsymbol{Z} = (Z_1, Z_2) \sim BP(\lambda_0, \lambda_1, \lambda_2)$. Covariates can be introduced into the model through the following schemes:

$$\lambda_j = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_j), \quad j = 0, 1, 2,$$

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:

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

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 $(\lambda_{10}, \lambda_{11}, \lambda_{12})$ and $(\lambda_{20}, \lambda_{21}, \lambda_{22})$ 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:

$$\lambda_{kj} = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_{kj}), \quad k = 1, 2, \quad j = 0, 1, 2,$$

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.

Appendix C

EM Algorithms for Two Type I Multivariate Zero-Truncated Models

C.1. Type I Multivariate Zero-Truncated Poisson Model

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

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

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

$$L^c(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\lambda_{ij}^{u_i z_{ij}} e^{-\lambda_{ij}}}{(u_i z_{ij})!}.$$

The complete data log-likelihood function is

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

where C_0 is a constant not related to $\boldsymbol{\beta}$.

The associated EM algorithm is given below.

- E-step: Replace u_i by

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

  where $\lambda_{ij} = \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta_j})$.

- M-step: Let

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

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

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

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

- E-step: Replace u_i by

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

- M-step: Update each λ_j through the following equation:

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

C.2. Type I Multivariate Zero-Truncated Negative Binomial Model

Suppose for each independent individual, $\boldsymbol{Z}_i \sim MZTNB(\boldsymbol{\lambda}_i, \boldsymbol{\phi})$, $i = 1, \ldots, n$, where $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{im})$. Similarly, the covariates could be introduced as $\lambda_{ij} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta_j})$. The likelihood function becomes

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}) = \prod_{i=1}^{n} \left[1 - \prod_{j=1}^{m} \left(\frac{\phi_j}{\lambda_{ij} + \phi_j} \right)^{\phi_j} \right]^{-1} \times \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\frac{\Gamma(z_{ij} + \phi_j)}{\Gamma(\phi_j) z_{ij}!} \left(\frac{\lambda_{ij}}{\lambda_{ij} + \phi_j} \right)^{z_{ij}} \left(\frac{\phi_j}{\lambda_{ij} + \phi_j} \right)^{\phi_j} \right].$$

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

$$L^c(\boldsymbol{\beta}, \boldsymbol{\phi}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\frac{\Gamma(u_i z_{ij} + \phi_j)}{\Gamma(\phi_j)(u_i z_{ij})!} \left(\frac{\lambda_{ij}}{\lambda_{ij} + \phi_j} \right)^{u_i z_{ij}} \left(\frac{\phi_j}{\lambda_{ij} + \phi_j} \right)^{\phi_j} \right].$$

The complete data log-likelihood function is

$$\ell^c(\boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_{j=1}^{m} \sum_{i=1}^{n} [\log(\Gamma(u_i z_{ij} + \phi_j)) - \log(\Gamma(\phi_j)) + u_i z_{ij} \log \lambda_{ij} + \phi_j \log \phi_j - (u_i z_{ij} + \phi_j) \log(\lambda_{ij} + \phi_j)] + C_1,$$

where C_1 is a constant not related to $(\boldsymbol{\beta}, \boldsymbol{\phi})$.

The EM algorithm can be described as follows.

- E-step:

- Replace u_i by

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

where $\lambda_{ij} = \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta_j})$.

- Replace $\log\Gamma(u_i z_{ij} + \phi_j)$ by

$$\bar{u}_i \log\left(\Gamma(z_{ij} + \phi_j)\right) + (1 - \bar{u}_i)\log(\Gamma(\phi_j)), \quad i = 1, \ldots, n.$$

- M-step: Let

$$\begin{aligned}\bar{\ell}_j^c(\boldsymbol{\beta}_j, \phi_j) = \sum_{i=1}^{n} [&\bar{u}_i \log\left(\Gamma(z_{ij} + \phi_j)\right) - \bar{u}_i \log\left(\Gamma(\phi_j)\right) + \bar{u}_i z_{ij} \log \lambda_{ij} \\ &+\phi_j \log \phi_j - (\bar{u}_i z_{ij} + \phi_j)\log(\lambda_{ij} + \phi_j)] .\end{aligned}$$

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

$$\begin{aligned}\frac{\partial \bar{\ell}_j^c}{\partial \boldsymbol{\beta_j}} &= \sum_{i=1}^{n} \frac{(\bar{u}_i z_{ij} - \lambda_{ij})\phi_j}{\lambda_{ij} + \phi_j} \boldsymbol{x_i}, \\ \frac{\partial \bar{\ell}_j^c}{\partial \phi_j} &= \sum_{i=1}^{n} \left[\bar{u}_i \psi(z_{ij} + \phi_j) - \bar{u}_i \psi(\phi_j) + \log \frac{\phi_j}{\lambda_{ij} + \phi_j} + \frac{\lambda_{ij} - \bar{u}_i z_{ij}}{\lambda_{ij} + \phi_j}\right], \\ \frac{\partial^2 \bar{\ell}_j^c}{\partial \boldsymbol{\beta_j} \partial \boldsymbol{\beta_j}^\top} &= -\sum_{i=1}^{n} \frac{(\bar{u}_i z_{ij} + \phi_j)\lambda_{ij}\phi_j}{(\lambda_{ij} + \phi_j)^2} \boldsymbol{x_i} \boldsymbol{x_i}^\top, \\ \frac{\partial^2 \bar{\ell}_j^c}{\partial \phi_j^2} &= \sum_{i=1}^{n} \left[\bar{u}_i \psi_1(z_{ij} + \phi_j) - \bar{u}_i \psi_1(\phi_j) + \frac{\lambda_{ij}^2 + \bar{u}_i z_{ij} \phi_j}{\phi_j (\lambda_{ij} + \phi_j)^2}\right], \\ \frac{\partial^2 \bar{\ell}_j^c}{\partial \boldsymbol{\beta_j} \partial \phi_j} &= \sum_{i=1}^{n} \frac{(\bar{u}_i z_{ij} - \lambda_{ij})\lambda_{ij}}{(\lambda_{ij} + \phi_j)^2} \boldsymbol{x_i}.\end{aligned}$$

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

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

- Replace $\log\Gamma(u_i z_{ij} + \phi_j)$ by

$$\bar{u}\log\left(\Gamma(z_{ij}+\phi_j)\right) + (1-\bar{u})\log(\Gamma(\phi_j)), \quad i = 1, \ldots, n.$$

- M-step: Let

$$\begin{aligned}\bar{\ell}_j^c(\lambda_j, \phi_j) = \sum_{i=1}^{n} [&\bar{u}\log\left(\Gamma(z_{ij}+\phi_j)\right) - \bar{u}\log\left(\Gamma(\phi_j)\right) + \bar{u} z_{ij}\log\lambda_j \\ &+\phi_j\log\phi_j - (\bar{u}z_{ij}+\phi_j)\log(\lambda_j+\phi_j)]\,.\end{aligned}$$

  - Update each λ_j through the following equation:

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

  - Substitute the value of λ_j obtained from the last step into $\bar{\ell}_j^c$, update ϕ_j using the one variable Newton-Raphson method.

Bibliography

Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76**(4), 643-653.

Akakpo, R. M., Xia, M. and Polansky, A. M. (2019). Frequentist inference in insurance ratemaking models adjusting for misrepresentation. *ASTIN Bulletin: The Journal of the IAA* **49**(1), 117-146.

Boucher, J. P., Denuit, M. and Guillén, M. (2009). 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.

Brooks, S., Gelman, A., Jones, G. L. and Meng, X. L. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

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

Chatelain, F., Lambert-Lacroix, S. and Tourneret, J. Y. (2009). Pairwise likelihood estimation for multivariate mixed Poisson models generated by Gamma intensities. *Statistics and Computing* **19**(3), 283-301.

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.

Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* **93**(1), 40-57.

Denuit, M., Maréchal, X., Pitrebois, S. and Walhin, J. F. (2007). *Actuarial Modeling of Claim Counts*. John Wiley and Sons.

Frees, E. W. (2009). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Frees, E. W., Lee, G. and Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks* **4**(1), 1-36.

Frees, E. W., Shi, P. and Valdez, E. A. (2009). Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin: The Journal of the IAA* **39**(1), 165-197.

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* **2**(1), 1-25.

Frees, E. W. and Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association* **103**(484), 1457-1469.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**(410), 398-409.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**(4), 457-472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA* **37**(2), 475-515.

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.

Grogger J. T. and Carson R. T. (1991) Models for truncated counts. *Journal of Applied Econometrics* **6**, 225-238.

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

Johnson, N., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.

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.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97-109.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**(1), 1593-1623.

Karlis D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics* **30**, 63-77.

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

Karlis, D. and Meligkotsidou, L. (2005). Multivariate Poisson regression with covariance structure. *Statistics and Computing* **15**(4), 255-265.

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review* **73**(1), 35-58.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Marcel Dekker, New York.

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.

McCullagh, P. and Nelder, J. A. (1989) *Generalized linear models, second edition.* Chapman and Hall.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**(6), 1087-1092.

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

Nelsen, R. (2006). *An Introduction to Copulas, second edition.* Springer.

Nikoloulopoulos, A. K. (2013). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. *Journal of Statistical Planning and Inference* **143**(11), 1923-1937.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS.* Hoboken, NJ: Wiley.

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

Pitt, M., Chan, D. and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93**(3), 537-554.

Rai, S. N. and Matthews, D. E. (1993). Improving the EM algorithm. *Biometrics* **49**, 587-591.

Shi, P. and Valdez, E. A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics* **55**, 18-29.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* **8**, 229-231.

Song, P. X. K., Li, M. and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65**(1), 60-68.

Sun, L., Xia, M., Tang, Y. and Jones, P. G. (2016). Bayesian adjustment for unidirectional misclassification in ordinal covariates. *Journal of statistical computation and simulation* **87**(18), 3440-3468.

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

Tian, G. L., Ding, X., Liu, Y. and Tang, M. L. (2019). Some new statistical methods for a class of zero-truncated discrete distributions with applications. *Computational Statistics* **34**, 1393-1426.

Tsionas E. G. (1999). Bayesian analysis of the multivariate Poisson distribution. *Communications in Statistics-Theory and Methods* **28**, 431-451.

Tsionas E. G. (2001). Bayesian multivariate Poisson regression. *Communications in Statistics-Theory and Methods* **30**, 243-255.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**(2), 307-333.

Winkelmann, R. (2000). Seemingly unrelated negative binomial regression. *Oxford Bulletin Economics and Statistics* **62** (4), 553-560.

Xia, M. (2018). Bayesian adjustment for insurance misrepresentation in heavy-tailed loss regression. *Risks* **6**(3), 83.

Xia, M., Gustafson, P. (2016). Bayesian regression models adjusting for unidirectional covariate misclassification. *The Canadian Journal of Statistics* **44**, 198-218.

Xia, M. and Gustafson, P. (2018). Bayesian inference for unidirectional misclassification of a binary response trait. *Statistics in Medicine* **37**(6), 933-947.

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.

Young, G., Valdez, E. A. and Kohn, R. (2009). Multivariate probit models for conditional claim-types. *Insurance: Mathematics and Economics* **44**(2), 214-228.

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.

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics* **33**(3), 335-356.