MULTIVARIATE ZERO-INFLATED DOUBLE POISSON DISTRIBUTION WITH APPLICATION

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

for the Degree of

Master of Science

In

Statistics

University of Regina

By

Luya Shi

Regina, Saskatchewan

March 2020

UNIVERSITY OF REGINA

FACULTY OF GRADUATE STUDIES AND RESEARCH

SUPERVISORY AND EXAMINING COMMITTEE

Luya Shi, candidate for the degree of Master of Science in Statistics**,** has presented a thesis titled, ***Multivariate Zero-Inflated Double Poisson Distribution With Application,*** in an oral examination held on November 20, 2019. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner:	Dr. Lisa Fan, Department of Computer Science
Supervisor:	Dr. DianLiang Deng, Department of Mathematics & Statistics
Committee Member:	Dr. Andrei Volodin, Department of Mathematics & Statistics
Chair of Defense:	Dr. Wei Peng, Faculty of Engineering & Applied Science

Abstract

In this thesis, a new multivariate zero-inflated double Poisson distribution is proposed, which is considered as the generalization of univariate zero-inflated double Poisson distribution. The statistical properties of new distribution, such as joint probability mass function, expectation, covariance matrix, marginal distribution and conditional distribution, are derived. The maximum likelihood estimates of parameters are obtained. The score test statistic is derived to test zero inflation of multivariate count data with many zeros by using score function and information matrix. The empirical powers of score test are gained by a simulation study. The application of real data is performed to test zero inflation.

Acknowledgements

My deepest gratitude goes first and foremost to Professor Dr. Dianliang Deng gratitude, my supervisor, for his care and help in my life and academics. During my two years of graduate studies, Dr. Deng's rigorous attitude towards knowledge and passion for research is always guiding me at all times. I would like to show my heartfelt gratitude to all professors in Department of Mathematics and Statistics. It is because of their professional guidance that I can improve my statistical knowledge in these two years. Finally, I would like to express gratitude to Department of Mathematic and Statistics and the Faculty of Graduate Studies and Research for the financial support.

Dedication

The thesis is dedicated to my family because of their support and true love. Their selfishness love will always support me in the future.

Contents

List of Tables

Chapter 1

Introduction

Counting data has been widely used in statistical areas. These kinds of data are often collected from varieties of fields which include medical, biology, economics, agriculture, insurance, biochemistry, ecology, environmental science and public health. The most considered model we used for analyzing the discrete data is the classic regression model such as Poisson, binomial and negative binomial regression model. However, the classic regression models all have restrictions when they are applied to fit the real data. Poisson regression model is only suitable for those data whose expected value approximately equals its variance. Obviously, this kind of situation seldom occurs in reality. Alternatively, count data shows extra variation for which the variance is much larger than expected value and thus negative binomial regression model is more flexible than Poisson regression model.

In order to get model accuracy, Efron [3] proposed a larger class of distributions

called double exponential families in 1986. Double Poisson(DP) distribution is a particular case included in the family which is applicable for both overdispersion and underdispersion. The approximate probability mass function of DP distribution is derived by Efron which does not sum to unity, even if the difference between exact probability mass function and approximate probability mass function is small.

However in the actual study, it often occurs that the counts of events involve excessive zeros compared to the double Poisson distribution, that is, the case of zero inflation. If the part of excessive zeros is not correctly explained, it will lead to very large deviation of parameter estimation and failure for inference. Therefore, a kind of zero-inflated double Poisson regression model which combines ordinary double Poisson distribution and Bernoulli distribution will be more suitable. The test for the existence of zero inflation is indispensable for the selection of models. To solve this problem, a score test method is proposed and test statistics under the null hypothesis is derived.

In 1960s, this study became noticeable by a number of scholars and statistician, such as Johnson and Kots in 1969 [5]. Later some scholars have proposed Hurdle model to apply the data in the field of economics [10] in 1986. Lambert (1992) [7] introduced zero-inflated Poisson model with covariates, establishing a mixed probability distribution for zero counts data and non-zero counts data, to apply in

electronics manufacturing quality control. In 1994, the zero-inflated negative binomial model was developed based on zero-inflated Poisson model by Greene [4]. The Berndt–Hall–Hall–Hausman (BHHH) algorithm was applied in the research, which was about bank consumers who had poor credit, to estimate the standard error of the model parameters. This zero-inflated negative binomial model is the development of Poisson model and negative binomial model which overcomes the shortcomings of Poisson model or negative binomial model in analyzing zero-inflated data. The zero-inflated negative binomial model can explain excessive zero values in count data and identify true zero values in the dependent variable. It also makes the estimates of parameters more effective and unbiased. Thereby researchers can obtain reliable hypothesis testing and parameter estimation in order to help researchers solve a series of practical problems which cannot be solved by traditional models. Xie presented a confidence interval test which is also based on zero-inflated Poisson (ZIP) regression model in 2001 [14]. The researchers compared the power of C test, R test, Likelihood Ratio test, Score test, confidence interval test and etc. through the simulation method. The findings showed that confidence interval test was less effective than other tests.

In the analysis of actual problems, it is crucial to assess whether the data has zero-inflation and to decide the model. In 1995, Broek proposed a score test based on the ZIP regression model [11]. Afterward Lee compared statistics in the score

test, the likelihood ratio test and the Wald test [8]. Deng and Paul [1] introduced score test in generalized linear models and conducted binomial and Poisson models in 2000. Later Deng and Paul (2005) [2] introduced a zero-inflated and over-dispersed generalized linear model. In 2017, Xie and other researchers applied score test into double Poisson regression model [13].

There are many studies based on univariate distributions for the zero-inflated data. However, with more complicated situations in many studies, researchers have extended univariate distribution to the multivariate distribution. Johnson and Kots proposed multivariate Poisson distribution to deal with a lot of defects [5]. In 1999, Li et al.[8] built up zero-inflated Poisson distribution using many possible approaches. Moreover, type I MZIP distribution was suggested by Liu and Tian [9] and compared with Li's MZIP model.

Generally, the studies included the Hurdle model, the zero-inflated Poisson model and the zero-inflated negative binomial model. Due to the restriction of the models above, they are not suitable for some particular types of data. Therefore, a new model, multivariate zero-inflated double Poisson distribution, is proposed to analyze multivariate count data in this thesis. The distribution is derived along with the approach proposed by Xie in 2017 [13]. We consider this new multivariate zero-inflated double Poisson distribution for a p-dimensional discrete variable. It is regarded as a mixed

distribution which combines a degenerate distribution with a mass of zero and p-dimensional double Poisson distribution. Degenerate distribution and p-dimensional double Poisson distribution are mutually independent.

The following is a synopsis of this thesis:

In Chapter 2, some other researchers' methodology as a background is provided in this paper. The double Poisson distribution is stated. The zero-inflated double Poisson model and its score test are given.

In Chapter 3, a new multivariate zero-inflated double Poisson distribution (MZIDP) was proposed, encouraged by great distribution characteristic of type I multivariate zero-inflated Poisson model of Liu and Tian [9] in 2015. Then some properties of MZIDP distribution such as joint probability mass function, expectation and variance, likelihood function are explored. Furthermore, the score test statistics and the likelihood ratio test statistics are utilized to test zero-inflation parameter in MZIDP model.

In Chapter 4, the simulation study for the zero-inflation by score test statistics are performed and the real data application is studied and the score test statistic is used to find out whether data exhabits zero-inflation.

In Chapter 5, the MZIDP studies are summarized and some future work is mentioned.

Chapter 2

Preliminary

2.1 Double Poisson Distribution

In 1986, Efron introduced double Poisson distribution [3] which is made up for the shortcomings of Poisson distribution. According to the distribution, the model has two parameters, μ and α. The exact probability mass function (PMF) is:

$$\tilde{f}_{\mu,\alpha}(y) = c(\mu,\alpha)(\alpha^{1/2}e^{-\alpha\mu})(\frac{e^{-y}y^y}{y!})(\frac{e\mu}{y})^{\alpha y}, \qquad y = 0,1,2,... \tag{2.1}$$

where the factor of $c(\mu,\alpha)$ is to ensure the density sum to 1 and it has the form that

$$\frac{1}{c(\mu,\alpha)} = \sum_{y=0}^{\infty}(\alpha^{1/2}e^{-\alpha\mu})(\frac{e^{-y}y^y}{y!})(\frac{e\mu}{y})^{\alpha y} \approx 1 + \frac{1-\alpha}{12\mu\alpha}(1+\frac{1}{\mu\alpha})$$

Efron derived the constant $c(\mu,\alpha)$ about equals to 1. Therefore, the approximate probability mass function of double Poisson distribution is in the following form

$$f_{\mu,\alpha}(y) = (\alpha^{1/2}e^{-\alpha\mu})(\frac{e^{-y}y^y}{y!})(\frac{e\mu}{y})^{\alpha y}, \qquad y = 0,1,2,... \tag{2.2}$$

The parameter μ is corresponding to the Poisson mean parameter. Parameter α is the dispersion parameter. When $\alpha = 1$, the double Poisson distribution simply becomes to Poisson distribution. Thus, double Poisson is overdispersed if $\alpha < 1$ and underdispersed if $\alpha > 1$.

Efron obtained the mean and variance of DP distribution based on exact probability density function [13].

$$E(Y) \approx \mu$$

$$Var(Y) \approx \frac{\mu}{\alpha}$$

2.2 Univariate Zero-inflated Double Poisson Distribution

Let $Z \sim$ Bernoulli $(1-\omega)$, $X \sim$ DP (μ, α), Z and X are mutually independent. The random variable $Y \sim$ ZIDP (ω, μ, α) has stochastic representation as following

$$Y \stackrel{d}{=} ZX = \begin{cases} 0 & \text{with probability } \omega \\ X & \text{with probability } 1-\omega \end{cases} \tag{2.3}$$

The indication "$\stackrel{d}{=}$" states the random variables in the left hand side and the random variables in the right hand side have the same distribution.

If $y = 0$,

$$\begin{aligned} P(Y=0) &= P(ZX=0) \\ &= P(Z=0) + P(Z=1, X=0) \\ &= \omega + (1-\omega) f(0; \mu, \alpha) \end{aligned} \tag{2.4}$$

If $y > 0$,

$$\begin{aligned} P(Y = y) &= P(ZX = y) \\ &= P(Z = 1, X = y) \\ &= (1 - \omega) f(y; \mu, \alpha) \end{aligned} \tag{2.5}$$

The zero-inflated density function is derived by combining equation (2.4) and (2.5)

$$P(Y = y) = \begin{cases} \omega + (1 - \omega) f(0; \mu, \alpha), & \text{if } y = 0, \\ (1 - \omega) f(y; \mu, \alpha), & \text{if } y = 1, 2, ..., \end{cases} \tag{2.6}$$

In the representation, ω is the zero-inflation or zero-deflation parameter. ω can be negative values. A zero deflated model appears when ω is in the range of $-\frac{f(0;\mu,\alpha)}{1-f(0;\mu,\alpha)} \leq \omega < 0$, otherwise a zero-inflated model arises when $\omega > 0$.

The expectation and variance of zero inflated double Poisson model will be simply collected

$$\begin{cases} E(Y) & = E(Z)E(X) \\ E(Y^2) & = E(Z)E(X^2) \\ Var(Y) & = E(Z)E(X^2) - (E(Z)E(X))^2, \end{cases}$$

where $E(Z) = 1 - \omega$ and $E(X) = \mu$. Further, we could obtain the expression in terms of the parameters ω, μ and α by plug in the representation into above expressions,

$$E(Y) = E(Z)E(X) = (1 - \omega)\mu$$

$$\begin{aligned} E(X^2) &= Var(X) + (E(X))^2 \\ &= \frac{\mu}{\alpha} + \mu^2 \end{aligned}$$

$$E(Y^2) = E(Z)E(X^2) = (1-\omega)(\frac{\mu}{\alpha} + \mu^2)$$

hence, the variance of random variable Y will be

$$\begin{aligned} Var(Y) &= E(Y^2) - (E(Y))^2 \\ &= (1-\omega)(\frac{\mu}{\alpha} + \mu^2) - [(1-\omega)\mu]^2 \\ &= (1-\omega)\mu(\frac{1}{\alpha} + \omega\mu). \end{aligned}$$

2.3 Score test for Univariate Double Poisson

Let Y_i , $i = 1, ..., n$ denote a sample of independent observations from (2.6). $Y_i \sim$ ZIDP (ω, α, μ_i). The linear predictor is denoted as $\mu_i = \exp(X_i^T\beta)$, and X_i is a q-dimensional vector of covariates. β is regression parameters which is a q-dimensional vector. Then the likelihood function of zero-inflated double Poisson distribution can be explained as

$$L(\omega, \alpha, \mu_1, ..., \mu_p; y) = \prod_{i=1}^{n} \{[\omega + (1-\omega)f(0; \mu_i, \alpha)]I(y_i = 0) + (1-\omega)f(y_i; \mu_i, \alpha)I(y_i > 0)\}. \tag{2.7}$$

Let $\gamma = \frac{\omega}{1-\omega}$, $\theta = (\alpha, \beta^T, \gamma)^T$, the log-likelihood with parameter θ is

$$l(\theta; y) = \sum_{i=1}^{n} \{-\ln(1+\gamma) + \boldsymbol{I}(y_i = 0)\ln[\gamma + f(0; \mu_i, \alpha)] + \boldsymbol{I}(y_i > 0)\ln f(y_i; \mu_i, \alpha)\}. \tag{2.8}$$

2.3.1 Testing for the Significance of Zero Inflation

Testing $H_0 : \omega = 0$ is equivalent to testing

$$H_0 : \gamma = 0 \qquad \text{against} \qquad H_1 : \gamma > 0 \tag{2.9}$$

Let $f_{0i} = f(0; \mu_i, \alpha)$, $T_i = \log f(y_i; \mu_i, \alpha)$, $f_{0\alpha} = (\frac{\partial f_{0i}}{\partial \alpha})^T$, $f_{0\beta} = (\frac{\partial f_{0i}}{\partial \beta})^T$, $f_{0\alpha\alpha} = (\frac{\partial^2 f_{0i}}{\partial \alpha^2})^T$, $f_{0\alpha\beta} = (\frac{\partial^2 f_{0i}}{\partial \alpha \partial \beta^T})^T$, $f_{0\beta\beta} = (\frac{\partial^2 f_{0i}}{\partial \beta \partial \beta^T})$, thus the score function is

$$U_\gamma(\theta) = \frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} \left\{ \frac{\boldsymbol{I}(y_i = 0)}{f_{0i} + \gamma} - \frac{1}{1 + \gamma} \right\} \tag{2.10}$$

The information matrix of parameter θ can be represented as

$$I(\theta) = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\beta} & I_{\alpha\gamma} \\ I'_{\alpha\beta} & I_{\beta\beta} & I_{\beta\gamma} \\ I'_{\alpha\gamma} & I'_{\beta\gamma} & I_{\gamma\gamma} \end{pmatrix} \tag{2.11}$$

For each part, $I_{\alpha\alpha}, I_{\alpha\gamma}, I_{\gamma\gamma}$ are matrices of size 1×1, $I_{\alpha\beta}$ is the matrix of size $1 \times p$, $I_{\beta\beta}$ is the matrix of size $p \times p$ and $I_{\beta\gamma}$ is the matrix of size $p \times 1$.

Respectively,

$$
\begin{aligned}
I_{\alpha\alpha} &= V_1 + \frac{f_{0\alpha}^T D f_{0\alpha} - \mathbf{1}^T f_{0\alpha\alpha}}{1+\gamma} \\
I_{\alpha\beta} &= V_2 + \frac{f_{0\alpha}^T D f_{0\beta} - \mathbf{1}^T f_{0\alpha\beta}}{1+\gamma} \\
I_{\alpha\gamma} &= \frac{\mathbf{1}^T D f_{0\alpha}}{1+\gamma} \\
I_{\beta\beta} &= V_3 + \frac{f_{0\beta}^T D f_{0\beta} - \mathbf{1}^T f_{0\beta\beta}}{1+\gamma} \\
I_{\beta\gamma} &= \frac{f_{0\beta}^T D \mathbf{1}}{1+\gamma} \\
I_{\gamma\gamma} &= \frac{-n}{(1+\gamma)^2} + \frac{\mathbf{1}^T D \mathbf{1}}{1+\gamma}
\end{aligned}
$$

where

$$
\begin{aligned}
& d_i = (\gamma + f_{0i})^{-1}, \qquad D = \mathrm{diag}(d_1, ..., d_n), \qquad \mathbf{1} = (1, ...1)^T, \\
& f_{0\alpha} = (\frac{1}{2\alpha} - \mu_i) f_{0i}, \qquad f_{0\beta} = -\alpha f_{0i} \mu_i X_i, \\
& f_{0\alpha\alpha} = (-\frac{1}{4\alpha^2} - \frac{\mu_i}{\alpha} + \mu_i^2) f_{0i}, \\
& f_{0\alpha\beta} = f_{0i}\mu_i(-\frac{3}{2} + \alpha\mu_i) X_i^T, \qquad f_{0\beta\beta} = f_{0i}(\alpha^2\mu_i^2 - \alpha\mu_i) X_i X_i^T, \\
& V_1 = \frac{1}{1+\gamma} \sum_{i=1}^{n} \frac{1 - f_{0i}}{2\alpha^2}, \qquad V_2 = -\frac{1}{1+\gamma} \sum_{i=1}^{n} f_{0i}\mu_i X_i^T, \\
& V_3 = \frac{1}{1+\gamma} \sum_{i=1}^{n} (1 - f_{0i}) \alpha \mu_i X_i X_i^T.
\end{aligned}
$$

Let $\hat{\theta}_\gamma = (\hat{\alpha}, \hat{\beta}^T, 0)^T$ be the maximum likelihood estimate (MLE) of θ under null hypothesis H_0. Thus the score statistics can be obtained by

$$
SC_\gamma = \left\{ (\frac{\partial l(\theta)}{\partial \gamma})^T J^{\gamma\gamma} (\frac{\partial l(\theta)}{\partial \gamma} \right\}_{\hat{\theta}_\gamma},
$$

let $\theta_1 = (\alpha, \beta^T)^T$, then the information matrix $I(\theta)$ can be partitioned with parameter θ_1 and γ as

$$I(\theta_1; \gamma) = \begin{pmatrix} I_{\theta_1\theta_1} & I_{\theta_1\gamma} \\ I_{\theta_1\gamma}^T & I_{\gamma\gamma} \end{pmatrix}. \tag{2.12}$$

Thus, $J^{\gamma\gamma}$ can be obtained by

$$J^{\gamma\gamma} = (I_{\gamma\gamma} - I_{\theta_1\gamma}^T I_{\theta_1\theta_1}^{-1} I_{\theta_1\gamma})^{-1}$$

Therefore the score test statistic under null hypothesis $H_0 : \gamma = 0$ is

$$SC_\gamma = \left\{ U_\gamma(\theta)^T \left[-\frac{n}{(1+\gamma)^2} + \frac{\mathbf{1}^T D \mathbf{1}}{1+\gamma} - I_{\theta_1\gamma}^T I_{\theta_1\theta_1}^{-1} I_{\theta_1\gamma} \right]^{-1} U_\gamma(\theta) \right\}_{\hat{\theta}_\gamma} \tag{2.13}$$

which follows the chi-squared distribution with one degree of freedom.

2.3.2 Testing for the Significance of Dispersion

The zero-inflated double Poisson simply decreases to zero-inflated Poisson distribution when the dispersion parameter $\alpha = 1$. Thus, it is important to test the subsequent hypothesis to check the ZIDP distribution is better than the ZIP distribution.

$$H_0 : \alpha = 1 \qquad \text{against} \qquad H_1 : \alpha \neq 1 \tag{2.14}$$

Let $\hat{\theta}_\alpha = (1, \hat{\beta}^T, \hat{\gamma})^T$ denotes the MLE of the parameter θ under the null hypothesis in (2.14). Thereby, we consider α for the interest parameter, then $\theta_2 = (\beta^T, \gamma)^T$ will tend to be nuisance parameter.

Based on the log-likehood function in (2.8), we will easily get the following score function,

$$U_\alpha(\theta) = \frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \left\{ \boldsymbol{I}(y_i = 0)(\frac{1}{2\alpha} - \mu_i)\frac{f_{0i}}{\gamma + f_{0i}} + \boldsymbol{I}(y_i > 0)(\frac{1}{2\alpha} - \mu_i + y_i + y_i \log(\frac{\mu_i}{y_i})) \right\}$$

Then we could establish the score test statistic as

$$SC_\alpha = \left\{ (\frac{\partial l(\theta)}{\partial \alpha})^T J^{\alpha\alpha} (\frac{\partial l(\theta)}{\partial \alpha} \right\}_{\hat{\theta}_\alpha}$$

Additionally, based on parameter α, θ_2, the fisher information matrix $I(\theta)$ should be partitioned into

$$I(\theta_2; \alpha) = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\theta_2} \\ I_{\alpha\theta_2}^T & I_{\theta_2\theta_2} \end{pmatrix},$$

then

$$J^{\alpha\alpha} = (I_{\alpha\alpha} - I_{\alpha\theta_2} I_{\theta_2\theta_2}^{-1} I_{\alpha\theta_2}^T)^{-1}$$

Therefore, the score test statistic is

$$SC_\alpha = \left\{ U_\alpha(\theta)^T \left[V_1 + \frac{f_{0\alpha}^T D f_{0\alpha} - \mathbf{1}^T f_{0\alpha\alpha}}{1 + \gamma} - I_{\alpha\theta_2} I_{\theta_2\theta_2}^{-1} I_{\alpha\theta_2}^T \right]^{-1} U_\alpha(\theta) \right\}_{\hat{\theta}_\alpha} \tag{2.15}$$

which has an approximate chi-squared distribution with one degree of freedom.

2.4 Type I Multivariate Zero-Inflated Poisson Distribution

Multivariate zero-inflated Poisson model was proposed by Yin Liu and Guo-Liang Tian [9] in 2015. The correlated theory and related properties are derived.

Let us consider an m-dimensional discrete random vector $\boldsymbol{Y} = (Y_1, ..., Y_m)^T$ which is tend to have a type I multivariate zero-inflated Poisson distribution with parameters $\phi \in [0, 1)$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)^T \in \mathbb{R}_+^m$ when

$$\boldsymbol{Y} \overset{d}{=} Z\boldsymbol{X} = \begin{cases} \boldsymbol{0} & \text{with probability } \phi \\ \boldsymbol{X} & \text{with probability } 1 - \phi \end{cases} \tag{2.16}$$

where $Z \sim \text{Bernoulli}(1 - \phi)$, $\boldsymbol{X} = (X_1, ..., X_m)^T$, $X_i \sim \text{Poisson}(\lambda_i)$ for $i = 1, ..., m$, and $(Z, X_1, ..., X_m)$ are mutually independent. We can note $\boldsymbol{Y} \sim \text{ZIP}^{(1)}(\phi; \lambda_1, ..., \lambda_m)$ or $\boldsymbol{Y} \sim \text{ZIP}_m^{(1)}(\phi, \boldsymbol{\lambda})$ as $\boldsymbol{Y}$ follows zero-inflated Poisson distribution where $\boldsymbol{X}$ is the base vector of the $\boldsymbol{Y}$. For any non-negative real vector $\boldsymbol{Y} = (Y_1, ..., Y_m)^T$, the joint probability mass function will be written as following form

$$\begin{aligned} f(\boldsymbol{y}|\phi, \lambda) &= [\phi + (1 - \phi)e^{-\lambda_+}]\boldsymbol{I}(\boldsymbol{y} = \boldsymbol{0}) + [(1 - \phi)e^{-\lambda_+} + \prod_{i=1}^{m} \frac{\lambda_i^{y_i}}{y_i!}]\boldsymbol{I}(\boldsymbol{y} \neq \boldsymbol{0}) \\ &= \phi Pr(\boldsymbol{\xi} = \boldsymbol{y}) + (1 - \phi)Pr(\boldsymbol{x} = \boldsymbol{y}) \end{aligned}$$

where $\lambda_+ = \sum_{i=1}^m \lambda_i$, $\boldsymbol{\xi} = (\xi_1, ..., \xi_m)^T$ and $\{\xi_i\}_{i=1}^m \overset{i.i.d}{\sim} \text{Degenerate}(0)$.

Let $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ be an independent random sample from type I m-dimensional ZIP distribution $\text{ZIP}^{(1)}(\phi; \lambda_1, ..., \lambda_m)$, where $\boldsymbol{y}_j = (Y_{1j}, ..., Y_{mj})^T$ for $j = 1, ..., n$ and $Y_{obs} = \boldsymbol{y}_1, ...\boldsymbol{y}_n$ are the observations. Therefore the likelihood function is

$$L(\phi, \boldsymbol{\lambda}|Y_{obs}) = [\phi + (1 - \phi)e^{-\lambda_+}]^{m_0} \times (1 - \phi)^{n - m_0} e^{-(n - m_0)\lambda_+} \prod_{i=1}^{m} \lambda_i^{N_i}$$

where $N_i = \sum_{j=1}^n y_{ij}$, $m_0 = \sum_{j=1}^n \boldsymbol{I}(\boldsymbol{y}_j = \boldsymbol{0})$, then the log-likelihood function is

$$\ell(\phi, \boldsymbol{\lambda}|Y_{obs}) = m_0 \log[\phi + (1 - \phi)e^{-\lambda_+}] + (n - m_0)[\log(1 - \phi) - \lambda_+] + \sum_{i=1}^{m} N_i \log \lambda_i$$

Assume we want to test

$$H_0 : \phi = 0 \qquad \text{against} \qquad H_1 : \phi \neq 0.$$

Let $\theta = \frac{\phi}{1-\phi}$, $\boldsymbol{\beta} = (\beta_1, ...\beta_m)^T = (\log \lambda_1, ..., \log \lambda_m)^T$, then testing H_0 is equivalent to testing $H_0^* : \theta = 0$. Now the log likelihood function is

$$\ell(\theta, \boldsymbol{\beta}|Y_{obs}) = -n \log(1+\theta) + \sum_{j=1}^{n} [\boldsymbol{I}(\boldsymbol{y}_j = \boldsymbol{0}) \log(\theta + e^{-\lambda_+}) + \boldsymbol{I}(\boldsymbol{y}_j \neq \boldsymbol{0}) \sum_{i=1}^{m} (y_{ij}\beta_i - \lambda_i)].$$

The score vector is

$$U(\boldsymbol{\beta}, \theta) = (\frac{\partial \ell(\boldsymbol{\beta}, \theta|Y_{obs})}{\partial \theta}, \frac{\partial \ell(\boldsymbol{\beta}, \theta|Y_{obs})}{\partial \beta_1}, ..., \frac{\partial \ell(\boldsymbol{\beta}, \theta|Y_{obs})}{\partial \beta_m})^\top.$$

Now let $J(\boldsymbol{\beta}, \theta)$ denote the Fisher information matrix which is the negative expectation of the second derivatives of log-likelihood. Then the score test statistic under H_0^* is

$$T = U^T(\hat{\boldsymbol{\beta}}_0, \hat{\theta}_0) J^{-1}(\hat{\boldsymbol{\beta}}_0, \hat{\theta}_0) U(\hat{\boldsymbol{\beta}}_0, \hat{\theta}_0) \sim \chi^2(1)$$

where $\hat{\theta}_0 = 0$, $\hat{\boldsymbol{\beta}}_0 = (\log(\sum_{j=1}^{n} y_{1j}/n), ..., \log(\sum_{j=1}^{n} y_{mj}/n))^T$ represent the MLEs of $\boldsymbol{\beta}$ under H_0^*.

$$U(\hat{\boldsymbol{\beta}}_0, \hat{\theta}_0) = \left(\frac{m_0}{\exp(-\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}/n} - n, \underbrace{0, ..., 0}_{\text{m}} \right)^\top$$

The corresponding p-value is as follows

$$P = Pr(T > t|H_0) = Pr(\chi^2(1) > t).$$

Chapter 3

Score Test for Multivariate Zero-inflated Double Poisson Distribution

Utilizing the representation (2.3) of the univariate zero-inflated double Poisson distribution, we could extend it to a multivariate situation. The components of multivariate zero-inflated double Poisson distribution is a vector form, which is represented with a common Z. The definition is showed as below.

Definition 1 A p-dimensional discrete random vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_p)^\top$ is assumed to follow a multivariate zero-inflated double Poisson distribution with parameters $\omega \in [0, 1)$, $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)^\top$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p)^\top$ if

$$\boldsymbol{Y} \overset{d}{=} Z\boldsymbol{X} = \begin{cases} \boldsymbol{0} & \text{with probability } \omega \\ \boldsymbol{X} & \text{with probability } 1-\omega \end{cases} \tag{3.1}$$

Where $Z \sim \text{Bernoulli}(1-\omega)$, $\boldsymbol{X} = (X_1, X_2, ..., X_p)^\top$, $X_j \sim \text{DP}(\mu_j, \alpha_j)$ for $j = 1, ..., p$, and Z and $X_1, X_2, ..., X_p$ are mutually independent. We could define the multivariate

zero-inflated double Poisson distribution in the form of $\boldsymbol{Y} \sim \text{ZIDP}(\omega; \mu_1, \alpha_1, ..., \mu_p, \alpha_p)$ or $\boldsymbol{Y} \sim \text{ZIDP}_p(\omega, \boldsymbol{\mu}, \boldsymbol{\alpha})$. $\boldsymbol{X}$ is the base vector of the $\boldsymbol{Y}$.

3.1 Joint Probability Mass Function

The joint probability mass function (pmf) of $\boldsymbol{Y} \sim \text{ZIDP}_p(\omega, \boldsymbol{\mu}, \boldsymbol{\alpha})$ is defined as

$$f(\boldsymbol{y}|\omega, \boldsymbol{\mu}, \boldsymbol{\alpha}) = \Pr(\boldsymbol{Y} = \boldsymbol{y}) = \Pr(ZX_1 = y_1, ZX_2 = y_2, ..., ZX_p = y_p)$$

If $\boldsymbol{y} = \boldsymbol{0}_p$, we have

$$\begin{aligned} f(\boldsymbol{y}|\omega, \boldsymbol{\mu}, \boldsymbol{\alpha}) &= \Pr(ZX_1 = 0, ZX_2 = 0, ..., ZX_p = 0) \\ &= \Pr(Z = 0) + \Pr(Z = 1, X_1 = 0, X_2 = 0, ..., X_p = 0) \\ &= \omega + (1 - \omega) \prod_{j=1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j} \end{aligned} \tag{3.2}$$

If $\boldsymbol{y} \neq \boldsymbol{0}_p$, we have

$$\begin{aligned} f(\boldsymbol{y}|\omega, \boldsymbol{\mu}, \boldsymbol{\alpha}) &= \Pr(ZX_1 = y_1, ZX_2 = y_2, ..., ZX_p = y_p) \\ &= \Pr(Z = 1, X_1 = y_1, X_2 = y_2, ..., X_p = y_p) \\ &= (1 - \omega) \prod_{j=1}^{p} (\alpha_j^{1/2} e^{-\alpha_j \mu_j}) (\frac{e^{-y_j} y_j^{y_j}}{y_j!}) (\frac{e\mu_j}{y_j})^{\alpha_j y_j} \end{aligned} \tag{3.3}$$

We can combine equation (3.2) and (3.3), therefore the joint probability mass function will be

$$\begin{aligned} f(\boldsymbol{y}|\omega, \boldsymbol{\mu}, \boldsymbol{\alpha}) &= \omega + (1-\omega)\prod_{j=1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j} \boldsymbol{I}(\boldsymbol{y} = \boldsymbol{0}) \\ &\quad + (1-\omega)\prod_{j=1}^{p} (\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j} \boldsymbol{I}(\boldsymbol{y} \neq \boldsymbol{0}) \qquad (3.4) \\ &= \omega \Pr(\boldsymbol{\xi} = \boldsymbol{y}) + (1-\omega)\Pr(\boldsymbol{X} = \boldsymbol{y}) \end{aligned}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_p)^\top$ and $\{\xi_i\}_{j=1}^{p} \overset{iid}{\sim}$ Degenerate(0).

3.2 Expectation and Covariance Matrix

According to equation (3.1), we can easily obtain the expectation of the model

$$\begin{aligned} E(\boldsymbol{Y}) &= E(Z\boldsymbol{X}) = E(Z)E(\boldsymbol{X}) \\ &= (1-\omega)\boldsymbol{\mu} \end{aligned}$$

where $\boldsymbol{X}$ and $\boldsymbol{\mu}$ are both p-dimensional vectors, also $\boldsymbol{\mu}$ is the expectation value of base vector $\boldsymbol{X}$.

Furthermore,

$$\begin{aligned} E(\boldsymbol{Y}\boldsymbol{Y}^\top) &= E(Z)E(\boldsymbol{X}\boldsymbol{X}^\top) \\ &= (1-\omega)[\mathrm{diag}(\frac{\boldsymbol{\mu}}{\boldsymbol{\alpha}}) + \boldsymbol{\mu}\boldsymbol{\mu}^\top] \end{aligned}$$

$$\begin{aligned} Cov(\boldsymbol{Y}) &= E(\boldsymbol{Y}\boldsymbol{Y}^\top) - E(\boldsymbol{Y})^2 \\ &= (1-\omega)[\mathrm{diag}(\frac{\boldsymbol{\mu}}{\boldsymbol{\alpha}}) + \omega\boldsymbol{\mu}\boldsymbol{\mu}^\top] \end{aligned}$$

where diag($\frac{\boldsymbol{\mu}}{\boldsymbol{\alpha}}$) is a diagonal matrix which has the following form

$$\begin{pmatrix} \frac{\mu_1}{\alpha_1} & 0 & \dots & 0 \\ 0 & \frac{\mu_2}{\alpha_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{\mu_p}{\alpha_p} \end{pmatrix}$$

3.3 Marginal Distributions

A marginal distribution is the distribution of one random variable without any sort of reference to the other random variable.

Let $\boldsymbol{Y} \sim \text{ZIDP}(\omega; \mu_1, \alpha_1, ..., \mu_p, \alpha_p)$. $\boldsymbol{Y}$ is partitioned into two parts.

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}^{(1)} \\ \boldsymbol{Y}^{(2)} \end{pmatrix}, \quad \text{where} \quad \boldsymbol{Y}^{(1)} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} \quad \text{and} \quad \boldsymbol{Y}^{(2)} = \begin{pmatrix} Y_{k+1} \\ \vdots \\ Y_p \end{pmatrix}$$

It is a good way to partition $\boldsymbol{X}$ in the same form as $\boldsymbol{Y}$. According to Definition 1, we could get the following equation.

$$\begin{cases} \boldsymbol{Y}^{(1)} \overset{d}{=} Z\boldsymbol{X}^{(1)} \sim \text{ZIDP}(\omega; \mu_1, \alpha_1, \cdots, \mu_k, \alpha_k) \quad \text{and} \\ \boldsymbol{Y}^{(2)} \overset{d}{=} Z\boldsymbol{X}^{(2)} \sim \text{ZIDP}(\omega; \mu_{k+1}, \alpha_{k+1}, \cdots, \mu_p, \alpha_p) \end{cases} \tag{3.5}$$

The following equality would be held for any positive integers $j_1, j_2, ..., j_k$ satisfying $1 \leqslant j_1 < j_2 < ... < j_k \leqslant p$

$$\begin{pmatrix} Y_{j_1} \\ \vdots \\ Y_{j_k} \end{pmatrix} \stackrel{d}{=} Z \begin{pmatrix} X_{j_1} \\ \vdots \\ X_{j_k} \end{pmatrix} \sim \text{ZIDP}(\omega; \mu_{j_1}, \alpha_{j_1}, ..., \mu_{j_k}, \alpha_{j_k}) \tag{3.6}$$

3.4 Conditional Distributions

A conditional distribution is a probability distribution for a sub-population. In other words, it shows the probability that a randomly selected item in a sub-population has the one characteristic of interest.

3.4.1 Conditional Distributions of $\boldsymbol{Y}^{(1)}|\boldsymbol{Y}^{(2)}$

We could get the conditional distribution of $\boldsymbol{Y}^{(1)}|\boldsymbol{Y}^{(2)}$ by utilizing (3.4) and (3.5).

$$\begin{aligned} \Pr\{\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}|\boldsymbol{Y}^{(2)} = \boldsymbol{y}^{(2)}\} &= \frac{f(\boldsymbol{y}|\omega, \boldsymbol{m}, \boldsymbol{p})}{Pr\{\boldsymbol{Y}^{(2)} = \boldsymbol{y}^{(2)}\}} \\ &= \frac{\omega \Pr(\boldsymbol{\xi} = \boldsymbol{y}) + (1-\omega)\Pr(\boldsymbol{X} = \boldsymbol{y})}{\omega \Pr(\boldsymbol{\xi} = \boldsymbol{y}^{(2)}) + (1-\omega)\Pr(\boldsymbol{X}^{(2)} = \boldsymbol{y}^{(2)})} \end{aligned} \tag{3.7}$$

where

$$\omega \Pr(\boldsymbol{\xi} = \boldsymbol{y}) + (1-\omega)\Pr(\boldsymbol{X} = \boldsymbol{y})$$
$$= [\omega + (1-\omega)\prod_{j=1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j}]\mathbf{I}(\boldsymbol{y} = \mathbf{0}) + [(1-\omega)\prod_{j=1}^{p} (\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}]\mathbf{I}(\boldsymbol{y} \neq \mathbf{0})$$

and

$$\omega \Pr(\boldsymbol{\xi} = \boldsymbol{y}^{(2)}) + (1-\omega)\Pr(\boldsymbol{X}^{(2)} = \boldsymbol{y}^{(2)})$$
$$= [\omega + (1-\omega) \prod_{j=k+1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j}]\mathbf{I}(\boldsymbol{y}^{(2)} = \mathbf{0})$$
$$+ [(1-\omega) \prod_{j=k+1}^{p} (\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}]\mathbf{I}(\boldsymbol{y}^{(2)} \neq \mathbf{0})$$

We should consider two cases.

$$\begin{cases} \text{case I}: \quad \boldsymbol{y}^{(2)} = \mathbf{0} \begin{cases} \boldsymbol{y}^{(1)} = \mathbf{0} \\ \boldsymbol{y}^{(1)} \neq \mathbf{0} \end{cases} \\ \text{case II}: \quad \boldsymbol{y}^{(2)} \neq \mathbf{0} \end{cases}$$

Case I:

If $\boldsymbol{y}^{(2)} = \mathbf{0}$, there are two possibilities. One is $\boldsymbol{y}^{(1)} = \mathbf{0}$. We could get the equation from (3.7).

$$Pr\{\boldsymbol{Y}^{(1)} = \mathbf{0} | \boldsymbol{Y}^{(2)} = \mathbf{0}\} = \frac{\omega + (1-\omega)\prod_{j=1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j}}{\omega + (1-\omega)\prod_{j=k+1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j}}$$

The other possibility is $\boldsymbol{y}^{(1)} \neq \mathbf{0}$. From (3.7) we have

$$Pr\{\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)} | \boldsymbol{Y}^{(2)} = \mathbf{0}\} = \frac{(1-\omega)\prod_{j=1}^{p} (\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}}{\omega + (1-\omega)\prod_{j=k+1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j}}$$

Under this situation, we construct a new parameter $\tau = \frac{\omega \prod_{j=k+1}^{p} \alpha_j^{-1/2} e^{\alpha_j \mu_j}}{\omega \prod_{j=k+1}^{p} \alpha_j^{-1/2} e^{\alpha_j \mu_j} + 1 - \omega}$, this will produce $Pr\{\boldsymbol{Y}^{(1)} = \mathbf{0} | \boldsymbol{Y}^{(2)} = \mathbf{0}\}$ with the same pattern as (3.2), and $Pr\{\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)} | \boldsymbol{Y}^{(2)} = \mathbf{0}\}$ with the same pattern as (3.3). Summarizing the situation indicated above, the new representation with parameter τ will be,

$$Pr\{\boldsymbol{Y}^{(1)} = \mathbf{0} | \boldsymbol{Y}^{(2)} = \mathbf{0}\} = \tau + (1 - \tau) \prod_{j=1}^{k} \alpha_j^{1/2} e^{-\alpha_j \mu_j} \tag{3.8}$$

$$\begin{aligned} Pr\{\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)} | \boldsymbol{Y}^{(2)} = \mathbf{0}\} &= \frac{(1 - \omega) \prod_{j=1}^{k} (\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e \mu_j}{y_j})^{\alpha_j y_j}}{\omega \prod_{j=k+1}^{p} \alpha_j^{-1/2} e^{\alpha_j \mu_j} + 1 - \omega} \\ &= (1 - \tau) \prod_{j=1}^{k} (\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e \mu_j}{y_j})^{\alpha_j y_j} \end{aligned} \tag{3.9}$$

By combining (3.8) and (3.9), we could get

$$\begin{aligned} \Pr\{\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)} | \boldsymbol{Y}^{(2)} = \mathbf{0}\} = [\tau + (1 - \tau) \prod_{j=1}^{k} \alpha_j^{1/2} e^{-\alpha_j \mu_j}] \boldsymbol{I}(\boldsymbol{Y}^{(1)} = \mathbf{0}) \\ + (1 - \tau) \prod_{j=1}^{k} (\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e \mu_j}{y_j})^{\alpha_j y_j} \boldsymbol{I}(\boldsymbol{Y}^{(1)} \neq \mathbf{0}) \end{aligned} \tag{3.10}$$

Thus the conditional distribution could be described as

$$\boldsymbol{Y}^{(1)} | (\boldsymbol{Y}^{(2)} = \mathbf{0}) \sim \text{ZIDP}(\tau; \mu_1, \alpha_1, ..., \mu_k, \alpha_k)$$

Case II:

As $\boldsymbol{y}^{(2)} \neq \mathbf{0}$, we could easily get $\boldsymbol{y} \neq \mathbf{0}$. Thus $Pr\{\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)} | \boldsymbol{Y}^{(2)} = \boldsymbol{y}^{(2)}\}$ will be

showed in the following representative by (3.7)

$$Pr\{\boldsymbol{Y}^{(1)}=\boldsymbol{y}^{(1)}|\boldsymbol{Y}^{(2)}=\boldsymbol{y}^{(2)}\}=\frac{(1-\omega)\prod_{j=1}^{p}(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_j}y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}}{(1-\omega)\prod_{j=k+1}^{p}(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_j}y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}}$$
$$=\prod_{j=1}^{k}(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_j}y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}$$

From equation above, we can point out that the conditional distribution of $Pr\{\boldsymbol{Y}^{(1)}=\boldsymbol{y}^{(1)}|\boldsymbol{Y}^{(2)}=\boldsymbol{y}^{(2)}\}$ has no relationship with Z, thereby $\boldsymbol{Y}^{(1)}|\boldsymbol{Y}^{(2)}\overset{d}{=}\boldsymbol{X}^{(1)}$. Particularly, when $\boldsymbol{y}^{(2)}\neq\boldsymbol{0}$, $(Y_1,...,Y_k)$ are independent with each other and $Y_j|\boldsymbol{Y}^{(2)}\overset{d}{=}X_j\sim$ DP(μ_j,α_j), is not depended on parameter ω.

3.4.2 Conditional Distributions of $Z|\boldsymbol{Y}$

Since the common component $Z\sim$ Bernoulli$(1-\omega)$, it has two results only: 1 or 0. Therefore,

$$Pr(Z=1|\boldsymbol{Y}=\boldsymbol{y})=\frac{Pr(Z=1,\boldsymbol{Y}=\boldsymbol{y})}{f(\boldsymbol{y}|\omega,\boldsymbol{m},\boldsymbol{p})}$$
$$=\frac{(1-\omega)\prod_{j=1}^{p}(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_j}y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}}{[\omega+(1-\omega)\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}]\mathbf{I}(\boldsymbol{y}=\boldsymbol{0})+[(1-\omega)\prod_{j=1}^{p}(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_j}y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}]\mathbf{I}(\boldsymbol{y}\neq\boldsymbol{0})}$$

Under the situation $\boldsymbol{y}\neq\boldsymbol{0}$, the equation can be simplify into

$$Pr(Z=1|\boldsymbol{Y}=\boldsymbol{y})=1$$

Under the situation of $\boldsymbol{y} = \mathbf{0}$, the next conditional distribution will be

$$Pr(Z = 1|\boldsymbol{Y} = \boldsymbol{y}) = \frac{(1-\omega)\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}}{\omega + (1-\omega)\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}}$$

Thus

$$Z|(\boldsymbol{Y} = \boldsymbol{y}) \sim \begin{cases} Bernoulli(\eta) & \text{if } \boldsymbol{y} = \mathbf{0} \\ Degenerate(1) & \text{if } \boldsymbol{y} \neq \mathbf{0} \end{cases} \tag{3.11}$$

The parameter η is denoted as,

$$\eta = \frac{(1-\omega)\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}}{\omega + (1-\omega)\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}} \tag{3.12}$$

3.4.3 Conditional Distribution of $\boldsymbol{X}|\boldsymbol{Y}$

There are two possibilities existing. One is $\boldsymbol{y} = \mathbf{0}$ and the other is $\boldsymbol{y} \neq \mathbf{0}$. Then the conditional distribution of $\boldsymbol{X}|\boldsymbol{Y}$ should be obtained.

When $\boldsymbol{y} = \mathbf{0}$, we obtain

$$\begin{aligned} Pr\{\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{Y} = \mathbf{0}\} =& \frac{Pr\{\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \mathbf{0}\}}{Pr\{\boldsymbol{Y} = \mathbf{0}\}} \\ =& \frac{Pr\{\boldsymbol{X} = \mathbf{0}, \boldsymbol{Y} = \mathbf{0}\}}{f(\mathbf{0}|\omega, \boldsymbol{m}, \boldsymbol{p})}\boldsymbol{I}(\boldsymbol{x} = \mathbf{0}) + \frac{Pr\{\boldsymbol{X} = \boldsymbol{x}, Z = 0\}}{f(\mathbf{0}|\omega, \boldsymbol{m}, \boldsymbol{p})}\boldsymbol{I}(\boldsymbol{x} \neq \mathbf{0}) \\ =& \frac{Pr\{\boldsymbol{X} = \mathbf{0}\}}{f(\mathbf{0}|\omega, \boldsymbol{m}, \boldsymbol{p})}\boldsymbol{I}(\boldsymbol{x} = \mathbf{0}) + \frac{\omega Pr\{\boldsymbol{X} = \boldsymbol{x}\}}{f(\mathbf{0}|\omega, \boldsymbol{m}, \boldsymbol{p})}\boldsymbol{I}(\boldsymbol{x} \neq \mathbf{0}) \end{aligned}$$

According to equation (3.4),

$$= \frac{\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}}{\omega + (1-\omega)\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}}\boldsymbol{I}(\boldsymbol{x} = \mathbf{0}) + \frac{\omega\prod_{j=1}^{p}(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_j}y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}}{\omega + (1-\omega)\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}}\boldsymbol{I}(\boldsymbol{x} \neq \mathbf{0})$$

Then substitute with η from (3.12),

$$= [\eta+(1-\eta)\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}]\boldsymbol{I}(\boldsymbol{x}=\boldsymbol{0})+[(1-\eta)\prod_{j=1}^{p}(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_j}y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}]\boldsymbol{I}(\boldsymbol{x}\neq\boldsymbol{0})$$

here, in accordance with η, we obtain

$$\boldsymbol{X}|(\boldsymbol{Y}=\boldsymbol{0}) \sim \text{ZIDP}(\eta;\mu_1,\alpha_1,...,\mu_p,\alpha_p) \tag{3.13}$$

When $\boldsymbol{y}\neq\boldsymbol{0}$, we have

$$\begin{aligned} Pr\{\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y}\} &= \frac{Pr\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}}{Pr\{\boldsymbol{Y}=\boldsymbol{y}\}} \\ &= \frac{Pr\{\boldsymbol{X}=\boldsymbol{y},Z=1\}}{f(\boldsymbol{y}|\omega,\boldsymbol{m},\boldsymbol{p})} \\ &= 1 \end{aligned}$$

When $\boldsymbol{Y}=\boldsymbol{y}\neq\boldsymbol{0}$ is specified, $X_1,...,X_p$ are independent with each other. Therefore we have

$$X_j|(\boldsymbol{Y}=\boldsymbol{y}\neq\boldsymbol{0}) \sim \text{Degenerate}(y_j), \quad j=1,...,p \tag{3.14}$$

3.4.4 Conditional Distribution of $X_j|(Y_j=y_j=0),\ j=1,...,p$

From (3.6), the property of $Y_j \sim \text{ZIDP}(\omega,\mu_j,\alpha_j)$ for $j=1,...,p$ is introduced. Then we have

$$\begin{aligned} Pr\{X_j=x_j|Y_j=0\} =& \frac{Pr\{X_j=x_j,Y_j=0\}}{Pr\{Y_j=0\}} \\ =& \frac{Pr\{X_j=0,Y_j=0\}}{f(0|\omega,\mu_j,\alpha_j)}\boldsymbol{I}(x_j=0)+\frac{Pr\{X_j=x_j,Z=0\}}{f(0|\omega,\mu_j,\alpha_j)}\boldsymbol{I}(x_j>0) \\ =& \frac{Pr\{X_j=0\}}{f(0|\omega,\mu_j,\alpha_j)}\boldsymbol{I}(x_j=0)+\frac{\omega Pr\{X_j=x_j\}}{f(0|\omega,\mu_j,\alpha_j)}\boldsymbol{I}(x_j>0) \end{aligned}$$

From (3.1) we have,

$$= \frac{\alpha_j^{1/2} e^{-\alpha_j \mu_j}}{\omega + (1-\omega)\alpha_j^{1/2} e^{-\alpha_j \mu_j}} \boldsymbol{I}(x_j = 0) + \frac{\omega(\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j}}{\omega + (1-\omega)\alpha_j^{1/2} e^{-\alpha_j \mu_j}} \boldsymbol{I}(x_j > 0)$$

Let $\eta_j^* = \frac{(1-\omega)\alpha_j^{1/2} e^{-\alpha_j \mu_j}}{\omega + (1-\omega)\alpha_j^{1/2} e^{-\alpha_j \mu_j}}$, then we get the equation below.

$$= [\eta_j^* + (1-\eta_j^*)\alpha_j^{1/2} e^{-\alpha_j \mu_j}] \boldsymbol{I}(x_j = 0) + (1-\eta_j^*)(\alpha_j^{1/2} e^{-\alpha_j \mu_j})(\frac{e^{-y_j} y_j^{y_j}}{y_j!})(\frac{e\mu_j}{y_j})^{\alpha_j y_j} \boldsymbol{I}(x_j > 0)$$

Thus, we have

$$X_j|(Y_j = 0) \sim \text{ZIDP}(\eta_j^*, \mu_j, \alpha_j) \quad \text{where} \quad \eta_j^* = \frac{(1-\omega)\alpha_j^{1/2} e^{-\alpha_j \mu_j}}{\omega + (1-\omega)\alpha_j^{1/2} e^{-\alpha_j \mu_j}} \tag{3.15}$$

3.4.5 Conditional Distribution of $X_j|(Y_j = y_j > 0),\ j = 1, ..., p$

Here we have

$$\begin{aligned} Pr\{X_j = x_j | Y_j = y_j\} &= \frac{Pr\{X_j = x_j, Y_j = y_j\}}{Pr\{Y_j = y_j\}} \\ &= \frac{Pr\{X_j = y_j, Z = 1\}}{f(y_j|\omega, \mu_j, \alpha_j)} \\ &= 1 \end{aligned}$$

Therefore, $X_j|(Y_j = y_j > 0) \sim \text{Degenerate}(y_j)$ could be achieved.

3.5 Likelihood Function

Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ are random sample of size n from a population of the p-dimensional zero-inflated double Poisson distribution $(\omega, \mu_1, \alpha_1, ..., \mu_p, \alpha_p)$, then $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{ip})$

for $i = 1, ..., n$. Each of y_{ij} corresponds with a double Poisson parameter μ_j, α_j. Now the realization of the random vector $\boldsymbol{Y}_i$ is denoted by $\boldsymbol{y}_i = (y_{i1}, ..., y_{ip})^\top$, then the observed data frame can be represented as $\boldsymbol{Y}_{obs} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_n\}^\top$.

Hence, the observed random variable of $\boldsymbol{Y}$ would has the following form

$$\boldsymbol{Y}_{(obs)} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}$$

Note: all R program of this thesis are depended on the data frame indicated above.

Since $\mathbf{Y_i}$=$(Y_{i1}, ..., Y_{ip})$ are the random sample of $\mathbf{Y}$=$(Y_1, ..., Y_p)$ = $\mathbf{Z}$ $(X_1, ..., X_p)$, where $\mathbf{Z} \sim$ Bernoulli$(1-\omega)$. $X_j \sim$ DP (μ_j, α_j), $j = 1, 2, ..., p$. From the equation (3.4), we define the likelihood function of multivariate zero-inflated double Poisson distribution with parameters $(\omega,\ \boldsymbol{\mu} = (\mu_1, ..., \mu_p),\ \boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p))$ as

$$\begin{aligned} \mathrm{L}(\omega, \boldsymbol{\mu}, \boldsymbol{\alpha}|Y_{obs}) = \prod_{i=1}^{n} \{ & [\omega + (1-\omega) \prod_{j=1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j}] \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) \\ & + [(1-\omega) \prod_{j=1}^{p} (\alpha_j^{1/2} e^{-\alpha_j \mu_j}) (\frac{e^{-y_{ij}} y_{ij}^{y_{ij}}}{y_{ij}!}) (\frac{e\mu_j}{y_{ij}})^{\alpha_j y_{ij}}] \boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0}) \} \end{aligned}$$

Now, for convenience let $\gamma = \frac{\omega}{1-\omega}$ and $\omega = \frac{\gamma}{1+\gamma}$, when $\gamma = 0 \Leftrightarrow \omega = 0$. Then the likelihood function in term of parameter θ is (where $\theta = (\gamma, \boldsymbol{\mu}, \boldsymbol{\alpha})$)

$$\begin{aligned} \mathrm{L}(\theta|Y_{obs}) = \prod_{i=1}^{n} \frac{1}{1+\gamma} \{ & [\gamma + \prod_{j=1}^{p} \alpha_j^{1/2} e^{-\alpha_j \mu_j}] \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) \\ & + \prod_{j=1}^{p} (\alpha_j^{1/2} e^{-\alpha_j \mu_j}) (\frac{e^{-y_{ij}} y_{ij}^{y_{ij}}}{y_{ij}!}) (\frac{e\mu_j}{y_{ij}})^{\alpha_j y_{ij}} \boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0}) \}. \end{aligned} \tag{3.16}$$

Hence by taking the logarithm on both sides, the log likelihood function with parameter θ is

$$\begin{aligned}
\ell(\theta|Y_{obs}) &= \ln(\frac{1}{1+\gamma})^n + \sum_{i=1}^{n}\{\ln[\gamma + \prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}]\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) \\
&\quad + \ln[\prod_{j=1}^{p}(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_{ij}}y_{ij}^{y_{ij}}}{y_{ij}!})(\frac{e\mu_j}{y_{ij}})^{\alpha_j y_{ij}}]\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\} \\
&= -n\,\ln(1+\gamma) + \sum_{i=1}^{n}\{\ln[\gamma + \prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}]\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) \\
&\quad + \sum_{j=1}^{p}[\ln[(\alpha_j^{1/2}e^{-\alpha_j\mu_j})(\frac{e^{-y_{ij}}y_{ij}^{y_{ij}}}{y_{ij}!})(\frac{e\mu_j}{y_{ij}})^{\alpha_j y_{ij}}]]\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\} \qquad (3.17) \\
&= -n\,\ln(1+\gamma) + \sum_{i=1}^{n}\{\ln[\gamma + \prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}]\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\} \\
&\quad + \sum_{i=1}^{n}\{\sum_{j=1}^{p}[\frac{1}{2}\ln\alpha_j - \alpha_j\mu_j - y_{ij} + y_{ij}\ln y_{ij} - \ln y_{ij}! \\
&\quad + \alpha_j y_{ij}(1 + \ln\mu_j - \ln y_{ij})]\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\}
\end{aligned}$$

3.6 Score Test For Zero-Inflation

In this section, the score statistic for testing the presence of zero-inflation in multivariate zero-inflated double Poisson model will be conducted.

Testing $H_0 : \omega = 0$ is equivalent to test

$$H_0 : \gamma = 0 \quad \text{against} \quad H_1 : \gamma > 0 \qquad (3.18)$$

Based on log likelihood function, the score function $U_\gamma(\theta)$ is the first order derivatives of $\ell(\theta)$ with respect to the parameter γ.

$$U_\gamma(\theta) = \frac{\partial\ell(\theta|Y_{obs})}{\partial\gamma}.$$

In addition, the fisher information matrix with the parameter θ is

$$\mathbb{I}(\theta) = \begin{pmatrix} -\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma^2}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma \partial \mu_1}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma \partial \mu_p}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma \partial \alpha_1}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma \partial \alpha_p}), \\ -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_1 \partial \gamma}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_1^2}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_p}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_1 \partial \alpha_1}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_1 \partial \alpha_p}) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots, \\ -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_p \partial \gamma}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_p \partial \mu_1}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_p^2}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_p \partial \alpha_1}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_p \partial \alpha_p}), \\ -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_1 \partial \gamma}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_1 \partial \mu_1}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_1 \partial \mu_p}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_1^2}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_1 \partial \alpha_p}), \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_p \partial \gamma}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_p \partial \mu_1}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_p \partial \mu_p}) & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_p \partial \alpha_1}) & \cdots & -\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_p^2}). \end{pmatrix}$$

For convenience, we define

$$\mathbf{F_{0i}} = \prod_{\mathbf{j=1}}^{\mathbf{p}} \alpha_{\mathbf{j}}^{\mathbf{1/2}} \mathbf{e}^{-\alpha_{\mathbf{j}} \mu_{\mathbf{j}}},$$

then we get

$$\frac{\partial \ell(\theta|Y_{obs})}{\partial \gamma} = -\frac{n}{1+\gamma} + \sum_{i=1}^{n} \frac{\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})}{\gamma + \mathbf{F_{0i}}},$$

$$\frac{\partial \ell(\theta|Y_{obs})}{\partial \mu_j} = \sum_{i=1}^{n} \{\frac{-\alpha_j \mathbf{F_{0i}}}{\gamma + \mathbf{F_{0i}}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) + (-\alpha_j + \frac{\alpha_j y_{ij}}{\mu_j}) \boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\},$$

$$\begin{aligned}\frac{\partial \ell(\theta|Y_{obs})}{\partial \alpha_j} = \sum_{i=1}^{n} \{&\frac{\mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{j}} - \mu_\mathbf{j})}{\gamma + \mathbf{F_{0i}}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) \\ &+ [\frac{1}{2\alpha_j} - \mu_j + y_{ij}(1 + \ln \mu_j - \ln y_{ij})] \boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\},\end{aligned}$$

$$\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \gamma^2} = \frac{n}{(1+\gamma)^2} + \sum_{i=1}^{n} \{-\frac{\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})}{(\gamma + \mathbf{F_{0i}})^\mathbf{2}}\},$$

$$\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \mu_j^2} = \sum_{i=1}^{n} \{\frac{\gamma \alpha_j^2 \mathbf{F_{0i}}}{(\gamma + \mathbf{F_{0i}})^\mathbf{2}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) - \frac{\alpha_j y_{ij}}{\mu_j^2} \boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\},$$

$$\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \alpha_j^2} = \sum_{i=1}^{n} \{\frac{\gamma \mathbf{F_{0i}}(\frac{\mathbf{-1}}{\mathbf{4}\alpha_\mathbf{j}^\mathbf{2}} + \mu_\mathbf{j}^\mathbf{2} - \frac{\mu_\mathbf{j}}{\alpha_\mathbf{j}}) - \frac{\mathbf{F_{0i}^2}}{\mathbf{2}\alpha_\mathbf{j}^\mathbf{2}}}{(\gamma + \mathbf{F_{0i}})^\mathbf{2}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) - \frac{1}{2}\alpha_j^{-2} \boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\},$$

$$\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \gamma \partial \mu_j} = \sum_{i=1}^{n} \{\frac{\alpha_j \mathbf{F_{0i}}}{(\gamma + \mathbf{F_{0i}})^\mathbf{2}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\},$$

$$\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \gamma \partial \alpha_j} = \sum_{i=1}^{n} \{\frac{-\mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{j}} - \mu_\mathbf{j})}{(\gamma + \mathbf{F_{0i}})^\mathbf{2}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\},$$

$$\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \mu_j \partial \mu_k}(j \neq k) = \sum_{i=1}^{n} \{\frac{(\gamma + \mathbf{F_{0i}})\alpha_\mathbf{j}\alpha_\mathbf{k}\mathbf{F_{0i}} - \alpha_\mathbf{j}\alpha_\mathbf{k}\mathbf{F_{0i}^2}}{(\gamma + \mathbf{F_{0i}})^\mathbf{2}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\},$$

$$\begin{aligned}\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \mu_j \partial \alpha_j} = \sum_{i=1}^{n} \{&\frac{(\gamma + \mathbf{F_{0i}})(-\frac{\mathbf{3}}{\mathbf{2}}\mathbf{F_{0i}} + \alpha_\mathbf{j}\mathbf{F_{0i}}\mu_\mathbf{j}) + \alpha_\mathbf{j}\mathbf{F_{0i}}^\mathbf{2}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{j}} - \mu_\mathbf{j})}{(\gamma + \mathbf{F_{0i}})^\mathbf{2}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\} \\ &+ \sum_{i=1}^{n} \{(-1 + \frac{y_{ij}}{\mu_j}) \boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\},\end{aligned}$$

$$\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \mu_j \partial \alpha_k}(j \neq k) = \sum_{i=1}^{n} \{\frac{-\alpha_j \mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{k}} - \mu_\mathbf{k})(\gamma + \mathbf{F_{0i}}) + \alpha_\mathbf{j}\mathbf{F_{0i}}^\mathbf{2}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{k}} - \mu_\mathbf{k})}{(\gamma + \mathbf{F_{0i}})^2} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\},$$

$$\frac{\partial^2 \ell(\theta|Y_{obs})}{\partial \alpha_j \partial \alpha_k}(j \neq k) = \sum_{i=1}^{n} \{\frac{\gamma \mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{j}} - \mu_\mathbf{j})(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{k}} - \mu_\mathbf{k})}{(\gamma + \mathbf{F_{0i}})^\mathbf{2}} \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\}.$$

We will get the expectation of indicator function of y_i.

$$\mathbf{E}[\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})] = \omega + (1-\omega)\mathbf{F_{0i}} = \frac{\gamma}{1+\gamma} + \frac{1}{1+\gamma}\mathbf{F_{0i}} = \frac{\gamma + \mathbf{F_{0i}}}{1+\gamma},$$
$$\mathbf{E}[\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})] = 1 - \frac{\gamma + \mathbf{F_{0i}}}{1+\gamma} = \frac{1 - \mathbf{F_{0i}}}{1+\gamma}.$$

Thus, the negative expectation of second-order derivatives are shown as given,

$$\begin{aligned}
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma^2}) &= -\frac{n}{(1+\gamma)^2} + \sum_{i=1}^{n} \frac{\mathbf{E}[\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})]}{(\gamma + \mathbf{F_{i0}})^{\mathbf{2}}} \\
&= -\frac{n}{(1+\gamma)^2} + \sum_{i=1}^{n} \{\frac{1}{1+\gamma}(\gamma + \mathbf{F_{0i}})^{\mathbf{-1}}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma \partial \mu_j}) &= \sum_{i=1}^{n} \{\frac{-\alpha_j \mathbf{F_{0i}}}{(1+\gamma)(\gamma + \mathbf{F_{0i}})}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma \partial \alpha_j}) &= \sum_{i=1}^{n} \{\frac{\mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{j}} - \mu_\mathbf{j})}{(1+\gamma)(\gamma + \mathbf{F_{0i}})}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_j^2}) &= \sum_{i=1}^{n} \{\frac{1}{1+\gamma}(-\alpha_j^2 \mathbf{F_{0i}} + \frac{\alpha_\mathbf{j}^\mathbf{2} \mathbf{F_{0i}}^\mathbf{2}}{\gamma + \mathbf{F_{0i}}} + \frac{\alpha_\mathbf{j}}{\mu_\mathbf{j}})\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_j \partial \mu_k}) &= \sum_{i=1}^{n} \{\frac{\alpha_j \alpha_k \mathbf{F_{0i}}}{1+\gamma}(-1 + \frac{\mathbf{F_{0i}}}{\gamma + \mathbf{F_{0i}}})\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_j \partial \alpha_j}) &= \sum_{i=1}^{n} \{\frac{1}{1+\gamma}(\frac{1}{2}\mathbf{F_{0i}} - \alpha_\mathbf{j}\mathbf{F_{0i}}\mu_\mathbf{j} - \frac{\mathbf{F_{0i}}^\mathbf{2}(\frac{\mathbf{1}}{\mathbf{2}} - \alpha_\mathbf{j}\mu_\mathbf{j})}{\gamma + \mathbf{F_{0i}}})\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_j \partial \alpha_k}) &= \sum_{i=1}^{n} \{\frac{\alpha_j \mathbf{F_{0i}}}{1+\gamma}(\frac{1}{2\alpha_k} - \mu_k)\frac{\gamma}{\gamma + \mathbf{F_{0i}}}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_j \partial \alpha_k}) &= \sum_{i=1}^{n} \{\frac{-\gamma \mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{j}} - \mu_\mathbf{j})(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{k}} - \mu_\mathbf{k})}{(1+\gamma)(\gamma + \mathbf{F_{0i}})}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_j^2}) &= \sum_{i=1}^{n} \{\frac{1}{1+\gamma}[\mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{4}\alpha_\mathbf{j}^\mathbf{2}} - \mu_\mathbf{j}^\mathbf{2} + \frac{\mu_\mathbf{j}}{\alpha_\mathbf{j}}) + \mathbf{F_{0i}}^\mathbf{2}(\frac{\mathbf{1}}{\mathbf{4}\alpha_\mathbf{j}^\mathbf{2}} + \mu_\mathbf{j}^\mathbf{2} - \frac{\mu_\mathbf{j}}{\alpha_\mathbf{j}})(\gamma + \mathbf{F_{0i}})^{\mathbf{-1}}] \\
&+ \frac{1}{2}\alpha_j^{-2}(\frac{1 - \mathbf{F_{0i}}}{1+\gamma})\}.
\end{aligned}$$

Under the null hypothesis, we should easily get

$$\begin{aligned}
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma_0^2}) &= -n + \sum_{i=1}^{n} \{\mathbf{F_{0i}^{-1}}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma_0 \partial \mu_{0_j}}) &= \sum_{i=1}^{n} \{-\alpha_{0_j}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \gamma_0 \partial \alpha_{0_j}}) &= \sum_{i=1}^{n} \{\frac{1}{2\alpha_{0_j}} - \mu_{0_j}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_{0_j}^2}) &= \sum_{i=1}^{n} \{\frac{\alpha_{0_j}}{\mu_{0_j}}\}, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_{0_j} \partial \mu_{0_k}}) &= 0, \qquad (j \neq k) \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_{0_j} \partial \alpha_{0_j}}) &= 0, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \mu_{0_j} \partial \alpha_{0_k}}) &= 0, \qquad (j \neq k) \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_{0_j} \partial \alpha_{0_k}}) &= 0, \\
-\mathbf{E}(\frac{\partial^2 \ell}{\partial \alpha_{0_j}^2}) &= \sum_{i=1}^{n} \{\frac{1}{2} \alpha_{0_j}^2\}.
\end{aligned}$$

Under the null hypothesis, let $\hat{\theta}_0 = (0, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\alpha}}_0)^T$ denotes the MLEs of θ, then the information matrix $\mathbb{I}(\hat{\theta}_0)$ could be partitioned as the following form.

$$\mathbb{I}(\hat{\theta}_0) = \begin{pmatrix} -n + \sum_{i=1}^{n}\{\mathbf{F_{0i}^{-1}}\} & \cdots & \sum_{i=1}^{n}\{-\alpha_{0j}\} & \cdots & \cdots & \sum_{i=1}^{n}\{\frac{1}{2\alpha_{0j}} - \mu_{0j}\} & \cdots \\ \hdashline \vdots & \ddots & & & & & \\ \sum_{i=1}^{n}\{-\alpha_{0j}\} & & \sum_{i=1}^{n}\{\frac{\alpha_{0j}}{\mu_{0_j}}\} & & & 0 & \\ \vdots & & & \ddots & & & \\ \vdots & & & & \ddots & & \\ \sum_{i=1}^{n}\{\frac{1}{2\alpha_{0j}} - \mu_{0j}\} & & 0 & & & \sum_{i=1}^{n}\{\frac{1}{2}\alpha_{0j}^{2}\} & \\ \vdots & & & & & & \ddots \end{pmatrix}_{(2p+1)\times(2p+1)}$$

By setting $\frac{\partial \ell}{\partial \mu_j} = 0$ and $\frac{\partial \ell}{\partial \alpha_j} = 0$ $(j = 1, 2, ..., p)$, we should get the maximum likelihood estimation.

For $\hat{\boldsymbol{\mu}}_0$,

$$\frac{\partial \ell}{\partial \mu_j} = 0,$$

$$\sum_{i=1}^{n}\{\frac{-\alpha_j \mathbf{F_{0i}}}{\gamma + \mathbf{F_{0i}}}\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) + (-\alpha_j + \frac{\alpha_j y_{ij}}{\mu_j})\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\} = 0,$$

$$\sum_{i=1}^{n}\{-\alpha_j \boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\} + (-\alpha_j + \frac{\alpha_j y_{ij}}{\mu_j})\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\} = 0,$$

$$\sum_{i=1}^{n}\{-\alpha_j + \frac{\alpha_j y_{ij}}{\mu_j}\} = 0,$$

$$-n\alpha_j + \frac{\alpha_j}{\mu_j}\sum_{i=1}^{n} y_{ij} = 0,$$

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} y_{ij}}{n}.$$

For $\hat{\boldsymbol{\alpha}}_0$,

$$\frac{\partial \ell}{\partial \alpha_j} = 0,$$

$$\sum_{i=1}^{n}\{\frac{\mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{2\alpha_j}} - \mu_\mathbf{j})}{\gamma + \mathbf{F_{0i}}}\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) + [\frac{1}{2\alpha_j} - \mu_j + y_{ij}(1 + \ln \mu_j - \ln y_{ij})]\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\} = 0,$$

$$\sum_{i=1}^{n}\{(\frac{1}{2\alpha_j} - \mu_j)\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) + [\frac{1}{2\alpha_j} - \mu_j + y_{ij}(1 + \ln \mu_j - \ln y_{ij})]\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\} = 0,$$

$$\sum_{i=1}^{n}\{\frac{1}{2\alpha_j} - \mu_j + y_{ij}(1 + \ln \mu_j - \ln y_{ij})\} = 0,$$

$$\hat{\alpha}_j = \frac{n}{2\{n\mu_j - \sum_{i=1}^{n}\{y_{ij} - y_{ij}\ln y_{ij} + y_{ij}\ln \mu_j\}\}}, \hat{\alpha}_j \qquad = \frac{n}{\sum_{i=1}^{n} 2(y_{ij}\ln y_i}$$

Then the score test statistic will be given.

$$SC = \left\{ U_\gamma^\top(\theta)\, \mathbb{J}^{-1}(\theta)\, U_\gamma(\theta) \right\} \big|_{\hat{\theta_0}}.$$

For the hypothesis, γ is the interest parameter. Let $\theta_1 = (\boldsymbol{\mu}, \boldsymbol{\alpha})$ is the nuisance parameter. Thus, the information matrix could be partitioned into

$$\mathbb{I}(\theta) = \begin{pmatrix} I_{\gamma\gamma} & I_{\gamma\theta_1} \\ I_{\theta_1\gamma} & I_{\theta_1\theta_1} \end{pmatrix}_{(2p+1)\times(2p+1)}$$

Therefore,

$$\mathbb{J}^{-1}(\theta) = (I_{\gamma\gamma} - I_{\gamma\theta_1} I_{\theta_1\theta_1}^{-1} I_{\theta_1\gamma})^{-1}.$$

Under H_0 (3.18), the score function will be

$$\begin{aligned} U_\gamma(\theta) &= \frac{\partial \ell(\theta|Y_{obs})}{\partial \gamma}\bigg|_{\hat{\theta_0}} \\ &= \left\{ -\frac{n}{1+\gamma} + \sum_{i=1}^{n} \frac{\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})}{\gamma + \mathbf{F_{0i}}} \right\} \bigg|_{\hat{\theta_0}} \\ &= -n + \sum_{i=1}^{n} \frac{\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})}{\mathbf{F_{0i}}}. \end{aligned} \tag{3.19}$$

$$\begin{aligned} I_{\gamma\gamma} &= -\mathbf{E}\left(\frac{\partial^2 \ell}{\partial \gamma^2}\right)\bigg|_{\hat{\theta_0}} \\ &= \left\{ -\frac{n}{(1+\gamma)^2} + \sum_{i=1}^{n} \{ \frac{1}{1+\gamma} (\gamma + \mathbf{F_{0i}})^{-\mathbf{1}} \} \right\} \bigg|_{\hat{\theta_0}} \\ &= -n + \sum_{i=1}^{n} \{ \mathbf{F_{0i}}^{-\mathbf{1}} \}. \end{aligned} \tag{3.20}$$

Using (3.19) and (3.20), the score test statistic is obtained.

$$\begin{aligned} SC &= \left\{U_\gamma^\top(\theta)\, \mathbb{J}^{-1}(\theta)\, U_\gamma(\theta)\right\}\Big|_{\hat{\theta_0}} \\ &= \left\{U_\gamma(\theta)^2 \mathbb{J}^{-1}(\theta)\right\}\Big|_{\hat{\theta_0}} \\ &= \left\{-n + \sum_{i=1}^{n} \frac{\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})}{\mathbf{F_{0i}}}\right\}^2 \left(-n + \sum_{i=1}^{n} {\mathbf{F_{0i}}}^{-1} - I_{\gamma\theta_1} I_{\theta_1\theta_1}^{-1} I_{\theta_1\gamma}\right)^{-1} \sim \chi^2(1). \end{aligned}$$

Thereby the corresponding p-value is as follows.

$$P_1 = \Pr(SC > z_1 | H_0) = \Pr\{\chi^2(1) > z_1\}. \tag{3.21}$$

When $P_1 < \alpha$, the null hypothesis H_0 can be rejected at the significance level of α.

3.7 Likelihood Ratio Test for Zero-Inflation

It is a good way to consider Likelihood Ratio Test to evaluate which is more applicable for current data analysis between two models. In large sample, if there is a nested relationship between two models, the difference between the log likelihood of two models times -2 approximates the chi-square distribution. If the result is small, it can be considered that the variable has little effect on the interpretation of the dependent variable, so we can refuse to introduce this variable into the model.

Under the same null hypothesis(3.18), the likelihood ratio test statistic can be expressed by following equation.

$$LR = -2\{\ell(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\alpha}}_0 | Y_{obs}) - \ell(\hat{\gamma}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}} | Y_{obs})\} \sim \chi^2(1), \tag{3.22}$$

where the MLEs of parameter $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$ under H_0 are $\hat{\boldsymbol{\mu}}_0 = (\frac{\sum_{i=1}^n y_{i1}}{n}, ..., \frac{\sum_{i=1}^n y_{ip}}{n})$, $\hat{\boldsymbol{\alpha}}_0 = (\frac{n}{\sum_{i=1}^n 2(y_{i1}\ln y_{i1} - y_{i1}\ln \mu_1)}, ..., \frac{n}{\sum_{i=1}^n 2(y_{ip}\ln y_{ip} - y_{ip}\ln \mu_p)})$, $(\hat{\gamma}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})$ are the unconstrained MLEs of $(\gamma, \boldsymbol{\mu}, \boldsymbol{\alpha})$. The log likelihood function is given respectively for two models.

$$\begin{aligned}\ell(\gamma, \boldsymbol{\mu}, \boldsymbol{\alpha}) = -n\,\ln(1+\gamma) + \sum_{i=1}^{n}\{\ln[\gamma + \prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}]\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\} \\ + \sum_{i=1}^{n}\{\sum_{j=1}^{p}[\frac{1}{2}\ln\alpha_j - \alpha_j\mu_j - y_{ij} + y_{ij}\ln y_{ij} - \ln y_{ij}! \\ + \alpha_j y_{ij}(1 + \ln\mu_j - \ln y_{ij})]\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\},\end{aligned}$$

where the $(\hat{\gamma}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})$ can be calculated by the maximum likelihood equations,

$$\begin{aligned}\frac{\partial\ell(\theta|Y_{obs})}{\partial\gamma} &= -\frac{n}{1+\gamma} + \sum_{i=1}^{n}\frac{\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})}{\gamma + \mathbf{F_{0i}}} = 0, \\ \frac{\partial\ell(\theta|Y_{obs})}{\partial\mu_j} &= \sum_{i=1}^{n}\{\frac{-\alpha_j\mathbf{F_{0i}}}{\gamma + \mathbf{F_{0i}}}\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) + (-\alpha_j + \frac{\alpha_j y_{ij}}{\mu_j})\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\} = 0, \\ \frac{\partial\ell(\theta|Y_{obs})}{\partial\alpha_j} &= \sum_{i=1}^{n}\{\frac{\mathbf{F_{0i}}(\frac{\mathbf{1}}{\mathbf{2}\alpha_\mathbf{j}} - \mu_\mathbf{j})}{\gamma + \mathbf{F_{0i}}}\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0}) \\ &\quad + [\frac{1}{2\alpha_j} - \mu_j + y_{ij}(1 + \ln\mu_j - \ln y_{ij})]\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\} = 0.\end{aligned}$$

While $\gamma = 0$,

$$\begin{aligned}\ell(\boldsymbol{\mu}_0, \boldsymbol{\alpha}_0) = \sum_{i=1}^{n}\{\ln[\prod_{j=1}^{p}\alpha_j^{1/2}e^{-\alpha_j\mu_j}]\boldsymbol{I}(\boldsymbol{Y}_i = \mathbf{0})\} \\ + \sum_{i=1}^{n}\{\sum_{j=1}^{p}[\frac{1}{2}\ln\alpha_j - \alpha_j\mu_j - y_{ij} + y_{ij}\ln y_{ij} - \ln y_{ij}! \\ + \alpha_j y_{ij}(1 + \ln\mu_j - \ln y_{ij})]\boldsymbol{I}(\boldsymbol{Y}_i \neq \mathbf{0})\},\end{aligned}$$

thereby the corresponding p-value is

$$P_2 = \Pr(LR > z_2|H_0) = Pr\{\chi^2(1) > z_2\} \tag{5.6}$$

When $P_2 > \alpha$, the null hypothesis H_0 cannot be rejected at the significance level of α.

Chapter 4

Simulation Study And Real Data Application

This chapter will give the simulation results of the score test mentioned in the previous chapter.

4.1 Test for Zero-Inflation

In this section, testing the significance of zero inflation in multivariate zero-inflated double Poisson will be derived by using score test and the results will be shown as well. The response variable $\boldsymbol{Y}$ is the observation from a population of p-dimensional multivariate zero-inflated double Poisson model that has the expressing as $\boldsymbol{Y} \sim \text{ZIDP}_p(\omega; \mu_1, \alpha_1, ..., \mu_p, \alpha_p)$. We choose $\omega = 0.0, 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.3$ as the zero inflation parameter under H_1. μ represents the mean and α is designed

to dispersion parameter. We consider $p = 2, 3, 4$ as the number of dimensions, $n = 100, 200, 300, 400, 500$ as the sample size for multivariate zero-inflated double Poisson model. We conduct 1000 replications at the significance level $\alpha = 0.05$ under each specific parameters for testing power. Covariates are not considered in this chapter for all simulations.

The parameter $\mu_1, ..., \mu_p, \alpha_1, ..., \alpha_p$ are randomly selected for double Poisson model. For a specific pair of parameters $(n, \omega, \mu_1, ..., \mu_p, \alpha_1, ..., \alpha_p)$, we can generate multivariate zero-inflated double Poisson data according to the following process.

We generate $Z_1, \ldots, Z_n \overset{iid}{\sim} \text{Bernoulli}(1-\omega)$ as first, then p-dimensional normal double Poisson data will be produced independently as follows

$$(x_{11}, \ldots, x_{n1}) \overset{iid}{\sim} \text{DP}(\mu_1, \alpha_1), \ldots, (x_{1p}, \ldots, x_{np}) \overset{iid}{\sim} \text{DP}(\mu_p, \alpha_p),$$

let,

$$Y_i^\top = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{ip} \end{pmatrix} \overset{d}{=} Z_i \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}, \quad i = 1, \ldots, n$$

therefore one set of simulated data is generated.

The results are simulated by R and shown in Table 4.1, 4.2, 4.3 and 4.4. From the results, if statistical power is high, the probability of making a Type II error goes down. Large samples offer greater test sensitivity than small samples. It can be seen from the simulation results in Table that score test hold the nominal level well at $\alpha =$

0.05. As ω or n increases, the empirical power of the score test methods has increased. That means as ω or n increases, the more accurate the score test for the existence of zero-inflation. Also as the dimension in multivariate zero-inflated double Poisson model enlarging, the power is increasing. Finally, power will approach to 1 at $\omega = 0.1$. That means the hypothesis test is very good at detecting a false null hypothesis. Furthermore, for multivariate zero-inflated double Poisson with dimensions p greater than 4, even a very small zero-inflated parameter can make the test very significant. Also we notice that the score test did not hold the nominal level well when dimension is more than 5. A reasonable sample size is important for score test when calculating the inverse of negative expected information matrix.

The results in Table 4.5 and Table 4.6 shows the estimators under null hypothesis $H_0 = 0$ for parameter $\mu_1, ..., \mu_p$ and $\alpha_1, ..., \alpha_p$. The parameters are estimated for three cases ($p = 2, p = 3, p = 4$). We note that the estimation for parameter $\mu_1, ..., \mu_p$ and $\alpha_1, ..., \alpha_p$ in each case are close to the given true value.

Table 4.1: Power simulation of the score test for testing zero-inflation parameter

		Score Test Power			
n	**p**	ω=0	ω=0.01	ω=0.02	ω=0.03
100	2	0.048	0.065	0.132	0.189
200		0.044	0.071	0.164	0.303
300		0.042	0.086	0.222	0.430
400		0.050	0.090	0.254	0.545
500		0.044	0.106	0.320	0.601
100	3	0.049	0.177	0.365	0.550
200		0.044	0.241	0.545	0.768
300		0.046	0.303	0.710	0.898
400		0.050	0.371	0.775	0.960
500		0.055	0.436	0.871	0.986
100	4	0.054	0.426	0.699	0.873
200		0.053	0.521	0.849	0.972
300		0.057	0.681	0.949	0.989
400		0.053	0.783	0.987	1.000
500		0.051	0.848	0.992	1.000

where p representative dimensions and n is the sample size

Table 4.2: Power simulation of the score test for testing zero-inflation parameter

		Score Test Power			
n	**p**	ω=0.05	ω=0.1	ω=0.2	ω=0.3
100	2	0.415	0.876	1.000	1.000
200		0.674	0.989	1.000	1.000
300		0.826	0.999	1.000	1.000
400		0.916	1.000	1.000	1.000
500		0.953	1.000	1.000	1.000
100	3	0.846	0.993	1.000	1.000
200		0.966	1.000	1.000	1.000
300		0.999	1.000	1.000	1.000
400		1.000	1.000	1.000	1.000
500		1.000	1.000	1.000	1.000
100	4	0.982	0.999	1.000	1.000
200		1.000	1.000	1.000	1.000
300		1.000	1.000	1.000	1.000
400		1.000	1.000	1.000	1.000
500		1.000	1.000	1.000	1.000

where p representative dimensions and n is the sample size

Table 4.3: Power simulation of the score test for testing zero-inflation parameter

		Score Test Power			
n	**p**	ω=0	ω=0.01	ω=0.02	ω=0.03
100	2	0.056	0.076	0.097	0.176
200		0.053	0.080	0.172	0.301
300		0.042	0.106	0.237	0.395
400		0.045	0.147	0.288	0.505
500		0.055	0.157	0.348	0.604
100	3	0.061	0.180	0.351	0.521
200		0.061	0.225	0.534	0.795
300		0.046	0.318	0.661	0.890
400		0.053	0.377	0.783	0.943
500		0.047	0.437	0.845	0.983
100	4	0.051	0.411	0.697	0.873
200		0.060	0.538	0.872	0.963
300		0.057	0.696	0.957	0.996
400		0.051	0.758	0.986	0.999
500		0.048	0.839	0.995	1.000

where p representative dimensions and n is the sample size

Table 4.4: Power simulation of the score test for testing zero-inflation parameter

		Score Test Power			
n	**p**	ω=0.05	ω=0.1	ω=0.2	ω=0.3
100	2	0.331	0.766	0.997	1.000
200		0.557	0.967	1.000	1.000
300		0.722	0.997	1.000	1.000
400		0.837	1.000	1.000	1.000
500		0.899	1.000	1.000	1.000
100	3	0.800	0.990	1.000	1.000
200		0.969	1.000	1.000	1.000
300		0.997	1.000	1.000	1.000
400		0.999	1.000	1.000	1.000
500		1.000	1.000	1.000	1.000
100	4	0.977	1.000	1.000	1.000
200		0.999	1.000	1.000	1.000
300		1.000	1.000	1.000	1.000
400		1.000	1.000	1.000	1.000
500		1.000	1.000	1.000	1.000

where p representative dimensions and n is the sample size

Table 4.5: The MLEs under $H_0 : \omega = 0$ if $p = 2, p = 3, p = 4$

True Value		μ_1=2.8	μ_2=3.2	α_1=2.0	α_2=2.5				
n	**p**								
100	2	2.815	3.228	1.994	2.404				
200		2.815	3.242	2.022	2.423				
300		2.809	3.228	2.035	2.436				
400		2.812	3.235	2.035	2.454				
500		2.817	3.242	2.066	2.476				
True Value		μ_1=2.8	μ_2=3.2	μ_3=3.8	α_1=2.0	α_2=2.5	α_3=1.5		
n	**p**								
100	3	2.815	3.232	3.771	1.992	2.382	1.555		
200		2.810	3.231	3.781	2.000	2.436	1.587		
300		2.804	3.233	3.781	2.047	2.437	1.611		
400		2.801	3.253	3.770	2.035	2.459	1.611		
500		2.810	3.238	3.773	2.060	2.485	1.628		
True Value		μ_1=2.4	μ_2=2.8	μ_3=3.2	μ_4=3.8	α_1=2.0	α_2=2.3	α_3=2.5	α_4=1.5
n	**p**								
100	4	2.420	2.836	3.212	3.770	1.963	2.193	2.414	1.565
200		2.427	2.841	3.251	3.783	1.981	2.240	2.448	1.583
300		2.432	2.841	3.251	3.776	1.996	2.249	2.447	1.592
400		2.414	2.831	3.239	3.773	1.988	2.258	2.453	1.599
500		2.433	2.829	3.235	3.767	2.011	2.265	2.466	1.630

where p representative dimensions and n is the sample size

Table 4.6: The MLEs under $H_0 : \omega = 0$ if $p = 2, p = 3, p = 4$

True Value		μ_1=1.5	μ_2=2	α_1=1.4	α_2=1.6				
n	**p**								
100	2	1.521	2.014	1.430	1.631				
200		1.515	2.012	1.438	1.624				
300		1.519	2.010	1.446	1.647				
400		1.511	2.010	1.454	1.658				
500		1.519	2.013	1.467	1.672				
True Value		μ_1=1.5	μ_2=2.0	μ_3=2.5	α_1=1.4	α_2=1.6	α_3=1.8		
n	**p**								
100	3	1.515	2.013	2.506	1.424	1.621	1.821		
200		1.517	2.017	2.506	1.437	1.657	1.846		
300		1.512	2.013	2.513	1.449	1.655	1.852		
400		1.519	2.020	2.506	1.453	1.647	1.856		
500		1.516	2.016	2.493	1.462	1.672	1.855		
True Value		μ_1=1.5	μ_2=2.0	μ_3=2.5	μ_4=3.0	α_1=1.4	α_2=1.6	α_3=1.8	α_4=2.2
n	**p**								
100	4	1.520	2.020	2.513	3.012	1.423	1.617	1.821	2.161
200		1.509	2.019	2.510	3.005	1.446	1.636	1.841	2.203
300		1.516	2.021	2.503	3.002	1.449	1.648	1.846	2.198
400		1.510	2.011	2.511	3.020	1.447	1.648	1.861	2.203
500		1.512	2.017	2.502	3.016	1.461	1.671	1.866	2.242

where p representative dimensions and n is the sample size

4.2 Real Data Application

The above test is further explained by an example of real data. The data set is collected from S Kocherlakota and K Kocherlakota (1992)[6]. Let y_1 and y_2 represent the number of plants of the species Lacistema aggregatum and Protium guianense in each of 100 systematically laid and contiguous quadrats.

Table 4.7: The Lacistema aggregatum and Protium guianense data[12]

	y_2					
y_1	0	1	2	3	4	Total
0	34	8	3	1	0	46
1	12	13	6	1	0	32
2	4	3	1	0	0	8
3	5	3	2	1	0	11
4	2	0	0	0	0	2
5	0	0	0	0	1	1
Total	57	27	12	3	1	100

In this case, we applied the multivariate zero-inflated double Poisson model to the data. The data is assumed to be 2-dimension. Let $\boldsymbol{Y}_1$,...,$\boldsymbol{Y}_n \overset{iid}{\sim} \text{MZIDP}(\omega, \mu_1, \mu_2, \alpha_1, \alpha_2)$, where $\boldsymbol{Y}_j = (y_{1j}, y_{2j})$ is one sample out of 100, $j = (1, ..., 100)$.

Assume we want to test the null hypothesis

$$H_0 : \omega = 0 \qquad \text{against} \qquad H_1 : \omega > 0$$

For this problem, apply the score test to detect if there are too many zero values. It can be calculated by (3.21) that the value of score statistic for this case is $SC = 10.3$.

Since $\chi^2(1) = 3.84$, $SC > \chi^2(1)$, thus we should reject H_0 at the significance level $\alpha = 0.05$ and conclude there exists the zero-inflation in the real data which means there are excessive zeros in plants of the species Lacistema aggregatum and Protium guianense in each of 100 systematically laid and contiguous quadrats.

Chapter 5

Concluding Remarks

This thesis mainly focuses on a new distribution, called the multivariate zero-inflated double Poisson distribution. A brief review of the zero-inflation model is mentioned. The MZIDP distribution is developed based on the ZIDP distribution to analyze multivariate count data with extra zeros. Statistical properties related to the distribution are studied, including joint probability mass function, expectation, variance, marginal distribution, conditional distribution, and so on. Likelihood function and maximum likelihood estimates are derived. To test whether there is zero-inflation in the count data, I propose score test and likelihood ratio test theory. Based on the calculation, the information matrix and score statistic are obtained.

From the simulation, the main criterion is based on the strength of power and the proportion of the type I error. According to R, the results show that score test can detect the excessive zeros well only for the data with lower dimensions, for example,

$p \leq 4$. As ω or sample size increases, the power of the score test methods increases. The more samples we have, the higher accuracy level of the score test for the existence of zero-inflation we obtain. In addition, the estimation of the parameter is close to the true value. Hence the simulation is reliable.

In Chapter 4, the data we use is from Walhin (2001) [12]. We apply the multivariate zero-inflated double Poisson distribution to the real data. The data is considered to follow 2-dimension MZIDP distribution, which y_1 and y_2 are following different ZIDP distributions independently. Based on the results from R, there is zero-inflation in the data.

In this thesis, the parameter estimation is not derived when multivariate zero-inflation exists. Beyond that, the comparison between likelihood ratio test and score test is not included. In addition, testing for the dispersion in multivariate zero-inflation double Poisson model is not given. These work will be conducted in future study.

Bibliography

[1] Dianliang Deng and Sudhir R Paul. Score tests for zero inflation in generalized linear models. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 563–570, 2000.

[2] Dianliang Deng and Sudhir R Paul. Score tests for zero-inflation and over-dispersion in generalized linear models. *Statistica Sinica*, pages 257–276, 2005.

[3] Bradley Efron. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81(395):709–721, 1986.

[4] William H Greene. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. 1994.

[5] Norman L Johnson and Samuel Kotz. Discrete distributions: Distributions in statistics. 1969.

[6] S Kocherlakota and K Kocherlakota. Bivariate discrete distributions, new york: Marcel and dekker, 1992.

[7] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[8] Andy H Lee, Kui Wang, Jane A Scott, Kelvin KW Yau, and Geoffrey J McLachlan. Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros. *Statistical methods in medical research*, 15(1):47–61, 2006.

[9] Yin Liu and Guo-Liang Tian. Type i multivariate zero-inflated poisson distribution with applications. *Computational Statistics & Data Analysis*, 83:200–222, 2015.

[10] John Mullahy. Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365, 1986.

[11] Jan Van den Broek. A score test for zero inflation in a poisson distribution. *Biometrics*, pages 738–743, 1995.

[12] Jean François Walhin. Bivariate zip models. *Biometrical Journal*, 43(2):147–160, 2001.

[13] Feng-chang Xie, Jin-guan Lin, and Bo-cheng Wei. Score tests for zero-inflated double poisson regression models. *Acta Mathematicae Applicatae Sinica, English Series*, 33(4):851–864, 2017.

[14] M Xie, B He, and TN Goh. Zero-inflated poisson model in statistical process control. *Computational statistics & data analysis*, 38(2):191–201, 2001.