

# **Model Selection and Regression Analysis for Sparse Discrete Data**

by

Hani Aldirawi

B.S., Islamic University of Gaza, 2011

M.S., University of Texas Pan American, 2015

THESIS

Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy in Mathematics  
in the Graduate College of the  
University of Illinois at Chicago, 2020

Chicago, Illinois

Defense Committee:

Jie Yang, Chair and Advisor

Min Yang

Jing Wang

Ping-Shou Zhong

Hua Yun Chen, Division of Epidemiology and Biostatistics

Copyright by

Hani Aldirawi

2020

This dissertation is dedicated to my wonderful parents, whom worked hard to raise, teach, and encourage me to pursue my dreams; to my wonderful lovely wife Israa, for her support, patience, and love; and to my brothers and sister for their love and support.

## ACKNOWLEDGMENT

First and foremost, I would like to thank Allah for giving me the blessing, strength, knowledge, ability, and opportunity to conduct research and complete my dissertation.

I would like to express my respect to my advisor, Dr. Jie Yang. Thank you Dr. Jie Yang, for your help, support, guidance, and encouragement. You are a great example of an expert, leader, and mentor. Thank you for your support during my PhD journey, especially the last two years.

I would like to thank my committee members for their valuable feedback and remarks in completing my thesis. Particularly, I would like to thank Dr. Jing Wang for her various valuable suggestions toward my presentation. I would like to thank Dr. Ping-Shou Zhong for sharing with me a relevant R-package and some other resources to improve my thesis. I also would like to thank Dr. Min Yang and Dr. Hua Yun Chen for their insightful feedback and comments about the Fisher information matrix.

I would like to thank my family members, friends, colleagues. I also want to thank Dr. Ahmed Metwally, Dr. Lei Wang, as well as my other coauthors. Special thanks to my friends Mohammed Merhi, Khaled Khasawneh, Eyad Massarwi, and Hashem Bani Aiyesh for their tremendous help and support.

This work is supported in part by LAS Award for Faculty of Science at UIC and NSF grant DMS-1924859.

## CONTRIBUTIONS OF AUTHORS

**Chapter 1** is partially based on published work in collaboration with Dr. Ahmed Metwally, and Dr. Jie Yang. This chapter is partially based on the following publication (The copyright permission is provided in the appendices).

Metwally, A. A., Aldirawi, H., Yang, J. (2018). A review on probabilistic models used in microbiome studies. *Communications in Information and Systems*, 18(3), 173-191. Copyright ©2018, International Press of Boston.

I wrote the first half of the paper, and Dr. Ahmed Metwally wrote the second half of the paper. Dr. Jie Yang advised the work, and helped for modifying the first draft especially the equations and formulas. All authors discussed the results and finalized the paper.

**Chapter 2** is based on published work in collaboration with Dr. Ahmed Metwally, Dr. Lei Wang, and Dr. Jie Yang. This chapter is partially based on the following publications (The copyright permission is provided in the appendices).

- Aldirawi, H., Yang, J., and Metwally, A. A. (2019). Identifying Appropriate Probabilistic Models for Sparse Discrete Omics Data. 2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 1-4). Copyright ©2019, IEEE.
- Wang, L., Aldirawi, H., and Yang, J. (2020). Identifying zero-inflated distributions with a new R package iZID. *Communications in Information and Systems*, 20(1), 23-44. Copyright ©2020, International Press of Boston.

## **CONTRIBUTIONS OF AUTHORS (Continued)**

In the first paper, I was the primary author; I wrote the draft paper, wrote all of the algorithms and R codes, and did the statistical analysis. Dr. Ahmed Metwally helped in data preparation, co-wrote the paper. Dr. Jie Yang advised the work, developed the ideas, and proof outline. All authors discussed the results and finalized the paper. In the second paper, I wrote the original R codes for all of the distributions, and co-wrote the draft paper. Dr. Lei Wang constructed the R package and wrote the draft paper. Dr. Jie Yang advised the work. All authors discussed the results and finalized the paper.

## TABLE OF CONTENTS

<u>CHAPTER</u>		<u>PAGE</u>
<b>1</b>	<b>INTRODUCTION . . . . .</b>	<b>1</b>
1.1	A Review on Probabilistic Models Used in Microbiome Studies . . .	1
1.1.1	Microbiome Data Resources and Data Representation . . . . .	2
1.1.2	Probabilistic Models for Single Feature . . . . .	4
1.1.3	Models Used in Snapshot Microbiome Studies . . . . .	4
1.2	A Review of Regression Models . . . . .	7
<b>2</b>	<b>IDENTIFYING APPROPRIATE PROBABILISTIC MODELS FOR SPARSE DATA . . . . .</b>	<b>12</b>
2.1	Introduction . . . . .	12
2.2	Kolmogorov-Smirnov Test . . . . .	13
2.3	Methods . . . . .	15
2.3.1	MLEs of zero-altered (Hurdle) models and zero-inflated models . . .	15
2.3.2	Consistency of Hurdle and zero-inflated distributions . . . . .	25
2.3.3	Sampling from general Hurdle and zero-inflated models . . . . .	31
2.3.4	Bootstrapped Monte Carlo estimate for the p-value of a discrete KS test . . . . .	32
2.3.5	Likelihood ratio test for selecting the best model . . . . .	35
2.4	Results . . . . .	36
2.5	iZID R Package . . . . .	39
2.5.1	Introduction . . . . .	39
2.5.2	Existing R packages for analyzing zero-inflated data . . . . .	39
2.5.3	Architecture of the package “iZID” . . . . .	42
2.5.4	Conclusion . . . . .	58
<b>3</b>	<b>REGRESSION MODELS . . . . .</b>	<b>60</b>
3.1	Probabilistic Model vs. Regression Model . . . . .	60
3.2	Hurdle Regression Models . . . . .	61
3.3	Zero-inflated Regression Model (ZIRM) . . . . .	64
3.4	Fisher Information Matrix . . . . .	66
3.4.1	Fisher information of zero-inflated regression models . . . . .	66
3.4.2	Fisher information of Hurdle regression models . . . . .	76
3.5	Significance Test for Model Coefficients via Bootstrap . . . . .	85
3.6	Comparison between Fisher-Information-Based and Bootstrapped Confidence Intervals . . . . .	89
<b>4</b>	<b>APPLICATION TO INSURANCE DATA . . . . .</b>	<b>92</b>

## TABLE OF CONTENTS (Continued)

<b><u>CHAPTER</u></b>		<b><u>PAGE</u></b>
	4.1 Introduction: Modeling Insurance Claim Data . . . . .	92
	4.2 Model Selection . . . . .	93
	4.3 Parameters Estimation . . . . .	93
	4.4 Data set . . . . .	94
<b>5</b>	<b>CONCLUSION . . . . .</b>	<b>97</b>
	<b>CITED LITERATURE . . . . .</b>	<b>99</b>



## LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I	A TYPICAL MICROBIAL COUNT TABLE . . . . .	3
II	A TOY EXAMPLE OF TAXONOMIC PROFILE COUNT TABLE . . . . .	3
III	NUMBER AND PERCENTAGE OF SPECIES OUT OF 229 SPECIES WITH NOT STATISTICALLY SIGNIFICANT P-VALUE ( P-VALUE > 0.05) . . . . .	38
IV	NUMBER OF CLAIMS IN DATASET “DATACAR” . . . . .	56
V	ZIP PROBABILISTIC KS-TEST APPLIED TO ZIP REGRESSION RE- SPONSES . . . . .	61
VI	ZIP REGRESSION: BOOTSTRAP CONFIDENCE INTERVAL VS. FISHER INFORMATION MATRIX CONFIDENCE INTERVAL . . . . .	91
VII	ZINB REGRESSION: BOOTSTRAP CONFIDENCE INTERVAL VS. FISHER INFORMATION MATRIX CONFIDENCE INTERVAL . . . . .	91
VIII	POSSIBLE LINK FUNCTIONS FOR $\phi$ . . . . .	94
IX	NUMBER OF CLAIMS IN THE DATASET . . . . .	95
X	AIC USING DIFFERENT LINK FUNCTIONS FOR $\phi$ . . . . .	96
XI	BIC USING DIFFERENT LINK FUNCTIONS FOR $\phi$ . . . . .	96

## LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	Quantile-quantile plot between six different theoretical distributions for a specific feature. Each figure represents a different discrete distribution: (a) ZIBB distribution, (b) ZIBNB distribution, (c) ZINB distribution, (d) ZIP distribution, (e) NB distribution, (f) Poisson distribution. The p-value above each figure represents the significance of the discrete KS test between the theoretical quantiles and the sample quantiles of the corresponding distribution. The ZIBB, and ZIBNB distribution are most appropriate distributions for modeling the OTU count among some other standard distributions . . . . .	37
2	Chart plot to compare the given data and the simulated data . . . . .	58
3	The percentage of total zero claims occurrences are 93.2%. . . . .	94

## LIST OF FIGURES (Continued)

**FIGURE**

**PAGE**

## SUMMARY

Sparse count data such as microbiome data, transcriptomics or RNA-seq data, or insurance claim data, are typically overdispersed and sparse with an exceeded number of zeros, which are often challenging to be modeled.

In this dissertation work, we aim to answer two questions: (1) How do we identify the most appropriate probabilistic model for a given sparse data? (2) With available covariates, how do we build the most appropriate regression model for predicting a sparse response?

In response to the first question, we propose a statistical procedure for identifying the most appropriate discrete probabilistic models for zero-inflated or Hurdle models based on the bootstrapped p-values of a sequence of discrete Kolmogorov-Smirnov (KS) test. We develop a general procedure for estimating the parameters for a large class of zero-inflated models and Hurdle models. We also develop a bootstrapped likelihood ratio testing procedure based on Neyman-Pearson theorem for selecting the best model when there are more than one probabilistic model candidates.

We develop a new R package “iZID” as a software tool to facilitate potential users to answer the first question as well. For zero-inflated count data, we use bootstrapped Monte Carlo procedure to control the bias issue in estimating the p-value of a KS Test, as well as bootstrapped likelihood ratio tests for zero-inflated model selection. Our package also provides some functions to simulate zero-inflated and hurdle count data and calculate maximum likelihood estimates of unknown parameters. Compared with other R packages available so far, our package covers more types of zero-inflated and hurdle distributions and provides adjusted p-value estimates after incorporating the influence of unknown model parameters.

## SUMMARY (Continued)

To answer the second question, we build a fairly general class of regression models, called Zero-Inflated Regression Models (ZIRM), which not only cover currently available zero-inflated regression models, such as ZIP, ZINB with fixed  $r$ , ZIBB with constant prior parameters, but also include new regression models, including ZINB with flexible  $r$ , ZIBB with flexible prior parameters, and ZIBNB. We also build the corresponding Hurdle Regression Models for zero-altered responses. With the enriched model candidates, we perform model selection based on AIC and BIC criteria. Our application to Insurance Claim Data shows that ZINB with flexible  $r$  is more appropriate than any others.

For general zero-inflated regression models, we derive and simplify its general form of Fisher information matrix and then perform significance tests for variable selection. We compare the confidence intervals based on the Fisher information matrix with the ones built by bootstrapping. The results are consistent with each other. Compared with the bootstrapping solutions, the variable selection based on Fisher information matrix is apparently more efficient. Nevertheless, we suggest the use of bootstrapping confidence intervals when the sample size is moderate or small.

# CHAPTER 1

## INTRODUCTION

Parts of this chapter were previously published as:

Metwally, A. A., Aldirawi, H., and Yang, J. (2018). A review on probabilistic models used in microbiome studies. *Communications in Information and Systems*, 18(3), 173-191. Copyright ©2018, International Press of Boston. (1)

### **1.1 A Review on Probabilistic Models Used in Microbiome Studies**

The microbiome, a dynamic ecosystem of microorganisms (bacteria, archaea, fungi, and viruses) that live in and on us, plays a vital role in host-immune responses resulting in significant effects on host health (2). The microbiome has been linked to some diseases such as diabetes, obesity, asthma, and transplant rejection (3; 4; 5; 6; 7). The human microbiome can be divided into the core microbiome and the variable microbiome (8). The core microbiome is the set of taxa or genes that present in a given body location (gut, kidney, skin, oral, etc.) in almost all humans. The variable microbiome arises from various factors such as host physiological status, host environment, host genotype, host lifestyle, and host pathobiology. Moreover, given the strong association between microbiome and various diseases, computational models have been built to predict phenotypes from microbial profiles (9; 10; 11).

In this chapter we briefly introduce microbiome data and the related probabilistic models to people who are interested in microbiome research and the corresponding analysis. We introduce the typical

format of microbiome data and we review the probabilistic models for modeling count data from each microbial feature independently.

### 1.1.1 Microbiome Data Resources and Data Representation

There are several initiatives to store and manage data from microbiome studies in order to make them available and free for everyone to use. The major public servers are MG-RAST (<https://www.mg-rast.org/>) and QIITA (<https://qiita.ucsd.edu/>). Also, National Center of Biotechnology Information (NCBI, <https://www.ncbi.nlm.nih.gov/>) is one of the most comprehensive resources that have a database of curated and updated microbial genomes and taxonomic tree.

To analyze these massive amount of sequence data, metagenomic reads are processed for each microbiome sample to construct taxonomic and/or functional profiles (12; 13; 14). The taxonomic profiles, functional profiles, or both for all samples, are then combined into one count table (see Table I and Table II as an example of a toy taxonomic profile) with a dimension of  $m \times n$ , where  $m$  denotes the number of microbial features  $F_1, \dots, F_m$  and  $n$  denotes the number of metagenomic samples  $S_1, \dots, S_n$ . The entry  $z_{ij}$  represents the number of reads from sample  $j$  that mapped to microbial feature  $i$ , while its capitalized version  $Z_{ij}$  represents the corresponding random variable. In the table,  $N_j = \sum_{i=1}^m z_{ij}$  denotes the total number of reads for the  $m$  features in sample  $S_j$ , and  $z_{i.} = \sum_{j=1}^n z_{ij}$  denotes the total number of reads mapped to features  $F_i$  in all samples. Since metagenomic samples may have different sequencing depths, the aggregated metagenomic counts need to be normalized among samples (15). There are several methods developed to tackle the normalization problem of a count table, such as centered log-ratio (CLR) transformation (16), cumulative sum scaling (17), median-of-ratios scaling factor (18), and trimmed mean of M values (19).

TABLE I: A TYPICAL MICROBIAL COUNT TABLE

Feature/Sample	$S_1$	$S_2$	$S_3$	...	$S_n$	Total
$F_1$	$z_{11}$	$z_{12}$	$z_{13}$	...	$z_{1n}$	$z_{1.}$
$F_2$	$z_{21}$	$z_{22}$	$z_{23}$	...	$z_{2n}$	$z_{2.}$
...	...	...	...	...	...	...
$F_m$	$z_{m1}$	$z_{m2}$	$z_{m3}$	...	$z_{mn}$	$z_{m.}$
Total	$N_1$	$N_2$	$N_3$	...	$N_n$	$N_{.}$

TABLE II: A TOY EXAMPLE OF TAXONOMIC PROFILE COUNT TABLE

Species/Sample	$S_1$	$S_2$	$S_3$	$S_4$	Total
<i>Streptococcus pneumoniae</i>	0	0	102	3	105
<i>Staphylococcus aureus</i>	0	75	0	0	75
<i>Escherichia coli</i>	14	0	278	0	292
Total	14	75	380	3	472



### 1.1.2 Probabilistic Models for Single Feature

In this dissertation, we focus on probabilistic models built for sequence read counts from a single microbiome feature, that is,  $Z_{ij}$  for feature  $F_i$  and subject  $S_j$ . Assuming  $Z_{ij}$  follows a probabilistic model with a few unknown parameters, statistical inference can be made based on estimated parameters from the data. In practice, there are two types of experimental design for microbiome studies: (1) snapshot studies, where each subject provides only one sample, (2) longitudinal studies, which include multiple samples per subject over time.

### 1.1.3 Models Used in Snapshot Microbiome Studies

- **Poisson model**

Poisson distribution has been widely used for modeling non-negative outcomes as a count. If a random feature count  $Z_{ij}$  follows a Poisson distribution with mean  $\theta > 0$ , it assigns the probability

$$P(Z_{ij} = k) = \frac{\theta^k}{k!} e^{-\theta}$$

for  $k = 0, 1, 2, \dots$ . As the mean count increases, the skewness diminishes, and the Poisson distribution becomes approximately a normal distribution (20). One property of Poisson distribution is that its variance equals the mean.

- **Negative binomial model**

The negative binomial (NB) distribution is an alternative probabilistic model for count data (21). It is especially useful when the sample variance exceeds the sample mean, known as over-dispersion.

Given a sequence of independent Bernoulli trials with probability  $p$  of success,  $Z_{ij}$  is the number of failures observed before the  $r^{th}$  success with the probability

$$P(Z_{ij} = k) = \binom{k+r-1}{k} p^r (1-p)^k$$

where  $r > 0$  and  $0 \leq p \leq 1$  are two parameters that can be estimated from the data.

- **Zero-inflated models**

For microbiome OTU counts, typically there are much more zeros than expected under the assumption of Poisson or negative binomial distributions. This phenomenon is known as zero-inflation. In order to solve this issue, zero-inflated models are used to model read counts that have an excess of zeros. A zero-inflated model assumes that the observed zeros are of two kinds; “sampling” or “structural”. The sampling zeros come from a Poisson, negative binomial, or some other distribution due to chance. Other observed zeros are due to some specific structure in the data (22). As a result, the combined probability under a zero-inflated model is

$$P_{ZI}(Z_{ij} = k) = \phi \mathbf{1}_{\{k=0\}} + (1 - \phi)P(Z_{ij} = k) \quad (1.1)$$

where  $\phi > 0$  is a parameter estimated from the data,  $P(Z_{ij} = k)$  stands for the probability determined by a Poisson, negative binomial, or other parametric distribution. Note that the zero-inflated model assigns the probability  $\phi + (1 - \phi)P(Z_{ij} = 0)$  to zero, which is larger than  $P(Z_{ij} = 0)$  itself. The corresponding distributions are known as zero-inflated Poisson (ZIP),

zero-inflated negative binomial (ZINB), zero-inflated beta binomial (ZIBB), zero-inflated Gaussian (ZIG) distributions, etc.

*Example 1.1. Zero-inflated beta binomial model (ZIBB)*

As a special kind of zero-inflated models introduced in Section 1.1.3, the zero-inflated beta binomial (ZIBB) model provides a flexible option for modeling  $Z_{ij}$ . In a ZIBB model, the probability  $P(Z_{ij} = k)$  in (Equation 1.1) is formulated by a beta-binomial distribution. It has two folds: (1) Given a probability  $p_{ij}$ ,  $Z_{ij}$  follows a binomial distribution with parameters  $N_j$  and  $p_{ij}$ ; (2) In order to make the model flexible, the probability  $p_{ij}$  itself is also random, which follows a beta distribution with parameters  $\alpha_1 > 0$ ,  $\alpha_2 > 0$ . As a result, the probability based on the beta-binomial distribution is

$$P(Z_{ij} = k) = \binom{N_j}{k} \frac{\text{Beta}(k + \alpha_1, N_j - k + \alpha_2)}{\text{Beta}(\alpha_1, \alpha_2)} \quad (1.2)$$

The probability  $P_{ZI}(Z_{ij} = k)$  based on the ZIBB model takes the same form as in (Equation 1.1).

- **Hurdle models**

Hurdle models, also known as zero-altered models, provide another way of dealing with the excess zeros in OTU counts (21). A hurdle model consists of two components, one generating the zeros and one generating the positive values. In contrast to zero-inflated models, a hurdle model assumes that all zeros are from the “structural” source. In order to make the comparison clearly, we define the hurdle models using a similar formula as in (Equation 1.1):

$$P_{ZA}(Z_{ij} = k) = \phi \mathbf{1}_{\{k=0\}} + (1 - \phi) P_{tr}(Z_{ij} = k) \quad (1.3)$$

where  $P_{tr}(Z_{ij} = k)$  is a truncated version of  $P(Z_{ij} = k)$  determined by  $P_{tr}(Z_{ij} = 0) = 0$  and  $P_{tr}(Z_{ij} = k) = P(Z_{ij} = k)/[1 - P(Z_{ij} = 0)]$  for  $k > 0$ . For example, if  $P(Z_{ij} = k)$  comes from a Poisson distribution, then  $P_{tr}(Z_{ij} = k)$  is known as a zero-truncated Poisson distribution (23).

The hurdle model  $P_{ZA}(Z_{ij} = k)$  collapses to the standard model  $P(Z_{ij} = k)$  if  $\phi = P(Z_{ij} = 0)$ . It clearly allows for excess zeros when  $\phi > P(Z_{ij} = 0)$ . Different from zero-inflated models, in principle, hurdle models can also model too few zeros when  $\phi < P(Z_{ij} = 0)$ . In other words, hurdle models are more flexible than zero-inflated models.

Similar to zero-inflated models, hurdle models include zero-altered Poisson (ZAP) or Poisson hurdle (PH), zero-altered negative binomial (ZANB) or negative binomial hurdle (NBH) models, etc.

## 1.2 A Review of Regression Models

In section 1.1 we discussed the probabilistic models without covariates. In this section we introduce the methods used for modeling sparse count data given some covariates.

Generalized linear models (GLM) can be used for modeling count data (24; 25). However, when the count data is sparse with a significant percentage of zeros, GLM is not recommended because the proportion of zeros ( $\phi_i$ ) must be linked to some distributions (26; 27).

Modified Poisson models that handle excess zeros without any covariates were described by Cohen (28). Allowing for covariates, the zero-altered Poisson or hurdle Poisson model was proposed (29; 30;

31). Based on this work, a zero-inflated Poisson (ZIP) model was proposed by Lambert (32; 33) with an application to defects in manufacturing as follows:

In our notations, the response  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  are independent and

$$Y_i \sim \begin{cases} 0 & \text{with probability } \phi_i \\ \text{Poisson}(\mu_i) & \text{with probability } 1 - \phi_i \end{cases}$$

$$\Pr(Y_i = y_i) = \begin{cases} \phi_i + (1 - \phi_i) e^{-\lambda_i} & \text{if } y_i = 0 \\ (1 - \phi_i) e^{-\lambda_i} \lambda_i^k / k! & \text{if } y_i > 0 \end{cases}$$

The parameters  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)^T$ , and  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)^T$  satisfy:  $\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta}$ , and  $\text{logit}(\boldsymbol{\phi}) = \log(\boldsymbol{\phi}/(1 - \boldsymbol{\phi})) = \mathbf{G}\boldsymbol{\gamma}$  for covariate matrices  $\mathbf{B}$  and  $\mathbf{G}$ .

Later on, Greene (34) used an extended version of the negative binomial model for sparse count data, the zero-inflated negative binomial model (ZINB) as follows:

$$\Pr(Y_i = y_i) = \begin{cases} \phi_i + (1 - \phi_i) \left(\frac{1}{1 + \kappa \lambda_i}\right)^{\kappa-1} & \text{if } y_i = 0 \\ (1 - \phi_i) \frac{\Gamma(\kappa-1+y_i)}{\Gamma(\kappa-1)(y_i!)} \left(\frac{\kappa \lambda_i}{1 + \kappa \lambda_i}\right)^{y_i} \left(\frac{1}{1 + \kappa \lambda_i}\right)^{\kappa-1} & \text{if } y_i > 0 \end{cases}$$

for  $i = 1, \dots, N$ . The mean and variance of the ZINB random variable are  $E(Y_i) = (1 - \phi_i) \lambda_i$  and  $\text{Var}(Y_i) = (1 - \phi_i) \lambda_i (1 + (\kappa + \phi_i) \lambda_i)$ , where  $\kappa$  is an overdispersion parameter.

The parameters  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)^T$ , and  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)^T$  satisfy:  $\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta}$ , and  $\text{logit}(\boldsymbol{\phi}) = \log(\boldsymbol{\phi}/(1 - \boldsymbol{\phi})) = \mathbf{G}\boldsymbol{\gamma}$  for covariate matrices  $\mathbf{B}$  and  $\mathbf{G}$ .

ZIP, ZINB, HP, and HNB regression models have lots of real life applications. We list some applications here:

- **Insurance claim data:** Using historical insurance data to predict the number of future claims is one of the main interests to insurance companies. Modeling insurance claim data is very challenging because the data are highly right skewed and sparse.
- **Health care data:** Majo (35) used ZIP for modeling health care data. In addition, Gilles (36) used ZIP and ZINB regression models to predict the number of doctor visits.
- **Ecology:** Sileshi et al. 2009 (37) compared four regression models (Poisson, negative binomial, ZIP, and ZINB) for modeling sparse ecological count data. They did the comparison using five data sets, and they concluded that the ZINB regression model fits better than the regular Poisson, and negative binomial regression models.
- **Security:** Chen (38) constructed ZINB regression model to predict number of bicycle thefts at either an intersection or a mid-block given some covariates such as number of street lights in the area, number of bus stops in the area, unemployed percentage in the area, and some other covariates.

Although GLMs have been widely used, they are largely confined to one-parameter distributions belong to the exponential family. Since there are many situations where the distribution is not a member of the exponential family, we need a method for more flexibility than GLMs.

Yee (2017) described a larger and more flexible statistical framework to extend GLMs, called vector generalized linear models (VGLMs) and vector generalized additive models (VGAMs) (39).

VGLMs model is based on assigning a link function  $g_i$  for each parameter  $\theta_i$ , where the parameter  $\theta_i$  is a linear combination of the explanatory variables.

$$g_i(\theta_i) = \eta_i = \boldsymbol{\beta}_i^T \mathbf{x} = \beta_{(i)1}x_1 + \cdots + \beta_{(i)p}x_p, \quad i = 1, \dots, b \quad (1.4)$$

VGAMs is an extension of (Equation 1.4). That is,

$$g_i(\theta_i) = \eta_i = \sum_{k=1}^d f_{(i)k}(x_k), \quad i = 1, \dots, b \quad (1.5)$$

where  $f_{(i)k}$  is a smooth function.

The models discussed in this dissertation work, including Hurdle Regression Models (HRM), and Zero-inflated Regression Models (ZIRM) are special cases of extended VGLMs.

We model HRM and ZIRM as follows:

$$g(\phi_i) = \mathbf{G}_i^T \boldsymbol{\gamma}, \quad i = 1, \dots, n \quad (1.6)$$

$$h_j(\theta_{ij}) = \mathbf{B}_{ij}^T \boldsymbol{\beta}_j, \quad i = 1, \dots, n; j = 1, \dots, b \quad (1.7)$$

where  $g$  and  $h_1, \dots, h_b$  are known link functions.  $\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_b$  are regression coefficients,  $\mathbf{G}_i = (r_1(\mathbf{x}_i), \dots, r_s(\mathbf{x}_i))^T \in \mathbb{R}^s$  and  $\mathbf{B}_{ij} = (q_{j1}(\mathbf{x}_i), \dots, q_{jt_j}(\mathbf{x}_i))^T \in \mathbb{R}^{t_j}$  are the corresponding predictors,  $r_i$ 's and  $q_{ji}$ 's are known functions. Examples include  $\mathbf{G}_i = \mathbf{B}_{ij} = (1, x_{i1}, \dots, x_{id})^T$  for main-effects model and  $\mathbf{G}_i = \mathbf{B}_{ij} = (1, x_{i1}, \dots, x_{id}, x_{i1}x_{i2}, \dots, x_{i,d-1}x_{id})^T$  for model with both main effects and order-2 interactions. We will review and discuss our approach with more details in 3.2, and 3.3.

The zero-Inflated regression models (ZIRM) here are a fairly general class of regression models, which not only cover currently available zero-inflated regression models, such as ZIP, ZINB with fixed

$r$ , ZIBB with constant prior parameters, but also include new regression models, including ZINB with flexible  $r$ , ZIBB with flexible prior parameters, and ZIBNB. We also build the corresponding Hurdle Regression Models for zero-altered responses. With the enriched model candidates, we perform model selection based on AIC and BIC criteria.



## CHAPTER 2

### IDENTIFYING APPROPRIATE PROBABILISTIC MODELS FOR SPARSE DATA

Previously published as:

- Aldirawi, H., Yang, J., and Metwally, A. A. (2019). Identifying Appropriate Probabilistic Models for Sparse Discrete Omics Data. 2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 1-4). Copyright ©2019, IEEE. (40)
- Wang, L., Aldirawi, H., and Yang, J. (2020). Identifying zero-inflated distributions with a new R package iZID. Communications in Information and Systems, 20(1), 23-44. Copyright ©2020, International Press of Boston. (41)

#### 2.1 Introduction

Sparse discrete count such as microbiome (1), transcriptomics (RNA-seq), insurance claim data (42), healthcare (35), and security (38) data are typically skewed, overdispersed, with an exceeded number of zeros. It is challenging to model this kind of data which are skewed and zero-inflated.

The selection of an appropriate probabilistic model is critical for sparse count data. For example, in order to determine if there is an association between an omic feature, such as gene or bacteria, and the disease, we may need to detect the significance of the difference between two groups of records. With appropriate probabilistic models identified successfully, we can improve the power of the statistical test significantly. Even when covariates are recorded with medical records, the correctly identified probabilistic model is critical for validating the model assumption for regression analysis.

In this chapter, we propose to use a bootstrapped Monte Carlo method to deal with all discrete KS tests, and estimating the p-value of the corresponding KS test with reduced biased. Besides the commonly used probabilistic models for sparse count data, in this chapter, we introduce new discrete models such as beta binomial (BB), beta negative binomial (BNB), zero-inflated beta binomial (ZIBB), beta binomial hurdle (BBH), zero-inflated beta negative binomial (ZIBNB), beta negative binomial hurdle (BNBH). The new models could be more flexible by attaching a beta prior distribution. We also introduce “iZID” R package, which contains all of the algorithms discussed in this chapter.

## 2.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test has been widely used for determining whether a random sample comes from a specific continuous distribution(43), known as “goodness-of-fit” in statistics. Suppose that we have a simple random sample  $X_1, X_2, \dots, X_n$  with its empirical distribution function defined as  $F_n(x) = n^{-1} \sum_{i=1}^n I_{(-\infty, x]}(X_i)$ . The KS test statistic  $D_n = \sup_x |F_n(x) - F_0(x)|$  is used for testing if the random sample follows the specified continuous cumulative distribution function (CDF)  $F_0(x)$ . The null hypothesis of KS test is  $X = \{X_1, \dots, X_n\} \sim F_0(X)$ .

The KS test can be used for any distribution with a continuous CDF  $F_0(x)$ . It’s one of the most popular goodness of fit tests, and used for lots of applications for continuous distributions. However, in many real applications, the distribution is either discrete (such as Poisson or NB) or mixed (such as zero-inflated half-normal) (44).

The following two strategies of goodness-of-fit tests have been practically used for testing if the data follows some discrete distributions.

*Strategy 1:* use the continuous KS test for discrete distributions. More specifically, in order to test if the data  $\mathbf{X} = \{X_1, \dots, X_n\}$  from a discrete distribution  $f_\theta(x)$  with unknown parameter  $\theta$ , one may first obtain an estimate  $\hat{\theta}$  from the data and simulate a random sample  $\mathbf{Y} = \{Y_1, \dots, Y_n\}$  from  $f_{\hat{\theta}}(x)$ , then calculate the p-value using `ks.test(X, Y)` in R. However, the p-value obtained by this way could be significantly biased.

As an illustrative example, we simulate a random sample  $X_1, \dots, X_{5000} \sim \text{NB}(2, 0.5)$ . By using Strategy 1, we obtain a p-value as small as  $8.8 \times 10^{-8}$ , along with a warning message saying p-value will be approximate in the presence of ties. These ties are mainly due to the discreteness of the distribution.

*Strategy 2:* would be the classical Chi-square goodness-of-fit test which could be used both for continuous and discrete distributions. Nevertheless, it is known that the KS test has greater power than Chi-Square test (45). Therefore KS test is more recommended than Chi-Square goodness of fit.

Recently, Dimitrova *et al.* (46) developed a feasible solution for extending the KS test for general discrete and mixed distributions. They expressed the complementary CDF of the distribution through the rectangle probability for uniform order statistic and compute it using Fast Fourier Transformation. They also provided an R package “KSgeneral” for calculating the p-value of the KS test when the distribution is discrete or mixed. Unfortunately, Dimitrova *et al.*’s method requires that the distribution parameters are known, which is not the case in general for real applications. Plugging in the parameter values estimated from the data tends to overestimate and thus biased the p-value, which would make the testing too conservative (47; 48).

In the statistical literature, Lilliefors (1967) considered the KS test for normal distribution with unknown parameters (47), Lilliefors (1969) explored the KS test for exponential distribution with unknown parameter (48), and Parsons (1982) discussed Weibull distributions with unknown parameters (49). In the R package “KScorrect”, Monte Carlo method was used for estimating the p-value for more general continuous distributions with unknown parameters (50). Unfortunately, many real data contain tied observations and thus a continuous distribution assumption is often not valid. All these results show that the p-value has to be adjusted when the model parameters are unknown. Nevertheless, they dealt with KS tests case by case. In this thesis, we propose to use a bootstrapped Monte Carlo method to deal with all discrete KS tests.

## 2.3 Methods

In order to use our bootstrapped KS test algorithm, some necessary derivations must be calculated. For example, we have to find the MLE for the Zero-inflated and Hurdle distributions. Also, we have to find a way to generate a random samples from Zero-inflated and Hurdle distributions.

### 2.3.1 MLEs of zero-altered (Hurdle) models and zero-inflated models

- **Zero-altered or Hurdle model and their MLEs**

We have discussed Zero-altered models, or *Hurdle models* in chapter 1. Given a baseline discrete distribution  $P_{\theta}(Y = y)$  with parameter  $\theta$ , the corresponding hurdle model can defined as follows:

$$P_{ZA}(Y = y) = \phi \mathbf{1}_{\{y=0\}} + (1 - \phi)P_{tr}(Y = y)$$

where  $\phi \geq 0$  is a weight parameter of zeros,  $P_{tr}(Y = y)$  is the zero-truncated version of the baseline distribution determined by  $P_{tr}(Y = 0) = 0$  and  $P_{tr}(Y = y) = P_{\theta}(Y = y)/[1 - P_{\theta}(Y = 0)]$  for  $y \neq 0$ .

The parameters of a Hurdle model include both  $\phi$  and  $\theta$ . Let  $Y_1, \dots, Y_n$  be a random sample from a Hurdle model. Suppose its baseline distribution has a probability mass function (pmf) or density function (pdf)  $f_{\theta}(y)$ . Then its distribution function can be written as

$$f_{ZA}(y | \phi, \theta) = \phi \cdot \mathbf{1}_{\{y=0\}} + \frac{1 - \phi}{1 - p_0(\theta)} f_{\theta}(y) \cdot \mathbf{1}_{\{y \neq 0\}} \quad (2.1)$$

and the likelihood function of  $(\phi, \theta)$  is

$$L(\phi, \theta) = \phi^{n-m} (1 - \phi)^m \cdot [1 - p_0(\theta)]^{-m} \prod_{i: Y_i \neq 0} f_{\theta}(Y_i) \quad (2.2)$$

Where  $m = \#\{i : Y_i \neq 0\}$  is the number of nonzero observations,  $p_0(\theta) = P_{\theta}(Y = 0)$ . Note that  $\phi$  and  $\theta$  are separable in the likelihood function. the log likelihood function of  $(\phi, \theta)$  is

$$l(\phi, \theta) = (n - m) \log \phi + m \log(1 - \phi) - m \log(1 - p_0(\theta)) + \sum_{i: Y_i \neq 0} \log f_{\theta}(Y_i) \quad (2.3)$$

Note that  $p_0(\theta) = f_{\theta}(0)$  for discrete cases or 0 for continuous cases.

$$\frac{\partial l(\phi, \theta)}{\partial \phi} = \frac{n - m}{\phi} - \frac{m}{1 - \phi} = 0 \quad (2.4)$$

Therefore, the maximum likelihood estimate (MLE) maximizing (Equation 2.2) is

$$\hat{\phi} = 1 - \frac{m}{n}, \quad \hat{\theta} = \operatorname{argmax}_{\theta} [1 - p_0(\theta)]^{-m} \prod_{i: Y_i \neq 0} f_{\theta}(Y_i)$$

That is,  $\hat{\theta}$  is simply the MLE for the truncated model with pmf or pdf  $f_{tr}(y; \theta) = f_{\theta}(y)/[1 - p_0(\theta)]$ ,  $y \neq 0$ . Note that  $f_{tr}(y; \theta) = f_{\theta}(y)$  for  $y \neq 0$  if the baseline distribution is continuous.

*Example 2.1.* For zero-altered Poisson or Hurdle Poisson distribution, the pmf of the baseline distribution is  $f_{\lambda}(y) = e^{-\lambda} \lambda^y / y!$  with  $p_0(\lambda) = e^{-\lambda}$ . The truncated pmf is

$$f_{tr}(y; \lambda) = \frac{e^{-\lambda}}{1 - e^{-\lambda}} \cdot \frac{\lambda^y}{y!}, \quad y = 1, 2, \dots$$

The loglikelihood for the zero-truncated Poisson is

$$\begin{aligned} l(\lambda) &= -m\lambda - m \log(1 - e^{-\lambda}) + \sum_{i: Y_i > 0} Y_i \cdot \log \lambda - \log \left( \prod_{i: Y_i > 0} Y_i! \right) \\ \frac{\partial l(\lambda)}{\partial \lambda} &= -m - m \frac{1}{e^{\lambda} - 1} + \sum_{i: Y_i > 0} Y_i / \lambda = 0 \end{aligned}$$

It's easy to verify that the MLE  $\hat{\lambda}$  of  $\lambda$  solves the likelihood equation  $\lambda = \bar{Y}(1 - e^{-\lambda})$  with  $\bar{Y} = m^{-1} \sum_{i: Y_i > 0} Y_i$ , which can be solved numerically.  $\square$

*Example 2.2.* For zero-altered negative binomial or negative binomial hurdle distribution, the pmf of the baseline distribution is

$$f(y, r, p) = \binom{y+r-1}{y} p^y (1-p)^r, y = 0, 1, 2, \dots$$

Where  $r > 0$  is the number of failures,  $k$  is the number of successes, and  $p$  is the probability of success.

The truncated pmf is  $\frac{f(y;r,p)}{1-p\theta(y_i=0)} = \frac{\binom{y+r-1}{y} p^y (1-p)^r}{1-(1-p)^r}$ , where  $\binom{y+r-1}{y} = \frac{\Gamma(y+r)}{y! \Gamma(r)}$

The likelihood for the zero-truncated negative binomial is given by:

$$L(r, p) = [1 - (1-p)^r]^m \prod_{i=1}^m \frac{\Gamma(y_i + r)}{\Gamma(y_i + 1) \Gamma(r)} p^{y_i} (1-p)^r$$

The loglikelihood for the zero-truncated negative binomial is

$$\begin{aligned} l(p, r) &= \sum_{i=1}^m \log \Gamma(y_i + r) - \sum_{i=1}^m \log \Gamma(y_i + 1) - m \log \Gamma(r) + \sum_{i=1}^m y_i \log p \\ &\quad + mr \log(1-p) - m \log[1 - (1-p)^r] \\ \frac{\partial l}{\partial p} &= \frac{\sum_{i=1}^m y_i}{p} - \frac{mr}{1-p} - \frac{mr(1-p)^{r-1}}{1-(1-p)^r} \\ \frac{\partial l}{\partial r} &= \sum_{i=1}^m \psi(y_i + r) - m\psi(r) + m \log(1-p) + \frac{m(1-p)^r \log(1-p)}{1-(1-p)^r} \end{aligned}$$

Where  $\psi(y) = \frac{\Gamma'(y)}{\Gamma(y)}$  is the digamma function. We can find the implicit solution to the above two differential equations numerically using Newton's method. □

*Example 2.3.* For zero-altered Beta Binomial or Beta Binomial Hurdle (BBH) distribution, the pmf of the baseline distribution is

Let  $\theta = (n, \alpha, \beta)$ . The pmf of beta-binomial distribution is

$$f_{\theta}(y) = \binom{n}{y} \frac{\text{Beta}(y + \alpha, n - y + \beta)}{\text{Beta}(\alpha, \beta)}$$

with  $y = 0, 1, \dots, n$  and

$$p_0(\theta) = \frac{\Gamma(n + \beta)\Gamma(\alpha + \beta)}{\Gamma(n + \alpha + \beta)\Gamma(\beta)}$$

Let  $L(\theta)$  be the likelihood of zero-truncated Beta Binomial distribution, then

$$L(n, \alpha, \beta) = \underset{\theta}{\operatorname{argmax}} \frac{\prod_{i: y_i \neq 0} f_{\theta}(y_i)}{[1 - p_0(\theta)]^m} = \left( \frac{\Gamma(\alpha + n + \beta)\Gamma(\beta)}{\Gamma(\alpha + n + \beta)\Gamma(\beta) - \Gamma(n + \beta)\Gamma(\alpha + \beta)} \right)^m \cdot \prod_{i=1}^m \left( \frac{\Gamma(n + 1)\Gamma(y_i + \alpha)\Gamma(n - y_i + \beta)\Gamma(\alpha + \beta)}{\Gamma(y_i + 1)\Gamma(n - y_i + 1)\Gamma(\alpha + n + \beta)\Gamma(\alpha)\Gamma(\beta)} \right)$$

The loglikelihood of zero-truncated Beta Binomial is given by:

$$\begin{aligned} l(n, \alpha, \beta) &= m \log \Gamma(n + 1) + m \log \Gamma(\alpha + \beta) - m \log \Gamma(\alpha) + \sum_{i=1}^n \log \Gamma(y_i + \alpha) \\ &\quad - m \log (\Gamma(\alpha + n + \beta)\Gamma(\beta) - \Gamma(n + \beta)\Gamma(\alpha + \beta)) + \sum_{i=1}^m \log \Gamma(n - y_i + \beta) \\ &\quad - \sum_{i=1}^m \log \Gamma(y_i + 1) - \sum_{i=1}^m \log \Gamma(n - y_i + 1) \end{aligned}$$

Numerically,  $\log[\Gamma(n + \alpha + \beta)\Gamma(\beta) - \Gamma(n + \beta)\Gamma(\alpha + \beta)]$  may be undefined for large  $n$  since both  $\Gamma(n + \alpha + \beta)$  and  $\Gamma(n + \beta)$  are numerical infinity. In this case, we use the fact  $\log(A - B) =$



$\log(1 - \exp(\log B - \log A)) + \log A$  if  $A \geq B$ , where  $A = \Gamma(n + \alpha + \beta)\Gamma(\beta)$ ,  $B = \Gamma(n + \beta)\Gamma(\alpha + \beta)$ .

To apply this fact, we have to verify that  $A > B$ , which means  $\Gamma(n + \alpha + \beta)\Gamma(\beta) > \Gamma(n + \beta)\Gamma(\alpha + \beta)$ .

To show  $\Gamma(\alpha + n + \beta)\Gamma(\beta) > \Gamma(\alpha + \beta)\Gamma(n + \beta)$ , it's sufficient to show  $\frac{\Gamma(\alpha + n + \beta)}{\Gamma(n + \beta)} > \frac{\Gamma(\alpha + \beta)}{\Gamma(\beta)}$ . Now,

$$\begin{aligned} \frac{\Gamma(\alpha + n + \beta)}{\Gamma(n + \beta)} &= \frac{(\alpha + \beta + n - 1)(\alpha + \beta + n - 2) \dots (\alpha + \beta)\Gamma(\alpha + \beta)}{(\beta + n - 1) \dots \beta\Gamma(\beta)} \\ &= \frac{(\alpha + \beta + n - 1)(\alpha + \beta + n - 2) \dots (\alpha + \beta)}{(\beta + n - 1)(\beta + n - 2) \dots \beta} \frac{\Gamma(\alpha + \beta)}{\Gamma(\beta)} \end{aligned}$$

The fraction  $\frac{(\alpha + \beta + n - 1)(\alpha + \beta + n - 2) \dots (\alpha + \beta)}{(\beta + n - 1)(\beta + n - 2) \dots \beta}$  contains the positive values of  $\beta$  in the numerator, which

makes it greater than 1. Therefore,  $\frac{\Gamma(\alpha + n + \beta)}{\Gamma(n + \beta)} > \frac{\Gamma(\alpha + \beta)}{\Gamma(\beta)}$

Let  $\Psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ , known as the *digamma* function. In order to apply Theorem 2.1, we need the formulae as follows:

$$\begin{aligned} \frac{\partial l(n, \alpha, \beta)}{\partial n} &= m \left( \frac{\exp(\log B - \log A)(\psi(n + \beta) - \psi(n + \alpha + \beta))}{1 - \exp(\log B - \log A)} \right) + m\psi(n + 1) \\ &\quad + \sum_{i=1}^m \psi(n - y_i + \beta) - \sum_{i=1}^m \psi(n - y_i + 1) - m\psi(\alpha + n + \beta) \\ \frac{\partial l(n, \alpha, \beta)}{\partial \alpha} &= m \left( \frac{\exp(\log B - \log A)(\psi(n + \beta) - \psi(n + \alpha + \beta))}{1 - \exp(\log B - \log A)} \right) \\ &\quad + \sum_{i=1}^m \psi(y_i + \alpha) + m\psi(\alpha + \beta) - m\psi(\alpha + n + \beta) - m\psi(\alpha) \\ \frac{\partial l(n, \alpha, \beta)}{\partial \beta} &= m \left( \frac{\exp(\log B - \log A)(\psi(n + \beta) + \psi(\alpha + \beta) - \psi(\alpha + n + \beta) - \psi(\beta))}{1 - \exp(\log B - \log A)} \right) \\ &\quad - m\psi(\beta) + \sum_{i=1}^m \psi(n - y_i + \beta) + m\psi(\alpha + \beta) - m\psi(\alpha + n + \beta) \end{aligned}$$

□

- **Zero-inflated models and their MLEs**

Unlike zero-altered models, a zero-inflated model always assume an excess of zeros. Besides zeros come from a baseline distribution, such as Poisson or negative binomial, there are additional zeros modeled by a weight parameter  $\phi \in [0, 1]$ .

When the baseline distribution is discrete with a pmf  $f_{\theta}(y)$ , the corresponding zero-inflated model has a pmf written as follows

$$f_{ZI}(y | \phi, \theta) = \phi \mathbf{1}_{\{y=0\}} + (1 - \phi) f_{\theta}(y) \quad (2.5)$$

When the baseline distribution is either continuous with a pdf  $f_{\theta}(y)$  or discrete but with  $p_0(\theta) = 0$ , the corresponding zero-inflated model is essentially the same as the corresponding zero-altered model.

Given a random sample  $y_1, \dots, y_n$  from a zero-inflated model  $f_{ZI}(y|\phi, \theta)$ , we aim to find the maximum likelihood estimate  $\hat{\phi}$  for  $\phi$  and  $\hat{\theta}$  for  $\theta$ . Similar as in Section 2.3.1, we denote  $m = \#\{i : y_i \neq 0\}$ .

If the baseline distribution satisfies  $P(Y = 0) = 0$ , the likelihood function

$$L(\phi, \theta) = \phi^{n-m} (1 - \phi)^m \cdot \prod_{i: y_i \neq 0} f_{\theta}(y_i)$$

Then  $\hat{\phi} = 1 - m/n$  and  $\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{i: y_i \neq 0} f_{\theta}(y_i)$ , which are the same as for Hurdle models.

If the baseline distribution has a pmf  $f_{\theta}(y)$ , the likelihood function is

$$L(\phi, \theta) = [\phi + p_0(\theta)(1 - \phi)]^{n-m} (1 - \phi)^m (1 - p_0(\theta))^m \prod_{i: y_i \neq 0} f_{tr}(y_i; \theta) \quad (2.6)$$

where  $p_0(\theta) = f_{\theta}(0)$ ,  $f_{tr}(y; \theta) = f_{\theta}(y)/[1 - p_0(\theta)]$ ,  $y \neq 0$ . By reparametrization, we let

$$\psi = 1 - [\phi + p_0(\theta)(1 - \phi)] = (1 - \phi)[1 - p_0(\theta)] \quad (2.7)$$

Then  $\phi = 1 - \psi/[1 - p_0(\theta)]$  and the likelihood of  $\psi$  and  $\theta$  is

$$L(\psi, \theta) = (1 - \psi)^{n-m} \psi^m \cdot \prod_{i: y_i \neq 0} f_{tr}(y_i; \theta)$$

which is separable for  $\psi$  and  $\theta$ .

**Theorem 2.1.** Let  $\theta_* = \operatorname{argmax}_{\theta} \prod_{i: y_i \neq 0} f_{tr}(y_i; \theta)$ . The maximum likelihood estimate  $(\hat{\phi}, \hat{\theta})$  maximizing (Equation 2.6) can be obtained as follows:

- (1) If  $m/n \leq 1 - p_0(\theta_*)$ , then  $\hat{\theta} = \theta_*$  and  $\hat{\phi} = 1 - m/n \cdot (1 - p_0(\theta_*))^{-1}$ .
- (2) Otherwise,  $\hat{\theta} = \operatorname{argmax}_{\theta} (1 - \psi(\theta))^{n-m} \psi(\theta)^m \prod_{i: y_i \neq 0} f_{tr}(y_i; \theta)$  and  $\hat{\phi} = 1 - \psi(\hat{\theta}) \cdot (1 - p_0(\hat{\theta}))^{-1}$ ,  
where  $\psi(\theta) = \min\{m/n, 1 - p_0(\theta)\}$ .

**Proof of Theorem 2.1:** First of all, we denote  $\psi_* = \operatorname{argmax}_{\psi} (1 - \psi)^{n-m} \psi^m$  and  $\theta_* = \operatorname{argmax}_{\theta} \prod_{i: y_i \neq 0} f_{tr}(y_i; \theta)$ .

It can be verified that  $\psi_* = m/n$ .

On the other hand,  $\psi = (1 - \phi)[1 - p_0(\theta)]$  with  $\phi \in [0, 1]$ , which implies  $\psi \in [0, 1 - p_0(\theta)]$ . If  $m/n \leq 1 - p_0(\theta_*)$ , then  $\hat{\psi} = m/n$ ,  $\hat{\theta} = \theta_*$  is the mle. In this case, the mle of  $\phi$  is  $\hat{\phi} = 1 - \hat{\psi}(1 - p_0(\theta_*))^{-1}$ .

Otherwise, we have  $m/n > 1 - p_0(\theta_*)$ . Then  $\hat{\psi} = \psi(\theta) = \min\{m/n, 1 - p_0(\theta)\}$  is the mle of  $\phi$  given  $\theta$ . In order to find the mle of  $\phi$  and  $\theta$ , we first find  $\theta^* = \operatorname{argmax}_{\theta} L(\psi(\theta), \theta)$ . Then  $\hat{\theta} = \theta^*$  and  $\hat{\psi} = \psi(\theta^*)$ .  $\square$

In order to apply Theorem 2.1, we need to deal with two maximization problems,  $\theta_* = \operatorname{argmax}_{\theta} L_{tr}(\theta)$  and  $\theta^* = \operatorname{argmax}_{\theta} L(\psi(\theta), \theta)$ , where  $L_{tr}(\theta) = \prod_{i:y_i \neq 0} f_{\theta}(y_i; \theta)$ . Note that

$$\begin{aligned} \frac{\partial \log L_{tr}(\theta)}{\partial \theta} &= \sum_{i:y_i \neq 0} \frac{\partial \log f_{\theta}(y_i)}{\partial \theta} - m \frac{\partial \log[1 - p_0(\theta)]}{\partial \theta} \\ \frac{\partial \log L(\psi(\theta), \theta)}{\partial \theta} &= \sum_{i:y_i \neq 0} \frac{\partial \log f_{\theta}(y_i)}{\partial \theta} - m \frac{\partial \log[1 - p_0(\theta)]}{\partial \theta}, \text{ if } 1 - p_0(\theta) > \frac{m}{n} \\ \frac{\partial \log L(\psi(\theta), \theta)}{\partial \theta} &= \sum_{i:y_i \neq 0} \frac{\partial \log f_{\theta}(y_i)}{\partial \theta} + (n - m) \frac{\partial \log p_0(\theta)}{\partial \theta}, \text{ if } 1 - p_0(\theta) < \frac{m}{n} \end{aligned}$$

Note that

$$\frac{\partial \log[1 - p_0(\theta)]}{\partial \theta} = -\frac{p_0(\theta)}{1 - p_0(\theta)} \cdot \frac{\partial \log p_0(\theta)}{\partial \theta}$$

Thus only  $\partial \log f_{\theta}(y)/\partial \theta$  and  $\partial \log p_0(\theta)/\partial \theta$ , or equivalently,  $\partial f_{\theta}(y)/\partial \theta$  and  $\partial p_0(\theta)/\partial \theta$ , are needed.

**Example 2.4. Zero-inflated negative binomial model (ZINB)** The pmf of negative-binomial distribution is

$$f(y, r, p) = \binom{y+r-1}{y} p^y (1-p)^r, y = 0, 1, 2, \dots$$

Where  $\binom{y+r-1}{y} = \frac{\Gamma(y+r)}{y!\Gamma(r)}$ , and  $r > 0$  is the number of failures,  $k$  is the number of successes, and  $p$  is the probability of success.

$$\begin{aligned}
p_0(\theta) &= (1-p)^r \\
\log(p(\theta)) &= \log \Gamma(y+r) - \log \Gamma(y+1) - \log \Gamma(r) + y \log(p) + r \log(1-p) \\
\log(p(\theta)) &= \log \Gamma(y+r) - \log \Gamma(y+1) - \log \Gamma(r) + y \log(p) + r \log(1-p) \\
\frac{\partial \log p_y(\theta)}{\partial r} &= \psi(y+r) - \psi(r) + \log(1-p) \\
\frac{\partial \log p_y(\theta)}{\partial p} &= \frac{y}{p} - \frac{r}{1-p} \\
\frac{\partial \log p_0(\theta)}{\partial r} &= \log(1-p) \\
\frac{\partial \log p_y(\theta)}{\partial p} &= \frac{-r}{1-p}
\end{aligned}$$

□

**Example 2.5. Zero-inflated beta-binomial model (ZIBB)** Let  $\theta = (n, \alpha, \beta)$ . The pmf of beta-binomial distribution is

$$f_{\theta}(y) = \binom{n}{y} \frac{\text{Beta}(y + \alpha, n - y + \beta)}{\text{Beta}(\alpha, \beta)}$$

with  $y = 0, 1, \dots, n$  and

$$\begin{aligned}
p_0(\theta) &= \frac{\Gamma(n + \beta)\Gamma(\alpha + \beta)}{\Gamma(n + \alpha + \beta)\Gamma(\beta)} \\
\frac{p_0(\theta)}{1 - p_0(\theta)} &= \frac{\Gamma(n + \beta)\Gamma(\alpha + \beta)}{\Gamma(n + \alpha + \beta)\Gamma(\beta) - \Gamma(n + \beta)\Gamma(\alpha + \beta)}
\end{aligned}$$

Let  $\Psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ , known as the *digamma* function. In order to apply Theorem 2.1, we need the formulae as follows

$$\begin{aligned}
\frac{\partial \log f_{\theta}(y)}{\partial n} &= \Psi(n+1) - \Psi(n-y+1) + \Psi(n-y+\beta) - \Psi(n+\alpha+\beta) \\
\frac{\partial \log f_{\theta}(y)}{\partial \alpha} &= \Psi(y+\alpha) - \Psi(n+\alpha+\beta) + \Psi(\alpha+\beta) - \Psi(\alpha) \\
\frac{\partial \log f_{\theta}(y)}{\partial \beta} &= \Psi(n-y+\beta) - \Psi(n+\alpha+\beta) + \Psi(\alpha+\beta) - \Psi(\beta) \\
\frac{\partial \log p_0(\theta)}{\partial n} &= \Psi(n+\beta) - \Psi(n+\alpha+\beta) \\
\frac{\partial \log p_0(\theta)}{\partial \alpha} &= \Psi(\alpha+\beta) - \Psi(n+\alpha+\beta) \\
\frac{\partial \log p_0(\theta)}{\partial \beta} &= \Psi(n+\beta) + \Psi(\alpha+\beta) - \Psi(n+\alpha+\beta) - \Psi(\beta)
\end{aligned}$$

### 2.3.2 Consistency of Hurdle and zero-inflated distributions

- **Consistency of Hurdle Distribution**

*Lemma 2.1.* Let  $Y_1, \dots, Y_n$  be a random sample from Hurdle model (Equation 2.1) and  $l(\phi, \theta)$  be the loglikelihood function. Then

$$E\left(\frac{\partial l}{\partial \phi}\right) = 0 \text{ and } E\left(\frac{\partial l}{\partial \theta}\right) = \frac{n(1-\phi)}{1-p_0(\theta)} \cdot E\left[\frac{\partial \log f_{\theta}(Y')}{\partial \theta}\right]$$

which is 0 if and only if  $E[\partial \log f_{\theta}(Y')/\partial \theta] = 0$ , where  $Y'$  follows the baseline distribution  $f_{\theta}(y)$ .

**Proof of Lemma 2.1:** The loglikelihood function of Hurdle model is

$$\begin{aligned} l(\phi, \theta) &= \log \phi \cdot \sum_{i=1}^n \mathbf{1}_{\{Y_i=0\}} + \log(1 - \phi) \cdot \sum_{i=1}^n \mathbf{1}_{\{Y_i \neq 0\}} \\ &\quad - \log[1 - p_0(\theta)] \sum_{i=1}^n \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i=1}^n \log f_\theta(Y_i) \mathbf{1}_{\{Y_i \neq 0\}} \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial l}{\partial \phi} &= \frac{1}{\phi} \cdot \sum_{i=1}^n \mathbf{1}_{\{Y_i=0\}} - \frac{1}{1 - \phi} \cdot \sum_{i=1}^n \mathbf{1}_{\{Y_i \neq 0\}} \\ \frac{\partial l}{\partial \theta} &= \frac{p_0(\theta)}{1 - p_0(\theta)} \cdot \frac{\partial \log p_0(\theta)}{\partial \theta} \sum_{i=1}^n \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i=1}^n \frac{\partial \log f_\theta(Y_i)}{\partial \theta} \mathbf{1}_{\{Y_i \neq 0\}} \end{aligned}$$

Since  $P(Y_i = 0) = \phi$ , then  $E(\partial l / \partial \phi) = 0$ . Let  $Y'_1, \dots, Y'_n$  be iid  $\sim f_\theta(y)$ . Then

$$\begin{aligned} E \left[ \frac{\partial \log f_\theta(Y_i)}{\partial \theta} \mathbf{1}_{\{Y_i \neq 0\}} \right] &= \frac{1 - \phi}{1 - p_0(\theta)} \cdot E \left[ \frac{\partial \log f_\theta(Y'_i)}{\partial \theta} \mathbf{1}_{\{Y'_i \neq 0\}} \right] \\ &= \frac{1 - \phi}{1 - p_0(\theta)} \cdot \left\{ E \left[ \frac{\partial \log f_\theta(Y'_i)}{\partial \theta} \right] - p_0(\theta) \cdot \frac{\partial \log p_0(\theta)}{\partial \theta} \right\} \end{aligned}$$

Then

$$E \left( \frac{\partial l}{\partial \theta} \right) = \frac{1 - \phi}{1 - p_0(\theta)} \cdot \sum_{i=1}^n E \left[ \frac{\partial \log f_\theta(Y'_i)}{\partial \theta} \right] = \frac{n(1 - \phi)}{1 - p_0(\theta)} \cdot E \left[ \frac{\partial \log f_\theta(Y'_1)}{\partial \theta} \right]$$

□

As a direct corollary of Theorem 17 in (51), the MLEs of Hurdle model have strong consistency under fairly general conditions.

*Theorem 2.2.* Let  $Y_1, \dots, Y_n$  be a random sample from Hurdle model (Equation 2.1) with true parameter value  $(\phi_0, \theta_0) \in (0, 1) \times \Theta$ , where  $\Theta$  is compact. Let

$$\hat{\phi} = \frac{\#\{i : Y_i = 0\}}{n}, \quad \hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \prod_{i: Y_i \neq 0} f_{tr}(Y_i; \theta)$$

Suppose (1)  $f_{\theta}(y)$  is continuous in  $\theta$  for all  $y$ ; (2)  $f_{\theta}(y) = f_{\theta_0}(y)$  for all  $y$  always implies  $\theta = \theta_0$ ; and (3) there exists a nonnegative function  $K(y)$  such that  $E[K(Y)] < \infty$  for  $Y \sim f_{tr}(y; \theta_0)$  and  $\log[f_{tr}(y; \theta)/f_{tr}(y; \theta_0)] \leq K(y)$  for all  $y \neq 0$  and  $\theta \in \Theta$ . Then

$$\hat{\phi} \xrightarrow{a.s.} \phi_0, \quad \hat{\theta}_n \xrightarrow{a.s.} \theta_0$$

as  $n$  goes to infinity.

*Example 2.6.* For zero-altered Poisson or Hurdle Poisson distribution, the pmf of the baseline distribution is  $f_{\lambda}(y) = e^{-\lambda} \lambda^y / y!$  with  $p_0(\lambda) = e^{-\lambda}$ . It can be verified that

$$E\left(\frac{\partial \log f_{\lambda}(Y')}{\partial \lambda}\right) = 0$$

if  $Y' \sim f_{\lambda}(y)$ . The truncated pmf is

$$f_{tr}(y; \lambda) = \frac{e^{-\lambda}}{1 - e^{-\lambda}} \cdot \frac{\lambda^y}{y!}, \quad y = 1, 2, \dots$$



The loglikelihood for the zero-truncated Poisson is

$$l(\lambda) = -m\lambda - m \log(1 - e^{-\lambda}) + \sum_{i:Y_i>0} Y_i \cdot \log \lambda - \log\left(\prod_{i:Y_i>0} Y_i!\right)$$

The MLE  $\hat{\lambda}$  of  $\lambda$  solves the likelihood equation  $\lambda = \bar{Y}(1 - e^{-\lambda})$  with  $\bar{Y} = m^{-1} \sum_{i:Y_i>0} Y_i$ , which can be solved numerically. If the true value  $\lambda_0 \in [\lambda_1, \lambda_2]$  for some  $0 < \lambda_1 < \lambda_2 < \infty$ , then  $K(y)$  in Theorem 2.2 can be chosen as

$$K(y) = \log \frac{\lambda_2}{\lambda_1} \cdot y + \log \frac{1 - e^{-\lambda_2}}{1 - e^{-\lambda_1}} + \lambda_2 - \lambda_1$$

Since there is no difference in practice as long as  $0 < \lambda_1 < \hat{\lambda} < \lambda_2 < \infty$ , we know  $\hat{\lambda} \xrightarrow{a.s.} \lambda_0$  as  $n$  goes to infinity.  $\square$

*Example 2.7.* For zero-altered negative binomial or Hurdle negative binomial distribution, the pmf of the baseline distribution with parameters  $\theta = (r, p) \in (0, \infty) \times [0, 1]$  is given by  $f_{\theta}(y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p^y (1-p)^r$ ,  $y \in \{0, 1, 2, \dots\}$ . Then  $p_0(\theta) = (1-p)^r$ . In order to apply Lemma 2.1, we obtain

$$\begin{aligned} \log f_{\theta}(y) &= \log \Gamma(y+r) - \log \Gamma(y+1) - \log \Gamma(r) + y \log p + r \log(1-p) \\ \frac{\partial \log f_{\theta}(y)}{\partial r} &= \Psi(y+r) - \Psi(r) + \log(1-p) \\ \frac{\partial \log f_{\theta}(y)}{\partial p} &= \frac{y}{p} - \frac{r}{1-p} \end{aligned}$$

where  $\Psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$  is known as the *digamma* function.

If  $Y' \sim f_\theta(y)$ , then  $E(Y') = pr/(1-p)$  and  $E(\partial \log f_\theta(Y')/\partial p) = 0$ . On the other hand, since

$\Gamma(y) = \int_0^\infty t^{y-1} e^{-t} dt$  and  $\Gamma'(y) = \int_0^\infty t^{y-1} e^{-t} \log t dt$  for  $y > 0$ , then

$$\begin{aligned}
E(\Psi(Y' + r)) &= \sum_{y=0}^{\infty} \frac{\Gamma'(y+r)}{\Gamma(y+r)} \cdot \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p^y (1-p)^r \\
&= \frac{(1-p)^r}{\Gamma(r)} \sum_{y=0}^{\infty} \frac{p^y}{y!} \Gamma'(y+r) \\
&= \frac{(1-p)^r}{\Gamma(r)} \sum_{y=0}^{\infty} \frac{p^y}{y!} \int_0^\infty t^{y+r-1} e^{-t} \log t dt \\
&= \frac{(1-p)^r}{\Gamma(r)} \int_0^\infty \left( \sum_{y=0}^{\infty} \frac{(pt)^y}{y!} e^{-pt} \right) \cdot t^{r-1} e^{-t(1-p)} \log t dt \\
&= \frac{(1-p)^r}{\Gamma(r)} \int_0^\infty t^{r-1} e^{-t(1-p)} \log t dt \quad (\text{let } s = (1-p)t) \\
&= \frac{(1-p)^r}{\Gamma(r)} \int_0^\infty s^{r-1} e^{-s} [\log s - \log(1-p)] ds \cdot (1-p)^{-r} \\
&= \frac{1}{\Gamma(r)} \left[ \int_0^\infty s^{r-1} e^{-s} \log s ds - \int_0^\infty s^{r-1} e^{-s} \log(1-p) ds \right] \\
&= \frac{1}{\Gamma(r)} [\Gamma'(r) - \log(1-p)\Gamma(r)] \\
&= \Psi(r) - \log(1-p)
\end{aligned}$$

Therefore,  $E(\partial \log f_\theta(Y')/\partial r) = E(\Psi(Y' + r)) - \Psi(r) + \log(1-p) = 0$ . By Lemma 2.1, we

know that Hurdle negative binomial distribution satisfies the regularity conditions.  $\square$

- **Consistency of Zero-inflated Distribution**

*Lemma 2.2.* Let  $Y_1, \dots, Y_n$  be a random sample from the zero-inflated model (Equation 2.5) and  $l(\phi, \theta)$  be the loglikelihood function. Suppose  $0 \leq \phi < 1$ . Then

$$E\left(\frac{\partial l}{\partial \phi}\right) = 0 \text{ and } E\left(\frac{\partial l}{\partial \theta}\right) = n(1 - \phi)E\left[\frac{\partial \log f_{\theta}(Y')}{\partial \theta}\right]$$

which is 0 if and only if  $E[\partial \log f_{\theta}(Y')/\partial \theta] = 0$ , where  $Y'$  follows the baseline distribution  $f_{\theta}(y)$ .

**Proof of Lemma 2.2:** The loglikelihood of the zero-inflated model is

$$\begin{aligned} l(\phi, \theta) &= \log[\phi + (1 - \phi)p_0(\theta)] \cdot \sum_{i=1}^n \mathbf{1}_{\{Y_i=0\}} \\ &\quad + \log(1 - \phi) \cdot \sum_{i=1}^n \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i=1}^n \log f_{\theta}(Y_i) \mathbf{1}_{\{Y_i \neq 0\}} \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial l}{\partial \phi} &= \frac{1 - p_0(\theta)}{\phi + (1 - \phi)p_0(\theta)} \cdot \sum_{i=1}^n \mathbf{1}_{\{Y_i=0\}} - \frac{1}{1 - \phi} \cdot \sum_{i=1}^n \mathbf{1}_{\{Y_i \neq 0\}} \\ \frac{\partial l}{\partial \theta} &= \frac{(1 - \phi)p_0(\theta)}{\phi + (1 - \phi)p_0(\theta)} \cdot \frac{\partial \log p_0(\theta)}{\partial \theta} \sum_{i=1}^n \mathbf{1}_{\{Y_i=0\}} + \sum_{i=1}^n \frac{\partial \log f_{\theta}(Y_i)}{\partial \theta} \mathbf{1}_{\{Y_i \neq 0\}} \end{aligned}$$

Since  $P(Y_i = 0) = \phi + (1 - \phi)p_0(\theta)$  and  $P(Y_i \neq 0) = (1 - \phi)[1 - p_0(\theta)]$ , then  $E(\partial l/\partial \phi) = 0$ .

Let  $Y'_1, \dots, Y'_n$  be iid  $\sim f_{\theta}(y)$ . Then

$$\begin{aligned} E\left[\frac{\partial \log f_{\theta}(Y_i)}{\partial \theta} \mathbf{1}_{\{Y_i \neq 0\}}\right] &= (1 - \phi) \cdot E\left[\frac{\partial \log f_{\theta}(Y'_i)}{\partial \theta} \mathbf{1}_{\{Y'_i \neq 0\}}\right] \\ &= (1 - \phi) \left\{ E\left[\frac{\partial \log f_{\theta}(Y'_i)}{\partial \theta}\right] - p_0(\theta) \cdot \frac{\partial \log p_0(\theta)}{\partial \theta} \right\} \end{aligned}$$

Then

$$E\left(\frac{\partial l}{\partial \theta}\right) = (1 - \phi) \sum_{i=1}^n E\left[\frac{\partial \log f_{\theta}(Y'_i)}{\partial \theta}\right] = n(1 - \phi)E\left[\frac{\partial \log f_{\theta}(Y'_1)}{\partial \theta}\right]$$

□

### 2.3.3 Sampling from general Hurdle and zero-inflated models

There are some functions in R for sampling data from some distributions such as “rpois” to generate a random samples from Poisson distribution, “rnbino” to generate a random samples from Negative Binomial distribution. To the best of our knowledge, there is no such ways for sampling from some probabilistic models such as zero-inflated beta binomial.

- **Sampling from general Hurdle models**

Here we propose an efficient algorithm based on the Central Limit Theorem to simulate data for from a general Hurdle model with reduced loops. The probability mass function of a general Hurdle model is defined by:  $P_H(Y = k) = \phi 1_{[k=0]} + (1 - \phi)P_{tr}(Y = k)$ , where  $\phi \geq 0$  is a weight parameter of zeros,  $P_{tr}(Y_i = k)$  is a zero-truncated version of the baseline distribution determined by  $P_{tr}(Y_i = 0) = 0$  and  $P_{tr}(Y_i = k) = P_{\theta}(Y_i = k)/[1 - P_{\theta}(Y = 0)]$  for  $k > 0$ . Given  $n, \phi$  and  $\theta$ , a general algorithm is described as follows:

- (i) Simulate  $X_1, X_2, \dots, X_n$  iid  $\sim \text{Bernoulli}(1 - \phi)$ .
- (ii) Let  $m = \{i : X_i \neq 0\}$  and simulate  $Y_1, Y_2, \dots, Y_m$  iid  $\sim P_{tr}(Y = k)$ , the zero-truncated distribution. More specifically, we first let  $M = (1 - p_0)^{-1} \left[ m + 2\sqrt{p_0(m + p_0)} + 2p_0 \right]$  and simulate  $Z_1, Z_2, \dots, Z_M$  iid  $\sim P(Y = k)$ , the original distribution before truncation, where  $p_0 = P(Y = 0)$ ; then we remove all zeros from  $Z_i$ 's and get  $U_1, U_2, \dots, U_t$ . If  $t \geq m$  we end with  $U_1, U_2, \dots, U_m$ ;

otherwise we use a reject-accept procedure to obtain  $U_{t+1}, U_{t+2}, \dots, U_m$  from the zero-truncated distribution. It can be verified that  $t$  is fairly closed to  $m$  if it is less than  $m$ .

Note that in step (ii), let  $T = \#\{i \mid Z_i \neq 0\}$ . Then  $T \sim \text{Binomial}(M, 1 - p_0)$ . Therefore,  $E(T) = M(1 - p_0)$  and  $\text{Var}(T) = Mp_0(1 - p_0)$ . According to the central limit theorem,  $(T - E(T))/\sqrt{\text{Var}(T)} \sim N(0, 1)$ . It can be verified that  $(m - M(1 - p_0))/\sqrt{Mp_0(1 - p_0)} = -2$ . Therefore,  $P(T < m) \approx 0.023$ .

- **Sampling From General Zero-Inflated Models**

Sampling from Zero inflated distribution is much easier than Hurdle distribution. The probability mass function of a general zero-inflated model is given by:  $P_{ZI}(Y_i = k) = \phi \mathbf{1}_{\{k=0\}} + (1 - \phi)P(Y_i = k)$ , where  $\phi \geq 0$  is a weight parameter of zeros. To generate a dataset from Zero inflated distribution, we (i) simulate  $Z_1, Z_2, \dots, Z_n$  iid Bernoulli( $\phi$ ), (ii) if  $Z_i = 1$ , let  $X_i = 0$ . If  $Z_i = 0$ , sample  $X_i \sim p$ .

### **2.3.4 Bootstrapped Monte Carlo estimate for the p-value of a discrete KS test**

For a continuous distribution with unknown parameters, Monte Carlo simulation based on the estimated parameters has been used to correct the biased p-values of the KS test (see R package “KScorrect”). Since the estimated parameters may not be reliable for small samples or inappropriate distribution assumptions, we propose a bootstrapped Monte Carlo estimate for estimating the p-value of discrete KS test.

More specifically, the goal is to test if the sample  $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$  comes from the discrete distribution with CDF  $F_\theta(x)$  where the parameter  $\theta$  is unknown. Algorithm 1 provides the details.

---

**Algorithm 1** Estimating p-value of Discrete KS test
 

---

- 1: Given  $\mathbf{X} = (X_1, X_2, \dots, X_n)$
  - 2: For  $b = 1, \dots, B$ , Resample  $X$  with replacement to get a bootstrapped sample  $\mathbf{X}^{(b)} = \{X_1^{(b)}, \dots, X_n^{(b)}\}$ .
  - 3: For each  $b$ , calculate the MLE  $\hat{\theta}^{(b)}$  of  $\theta$ .
  - 4: Simulate  $\mathbf{X}^{(c)} = \{X_1^{(c)}, \dots, X_n^{(c)}\}$  iid from  $F_{\hat{\theta}^{(b)}}$ , which is the CDF  $F_{\theta}(x)$  with parameter  $\theta = \hat{\theta}^{(b)}$ .
  - 5: Calculate the KS statistic  $D_n^{(b)} = \sup_x |\hat{F}_n^{(c)}(x) - F_{\hat{\theta}^{(b)}}(x)|$ , where  $\hat{F}_n^{(c)}(x)$  is the empirical distribution function of  $\mathbf{X}^{(c)}$ .
  - 6: Estimate the p-value by  $\frac{\#\{b | D_n^{(b)} > D_n\} + 1}{B + 1}$  where  $D_n = \sup_x |\hat{F}_n(x) - F_{\hat{\theta}}(x)|$  is the KS statistic based on the original data and its MLE  $\hat{\theta}$ .
- 

In step (6), we add 1 to both the numerator and denominator to avoid zero p-value (52).

For step (4) above, we can simulate data easily from the regular distribution using standard R functions. For example, `rpois` function is used to generate data from Poisson distribution, `rnbinom` to generate Negative Binomial dataset. In addition, there are a few other R packages for generating a dataset from some Hurdle distribution. For example, “countreg” package provides a `hpois` function for generating dataset from PH distribution.

In our study, we allow the use of some new distributions such as BB and BNB distributions, as well as the corresponding zero-inflated and Hurdle models.

• **An Illustrative Example**

Reconsider the example of  $X_1, \dots, X_{5000} \sim \text{NB}(2, 0.5)$  in Section 2.1 (see the example in Strategy 1). We are interested in testing whether the data follows a NB distribution or not. By using our discrete KS test with unknown parameters, we obtain a large p-value (0.867) based on the bootstrapped Monte Carlo simulation, which means the data follows a NB distribution. As a

comparison, by using the regular KS test mentioned in Strategy 1 of Section 2.1, we obtain a very small p-value ( $8.8 \times 10^{-8}$ ). It shows that our estimate is more reliable.

*Example 2.8.* “dgof” is another existing R package for computing the exact p-value for some discrete distributions. In this example, we consider a random samples  $X_1, \dots, X_{100} \sim \text{NB}(2, 0.01)$ . We are interested in testing whether the data follows a NB distribution or not. By using our discrete KS test with unknown parameters, we obtain a non-significant p-value (0.23) based on the bootstrapped Monte Carlo simulation, which means the data follows a NB distribution. As a comparison, by using the KS.test in “dgof” R package, we obtain a very small p-value ( $2.2 \times 10^{-16}$ ). It shows that our estimate is more reliable when  $p$  is small.

```
> set.seed(2310)

> x=rnbinom(100, 2, 0.01)

> ks.test(x, ecdf(rnbinom(100, 2, 0.01)), exact=TRUE) #using "dgof" package

One-sample Kolmogorov-Smirnov test

data:  x

D = 0.22, p-value < 2.2e-16

alternative hypothesis: two-sided

> dis.kstest(x, nsim=100, bootstrap=TRUE, distri="nb")$pvalue

[1] 0.23
```

### 2.3.5 Likelihood ratio test for selecting the best model

In case two or more distributions pass the KS test, we perform a likelihood ratio test for selecting one model against another. According to the Neyman-Pearson theorem, the test based on likelihood ratio is the most powerful one. More specifically, suppose we are interested in testing the hypothesis  $H_0 : X_1, X_2, \dots, X_n \text{ iid} \sim f(x; \theta)$  with unknown parameter  $\theta$  against  $H_1 : X_1, X_2, \dots, X_n \text{ iid} \sim g(x; \phi)$  with unknown parameter  $\phi$ . The likelihood ratio test statistic is given by:

$$\Lambda = \log \frac{\prod_{i=1}^n f(x_i, \hat{\theta})}{\prod_{i=1}^n g(x_i, \hat{\phi})} \quad (2.8)$$

where  $\hat{\phi}$  and  $\hat{\theta}$  are the the corresponding MLE's. The Neyman-Pearson theorem (53) guarantees that the test based on  $\lambda$  is the most powerful test given a significance level.

To estimate the p-value of the likelihood ratio test, we use the following Bootstrapped Monte Carlo algorithm:

---

**Algorithm 2** Estimating p-value of Likelihood Ratio Test

---

- 1: Given  $\mathbf{X} = (X_1, X_2, \dots, X_n)$
  - 2: For  $b = 1, \dots, B$ , resample  $X$  with replacement to get a bootstrapped sample  $\mathbf{X}^{(b)} = \{X_1^{(b)}, \dots, X_n^{(b)}\}$ .
  - 3: For each  $b$ , calculate the MLE  $\hat{\theta}^{(b)}$  and  $\hat{\phi}^{(b)}$ .
  - 4: Simulate  $Z_1^{(b)}, \dots, Z_n^{(b)}$  from the null distribution  $f(x; \theta^{(b)})$ .
  - 5: Calculate  $\Lambda^{(b)} = \Lambda(Z_1^{(b)}, \dots, Z_n^{(b)})$ .
  - 6: Estimate the p-value by  $\frac{\#\{b: \Lambda^{(b)} < \Lambda\} + 1}{B + 1}$ .
-



Small p-value indicates that the two distributions are different, and  $g(x, \phi)$  is the more appropriate distribution. Large p-value means that there is no statistical difference between the two distributions.

For example, Based on KS discrete test, one feature follows both ZIBB and ZINB distributions. To see which distribution is more appropriate, we use the Nerman Pearson lemma as have discussed above. We test  $H_0 : X_1, X_2, \dots, X_n \sim \text{ZIBB}$  vs.  $H_1 : X_1, X_2, \dots, X_n \sim \text{ZINB}$  with unknown parameter  $\phi$ . The test p-value is significant (p-value = 0.0341). We conclude that ZINB is more appropriate for modeling that feature than ZIBB.

## 2.4 Results

We applied the proposed discrete KS test with unknown parameters to a list of 229 bacterial and fungal OTUs from (Tipton *et al.* study) (54). We are interested in knowing how many of the 229 OTU follows each of the following distributions: Poisson, Negative Binomial, Beta-Binomial, Beta Negative Binomials, and the corresponding zero-inflated and Hurdle models.

Table III summarizes the number of species with not statistically significant p-value (KS p-value > 0.05).

As shown, Poisson, zero-inflated Poisson, and Poisson Hurdle are not appropriate distributions to model sparse microbial features as only 0.4%, 2%, 1% out of 229 the features were able to be appropriately fitted using these distributions, respectively. On the other hand, binomial and negative binomial families can be used to approximate sparse microbial data, with BNBH as the best distribution to model such dataset (being able to appropriately fit 53% of the 229 features) using the proposed conservative method.

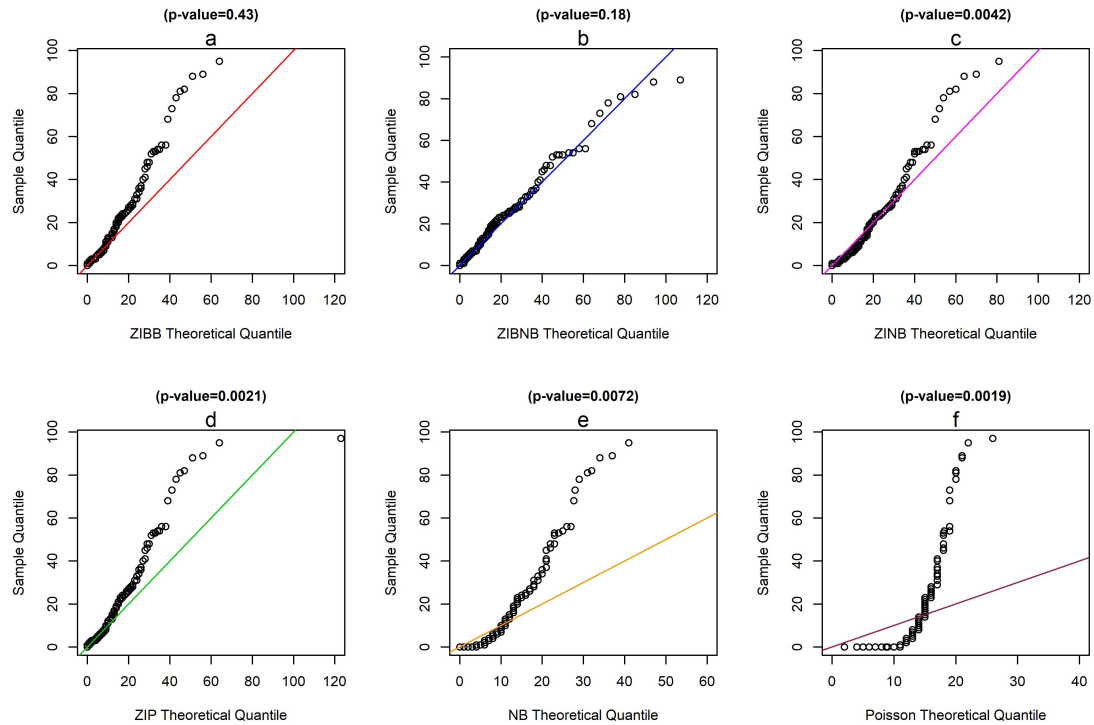


Figure 1: Quantile-quantile plot between six different theoretical distributions for a specific feature. Each figure represents a different discrete distribution: (a) ZIBB distribution, (b) ZIBNB distribution, (c) ZINB distribution, (d) ZIP distribution, (e) NB distribution, (f) Poisson distribution. The p-value above each figure represents the significance of the discrete KS test between the theoretical quantiles and the sample quantiles of the corresponding distribution. The ZIBB, and ZIBNB distribution are most appropriate distributions for modeling the OTU count among some other standard distributions

Figure 2 shows the quantile-quantile or q-q plot for modeling one arbitrary read count using the following different distributions (ZIBB, ZIBNB, ZINB, ZIP, NB, and Poisson). The vertical line represents the Sample quantiles, and the horizontal line represents the theoretical quantiles which are calculated by generated a random numbers from each of the distributions. If the data follow the assumed distribution, then the q-q plot points will fall approximately on a straight line.

TABLE III: NUMBER AND PERCENTAGE OF SPECIES OUT OF 229 SPECIES WITH NOT STATISTICALLY SIGNIFICANT P-VALUE ( P-VALUE > 0.05)

Distribution	Number	Percentage
Poisson	1	0.4%
NB	23	10%
BB	76	33%
BNB	60	26%
ZIP	3	2%
ZINB	25	11%
ZIBB	89	39%
ZIBNB	110	48%
PH	2	1%
NBH	56	24%
BBH	92	40%
BNBH	121	53%

The p-value above each sub-figure of Figure 2 represents the discrete KS test p-value. High p-value indicates that the feature follows a specific discrete distribution, and small p-value indicates that the feature does not follow that distributions. Most of q-q plot points of graphs (a), and (b) lie on the straight line, which means that the feature follows ZIBB, and ZIBNB. This result is consistent with their large p-value.

We conclude that zero-inflated and Hurdle models for Beta-Binomial and Beta Negative Binomial are the most appropriate models for Tipton *et al.* dataset.

## **2.5 iZID R Package**

### **2.5.1 Introduction**

In this chapter we have seen that when conducting one-sample Kolmogorov-Smirnov (KS) test for count data, the estimated p-value is biased due to plugging in sample estimates of unknown parameters. To overcome the bias issue induced by plugging in estimated parameters, We proposed a bootstrapped Monte Carlo procedure to estimate the p-value of a KS test for discrete probabilistic models (40). In the circumstance that more than one models pass the KS tests, We proposed a bootstrapped procedure for estimating the p-values of the likelihood ratio tests for pairwise comparisons of candidate models (40). We develop a new R package named “iZID” for identifying Zero-Inflated and Hurdle Distributions, available from the Comprehensive R Archive Network (CRAN, <https://cran.r-project.org/>) (41). For user’s convenience, we cover regular Poisson, negative binomial, beta binomial, and beta negative binomial distributions as well. Using “iZID”, the p-value is estimated by counting the number of random samples whose KS test statistics are greater than the KS statistic derived from the original data. Since the random samples are generated using the maximum likelihood estimates obtained from the bootstrapped or original data, the resulting p-value is automatically adjusted for the influence of plugging in sample estimates.

### **2.5.2 Existing R packages for analyzing zero-inflated data**

Several packages are currently available from the Comprehensive R Archive Network (CRAN) for analyzing zero-inflated data, including “bzinb”, “hurdlr”, “mazeinda”, “mhurdle”, “rbtt”, “ZIBBSeqDiscovery”, “ZIBseq”, “zic”, “ZIM”, “ziphsmm”, etc.

Package “bzinb” provides tools for random sample generation, maximum likelihood estimation and log likelihood computation for bivariate zero-inflated negative binomial and Poisson distributions. With “hurdlr”, users are able to fit hurdle or zero-inflated negative binomial and Poisson regression models using Bayesian strategy. Package “mazeinda” is tailored to compute and test the significance of pairwise monotonic association for count data with any degree of zero-inflation. The creation of “mhurdle” is inspired by the households’ expenditure survey data where many zeros exist in predictors recording the cost of some goods or activities. The function `mhurdle` in package “mhurdle” enables the estimation of a large class of regression models with up to three hurdles, which allows that zero observations in predictors occur by up to three structural reasons. Package “rbtt” tries to tackle the inflation of type I error in two-sample t-tests comparing two groups of zero-inflated data via robust bootstrapped test. Package “ZIBBSeqDiscovery” models the relationship between the count data and some covariates of interest by zero-inflated beta-binomial models. Package “ZIBseq” regresses the counts on categorical clinical conditions in zero-inflated beta models. Package “zic” outputs the Bayesian estimate of zero-inflated count models while assuming that the parameters follow certain prior distributions. Package “ZIM” enables both observation-driven and parameter-driven modeling for time series with excess zeros. Package “ziphsmm” analyses longitudinal continuous-time data via zero-inflated Poisson hidden (semi-)Markov models.

Except for packages “mazeinda” and “rbtt”, the rest fit count data to specific models. To the best of our knowledge, our package “iZID” is the first one to conduct KS test for count data with p-values adjusted for the influence of sample estimate of unknown parameters. Example 2.9 below shows that our function `dis.kstest` is more reliable than the basic R function `ks.test` in estimating p-values.

*Example 2.9.* In this experiment, we simulate  $N = 100$  random numbers from a zero-inflated negative binomial (ZINB) distribution with parameters  $\phi = 0.6, r = 2, p = 0.01$ . The maximum likelihood estimates  $\hat{\phi} = 0.590, \hat{r} = 2.06, \hat{p} = 0.011$  are fairly accurate. Nevertheless, if one wants to test if the original sample from a ZINB distribution by simulating another random sample using the estimated parameter values, the classical R function `ks.test` rejects ZINB model with p-value 0.01 and a warning message. If we use our function `dis.kstest` in package “iZID”, the adjusted p-value is 0.12 which passes the ZINB model. For readers’ reference, we provide the R code and output below:

```
> set.seed(343)

> nsimu=100

> x=sample.zi(N=nsimu, phi=0.6, distri = "nb", r=2, p=0.01)

> mle=nb.zihmle(x, r=5, p=0.5, type="zi")

> mle
```

	r	p	phi	loglik
[1,]	2.058397	0.0112907	0.5899598	-316.7666

```
> y=sample.zi(N=nsimu, phi=mle[3], distri = "nb", r=mle[1], p=mle[2])

> ks.test(x,y)
```

Two-sample Kolmogorov-Smirnov test

data: x and y

D = 0.23, p-value = 0.01008

alternative hypothesis: two-sided

Warning message:

In `ks.test(x, y)` : p-value will be approximate in the presence of ties

```
> dis.kstest(x, nsim=200, bootstrap = TRUE, distri = "zinb")$pvalue
```

```
[1] 0.12
```

### 2.5.3 Architecture of the package “iZID”

The package “iZID” contains four main functions: `dis.kstest`, `model.lrt`, `sample.zi` and `sample.h`. Function `dis.kstest` computes bootstrapped or Monte Carlo p-value of one-sample KS test under a specific discrete distribution. Function `model.lrt` implements a likelihood ratio test to select between two candidate models, in the case that more than one models have p-values greater than the pre-specified significance level. Functions `sample.zi` and `sample.h` are random sample generators, where the former outputs random deviates of zero-inflated models and the latter generates random counts from hurdle models. This package also provides some miscellaneous functions to calculate maximum likelihood estimate and the corresponding log likelihood value for a large set of models modeling count data. To accelerate the calculation process, we parallelize the computation of bootstrapped Monte Carlo estimates using R package “doParallel” and “foreach”.

- **dis.kstest**

To estimate the p-value of a KS test given a pre-specified distribution as null hypothesis, the user may call the function `dis.kstest` with the syntax:

```
dis.kstest(x, nsim=100, bootstrap=TRUE, distri='Poisson', r=NULL, p=NULL,
alpha1=NULL, alpha2=NULL, n=NULL, lowerbound=0.01, upperbound=100000, parallel=FALSE)
```

<code>x</code>	Independent non-negative integers which stands for counts. Can be a vector or a matrix.
<code>nsim</code>	Number of bootstrapped samples generated for computing maximum likelihood estimate of unknown parameters.
<code>distri</code>	The distribution under null hypothesis. Currently, standard Poisson, negative binomial, beta binomial, beta negative binomial distributions as well as their zero-inflated and hurdle versions are available in the package. Accordingly, <code>distri</code> can be set to be one of {Poisson, nb, bb, bnb, zip, zinb, zibb, zibnb, ph, nbh, bbh, bnbh}. Note that users do not need to provide an estimate for unknown parameters. Instead, <code>dis.kstest</code> automatically carries out the task.
<code>r, p</code>	Optional arguments for assigning initial values of unknown parameters of standard, zero-inflated and hurdle negative binomial distributions.
<code>alpha(1,2)</code> and <code>n</code>	Optional arguments for assigning initial values of unknown parameters of standard, zero-inflated and hurdle beta binomial distributions.
<code>alpha(1,2)</code> and <code>r</code>	Optional arguments for assigning initial values of unknown parameters of standard, zero-inflated and hurdle beta negative binomial distributions.
<code>lowerbound</code>	The lower searching bound.
<code>upperbound</code>	The upper searching bound.



**Details:**

- `dis.kstest` will be initialized with naive sample estimates if initial values are not given. The negative log likelihood function is minimized via basic R function `optim` with the searching interval decided by `lowerbound` and `upperbound`, except that the optimization of `p` takes `1-lowerbound` as the upper searching bound.
- The way to calculate p-value of KS test is illustrated in Algorithm 1. Given a random sample  $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$ , `nsim` bootstrapped samples are obtained by resampling  $\mathbf{x}$  with replacement if setting `bootstrap=TRUE` (by default). If setting `bootstrap=FALSE`, `nsim` new samples will be simulated with mle of original data  $\mathbf{x}$ , and KS statistics of the new samples will be computed.
- `dis.kstest` returns an object of class “dis.kstest” which contains all the input values, maximum likelihood estimate of the bootstrapped samples and original data  $\mathbf{x}$ , and most importantly, the p-value.

- **model.lrt**

If the p-values returned by `dis.kstest` are not significant for more than one distributions, a likelihood ratio test can be performed to select a relatively “better” model for the data on hand.

The way to call `model.lrt` is as follows:

```
model.lrt(d1,d2,parallel = FALSE)
```

where `d1` and `d2` are two objects of class “dis.kstest” under different distributions. The likelihood ratio test statistic is the difference between log likelihood of the alternative and the null distribu-

tion decided by  $d_2$  and  $d_1$ , respectively. The algorithm is given in Algorithm 2. One may simulate  $nsim$  new samples under the null distribution using  $nsim$  mles inherited from  $d1\$mle\_new$ , and then calculate the differences between log likelihood of new samples under the alternative and the null hypotheses as statistics of the likelihood ratio tests. Function `dis.kstest` returns the proportion of test statistics of new samples that is greater than the statistic of original data  $\mathbf{x}$ . A small p-value indicates that the data on hand is more likely to come from the alternative distribution. Otherwise, the null distribution shows no significant difference to the alternative one.

- **Generate random samples from zero-inflated and hurdle distributions**

Random deviates from standard Poisson and negative binomial distributions can be generated by basic R functions `rpois` and `rnbinom`, respectively. With R package “ExtraDistr”, functions `rbbinom` and `rnbibinom` are available for standard beta binomial and beta negative binomial distributions, respectively. In addition, there are a few other R packages for generating a dataset from some hurdle distributions. For example, package “countreg” provides function `hpois` for generating dataset from Poisson Hurdle distribution.

In our package “iZID”, we allow the use of new distributions including beta binomial and beta negative binomial distributions, and more importantly, their corresponding zero-inflated and hurdle models. We introduced a procedure grounded upon the central limit theorem to produce random values from zero-inflated and hurdle models. In package “iZID”, we implement the procedure to the following two functions:

```
sample.zi(N,phi,distri='poisson',lambda=NA,r=NA,p=NA,
```

```

alpha1=NA,alpha2=NA, n=NA)}

sample.h(N,phi,distri='poisson',lambda=NA,r=NA,p=NA,

alpha1=NA,alpha2=NA, n=NA)

```

These two functions have exactly the same arguments. Here  $N$  represents the size of random sample to return. Argument `phi` stands for the value of structural parameter  $\phi$  in zero-inflated and hurdle models, e.g., formulae (Equation 2.5). The input `distri` currently belongs to the set of four standard distributions {Poisson, nb, bb, bnb}. For example, by setting `distri=nb`, `sample.zi` and `sample.h` return zero-inflated and hurdle negative binomial distributed random deviates, respectively. Arguments `lambda`, `r`, `p`, `alpha1`, `alpha2` and `n` are parameter values for different distributions, which must be specified. For instance, with `distri=nb`, users need to provide values for `r` and `p`.

- **Calculate maximum likelihood estimate and log likelihood**

In order to calculate the maximum likelihood estimate as well as the value of log likelihood of the aforementioned four standard distributions and their zero-inflated and hurdle versions, one may simply use the following lines of code with package “iZID”:

```

poisson.mle(x)

bb.mle(x,n,alpha1,alpha2,lowerbound = 0.01, upperbound = 10000)

nb.mle(x,r,p,lowerbound = 0.01, upperbound = 10000)

bnb.mle(x,r,alpha1,alpha2,lowerbound = 0.01, upperbound = 10000)

```

```

poisson.zihmle(x,type=c('zi','h'),lowerbound = 0.01, upperbound = 10000)

bb.zihmle(x,n,alpha1,alpha2,type=c('zi','h'),lowerbound = 0.01,
          upperbound = 10000)

nb.zihmle(x,r,p,type=c('zi','h'),lowerbound = 0.01, upperbound = 10000)

bnb.zihmle(x,r,alpha1,alpha2,type=c('zi','h'),lowerbound = 0.01,
           upperbound = 10000)

```

The first four functions are designed for standard distributions. The rest are for zero-inflated models with setting `type='zi'` and hurdle models with setting `type='h'`. Note that the value of arguments will not be checked within the functions. Thus, results could be misleading with improper inputs. When calling `nb.zihmle` and `bnb.zihmle`, the users may receive warning messages such as “...cannot obtain mle with the current model type...” if the optimization procedure by R function `optim` does not converge. In this case, the output will be identical to the maximum likelihood estimates for standard negative binomial or beta negative binomial distribution.

#### Illustration

- **Quick start**

In order to utilize the “iZID” package, one may start with simulating random samples given appropriate arguments:

```

library(iZID) ##load the package

##generate random deviates from zero-inflated negative binomial distr

```

```
sample.zi(N=28,phi=0.3,distri='nb',r=6,p=0.4)
```

```
[1] 8 11 10 12 13 0 6 0 10 7 0 5 8 11 11 5 7 7 6 15 0 5 9 14 10 5 10 12
```

```
##generate random deviates from hurdle beta negative binomial distr
```

```
sample.h(N=28,phi=0.6,distri='bnb',r=6,alpha1=3,alpha2=7)
```

```
[1] 0 14 0 20 0 0 17 0 0 0 18 0 0 0 0 0 36 19 14 20 0 24 0 7 2 10 0 0
```

One may test if the maximum likelihood estimates of parameters are close to the truth.

```
temp1=sample.zi(N=300,phi=0.3,distri='poisson',lambda=5)
```

```
poisson.zihmle(temp1,type='zi')
```

```
lambda    phi    loglik
```

```
[1,] 5.058126 0.2955213 -640.1416
```

From the above output, the estimates of  $\lambda$  and  $\phi$  approximate the true values. In the circumstances when the underlying distribution of data `temp1` is unknown, one may fit other models as follows:

```
nb.zihmle(temp1,type='zi',r=3,p=0.5)
```

```
r          p          phi    loglik
```

```
[1,] 340.5231 0.9853779 0.295327 -640.1886
```

```
bb.zihmle(temp1,type='zi',n=3,alpha1=3,alpha2=5)
```

```
n    alpha1  alpha2  phi    loglik
```

```
[1,] 637.37 28.23 178.30 0.3 7120.57
```

```
bnb.zihmle(temp1,type='zi',r=3,alpha1=3,alpha2=5)
```

```

      r      alpha1      alpha2      phi      loglik
[1,] 10000    1614.465    541.4786    0    30367934

```

Warning message:

```
In bnb.zihmle(temp1, type = "zi", r = 3, alpha1 = 3, alpha2 = 5, :
cannot obtain mle with the current model type, the output estimate is
derived from general beta negative binomial distribution.
```

Note that the log likelihood of beta binomial distribution for data `temp1` exceeds that of zero-inflated Poisson distribution, though the latter is the true underlying model. It suggests the need of conducting KS tests to identify an “appropriate” model before estimating model parameters. Without specifying any initial guess on parameters, the procedure of obtaining p-values works as follows:

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='Poisson')$pvalue
```

```
[1] 0
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='nb')$pvalue
```

```
[1] 0
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='bb')$pvalue
```

```
[1] 0
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='bnb')$pvalue
```

```
[1] 0
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='zip')$pvalue
```

```
[1] 0.97
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='zinb')$pvalue
```

```
[1] 0.97
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='zibb')$pvalue
```

```
[1] 0
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='zibnb')$pvalue
```

```
[1] 0
```

Warning message:

```
In bnb.zihmle(x, r, alpha1, alpha2, type = "zi") :
```

```
cannot obtain mle with the current model type, the output estimate is
derived from general beta negative binomial distribution.
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='ph')$pvalue
```

```
[1] 0.98
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='nbh')$pvalue
```

```
[1] 0.94
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='bbh')$pvalue
```

```
[1] 0
```

```
dis.kstest(temp1,nsim=100,bootstrap=TRUE,distri='bnbh')$pvalue
```

```
[1] 0
```

Warning message:

```
In bnb.zihmle(x, r, alpha1, alpha2, type = "h") :
```

```
cannot obtain mle with the current model type, the output estimate is
```

```
derived from general beta negative binomial distribution.
```

The divergence of empirical distribution of `temp1` from zero-inflated Poisson and negative binomial distributions and their hurdle versions is not significant with p-values close to 1. Since a zero-inflated model and its hurdle version are closely related, we are more interested in distinguishing two types of distributions, say, zero-inflated Poisson or negative binomial, which can be done by using the function `model.lrt`. Define the two “dis.kstest” objects returned from zero-inflated Poisson and negative binomial as “d1” and “d2”, respectively.

```
model.lrt(d1,d2)
```

```
[1] 0.5
```

With the current sample size of data `temp1`, the likelihood ratio test, which is the most powerful test, does not tell the difference between zero-inflated Poisson and negative binomial distribution.

In this case, a larger sample size would be needed.



- **Comparison with R package “KSgeneral”**

Package “KSgeneral” (46) supports the computation of p-value for discrete KS test, assuming that parameter values in the null distribution are already known. To conduct a KS test via “KSgeneral”, we need to substitute the unknown parameters with their maximum likelihood estimates. Suppose that a random sample  $\{X_1, \dots, X_{1000}\}$  is generated from a zero-inflated negative binomial distribution as below:

```
library(iZID)

library(extraDistr)

library(KSgeneral)

##generate random deviates from zero-inflated negative binomial distr

set.seed(10086)

x=sample.zi(N=1000,phi=0.7,distri='nb',p=0.6,r=5)

table(x)

x

0    1    2    3    4    5    6    7    8    9    10   11   12

726 52 58 59 40 24 22 10 4  1  2  2

## some naive initial estimates of unknown parameters

r=max(x)

p=sum(x>0)/length(x)
```

```
n=max(x)+1
```

```
alpha1=abs(mean(x)*(mean(x)*(1-mean(x))/var(x)-1))
```

```
alpha2=abs((1-mean(x))*(mean(x)*(1-mean(x))/var(x)-1))
```

To test if the simulated data follows from zero-inflated negative binomial distribution:

```
## maximum likelihood estimates of unknown parameters
```

```
temp1=nb.zihmle(x,type='zi',r=r,p=p)
```

```
temp1
```

```
      r      p      phi    loglik
```

```
[1,] 5.477278 0.6428991 0.6992482 -1127.7
```

```
y1=stepfun(0:max(x), c(0, temp1[3]+(1-temp1[3])*pnbinom(0:max(x),
```

```
  size=ceiling(temp1[1]),p=temp1[2])))
```

```
## conduct discrete KS test with function disc_ks_test in "KSgeneral"
```

```
disc_ks_test(x=x, y=y1, exact=T, tol=1e-08)$p
```

```
[1] 0.6051321
```

```
## conduct discrete KS test with function dis.kstest in "iZID"
```

```
dis.kstest(x,nsim=100,bootstrap=TRUE,distri='zinb',r=r,p=p)$pvalue
```

```
[1] 0.27
```

From the results above, there is no significant evidence showing that the simulated data comes from distributions other than ZINB. However, a more realistic scenario is that we may also testify other null distributions like ZIBB, ZIP or ZIBNB.

```
## when the null distribution is ZIBB
```

```
templ=bb.zihmle(x,type='zi',n=n,alpha1=alpha1,alpha2=alpha2)
```

```
y1=stepfun(0:max(x), c(0, templ[4]+(1-templ[4])*pbbinom(0:max(x),
```

```
size=round(templ[1]), alpha=templ[2], beta=templ[3])))
```

```
disc_ks_test(x=x, y=y1, exact=T, tol=1e-08)$p
```

```
[1] 1
```

```
dis.kstest(x,bootstrap=TRUE,distri='zibb',n=n,alpha1=alpha1,
```

```
alpha2=alpha2)$pvalue
```

```
[1] 0
```

```
## when the null distribution is ZIP
```

```
templ=poisson.zihmle(x,type='zi')
```

```
y1=stepfun(0:max(x), c(0, templ[2]+(1-templ[2])*ppois(0:max(x),
```

```
lambda=templ[1])))
```

```
disc_ks_test(x=x, y=y1, exact=T, tol=1e-08)$p
```

```
[1] 0.4722135
```

```
dis.kstest(x,nsim=100,bootstrap=TRUE,distri='zip')$pvalue
```

```
[1] 0.47
```

```
## when the null distribution is ZIBNB

temp1=bnb.zihmle(x,type='zi',r=r,alpha1=alpha1,alpha2=alpha2)

y1=stepfun(0:max(x), c(0, temp1[4]+(1-temp1[4])*pnbibinom(0:max(x),

  size=round(temp1[1]), alpha=temp1[2], beta=temp1[3])))

disc_ks_test(x=x, y=y1, exact=T, tol=1e-08)$p
```

```
[1] 1
```

```
dis.kstest(x,bootstrap=TRUE,distri='zibnb',r=r,alpha1=alpha1,

  alpha2=alpha2)$pvalue
```

```
[1] 0
```

Neither function `disc_ks_test` in package “KSgeneral” nor our function `dis.kstest` could distinguish between ZINB and ZIP distributions with the current sample size. As for ZIBB and ZIBNB distributions, the p-value 1 obtained by `disc_ks_test` is apparently misleading, while our `dis.kstest` correctly rejects the two null hypotheses with p-values equal to 0.

- **A real data example**

In this subsection, we use the real dataset “dataCar” from R package “insuranceData” for illustration. The data consists of 67,856 one-year vehicle insurance policies issued in 2014-2015. The variable `number` of claims is a sparse count variable. The goal is to identify the distribution of the variable. Table IV shows the numbers of claims as well as percentages.

TABLE IV: NUMBER OF CLAIMS IN DATASET “DATACAR”

Occurrence	Frequency	Percentage
0	63,232	<b>93.18%</b>
1	4,333	6.39%
2	271	0.40%
3	18	0.03%
4	2	0.00%
Total	67,856	100.00%

To check if the data follows any specific discrete distribution, we use `dis.kstest` in our package.

A large p-value implies that the data may follow the pre-specified discrete distribution. The following R codes show how to test if the variable `number of claims` follows Poisson, negative binomial, ZIP, or ZINB distribution.

```
library(insuranceData)

library(car)

data(dataCar)

attach(dataCar)

X=dataCar[,4] #Number of claims variable

dis.kstest(X,nsim=200,bootstrap=TRUE,distri='Poisson')$pvalue

[1] 0.035

dis.kstest(X,nsim=200,bootstrap=TRUE,distri='nb')$pvalue

[1] 0
```

```
dis.kstest(X,nsim=200,bootstrap=TRUE,distri='zip')$pvalue
```

```
[1] 0.955
```

```
dis.kstest(X,nsim=200,bootstrap=TRUE,distri='zinb')$pvalue
```

```
[1] 0
```

```
dis.kstest(X,nsim=200,bootstrap=TRUE,distri='zip')$mle.ori
```

```
lambda      phi    loglik
```

```
[1,] 0.1324475 0.4506756 -18052.2
```

The above output implies that the data follows ZIP distribution with estimated parameters  $\hat{\phi} = 0.451$ , and  $\hat{\lambda} = 0.132$ . To confirm this conclusion, we simulate a random sample from the ZIP distribution with  $\phi = 0.451$ , and  $\lambda = 0.132$  as follows:

```
Y=sample.zi(N=length(X),phi=0.4506756,distri='Poisson',lambda=0.1324475)
```

Using R function `table`, we can see that the distributions of the original data  $X$  and the simulated data  $Y$  match each other very well (see also Figure Figure 2).

```
table(X)
```

```
0    1    2    3    4
```

```
63232 4333 271 18  2
```

```
table(Y)
```

```
0    1    2    3    4
```

```
63172 4371 298 14  1
```

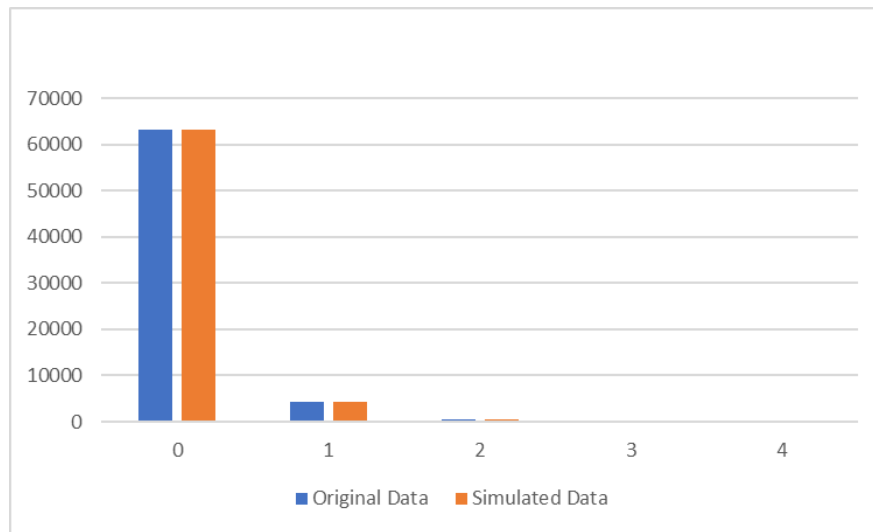


Figure 2: Chart plot to compare the given data and the simulated data

#### 2.5.4 Conclusion

In this chapter, we introduce a new R package “iZID” which provides the bootstrapped Monte Carlo estimates of p-values of discrete KS tests, as well as a function `model.lrt` to perform a likelihood ratio test when two or more distributions pass the KS test. Besides, “iZID” supports the generation of random deviates from zero-inflated distributions as well as hurdle models, and the computation of maximum likelihood estimates of a large class of models. The implementation of functions `dis.kstest` and `model.lrt` are speeded up by parallel computing via packages “foreach” and “doParallel”.

Due to the nature of gamma functions, the optimization of the likelihood function of zero-inflated and hurdle beta binomial and beta negative binomial distributions may not converge. In this circumstance, the results of corresponding standard distributions are returned. We plan to further improve and

update the functions in the package for obtaining more robust and reliable sample estimates of parameters.



## CHAPTER 3

### REGRESSION MODELS

#### 3.1 Probabilistic Model vs. Regression Model

In order to test whether the data follow a specific discrete distribution with unknown parameters, we developed a bootstrapped procedure for estimating the p-values of Kolmogorov–Smirnov (KS) tests and applied it to a list of 229 bacterial and fungal OTUs. Their 12 candidate distributions include Poisson, negative binomial (NB), beta binomial (BB), beta negative binomial (BNB), and the corresponding zero-inflated and Hurdle models.

The model selection performed in was for probabilistic models without covariates. Nevertheless, the following simulation study shows that it may still be informative for selecting regression models when the coefficients of covariates are relatively small.

*Example 3.1.* Suppose the covariates are  $\mathbf{x}_i \in \mathbb{R}^d$ ,  $i = 1, \dots, N$  and the assumed parameters are  $\gamma_0, \beta_0 \in \mathbb{R}$ ,  $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^d$ . For  $c = 0.5, 0.2, 0.1$  or  $0.01$ ,  $b = 1, \dots, 100$ , we simulate the responses  $Y_i^{(b)} \sim \text{ZIP}(\phi_i, \boldsymbol{\theta}_i)$  with  $\phi_i = g^{-1}(\gamma_0 + c \cdot \boldsymbol{\gamma}^T \mathbf{x}_i)$  and  $\boldsymbol{\theta}_i = h^{-1}(\beta_0 + c \cdot \boldsymbol{\beta}^T \mathbf{x}_i)$ . Then we apply the bootstrapped KS-test in to  $Y_1^{(b)}, \dots, Y_N^{(b)}$  and check if it follows a probabilistic ZIP model. The results are summarized into Table V. It shows that  $Y_1^{(b)}, \dots, Y_N^{(b)}$  passes the ZIP test with a high chance when  $c$  is relatively small.

TABLE V: ZIP PROBABILISTIC KS-TEST APPLIED TO ZIP REGRESSION RESPONSES

$c$	Number of Passed KS Tests	Percentage
0	100	100%
0.01	96	96%
0.05	88	88%
0.10	47	47%
0.20	8	8%
0.50	1	1%

### 3.2 Hurdle Regression Models

In this section, we consider fairly general zero-altered or Hurdle regression models for independent observations  $(Y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , where  $Y_i$  is a univariate response following a zero-altered model  $f_{ZA}(y|\phi_i, \boldsymbol{\theta}_i)$  with parameters  $\phi_i \in [0, 1]$ ,  $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{ib})^T \in \mathbb{R}^b$ , and covariates  $\mathbf{x}_i = (x_{i1}, \dots, x_{id})^T \in \mathbb{R}^d$ .

A zero-altered or Hurdle regression model considered here assumes the existence of link functions  $g$  and  $h_1, \dots, h_b$  such that  $g(\phi_i) = \mathbf{G}_i^T \boldsymbol{\gamma}$  and  $h_j(\theta_{ij}) = \mathbf{B}_{ij}^T \boldsymbol{\beta}_j$ ,  $i = 1, \dots, n$ ;  $j = 1, \dots, b$ , where  $\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_b$  are regression coefficients,  $\mathbf{G}_i = (r_1(\mathbf{x}_i), \dots, r_s(\mathbf{x}_i))^T \in \mathbb{R}^s$  and  $\mathbf{B}_{ij} = (q_{j1}(\mathbf{x}_i), \dots, q_{jt_j}(\mathbf{x}_i))^T \in \mathbb{R}^{t_j}$  are the corresponding predictors,  $r_i$ 's and  $q_{ji}$ 's are known functions. Examples include  $\mathbf{G}_i = \mathbf{B}_{ij} = (1, x_{i1}, \dots, x_{id})^T$  for main-effects model and  $\mathbf{G}_i = \mathbf{B}_{ij} = (1, x_{i1}, \dots, x_{id}, x_{i1}x_{i2}, \dots, x_{i,d-1}x_{id})^T$  for model with both main effects and order-2 interactions.

Recall that with the baseline distribution function (pmf or pdf)  $f_{\theta}(y)$ , the distribution function of  $Y_i$  given  $\phi_i$  and  $\theta_i$  can be written as

$$f_{ZA}(y_i | \phi_i, \theta_i) = \begin{cases} \phi_i & \text{if } y_i = 0 \\ \frac{1-\phi_i}{1-p_0(\theta_i)} f_{\theta_i}(y_i) & \text{if } y_i \neq 0 \end{cases}$$

where  $p_0(\theta_i) = f_{\theta_i}(0)$  for discrete case or 0 for continuous case.

Given the link functions  $g, h_1, \dots, h_b$ , we have  $\phi_i = g^{-1}(\mathbf{G}_i^T \boldsymbol{\gamma})$ , and  $\theta_{ij} = h_j^{-1}(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j)$ . Then the likelihood function for the regression coefficients is

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_b) = \prod_{i: y_i=0} \phi_i \prod_{i: y_i \neq 0} (1 - \phi_i) \cdot \prod_{i: y_i \neq 0} \frac{f_{\theta_i}(y_i)}{1 - p_0(\theta_i)} \quad (3.1)$$

Since  $\boldsymbol{\gamma}$  is separable from  $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_b$  in (Equation 3.1), the mle of  $\boldsymbol{\gamma}$  can be obtained by fitting a generalized linear model (GLM) with responses  $\mathbf{1}_{Y_i=0}$ 's.

*Theorem 3.1.* The MLEs for the Hurdle regression model coefficients can be obtained as follows:

(1) By fitting a GLM with binary response  $z_i = \mathbf{1}_{y_i=0}$ , predictors  $\mathbf{G}_i$  and link function  $g$ , we obtain the MLE  $\hat{\boldsymbol{\gamma}}$ .

(2) By fitting the zero-truncated regression model on the nonzero observations, we obtain the MLEs  $\hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_b$  that maximize  $\prod_{i: y_i \neq 0} f_{\theta_i}(y_i) / [1 - p_0(\theta_i)]$ .

**Proof of Theorem 3.1:** Recall that

$$P_{ZA}(Y = y) = \phi_i \mathbf{1}_{\{y=0\}} + (1 - \phi_i) P_{tr}(y | \theta_i)$$

Assign link functions for the parameters  $\phi_i$ , and  $\theta_i$ . Therefore,  $g(\phi_i) = x_i^T \gamma$ , and  $h_j(\theta_{ij}) = x_{ij}^T \beta_j$

Then

$$l(\gamma, \beta_1, \dots, \beta_d) = \sum_{i=1}^n \log \phi_i \mathbf{1}_{\{y_i=0\}} + \sum_{i=1}^n \log(1 - \phi_i) \mathbf{1}_{\{y_i \neq 0\}} + \sum_{i=1}^n \log f_{tr}(y|\theta_i) \mathbf{1}_{\{y_i \neq 0\}}$$

Hence,

$$\hat{\gamma} = \sum_{i=1}^n \log \phi_i \mathbf{1}_{\{y_i=0\}} + \sum_{i=1}^n \log(1 - \phi_i) \mathbf{1}_{\{y_i \neq 0\}}$$

Which is a regular GLM with binary response.

$$(\hat{\beta}_1, \dots, \hat{\beta}_d) = \arg \sum_{i=1}^n \log f_{tr}(y|\theta_i) \mathbf{1}_{\{y_i \neq 0\}}$$

□

In order to find the MLEs of  $\beta_j$ 's, we need the formulae of the first-order derivatives of the log-likelihood of the zero-truncated regression model:

$$l(\beta_1, \dots, \beta_b) = \sum_{i: y_i \neq 0} \log f_{\theta_i}(y_i) - \sum_{i: y_i \neq 0} \log[1 - p_0(\theta_i)]$$

For  $j = 1, \dots, b$ , the first-order derivatives are

$$\frac{\partial l}{\partial \beta_j} = \sum_{i: y_i \neq 0} \left[ \frac{\partial \log f_{\theta_i}(y_i)}{\partial \theta_{ij}} + \frac{p_0(\theta_i)}{1 - p_0(\theta_i)} \cdot \frac{\partial \log p_0(\theta_i)}{\partial \theta_{ij}} \right] (h_j^{-1})'(\mathbf{B}_{ij}^T \beta_j) \mathbf{B}_{ij}$$

Thus only  $\partial \log f_{\theta_i}(y_i)/\partial \theta_{ij}$  and  $\partial \log p_0(\theta_i)/\partial \theta_{ij}$  are needed.

**Example 3.2. Hurdle beta negative binomial (HBNB) regression model** In a HBNB model,  $Y_i$  follows a beta-negative binomial distribution with the pmf

$$f_{\theta_i}(y_i) = \binom{r_i + y_i - 1}{y_i} \frac{\text{Beta}(r_i + \alpha_i, y_i + \beta_i)}{\text{Beta}(\alpha_i, \beta_i)}$$

at  $y_i = 0, 1, \dots$  with  $\theta_i = (r_i, \alpha_i, \beta_i)^T$ . Then

$$\begin{aligned} p_0(\theta_i) &= \frac{\Gamma(r_i + \alpha_i)\Gamma(\alpha_i + \beta_i)}{\Gamma(r_i + \alpha_i + \beta_i)\Gamma(\alpha_i)} \\ \frac{p_0(\theta_i)}{1 - p_0(\theta_i)} &= \frac{\Gamma(r_i + \alpha_i)\Gamma(\alpha_i + \beta_i)}{\Gamma(r_i + \alpha_i + \beta_i)\Gamma(\alpha_i) - \Gamma(r_i + \alpha_i)\Gamma(\alpha_i + \beta_i)} \\ \frac{\partial \log f_{\theta_i}(y_i)}{\partial r_i} &= \Psi(r_i + y_i) - \Psi(r_i) + \Psi(r_i + \alpha_i) - \Psi(r_i + y_i + \alpha_i + \beta_i) \\ \frac{\partial \log f_{\theta_i}(y_i)}{\partial \alpha_i} &= \Psi(r_i + \alpha_i) - \Psi(r_i + y_i + \alpha_i + \beta_i) + \Psi(\alpha_i + \beta_i) - \Psi(\alpha_i) \\ \frac{\partial \log f_{\theta_i}(y_i)}{\partial \beta_i} &= \Psi(y_i + \beta_i) - \Psi(r_i + y_i + \alpha_i + \beta_i) + \Psi(\alpha_i + \beta_i) - \Psi(\beta_i) \\ \frac{\partial \log p_0(\theta_i)}{\partial r_i} &= \Psi(r_i + \alpha_i) - \Psi(r_i + \alpha_i + \beta_i) \\ \frac{\partial \log p_0(\theta_i)}{\partial \alpha_i} &= \Psi(r_i + \alpha_i) + \Psi(\alpha_i + \beta_i) - \Psi(r_i + \alpha_i + \beta_i) - \Psi(\alpha_i) \\ \frac{\partial \log p_0(\theta_i)}{\partial \beta_i} &= \Psi(\alpha_i + \beta_i) - \Psi(r_i + \alpha_i + \beta_i) \end{aligned}$$

If we define the link functions  $h_1 = h_2 = h_3 = \log$ , then  $(h_j^{-1})' = \exp$ . □

### 3.3 Zero-inflated Regression Model (ZIRM)

In this section, we consider a fairly general class of zero-inflated regression models for independent observations  $(Y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , where  $Y_i$  is a univariate response following a zero-inflated model  $f_{ZI}(y|\phi_i, \theta_i)$  with parameters  $\phi_i \in [0, 1]$ ,  $\theta_i = (\theta_{i1}, \dots, \theta_{ib})^T \in \mathbb{R}^b$ , and covariates  $\mathbf{x}_i = (x_{i1}, \dots, x_{id})^T$ .

With the baseline distribution function (pmf or pdf)  $f_{\theta}(y)$ , the distribution function of  $Y_i$  given  $\phi_i$  and  $\theta_i$  can be written as

$$f_{ZI}(y_i | \phi_i, \theta_i) = \begin{cases} \phi_i + (1 - \phi_i)p_0(\theta_i) & \text{if } y_i = 0 \\ (1 - \phi_i)f_{\theta_i}(y_i) & \text{if } y_i \neq 0 \end{cases}$$

where  $p_0(\theta_i) = f_{\theta_i}(0)$  for discrete case or 0 for continuous case.

Similar as in Hurdle regression models, a general zero-inflated regression model requires link functions  $g$  and  $h_1, \dots, h_b$  such that  $\phi_i = g^{-1}(\mathbf{G}_i^T \boldsymbol{\gamma})$  and  $\theta_{ij} = h_j^{-1}(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j)$ ,  $i = 1, \dots, n; j = 1, \dots, b$ , with regression coefficients  $\boldsymbol{\gamma} \in \mathbb{R}^s, \boldsymbol{\beta}_1 \in \mathbb{R}^{t_1}, \dots, \boldsymbol{\beta}_b \in \mathbb{R}^{t_b}$ . The likelihood and log-likelihood functions for the regression coefficients are

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_b) &= \prod_{i: y_i \neq 0} (1 - \phi_i) f_{\theta_i}(y_i) \cdot \prod_{i: y_i = 0} [\phi_i + (1 - \phi_i) p_0(\theta_i)] \\ l(\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_b) &= \sum_{i: y_i \neq 0} \log(1 - \phi_i) + \sum_{i: y_i \neq 0} \log f_{\theta_i}(y_i) \\ &\quad + \sum_{i: y_i = 0} \log[\phi_i + (1 - \phi_i) p_0(\theta_i)] \end{aligned}$$

The first-order derivatives are

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\gamma}} &= - \sum_{i: y_i \neq 0} \frac{(g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma})}{1 - \phi_i} \mathbf{G}_i + \sum_{i: y_i = 0} \frac{[1 - p_0(\theta_i)](g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma})}{\phi_i + (1 - \phi_i) p_0(\theta_i)} \mathbf{G}_i \\ \frac{\partial l}{\partial \boldsymbol{\beta}_j} &= \sum_{i: y_i \neq 0} \frac{\partial \log f_{\theta_i}(y_i)}{\partial \theta_{ij}} \cdot (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \\ &\quad + \sum_{i: y_i = 0} \frac{(1 - \phi_i) p_0(\theta_i)}{\phi_i + (1 - \phi_i) p_0(\theta_i)} \cdot \frac{\partial \log p_0(\theta_i)}{\partial \theta_{ij}} \cdot (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \end{aligned}$$

Thus only  $\partial \log f_{\theta_i}(y_i) / \partial \theta_{ij}$  and  $\partial \log p_0(\theta_i) / \partial \theta_{ij}$  are needed.

### 3.4 Fisher Information Matrix

Let  $\boldsymbol{\pi} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_b^T)^T \in \mathbb{R}^p$  be the vector of regression coefficients, where  $p = s + t_1 + \dots + t_b$ . Under regularity conditions (see, for example, Ferguson (1996, Theorem 18))(51), the Fisher information matrix

$$\mathbf{F}(\boldsymbol{\pi}) = E \left( \frac{\partial l}{\partial \boldsymbol{\pi}} \cdot \frac{\partial l}{\partial \boldsymbol{\pi}^T} \right) \quad (3.2)$$

where  $\partial l / \partial \boldsymbol{\pi} = (\partial l / \partial \boldsymbol{\gamma}^T, \partial l / \partial \boldsymbol{\beta}_1^T, \dots, \partial l / \partial \boldsymbol{\beta}_b^T)^T$ . Note that  $E(\partial l / \partial \boldsymbol{\pi}) = 0$  under the regularity conditions.

#### 3.4.1 Fisher information of zero-inflated regression models

*Theorem 3.2.* Under regularity conditions, the Fisher information matrix of Zero-inflated Regression Model (ZIRM) defined in section 3.3 is given by:

$$\mathbf{F}(\boldsymbol{\pi}) = E \begin{bmatrix} \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \dots & \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \\ \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \dots & \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \dots & \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \end{bmatrix} \quad (3.3)$$

where

$$\begin{aligned}
E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\gamma}^T}\right) &= \sum_{i=1}^n \frac{1-P_i}{P_i} \mathbf{A}_i \mathbf{A}_i^T \\
E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_i^T}\right) &= \sum_{i=1}^n (1-\phi_i) \left\{ \frac{p_0(\boldsymbol{\theta}_i)}{P_i} \cdot \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{il}} - E\left[\frac{\partial \log f_{\boldsymbol{\theta}_i}(Y'_i)}{\partial \theta_{il}}\right] \right\} \mathbf{A}_i \mathbf{C}_{il}^T \\
E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_s} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_t^T}\right) &= \left[ \sum_{i=1}^n (1-\phi_i) e_{is} \mathbf{C}_{is} \right] \cdot \left[ \sum_{i=1}^n (1-\phi_i) e_{it} \mathbf{C}_{it} \right]^T \\
&\quad - \sum_{i=1}^n (1-\phi_i)^2 e_{is} e_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T + \sum_{i=1}^n (1-\phi_i) e_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T \\
&\quad - \sum_{i=1}^n \frac{\phi_i(1-\phi_i)p_0(\boldsymbol{\theta}_i)}{\phi_i + (1-\phi_i)p_0(\boldsymbol{\theta}_i)} \cdot \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{is}} \cdot \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{it}} \mathbf{C}_{is} \mathbf{C}_{it}^T
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A}_i &= \frac{1}{1-\phi_i} \cdot (g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma}) \mathbf{G}_i \in \mathbb{R}^s \\
\mathbf{B}_i &= \frac{1-p_0(\boldsymbol{\theta}_i)}{\phi_i + (1-\phi_i)p_0(\boldsymbol{\theta}_i)} \cdot (g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma}) \mathbf{G}_i \in \mathbb{R}^s \\
P_i &= P(Y_i = 0) = \phi_i + (1-\phi_i)p_0(\boldsymbol{\theta}_i) \\
\mathbf{C}_{ij} &= (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \in \mathbb{R}^{t_j} \\
e_{ist} &= E\left[\frac{\partial \log f_{\boldsymbol{\theta}_i}(Y'_i)}{\partial \theta_{is}} \cdot \frac{\partial \log f_{\boldsymbol{\theta}_i}(Y'_i)}{\partial \theta_{it}}\right]
\end{aligned}$$

Note that under regularity conditions,  $E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_j}\right) = \mathbf{0}$  and  $e_{is} = 0$ , while in general,

$$E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_j}\right) = \sum_{i=1}^n (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} (1-\phi_i) E\left[\frac{\partial \log f_{\boldsymbol{\theta}_i}(Y'_i)}{\partial \theta_{ij}}\right]$$



**Proof of Theorem 3.2:** For zero-inflated regression models,

$$\begin{aligned}
 l(\boldsymbol{\pi}) &= \sum_{i=1}^n \log(1 - \phi_i) \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i=1}^n \log f_{\boldsymbol{\theta}_i}(Y_i) \mathbf{1}_{\{Y_i \neq 0\}} \\
 &\quad + \sum_{i=1}^n \log [\phi_i + (1 - \phi_i) p_0(\boldsymbol{\theta}_i)] \mathbf{1}_{\{Y_i = 0\}}
 \end{aligned}$$

In order to simplify the notations, we let

$$\begin{aligned}
 \mathbf{A}_i &= \frac{1}{1 - \phi_i} \cdot (g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma}) \mathbf{G}_i \in \mathbb{R}^s \\
 \mathbf{B}_i &= \frac{1 - p_0(\boldsymbol{\theta}_i)}{\phi_i + (1 - \phi_i) p_0(\boldsymbol{\theta}_i)} \cdot (g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma}) \mathbf{G}_i \in \mathbb{R}^s \\
 h_{ij}(Y_i) &= \frac{\partial \log f_{\boldsymbol{\theta}_i}(Y_i)}{\partial \theta_{ij}} \in \mathbb{R} \\
 \mathbf{C}_{ij} &= (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \in \mathbb{R}^{t_j} \\
 \mathbf{D}_{ij} &= \frac{(1 - \phi_i) p_0(\boldsymbol{\theta}_i)}{\phi_i + (1 - \phi_i) p_0(\boldsymbol{\theta}_i)} \cdot \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{ij}} \cdot (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \in \mathbb{R}^{t_j} \\
 P_i &= P(Y_i = 0) = \phi_i + (1 - \phi_i) p_0(\boldsymbol{\theta}_i) \\
 1 - P_i &= P(Y_i \neq 0) = (1 - \phi_i)[1 - p_0(\boldsymbol{\theta}_i)]
 \end{aligned}$$

Then it can be verified that

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\gamma}} &= -\sum_{i=1}^n \mathbf{A}_i \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i=1}^n \mathbf{B}_i \mathbf{1}_{\{Y_i = 0\}} \\
\frac{\partial l}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^n h_{ij}(Y_i) \mathbf{C}_{ij} \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i=1}^n \mathbf{D}_{ij} \mathbf{1}_{\{Y_i = 0\}} \\
E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}}\right) &= -\sum_{i=1}^n \mathbf{A}_i (1 - P_i) + \sum_{i=1}^n \mathbf{B}_i P_i = \mathbf{0} \\
E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_j}\right) &= \sum_{i=1}^n \mathbf{C}_{ij} E[h_{ij}(Y_i) \mathbf{1}_{\{Y_i \neq 0\}}] + \sum_{i=1}^n \mathbf{D}_{ij} P_i \\
&= \sum_{i=1}^n (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \left\{ E\left[\frac{\partial \log f_{\theta_i}(Y_i)}{\partial \theta_{ij}} \mathbf{1}_{\{Y_i \neq 0\}}\right] + \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{ij}} (1 - \phi_i) p_0(\boldsymbol{\theta}_i) \right\} \\
&= \sum_{i=1}^n (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} (1 - \phi_i) \left\{ E\left[\frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{ij}} \mathbf{1}_{\{Y'_i \neq 0\}}\right] + \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{ij}} p_0(\boldsymbol{\theta}_i) \right\} \\
&= \sum_{i=1}^n (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} (1 - \phi_i) E\left[\frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{ij}}\right]
\end{aligned}$$

where  $Y'_i \sim f_{\theta_i}$ , a baseline distribution. Note that if  $E\left[\partial \log f_{\theta_i}(Y'_i)/\partial \theta_{ij}\right] = 0$  for  $i = 1, \dots, n$ , then

$E\left(\partial l / \partial \boldsymbol{\beta}_j\right) = \mathbf{0}$  for  $j = 1, \dots, b$ .

The leading term of the Fisher information matrix (Equation 3.5) can be written as follows

$$E \begin{bmatrix} \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \cdots & \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \\ \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \cdots & \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \cdots & \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \end{bmatrix} \quad (3.4)$$

We calculate these entries one by one. Note that matrix (Equation 3.6) is symmetric if  $l \in C^2$ , that is, twice continuously differentiable.

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\gamma}^T} &= \sum_{i=1}^n \mathbf{A}_i \mathbf{A}_i^T \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i \neq j} \mathbf{A}_i \mathbf{A}_j^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\
&\quad - \sum_{i \neq j} \mathbf{A}_i \mathbf{B}_j^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j = 0\}} - \sum_{i \neq j} \mathbf{B}_i \mathbf{A}_j^T \mathbf{1}_{\{Y_i = 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\
&\quad + \sum_{i=1}^n \mathbf{B}_i \mathbf{B}_i^T \mathbf{1}_{\{Y_i = 0\}} + \sum_{i \neq j} \mathbf{B}_i \mathbf{B}_j^T \mathbf{1}_{\{Y_i = 0\}} \mathbf{1}_{\{Y_j = 0\}}
\end{aligned}$$

$$\begin{aligned}
E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\gamma}^T}\right) &= \sum_{i=1}^n \mathbf{A}_i \mathbf{A}_i^T (1 - P_i) + \sum_{i \neq j} \mathbf{A}_i \mathbf{A}_j^T (1 - P_i)(1 - P_j) \\
&\quad - \sum_{i \neq j} \mathbf{A}_i \mathbf{B}_j^T (1 - P_i) P_j - \sum_{i \neq j} \mathbf{B}_j \mathbf{A}_i^T (1 - P_i) P_j \\
&\quad + \sum_{i=1}^n \mathbf{B}_i \mathbf{B}_i^T P_i + \sum_{i \neq j} \mathbf{B}_i \mathbf{B}_j^T P_i P_j \\
&= \left[ \sum_{i=1}^n \mathbf{A}_i (1 - P_i) \right] \cdot \left[ \sum_{i=1}^n \mathbf{A}_i (1 - P_i) \right]^T + \sum_{i=1}^n \mathbf{A}_i \mathbf{A}_i^T P_i (1 - P_i) \\
&\quad + \left[ \sum_{i=1}^n \mathbf{B}_i P_i \right] \cdot \left[ \sum_{i=1}^n \mathbf{B}_i P_i \right]^T + \sum_{i=1}^n \mathbf{B}_i \mathbf{B}_i^T P_i (1 - P_i) \\
&\quad - \left[ \sum_{i=1}^n \mathbf{A}_i (1 - P_i) \right] \cdot \left[ \sum_{i=1}^n \mathbf{B}_i P_i \right]^T + \sum_{i=1}^n \mathbf{A}_i \mathbf{B}_i^T P_i (1 - P_i) \\
&\quad - \left[ \sum_{i=1}^n \mathbf{B}_i P_i \right] \cdot \left[ \sum_{i=1}^n \mathbf{A}_i (1 - P_i) \right]^T + \sum_{i=1}^n \mathbf{B}_i \mathbf{A}_i^T P_i (1 - P_i)
\end{aligned}$$

Since  $(1 - P_i)\mathbf{A}_i - P_i\mathbf{B}_i = \mathbf{0}$  for each  $i = 1, \dots, n$ , it can be verified that

$$\begin{aligned}
 E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\gamma}^T}\right) &= \left[ \sum_{i=1}^n (1 - P_i)\mathbf{A}_i - \sum_{i=1}^n P_i\mathbf{B}_i \right] \cdot \left[ \sum_{i=1}^n (1 - P_i)\mathbf{A}_i - \sum_{i=1}^n P_i\mathbf{B}_i \right]^T \\
 &\quad + \sum_{i=1}^n P_i(1 - P_i)(\mathbf{A}_i + \mathbf{B}_i)(\mathbf{A}_i + \mathbf{B}_i)^T \\
 &= \sum_{i=1}^n P_i(1 - P_i)(\mathbf{A}_i + \mathbf{B}_i)(\mathbf{A}_i + \mathbf{B}_i)^T \\
 &= \sum_{i=1}^n \frac{1 - P_i}{P_i} \mathbf{A}_i \mathbf{A}_i^T
 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_l^T} &= - \sum_{i=1}^n \mathbf{A}_i \mathbf{C}_{il}^T h_{il}(Y_i) \mathbf{1}_{\{Y_i \neq 0\}} - \sum_{i \neq j} \mathbf{A}_i \mathbf{C}_{jl}^T h_{jl}(Y_j) \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\
 &\quad - \sum_{i \neq j} \mathbf{A}_i \mathbf{D}_{jl}^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j = 0\}} + \sum_{i \neq j} \mathbf{B}_i \mathbf{C}_{jl}^T h_{jl}(Y_j) \mathbf{1}_{\{Y_i = 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\
 &\quad + \sum_{i=1}^n \mathbf{B}_i \mathbf{D}_{il}^T \mathbf{1}_{\{Y_i = 0\}} + \sum_{i \neq j} \mathbf{B}_i \mathbf{D}_{jl}^T \mathbf{1}_{\{Y_i = 0\}} \mathbf{1}_{\{Y_j = 0\}}
 \end{aligned}$$

Note that for zero-inflated models

$$\begin{aligned}
 E\left[h_{jl}(Y_j) \mathbf{1}_{\{Y_j \neq 0\}}\right] &= E\left(\frac{\partial \log f_{\theta_j}(Y_j)}{\partial \theta_{jl}} \mathbf{1}_{\{Y_j \neq 0\}}\right) \\
 &= (1 - \phi_j) E\left(\frac{\partial \log f_{\theta_j}(Y'_j)}{\partial \theta_{jl}} \mathbf{1}_{\{Y'_j \neq 0\}}\right) \\
 &= (1 - \phi_j) E\left(\frac{\partial \log f_{\theta_j}(Y'_j)}{\partial \theta_{jl}}\right) - (1 - \phi_j) p_0(\boldsymbol{\theta}_j) \frac{\partial \log p_0(\boldsymbol{\theta}_j)}{\partial \theta_{jl}} \\
 &\triangleq E_{jl}
 \end{aligned}$$

where  $Y'_j$  follows the baseline distribution  $f_{\theta_j}(y)$ . Then

$$\begin{aligned}
& E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_i^T}\right) \\
&= -\sum_{i=1}^n E_{il} \mathbf{A}_i \mathbf{C}_{il}^T - \sum_{i \neq j} (1 - P_i) E_{jl} \mathbf{A}_i \mathbf{C}_{jl}^T - \sum_{i \neq j} (1 - P_i) P_j \mathbf{A}_i \mathbf{D}_{jl}^T \\
&\quad + \sum_{i \neq j} P_i E_{jl} \mathbf{B}_i \mathbf{C}_{jl}^T + \sum_{i=1}^n P_i \mathbf{B}_i \mathbf{D}_{il}^T + \sum_{i \neq j} P_i P_j \mathbf{B}_i \mathbf{D}_{jl}^T \\
&= -\left[\sum_{i=1}^n (1 - P_i) \mathbf{A}_i\right] \cdot \left[\sum_{i=1}^n E_{il} \mathbf{C}_{il}^T\right] + \sum_{i=1}^n (1 - P_i) E_{il} \mathbf{A}_i \mathbf{C}_{il}^T - \sum_{i=1}^n E_{il} \mathbf{A}_i \mathbf{C}_{il}^T \\
&\quad - \left[\sum_{i=1}^n (1 - P_i) \mathbf{A}_i\right] \cdot \left[\sum_{i=1}^n P_i \mathbf{D}_{il}^T\right] + \sum_{i=1}^n P_i (1 - P_i) \mathbf{A}_i \mathbf{D}_{il}^T \\
&\quad + \left[\sum_{i=1}^n P_i \mathbf{B}_i\right] \cdot \left[\sum_{i=1}^n E_{il} \mathbf{C}_{il}^T\right] - \sum_{i=1}^n P_i E_{il} \mathbf{B}_i \mathbf{C}_{il}^T \\
&\quad + \sum_{i=1}^n P_i \mathbf{B}_i \mathbf{D}_{il}^T + \left[\sum_{i=1}^n P_i \mathbf{B}_i\right] \cdot \left[\sum_{i=1}^n P_i \mathbf{D}_{il}^T\right] - \sum_{i=1}^n P_i^2 \mathbf{B}_i \mathbf{D}_{il}^T \\
&= \left[\sum_{i=1}^n P_i \mathbf{B}_i - \sum_{i=1}^n (1 - P_i) \mathbf{A}_i\right] \cdot \left[\sum_{i=1}^n E_{il} \mathbf{C}_{il}^T + \sum_{i=1}^n P_i \mathbf{D}_{il}^T\right] \\
&\quad - \sum_{i=1}^n P_i E_{il} \mathbf{A}_i \mathbf{C}_{il}^T + \sum_{i=1}^n P_i (1 - P_i) \mathbf{A}_i \mathbf{D}_{il}^T \\
&\quad - \sum_{i=1}^n P_i E_{il} \mathbf{B}_i \mathbf{C}_{il}^T + \sum_{i=1}^n P_i (1 - P_i) \mathbf{B}_i \mathbf{D}_{il}^T \\
&= -\sum_{i=1}^n P_i E_{il} (\mathbf{A}_i + \mathbf{B}_i) \mathbf{C}_{il}^T + \sum_{i=1}^n P_i (1 - P_i) (\mathbf{A}_i + \mathbf{B}_i) \mathbf{D}_{il}^T \\
&= -\sum_{i=1}^n E_{il} \mathbf{A}_i \mathbf{C}_{il}^T + \sum_{i=1}^n (1 - P_i) \mathbf{A}_i \mathbf{D}_{il}^T
\end{aligned}$$

since  $\mathbf{A}_i = P_i(\mathbf{A}_i + \mathbf{B}_i)$ . On the other hand,

$$\begin{aligned} E_{il} &= (1 - \phi_i)E\left(\frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{il}}\right) - (1 - \phi_i)p_0(\theta_i)\frac{\partial \log p_0(\theta_i)}{\partial \theta_{il}} \\ \mathbf{D}_{il} &= \frac{(1 - \phi_i)p_0(\theta_i)}{P_i} \cdot \frac{\partial \log p_0(\theta_i)}{\partial \theta_{il}} \mathbf{C}_{il} \end{aligned}$$

Then

$$E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_t^T}\right) = \sum_{i=1}^n (1 - \phi_i) \left\{ \frac{p_0(\theta_i)}{P_i} \cdot \frac{\partial \log p_0(\theta_i)}{\partial \theta_{il}} - E\left[\frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{il}}\right] \right\} \mathbf{A}_i \mathbf{C}_{il}^T$$

where  $Y'_i$  follows the baseline distribution  $f_{\theta_i}(y)$ .

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}_s} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_t^T} &= \sum_{i=1}^n h_{is}(Y_i)h_{it}(Y_i)\mathbf{C}_{is}\mathbf{C}_{it}^T\mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i \neq j} h_{is}(Y_i)h_{jt}(Y_j)\mathbf{C}_{is}\mathbf{C}_{jt}^T\mathbf{1}_{\{Y_i \neq 0\}}\mathbf{1}_{\{Y_j \neq 0\}} \\ &\quad + \sum_{i \neq j} h_{is}(Y_i)\mathbf{C}_{is}\mathbf{D}_{jt}^T\mathbf{1}_{\{Y_i \neq 0\}}\mathbf{1}_{\{Y_j = 0\}} + \sum_{i \neq j} h_{jt}(Y_j)\mathbf{D}_{is}\mathbf{C}_{jt}^T\mathbf{1}_{\{Y_i = 0\}}\mathbf{1}_{\{Y_j \neq 0\}} \\ &\quad + \sum_{i=1}^n \mathbf{D}_{is}\mathbf{D}_{it}^T\mathbf{1}_{\{Y_i = 0\}} + \sum_{i \neq j} \mathbf{D}_{is}\mathbf{D}_{jt}^T\mathbf{1}_{\{Y_i = 0\}}\mathbf{1}_{\{Y_j = 0\}} \end{aligned}$$

Recall that  $E_{is} = E[h_{is}(Y_i)\mathbf{1}_{\{Y_i \neq 0\}}]$ . Denote  $E_{ist} = E[h_{is}(Y_i)h_{it}(Y_i)\mathbf{1}_{\{Y_i \neq 0\}}]$ . Then

$$\begin{aligned}
E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_s} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_t^T}\right) &= \sum_{i=1}^n E_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T + \sum_{i \neq j} E_{is} E_{jt} \mathbf{C}_{is} \mathbf{C}_{jt}^T \\
&\quad + \sum_{i \neq j} P_j E_{is} \mathbf{C}_{is} \mathbf{D}_{jt}^T + \sum_{i \neq j} P_i E_{jt} \mathbf{D}_{is} \mathbf{C}_{jt}^T \\
&\quad + \sum_{i=1}^n P_i \mathbf{D}_{is} \mathbf{D}_{it}^T + \sum_{i \neq j} P_i P_j \mathbf{D}_{is} \mathbf{D}_{jt}^T \\
&= \sum_{i=1}^n E_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T + \left[ \sum_{i=1}^n E_{is} \mathbf{C}_{is} \right] \left[ \sum_{i=1}^n E_{it} \mathbf{C}_{it} \right]^T - \sum_{i=1}^n E_{is} E_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T \\
&\quad + \left[ \sum_{i=1}^n E_{is} \mathbf{C}_{is} \right] \left[ \sum_{i=1}^n P_i \mathbf{D}_{it} \right]^T - \sum_{i=1}^n P_i E_{is} \mathbf{C}_{is} \mathbf{D}_{it}^T \\
&\quad + \left[ \sum_{i=1}^n P_i \mathbf{D}_{is} \right] \left[ \sum_{i=1}^n E_{it} \mathbf{C}_{it} \right]^T - \sum_{i=1}^n P_i E_{it} \mathbf{D}_{is} \mathbf{C}_{it}^T + \sum_{i=1}^n P_i \mathbf{D}_{is} \mathbf{D}_{it}^T \\
&\quad + \left[ \sum_{i=1}^n P_i \mathbf{D}_{is} \right] \cdot \left[ \sum_{i=1}^n P_i \mathbf{D}_{it} \right]^T - \sum_{i=1}^n P_i^2 \mathbf{D}_{is} \mathbf{D}_{it}^T \\
&= \left[ \sum_{i=1}^n E_{is} \mathbf{C}_{is} + \sum_{i=1}^n P_i \mathbf{D}_{is} \right] \left[ \sum_{i=1}^n E_{it} \mathbf{C}_{it} + \sum_{i=1}^n P_i \mathbf{D}_{it} \right]^T \\
&\quad + \sum_{i=1}^n E_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T - \sum_{i=1}^n E_{is} E_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T \\
&\quad - \sum_{i=1}^n P_i E_{is} \mathbf{C}_{is} \mathbf{D}_{it}^T - \sum_{i=1}^n P_i E_{it} \mathbf{D}_{is} \mathbf{C}_{it}^T + \sum_{i=1}^n P_i (1 - P_i) \mathbf{D}_{is} \mathbf{D}_{it}^T
\end{aligned}$$

Denote  $e_{is} = E[\partial \log f_{\theta_i}(Y'_i) / \partial \theta_{is}]$  and

$$e_{ist} = E\left[\frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{is}} \cdot \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{it}}\right]$$

It can be verified that

$$E_{ist} = (1 - \phi_i)e_{ist} - (1 - \phi_i)p_o(\theta_i) \cdot \frac{\partial \log p_o(\theta_i)}{\partial \theta_{is}} \cdot \frac{\partial \log p_o(\theta_i)}{\partial \theta_{it}}$$

Recall that

$$\begin{aligned} E_{is} &= (1 - \phi_i)e_{is} - (1 - \phi_i)p_o(\theta_i) \frac{\partial \log p_o(\theta_i)}{\partial \theta_{is}} \\ \mathbf{D}_{is} &= \frac{(1 - \phi_i)p_o(\theta_i)}{P_i} \cdot \frac{\partial \log p_o(\theta_i)}{\partial \theta_{is}} \mathbf{C}_{is} \end{aligned}$$

It can be verified that  $E_{is}\mathbf{C}_{is} + P_i\mathbf{D}_{is} = (1 - \phi_i)e_{is}\mathbf{C}_{is}$  and

$$\begin{aligned} & -E_{is}E_{it}\mathbf{C}_{is}\mathbf{C}_{it}^T - P_iE_{is}\mathbf{C}_{is}\mathbf{D}_{it}^T - P_iE_{it}\mathbf{D}_{is}\mathbf{C}_{it}^T + P_i(1 - P_i)\mathbf{D}_{is}\mathbf{D}_{it}^T \\ &= (1 - \phi_i)^2 \left[ -e_{is}e_{it} + \frac{p_o^2(\theta_i)}{P_i} \frac{\partial \log p_o(\theta_i)}{\partial \theta_{is}} \cdot \frac{\partial \log p_o(\theta_i)}{\partial \theta_{it}} \right] \mathbf{C}_{is}\mathbf{C}_{it}^T \end{aligned}$$

Therefore we obtain the simplified formula:

$$\begin{aligned} E\left(\frac{\partial l}{\partial \beta_s} \cdot \frac{\partial l}{\partial \beta_t^T}\right) &= \left[ \sum_{i=1}^n (1 - \phi_i)e_{is}\mathbf{C}_{is} \right] \cdot \left[ \sum_{i=1}^n (1 - \phi_i)e_{it}\mathbf{C}_{it} \right]^T \\ &\quad - \sum_{i=1}^n (1 - \phi_i)^2 e_{is}e_{it}\mathbf{C}_{is}\mathbf{C}_{it}^T + \sum_{i=1}^n (1 - \phi_i)e_{ist}\mathbf{C}_{is}\mathbf{C}_{it}^T \\ &\quad - \sum_{i=1}^n \frac{\phi_i(1 - \phi_i)p_o(\theta_i)}{\phi_i + (1 - \phi_i)p_o(\theta_i)} \cdot \frac{\partial \log p_o(\theta_i)}{\partial \theta_{is}} \cdot \frac{\partial \log p_o(\theta_i)}{\partial \theta_{it}} \mathbf{C}_{is}\mathbf{C}_{it}^T \end{aligned}$$

Note that there is no  $\mathbf{D}_{is}$  in the simplified formula. For many cases,  $e_{is} = 0$  which further simplifies the formula. Note that  $s = t$  is allowed in the above formula. □



### 3.4.2 Fisher information of Hurdle regression models

*Theorem 3.3.* Under regularity conditions, the Fisher information matrix of Hurdle Regression Mode defined in section 3.2 is given by:

$$\mathbf{F}(\boldsymbol{\pi}) = E \begin{bmatrix} \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \cdots & \frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \\ \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \cdots & \frac{\partial l}{\partial \boldsymbol{\beta}_1} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\gamma}^T} & \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\beta}_1^T} & \cdots & \frac{\partial l}{\partial \boldsymbol{\beta}_b} \frac{\partial l}{\partial \boldsymbol{\beta}_b^T} \end{bmatrix} \quad (3.5)$$

where

$$\begin{aligned} E \left( \frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\gamma}^T} \right) &= \sum_{i=1}^n \frac{(1 - \phi_i)}{\phi_i} \mathbf{A}_i \mathbf{A}_i^T \\ E \left( \frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_l^T} \right) &= - \sum_{i=1}^n \frac{(1 - \phi_i)}{1 - p_0(\boldsymbol{\theta}_i)} E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{il}} \right] \mathbf{A}_i \mathbf{C}_{il}^T \\ E \left( \frac{\partial l}{\partial \boldsymbol{\beta}_s} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_t^T} \right) &= \left[ \sum_{i=1}^n \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} e_{is} \mathbf{C}_{is} \right] \cdot \left[ \sum_{i=1}^n \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} e_{it} \mathbf{C}_{it} \right]^T \\ &+ \sum_{i=1}^n \frac{(1 - \phi_i) p_0(\boldsymbol{\theta}_i)}{[1 - p_0(\boldsymbol{\theta}_i)]^2} \left[ e_{is} \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{it}} + e_{it} \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{is}} \right. \\ &\quad \left. + \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{is}} \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{it}} \right] \mathbf{C}_{is} \mathbf{C}_{it}^T \\ &+ \sum_{i=1}^n \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} e_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T - \sum_{i=1}^n \left[ \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} \right]^2 e_{is} e_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T \end{aligned}$$

where

$$\begin{aligned}
P_i &= P(Y_i = 0) = \phi_i \\
\mathbf{A}_i &= \frac{1}{1 - \phi_i} \cdot (g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma}) \mathbf{G}_i \in \mathbb{R}^s \\
\mathbf{K}_i &= \frac{1}{\phi_i} (g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma}) \mathbf{G}_i = \frac{1 - \phi_i}{\phi_i} \mathbf{A}_i \\
h_{ij}(Y_i) &= \frac{\partial \log f_{\theta_i}(Y_i)}{\partial \theta_{ij}} \in \mathbb{R} \\
\mathbf{C}_{ij} &= (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \in \mathbb{R}^{t_j} \\
\mathbf{T}_{ij} &= \frac{p_0(\boldsymbol{\theta}_i)}{1 - p_0(\boldsymbol{\theta}_i)} \cdot \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{ij}} \cdot (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \in \mathbb{R}^{t_j}
\end{aligned}$$

Note that under regularity conditions,  $E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_j}\right) = \mathbf{0}$  and  $e_{is} = 0$ , while in general,

$$E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_j}\right) = \sum_{i=1}^n (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \frac{(1 - \phi_i)}{1 - p_0(\boldsymbol{\theta}_i)} E\left[\frac{\partial \log f_{\theta_i}(Y_i)}{\partial \theta_{ij}}\right]$$

**Proof of Theorem 3.3:** For Hurdle regression models,

$$\begin{aligned}
l(\boldsymbol{\pi}) &= \sum_{i=1}^n \log(\phi_i) \mathbf{1}_{\{Y_i=0\}} + \sum_{i=1}^n \log(1 - \phi_i) \mathbf{1}_{\{Y_i \neq 0\}} \\
&\quad + \sum_{i=1}^n \log f_{\theta_i}(Y_i) \mathbf{1}_{\{Y_i \neq 0\}} - \sum_{i=1}^n \log [1 - p_0(\boldsymbol{\theta}_i)] \mathbf{1}_{\{Y_i \neq 0\}}
\end{aligned}$$

In order to simplify the notations, we let

$$\begin{aligned}
P_i &= P(Y_i = 0) = \phi_i \\
\mathbf{A}_i &= \frac{1}{1 - \phi_i} \cdot (g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma}) \mathbf{G}_i \in \mathbb{R}^s \\
\mathbf{K}_i &= \frac{1}{\phi_i} (g^{-1})'(\mathbf{G}_i^T \boldsymbol{\gamma}) \mathbf{G}_i = \frac{1 - \phi_i}{\phi_i} \mathbf{A}_i \\
h_{ij}(Y_i) &= \frac{\partial \log f_{\theta_i}(Y_i)}{\partial \theta_{ij}} \in \mathbb{R} \\
\mathbf{C}_{ij} &= (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \in \mathbb{R}^{t_j} \\
\mathbf{T}_{ij} &= \frac{p_0(\boldsymbol{\theta}_i)}{1 - p_0(\boldsymbol{\theta}_i)} \cdot \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{ij}} \cdot (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \in \mathbb{R}^{t_j}
\end{aligned}$$

Then it can be verified that

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^n \mathbf{K}_i \mathbf{1}_{\{Y_i=0\}} - \sum_{i=1}^n \mathbf{A}_i \mathbf{1}_{\{Y_i \neq 0\}} \\
\frac{\partial l}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^n h_{ij}(Y_i) \mathbf{C}_{ij} \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i=1}^n \mathbf{T}_{ij} \mathbf{1}_{\{Y_i \neq 0\}} \\
E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}}\right) &= \sum_{i=1}^n \frac{1 - \phi_i}{\phi_i} \mathbf{A}_i P_i - \sum_{i=1}^n \mathbf{A}_i (1 - P_i) = \mathbf{0} \\
E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_j}\right) &= \sum_{i=1}^n \mathbf{C}_{ij} E\left[h_{ij}(Y_i) \mathbf{1}_{\{Y_i \neq 0\}}\right] + \sum_{i=1}^n \mathbf{T}_{ij} (1 - P_i) \\
&= \sum_{i=1}^n (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \left\{ E\left[\frac{\partial \log f_{\theta_i}(Y_i)}{\partial \theta_{ij}} \mathbf{1}_{\{Y_i \neq 0\}}\right] + \frac{p_0(\boldsymbol{\theta}_i)}{1 - p_0(\boldsymbol{\theta}_i)} \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{ij}} (1 - \phi_i) \right\} \\
&= \sum_{i=1}^n (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \frac{(1 - \phi_i)}{1 - p_0(\boldsymbol{\theta}_i)} \left\{ E\left[\frac{\partial \log f_{\theta_i}(Y_i)}{\partial \theta_{ij}} \mathbf{1}_{\{Y_i' \neq 0\}}\right] + \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{ij}} p_0(\boldsymbol{\theta}_i) \right\} \\
&= \sum_{i=1}^n (h_j^{-1})'(\mathbf{B}_{ij}^T \boldsymbol{\beta}_j) \mathbf{B}_{ij} \frac{(1 - \phi_i)}{1 - p_0(\boldsymbol{\theta}_i)} E\left[\frac{\partial \log f_{\theta_i}(Y_i')}{\partial \theta_{ij}}\right]
\end{aligned}$$

where  $Y'_i \sim f_{\theta_i}$ , a baseline distribution. Note that if  $E\left[\partial \log f_{\theta_i}(Y'_i)/\partial \theta_{ij}\right] = 0$  for  $i = 1, \dots, n$ , then  $E\left(\partial l/\partial \beta_j\right) = 0$  for  $j = 1, \dots, b$ .

The leading term of the Fisher information matrix (Equation 3.5) can be written as follows

$$E \begin{bmatrix} \frac{\partial l}{\partial \gamma} \frac{\partial l}{\partial \gamma^T} & \frac{\partial l}{\partial \gamma} \frac{\partial l}{\partial \beta_1^T} & \cdots & \frac{\partial l}{\partial \gamma} \frac{\partial l}{\partial \beta_b^T} \\ \frac{\partial l}{\partial \beta_1} \frac{\partial l}{\partial \gamma^T} & \frac{\partial l}{\partial \beta_1} \frac{\partial l}{\partial \beta_1^T} & \cdots & \frac{\partial l}{\partial \beta_1} \frac{\partial l}{\partial \beta_b^T} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial l}{\partial \beta_b} \frac{\partial l}{\partial \gamma^T} & \frac{\partial l}{\partial \beta_b} \frac{\partial l}{\partial \beta_1^T} & \cdots & \frac{\partial l}{\partial \beta_b} \frac{\partial l}{\partial \beta_b^T} \end{bmatrix} \quad (3.6)$$

We calculate these entries one by one. Note that matrix (Equation 3.6) is symmetric if  $l \in C^2$ , that is, twice continuously differentiable.

$$\begin{aligned} \frac{\partial l}{\partial \gamma} \cdot \frac{\partial l}{\partial \gamma^T} &= \sum_{i=1}^n \mathbf{A}_i \mathbf{A}_i^T \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i \neq j} \mathbf{A}_i \mathbf{A}_j^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\ &\quad - \sum_{i \neq j} \mathbf{A}_i \mathbf{K}_j^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j = 0\}} - \sum_{i \neq j} \mathbf{K}_i \mathbf{A}_j^T \mathbf{1}_{\{Y_i = 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\ &\quad + \sum_{i=1}^n \mathbf{K}_i \mathbf{K}_i^T \mathbf{1}_{\{Y_i = 0\}} + \sum_{i \neq j} \mathbf{K}_i \mathbf{K}_j^T \mathbf{1}_{\{Y_i = 0\}} \mathbf{1}_{\{Y_j = 0\}} \end{aligned}$$

Since  $(1 - P_i)\mathbf{A}_i - P_i\mathbf{K}_i = \mathbf{0}$  for each  $i = 1, \dots, n$ , it can be verified that

$$\begin{aligned}
 E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\gamma}^T}\right) &= \left[ \sum_{i=1}^n (1 - P_i)\mathbf{A}_i - \sum_{i=1}^n P_i\mathbf{K}_i \right] \cdot \left[ \sum_{i=1}^n (1 - P_i)\mathbf{A}_i - \sum_{i=1}^n P_i\mathbf{K}_i \right]^T \\
 &\quad + \sum_{i=1}^n P_i(1 - P_i)(\mathbf{A}_i + \mathbf{K}_i)(\mathbf{A}_i + \mathbf{K}_i)^T \\
 &= \sum_{i=1}^n P_i(1 - P_i)(\mathbf{A}_i + \mathbf{K}_i)(\mathbf{A}_i + \mathbf{K}_i)^T \\
 &= \sum_{i=1}^n \frac{(1 - \phi_i)}{\phi_i} \mathbf{A}_i \mathbf{A}_i^T
 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_l^T} &= - \sum_{i=1}^n \mathbf{A}_i \mathbf{C}_{il}^T h_{il}(Y_i) \mathbf{1}_{\{Y_i \neq 0\}} - \sum_{i \neq j} \mathbf{A}_i \mathbf{C}_{jl}^T h_{jl}(Y_j) \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\
 &\quad - \sum_{i=1}^n \mathbf{A}_i \mathbf{T}_{il}^T \mathbf{1}_{\{Y_i \neq 0\}} - \sum_{i \neq j} \mathbf{A}_i \mathbf{T}_{jl}^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\
 &\quad + \sum_{i \neq j} \mathbf{K}_i h_{jl}(Y_j) \mathbf{C}_{jl}^T \mathbf{1}_{\{Y_i = 0\}} \mathbf{1}_{\{Y_j \neq 0\}} + \sum_{i \neq j} \mathbf{K}_i \mathbf{T}_{jl}^T \mathbf{1}_{\{Y_i = 0\}} \mathbf{1}_{\{Y_j \neq 0\}}
 \end{aligned}$$

Note that for Hurdle models

$$\begin{aligned}
 E[h_{jl}(Y_j) \mathbf{1}_{\{Y_j \neq 0\}}] &= E\left(\frac{\partial \log f_{\theta_j}(Y_j)}{\partial \theta_{jl}} \mathbf{1}_{\{Y_j \neq 0\}}\right) \\
 &= \frac{1 - \phi_j}{1 - p_0(\boldsymbol{\theta}_j)} E\left(\frac{\partial \log f_{\theta_j}(Y'_j)}{\partial \theta_{jl}} \mathbf{1}_{\{Y'_j \neq 0\}}\right) \\
 &= \frac{1 - \phi_j}{1 - p_0(\boldsymbol{\theta}_j)} E\left(\frac{\partial \log f_{\theta_j}(Y'_j)}{\partial \theta_{jl}}\right) - \frac{1 - \phi_j}{1 - p_0(\boldsymbol{\theta}_j)} p_0(\boldsymbol{\theta}_j) \frac{\partial \log p_0(\boldsymbol{\theta}_j)}{\partial \theta_{jl}} \\
 &\triangleq E_{jl}
 \end{aligned}$$

where  $Y'_j$  follows the baseline distribution  $f_{\theta_j}(y)$ . Then

$$\begin{aligned}
& E\left(\frac{\partial l}{\partial \gamma} \cdot \frac{\partial l}{\partial \beta_l^T}\right) \\
&= -\sum_{i=1}^n E_{il} \mathbf{A}_i \mathbf{C}_{il}^T - \sum_{i \neq j} (1 - P_i) E_{jl} \mathbf{A}_i \mathbf{C}_{jl}^T - \sum_{i=1}^n (1 - P_i) \mathbf{A}_i \mathbf{T}_{il}^T \\
&\quad - \sum_{i \neq j} (1 - P_i)(1 - P_j) \mathbf{A}_i \mathbf{T}_{jl}^T + \sum_{i \neq j} P_i \mathbf{k}_i E_{jl} \mathbf{C}_{jl}^T + \sum_{i \neq j} P_i (1 - P_j) \mathbf{K}_i \mathbf{T}_{jl}^T \\
&= -\left[\sum_{i=1}^n (1 - P_i) \mathbf{A}_i\right] \cdot \left[\sum_{i=1}^n E_{il} \mathbf{C}_{il}^T\right] + \sum_{i=1}^n (1 - P_i) E_{il} \mathbf{A}_i \mathbf{C}_{il}^T - \sum_{i=1}^n E_{il} \mathbf{A}_i \mathbf{C}_{il}^T \\
&\quad - \left[\sum_{i=1}^n (1 - P_i) \mathbf{A}_i\right] \cdot \left[\sum_{i=1}^n (1 - P_i) \mathbf{T}_{il}^T\right] + \sum_{i=1}^n (1 - P_i)^2 \mathbf{A}_i \mathbf{T}_{il}^T - \sum_{i=1}^n (1 - P_i) \mathbf{A}_i \mathbf{T}_{il}^T \\
&\quad + \left[\sum_{i=1}^n P_i \mathbf{K}_i\right] \cdot \left[\sum_{i=1}^n E_{il} \mathbf{C}_{il}^T\right] - \sum_{i=1}^n P_i E_{il} \mathbf{k}_i \mathbf{C}_{il}^T \\
&\quad + \left[\sum_{i=1}^n P_i \mathbf{K}_i\right] \cdot \left[\sum_{i=1}^n (1 - P_i) \mathbf{T}_{il}^T\right] - \sum_{i=1}^n P_i (1 - P_i) \mathbf{K}_i \mathbf{T}_{il}^T \\
&= \left[\sum_{i=1}^n P_i \mathbf{K}_i - \sum_{i=1}^n (1 - P_i) \mathbf{A}_i\right] \cdot \left[\sum_{i=1}^n E_{il} \mathbf{C}_{il}^T + \sum_{i=1}^n (1 - P_i) \mathbf{T}_{il}^T\right] \\
&\quad - \sum_{i=1}^n P_i E_{il} \mathbf{A}_i \mathbf{C}_{il}^T - \sum_{i=1}^n P_i (1 - P_i) \mathbf{A}_i \mathbf{T}_{il}^T \\
&\quad - \sum_{i=1}^n P_i E_{il} \mathbf{K}_i \mathbf{C}_{il}^T - \sum_{i=1}^n P_i (1 - P_i) \mathbf{K}_i \mathbf{T}_{il}^T \\
&= -\sum_{i=1}^n P_i E_{il} (\mathbf{A}_i + \mathbf{K}_i) \mathbf{C}_{il}^T - \sum_{i=1}^n P_i (1 - P_i) (\mathbf{A}_i + \mathbf{K}_i) \mathbf{T}_{il}^T \\
&= -\sum_{i=1}^n E_{il} \mathbf{A}_i \mathbf{C}_{il}^T - \sum_{i=1}^n (1 - P_i) \mathbf{A}_i \mathbf{T}_{il}^T
\end{aligned}$$

since  $\mathbf{A}_i = P_i(\mathbf{A}_i + \mathbf{K}_i)$ . On the other hand,

$$\begin{aligned} E_{il} &= \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} E\left(\frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{il}}\right) - \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} p_0(\boldsymbol{\theta}_i) \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{il}} \\ \mathbf{T}_{il} &= \frac{p_0(\boldsymbol{\theta}_i)}{1 - p_0(\boldsymbol{\theta}_i)} \cdot \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{il}} \cdot (h_l^{-1})'(\mathbf{B}_{il}^T \boldsymbol{\beta}_j) \mathbf{B}_{il} \end{aligned}$$

Then

$$E\left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_l^T}\right) = - \sum_{i=1}^n \frac{(1 - \phi_i)}{1 - p_0(\boldsymbol{\theta}_i)} E\left[\frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{il}}\right] \mathbf{A}_i \mathbf{C}_{il}^T$$

where  $Y'_i$  follows the baseline distribution  $f_{\theta_i}(y)$ .

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}_s} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_t^T} &= \sum_{i=1}^n h_{is}(Y_i) h_{it}(Y_i) \mathbf{C}_{is} \mathbf{C}_{it}^T \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i \neq j} h_{is}(Y_i) h_{jt}(Y_j) \mathbf{C}_{is} \mathbf{C}_{jt}^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\ &\quad + \sum_{i=1}^n h_{is}(Y_i) \mathbf{C}_{is} \mathbf{T}_{it}^T \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i \neq j} h_{is}(Y_i) \mathbf{C}_{is} \mathbf{T}_{jt}^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\ &\quad + \sum_{i=1}^n h_{it}(Y_i) \mathbf{T}_{is} \mathbf{C}_{it}^T \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i \neq j} h_{it}(Y_i) \mathbf{T}_{js} \mathbf{C}_{it}^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \\ &\quad + \sum_{i=1}^n \mathbf{T}_{is} \mathbf{T}_{it}^T \mathbf{1}_{\{Y_i \neq 0\}} + \sum_{i \neq j} \mathbf{T}_{is} \mathbf{T}_{jt}^T \mathbf{1}_{\{Y_i \neq 0\}} \mathbf{1}_{\{Y_j \neq 0\}} \end{aligned}$$

Recall that  $E_{is} = E[h_{is}(Y_i)\mathbf{1}_{\{Y_i \neq 0\}}]$ . Denote  $E_{ist} = E[h_{is}(Y_i)h_{it}(Y_i)\mathbf{1}_{\{Y_i \neq 0\}}]$ . Then

$$\begin{aligned}
E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_s} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_t^T}\right) &= \sum_{i=1}^n E_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T + \sum_{i \neq j} E_{is} E_{jt} \mathbf{C}_{is} \mathbf{C}_{jt}^T \\
&\quad + \sum_{i=1}^n E_{is} \mathbf{C}_{is} \mathbf{T}_{it}^T + \sum_{i \neq j} (1 - P_j) E_{is} \mathbf{C}_{is} \mathbf{T}_{jt}^T \\
&\quad + \sum_{i=1}^n E_{it} \mathbf{T}_{is} \mathbf{C}_{it}^T + \sum_{i \neq j} (1 - P_j) E_{it} \mathbf{T}_{js} \mathbf{C}_{it}^T \\
&\quad + \sum_{i=1}^n (1 - P_i) \mathbf{T}_{is} \mathbf{T}_{it}^T + \sum_{i \neq j} (1 - P_i)(1 - P_j) \mathbf{T}_{is} \mathbf{T}_{jt}^T \\
&= \sum_{i=1}^n E_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T + \left[ \sum_{i=1}^n E_{is} \mathbf{C}_{is} \right] \left[ \sum_{i=1}^n E_{it} \mathbf{C}_{it} \right]^T - \sum_{i=1}^n E_{is} E_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T \\
&\quad + \sum_{i=1}^n E_{is} \mathbf{C}_{is} \mathbf{T}_{it}^T + \left[ \sum_{i=1}^n E_{is} \mathbf{C}_{is} \right] \left[ \sum_{i=1}^n (1 - P_i) \mathbf{T}_{it} \right]^T - \sum_{i=1}^n (1 - P_i) E_{is} \mathbf{C}_{is} \mathbf{T}_{it}^T \\
&\quad + \sum_{i=1}^n E_{it} \mathbf{T}_{is} \mathbf{C}_{it}^T + \left[ \sum_{i=1}^n (1 - P_i) \mathbf{T}_{is} \right] \left[ \sum_{i=1}^n E_{it} \mathbf{C}_{it} \right]^T \\
&\quad - \sum_{i=1}^n (1 - P_i) E_{it} \mathbf{T}_{is} \mathbf{C}_{it}^T + \sum_{i=1}^n (1 - P_i) \mathbf{T}_{is} \mathbf{T}_{it}^T \\
&\quad + \left[ \sum_{i=1}^n (1 - P_i) \mathbf{T}_{is} \right] \cdot \left[ \sum_{i=1}^n (1 - P_i) \mathbf{T}_{it} \right]^T - \sum_{i=1}^n (1 - P_i)^2 \mathbf{T}_{is} \mathbf{T}_{it}^T \\
&= \left[ \sum_{i=1}^n E_{is} \mathbf{C}_{is} + \sum_{i=1}^n (1 - P_i) \mathbf{T}_{is} \right] \left[ \sum_{i=1}^n E_{it} \mathbf{C}_{it} + \sum_{i=1}^n (1 - P_i) \mathbf{T}_{it} \right]^T \\
&\quad + \sum_{i=1}^n E_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T - \sum_{i=1}^n E_{is} E_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T + \sum_{i=1}^n P_i E_{is} \mathbf{C}_{is} \mathbf{T}_{it}^T \\
&\quad + \sum_{i=1}^n P_i E_{it} \mathbf{T}_{is} \mathbf{C}_{it}^T + \sum_{i=1}^n P_i (1 - P_i) \mathbf{T}_{is} \mathbf{T}_{it}^T
\end{aligned}$$



Denote  $e_{is} = E \left[ \partial \log f_{\theta_i}(Y'_i) / \partial \theta_{is} \right]$  and

$$e_{ist} = E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{is}} \cdot \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{it}} \right]$$

It can be verified that

$$E_{ist} = \frac{1 - \phi_i}{1 - p_0(\theta_i)} e_{ist} - \frac{1 - \phi_i}{1 - p_0(\theta_i)} p_0(\theta_i) \cdot \frac{\partial \log p_0(\theta_i)}{\partial \theta_{is}} \cdot \frac{\partial \log p_0(\theta_i)}{\partial \theta_{it}}$$

Recall that

$$\begin{aligned} E_{is} &= \frac{1 - \phi_i}{1 - p_0(\theta_i)} e_{is} - \frac{1 - \phi_i}{1 - p_0(\theta_i)} p_0(\theta_i) \frac{\partial \log p_0(\theta_i)}{\partial \theta_{is}} \\ \mathbf{T}_{is} &= \frac{p_0(\theta_i)}{1 - p_0(\theta_i)} \cdot \frac{\partial \log p_0(\theta_i)}{\partial \theta_{is}} \mathbf{C}_{is} \end{aligned}$$

It can be verified that  $E_{is} \mathbf{C}_{is} + (1 - P_i) \mathbf{T}_{is} = \frac{1 - \phi_i}{1 - p_0(\theta_i)} e_{is} \mathbf{C}_{is}$  and

$$\begin{aligned} & E_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T - E_{is} E_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T + P_i E_{is} \mathbf{C}_{is} \mathbf{T}_{it}^T + P_i E_{it} \mathbf{T}_{is} \mathbf{C}_{it}^T + P_i (1 - P_i) \mathbf{T}_{is} \mathbf{T}_{it}^T \\ &= \frac{1 - \phi_i}{[1 - p_0(\theta_i)]^2} p_0(\theta_i) \left[ e_{is} \frac{\partial \log p_0(\theta_i)}{\partial \theta_{it}} + e_{it} \frac{\partial \log p_0(\theta_i)}{\partial \theta_{is}} + \frac{\partial \log p_0(\theta_i)}{\partial \theta_{is}} \cdot \frac{\partial \log p_0(\theta_i)}{\partial \theta_{it}} \right] \mathbf{C}_{is} \mathbf{C}_{it}^T \\ &+ \frac{1 - \phi_i}{1 - p_0(\theta_i)} e_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T - \left[ \frac{1 - \phi_i}{1 - p_0(\theta_i)} \right]^2 e_{is} e_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T \end{aligned}$$

Therefore we obtain the simplified formula:

$$\begin{aligned}
E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_s} \cdot \frac{\partial l}{\partial \boldsymbol{\beta}_t^T}\right) &= \left[ \sum_{i=1}^n \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} e_{is} \mathbf{C}_{is} \right] \cdot \left[ \sum_{i=1}^n \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} e_{it} \mathbf{C}_{it} \right]^T \\
&+ \sum_{i=1}^n \frac{(1 - \phi_i) p_0(\boldsymbol{\theta}_i)}{[1 - p_0(\boldsymbol{\theta}_i)]^2} \left[ e_{is} \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{it}} + e_{it} \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{is}} + \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{is}} \frac{\partial \log p_0(\boldsymbol{\theta}_i)}{\partial \theta_{it}} \right] \mathbf{C}_{is} \mathbf{C}_{it}^T \\
&+ \sum_{i=1}^n \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} e_{ist} \mathbf{C}_{is} \mathbf{C}_{it}^T - \sum_{i=1}^n \left[ \frac{1 - \phi_i}{1 - p_0(\boldsymbol{\theta}_i)} \right]^2 e_{is} e_{it} \mathbf{C}_{is} \mathbf{C}_{it}^T
\end{aligned}$$

Note that there is no  $\mathbf{T}_{is}$  in the simplified formula. For many cases,  $e_{is} = 0$  which further simplifies the formula. Note that  $s = t$  is allowed in the above formula.  $\square$

*Example 3.3.* For zero-inflated Poisson (ZIP) regression, the pmf of the baseline distribution is  $f_\theta(y) = e^{-\theta} \theta^y / y!$  with  $y = 0, 1, \dots$ . Then  $\partial \log f_\theta(y) / \partial \theta = y / \theta - 1$ . If  $Y' \sim f_\theta$ , then

$$E\left(\frac{\partial \log f_\theta(Y')}{\partial \theta}\right) = \frac{E(Y')}{\theta} - 1 = 0$$

Therefore,  $E(\partial l / \partial \boldsymbol{\beta}) = 0$  for ZIP regressions.  $\square$

### 3.5 Significance Test for Model Coefficients via Bootstrap

Let  $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

1. Resample  $S$  with replacement to get  $S_b^* = \{(x_1^{(b)}, y_1^{(b)}), (x_2^{(b)}, y_2^{(b)}), \dots, (x_n^{(b)}, y_n^{(b)})\}$

2. Refit the model to get  $\hat{\boldsymbol{\theta}}^{(b)} = (\hat{\theta}_0^{(b)}, \hat{\theta}_1^{(b)}, \dots, \hat{\theta}_p^{(b)})$

3. The  $(1 - \alpha)\%$  Bootstrap confidence intervals of  $\theta_j$  are  $(\xi_{\alpha/2}^*, \xi_{1-\alpha/2}^*)$ , where  $\xi_\alpha^*$  is the  $\alpha$ 'th sample quantile of  $(\hat{\theta}_j^{(1)}, \hat{\theta}_j^{(2)}, \dots, \hat{\theta}_j^{(B)})$ .

Here we used simple bootstrapped percentiles. More advanced bootstrapped confidence intervals could be used such as in (Wu, 1986) (55).

**Example 3.4. ZIP regression model (continued)** The baseline distribution of Poisson distribution is  $f_\lambda(y) = e^{-\lambda}\lambda^y/y!$  with  $p_0(\lambda) = e^{-\lambda}$ . That is, it has only one parameter  $\theta = \lambda$  with  $b = 1$ . In order to calculate the Fisher information matrix of ZIP regression model, the needed quantities are listed below. Note that  $s = t \equiv 1$  in this case.

$$\begin{aligned}\frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{is}} &= \frac{Y'_i}{\lambda_i} - 1 \\ \frac{\partial \log p_0(\theta_i)}{\partial \theta_{is}} &= -1 \\ e_{is} &= E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{is}} \right] = 0 \\ e_{ist} &= E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{is}} \cdot \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial \theta_{it}} \right] = \text{Var} \left( \frac{Y'_i}{\lambda_i} - 1 \right) = \frac{1}{\lambda_i}\end{aligned}$$

□

**Lemma 3.1.** If  $Y'$  follows Negative Binomial( $r, p$ ) with pmf  $\frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}p^y(1-p)^r$ ,  $y \in \{0, 1, 2, \dots\}$ , then

$$\begin{aligned}E[\Psi(Y' + r)] &= \Psi(r) - \log(1-p) \\ E[Y'\Psi(Y' + r)] &= \frac{pr}{1-p} [\Psi(r+1) - \log(1-p)]\end{aligned}$$

where  $\Psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$  is the digamma function.

**Proof of Lemma 3.1:** From Example 2.7, we get  $E[\Psi(Y' + r)] = \Psi(r) - \log(1 - p)$ . Recall that

$\Gamma(y) = \int_0^\infty t^{y-1} e^{-t} dt$  and  $\Gamma'(y) = \int_0^\infty t^{y-1} e^{-t} \log t dt$  for  $y > 0$ . Then

$$\begin{aligned}
E[Y' \Psi(Y' + r)] &= \sum_{y=0}^{\infty} y \frac{\Gamma'(y+r)}{\Gamma(y+r)} \cdot \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p^y (1-p)^r \\
&= \frac{(1-p)^r}{\Gamma(r)} \sum_{y=1}^{\infty} \frac{p^y}{(y-1)!} \Gamma'(y+r) \quad (\text{replace } y \text{ with } y+1) \\
&= \frac{p(1-p)^r}{\Gamma(r)} \sum_{y=0}^{\infty} \frac{p^y}{y!} \Gamma'(y+r+1) \\
&= \frac{p(1-p)^r}{\Gamma(r)} \sum_{y=0}^{\infty} \frac{p^y}{y!} \int_0^\infty t^{y+r} e^{-t} \log t dt \\
&= \frac{p(1-p)^r}{\Gamma(r)} \int_0^\infty \left( \sum_{y=0}^{\infty} \frac{(pt)^y}{y!} e^{-pt} \right) \cdot t^r e^{-t(1-p)} \log t dt \\
&= \frac{p(1-p)^r}{\Gamma(r)} \int_0^\infty t^r e^{-t(1-p)} \log t dt \quad (\text{let } s = (1-p)t) \\
&= \frac{p}{(1-p)\Gamma(r)} \int_0^\infty s^r e^{-s} [\log s - \log(1-p)] ds \\
&= \frac{p}{(1-p)\Gamma(r)} \left[ \int_0^\infty s^r e^{-s} \log s ds - \log(1-p) \int_0^\infty s^r e^{-s} ds \right] \\
&= \frac{p}{(1-p)\Gamma(r)} [\Gamma'(r+1) - \log(1-p)\Gamma(r+1)] \\
&= \frac{pr}{1-p} [\Psi(r+1) - \log(1-p)] \quad (\text{since } \Gamma(r+1) = r\Gamma(r))
\end{aligned}$$

□

**Example 3.5. ZINB regression model (with flexible  $r$ )** For zero-inflated negative binomial distributions, the pmf of the baseline distribution with parameters  $\theta = (r, p) \in (0, \infty) \times [0, 1]$  is given by

$f_{\theta}(y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p^y (1-p)^r$ ,  $y \in \{0, 1, 2, \dots\}$  with  $p_0(\theta) = (1-p)^r$ . Recall that  $E(Y') = \frac{pr}{1-p}$  and  $Var(Y') = \frac{pr}{(1-p)^2}$  if  $Y' \sim f_{\theta}(y)$ . For  $i = 1, \dots, n$  and  $\theta_i = (r_i, p_i)$ ,

$$\begin{aligned} \frac{\partial \log f_{\theta_i}(y)}{\partial r_i} &= \Psi(y + r_i) - \Psi(r_i) + \log(1 - p_i) \\ \frac{\partial \log f_{\theta_i}(y)}{\partial p_i} &= \frac{y}{p_i} - \frac{r_i}{1 - p_i} \\ \frac{\partial \log p_0(\theta_i)}{\partial r_i} &= \log(1 - p_i) \\ \frac{\partial \log p_0(\theta_i)}{\partial p_i} &= -\frac{r_i}{1 - p_i} \end{aligned}$$

where  $\Psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$  is known as the *digamma* function. Recall that  $\Gamma(r+1) = r\Gamma(r)$  and  $\Psi(r+1) = \Psi(r) + r^{-1}$ . According to Lemma 3.1, it can be verified that if  $Y'_i \sim NB(r_i, p_i)$ , then

$$\begin{aligned} e_{i1} &= E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial r_i} \right] = 0 \\ e_{i2} &= E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial p_i} \right] = 0 \\ e_{i11} &= E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial r_i} \cdot \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial r_i} \right] = E [\Psi^2(Y'_i + r_i)] - [\log(1 - p_i) - \Psi(r_i)]^2 \\ e_{i12} &= e_{i21} = E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial r_i} \cdot \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial p_i} \right] = \frac{1}{1 - p_i} \\ e_{i22} &= E \left[ \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial p_i} \cdot \frac{\partial \log f_{\theta_i}(Y'_i)}{\partial p_i} \right] = \frac{r_i}{p_i(1 - p_i)^2} \end{aligned}$$

Since an explicit formula of  $E[\Psi^2(Y'_i + r_i)]$  is not available, we calculate it numerically. That is, for each  $i = 1, \dots, n$ , we simulate  $Y'_{i1}, \dots, Y'_{im}$  iid  $\sim NB(r_i, p_i)$ , where  $m$  is a predetermined sample size, for example,  $m = 1000$ . Then  $E[\Psi^2(Y'_i + r_i)]$  is estimated by  $m^{-1} \sum_{l=1}^m \Psi^2(Y'_{il} + r_i)$ .  $\square$

### 3.6 Comparison between Fisher-Information-Based and Bootstrapped Confidence Intervals

In this section, we perform two simulation studies to examine the parameter estimation of ZIP and ZINB ( $r$  is flexible) regression models, implemented in R Studio 1.3.959.

- **ZIP Regression:** Let  $Y_i$  be the ZIP outcome of interest for the  $i^{th}$  participant. Also, let  $x_{i1} = \mathbf{1}$  and let  $x_{i2}, x_{i3}, x_{i4}$  be additional covariate desired in regression model where they generated from a standard normal distribution. The true parameters are  $\gamma = (-2.1, 0.5, -0.2, 0.7)^T$ , and  $\beta = (0, -0.9, 0.11, 1.19)^T$ . We assign a *logit* link function  $g(\phi_i) = \text{logit}(\phi_i) = \log \frac{\phi_i}{1-\phi_i} = X_i^T \gamma$ , and a log link function to  $\lambda_i > 0; h(\lambda_i) = \log(\lambda_i) = X_i^T \beta$ .

Based on 500 simulations in section 3.5, the 95% bootstrap confidence intervals and Fisher information matrix confidence intervals are given in Table VI. Based on the simulation study, the bootstrap and Fisher-information confidence intervals are very similar. However, the Fisher-information confidence intervals are more symmetric about the true value in general especially for  $\gamma_2$ . The true value of  $\gamma_2$  is -0.20 and the bootstrap confidence interval is (-0.0131, -0.2205), it's clear that -0.20 is not in the center of the Bootstrap confidence interval. On the other hand, the Fisher-information confidence interval is (-0.1225, -0.3358) where -0.20 is exactly in the center since the mle in this case is fairly accurate.

- **ZINB Regression:** Let  $Y_i$  be the ZINB outcome of interest for the  $i^{th}$  participant. Also, let  $x_{i1} = \mathbf{1}$  and let  $x_{i2}, x_{i3}, x_{i4}$  be additional covariate desired in regression model where they generated from a standard normal distribution. The true parameters are  $\gamma = (-2.5, 0.30, -0.20, 0.50)^T$ ,  $\beta_1 = (-0.20, -0.85, 0.11, -0.16)^T$ ,  $\beta_2 = (-0.30, -0.53, 0.15, 0.30)^T$ .

We assign a *logit* link function  $g(\phi_i) = \text{logit}(\phi_i) = \log \frac{\phi_i}{1-\phi_i} = \mathbf{X}_i^T \boldsymbol{\gamma}$ , a log link function  $h_1(r_i) = \log(r_i) = \mathbf{X}_i^T \boldsymbol{\beta}_1$ , and *logit* link function  $h_2(p_i) = \text{logit}(p_i) = \log \frac{p_i}{1-p_i} = \mathbf{X}_i^T \boldsymbol{\beta}_2$ .

Based on 500 simulations in section 3.5, the 95% bootstrap confidence intervals and Fisher information matrix confidence intervals are given in Table VII. Based on the simulation study, the Fisher-information confidence intervals are shorter than the bootstrap confidence intervals in general.

The true value of  $\beta_{01}$  is -0.20 which is a significant negative value. However, the bootstrap confidence interval of  $\beta_{01}$  is (-0.311, 0.040) which includes 0. The Fisher-information confidence interval is (-0.259, -0.177) which is more accurate than the bootstrap confidence interval.

In conclusion, we prefer Fisher-information-based confidence intervals when it is available. Nevertheless, when Fisher information matrix is not available due to intensive computation involved or the lack of explicit formula, or when sample size is relatively small, bootstrapped confidence intervals provide us a feasible solution.

TABLE VI: ZIP REGRESSION: BOOTSTRAP CONFIDENCE INTERVAL VS. FISHER INFORMATION MATRIX CONFIDENCE INTERVAL

Parameter	True Value	Bootstrap CI		Fisher-information CI	
		Confidence Interval	Length	Confidence Interval	Length
$\gamma_0$	-2.1	(-2.419, -1.959)	0.46	(-2.385, -1.951)	0.434
$\gamma_1$	0.50	(0.395, 0.614)	0.219	(0.400, 0.668)	0.268
$\gamma_2$	-0.20	<b>(-0.221, -0.013)</b>	0.208	(-0.336, -0.123)	0.213
$\gamma_3$	0.70	(0.653, 0.996)	0.343	(0.613, 0.962)	0.349
$\beta_0$	0	(-0.012, 0.046)	0.058	(-0.029, 0.039)	0.068
$\beta_1$	-0.90	(-0.910, -0.879)	0.031	(-0.913, -0.882)	0.031
$\beta_2$	0.11	(0.099, 0.129)	0.03	(0.097, 0.127)	0.03
$\beta_3$	1.19	(1.170, 1.201)	0.031	(1.175, 1.208)	0.033

TABLE VII: ZINB REGRESSION: BOOTSTRAP CONFIDENCE INTERVAL VS. FISHER INFORMATION MATRIX CONFIDENCE INTERVAL

Parameter	True Value	Bootstrap CI		Fisher-information CI	
		Confidence Interval	Length	Confidence Interval	Length
$\gamma_0$	-2.5	(-3.231, -1.881)	1.35	(-2.997, -2.141)	<b>0.856</b>
$\gamma_1$	0.30	(0.198, 0.381)	0.183	(0.246, 0.513)	0.267
$\gamma_2$	-0.20	(-0.391, -0.059)	0.332	(-0.403, -0.064)	0.339
$\gamma_3$	0.50	(0.431, 0.620)	0.189	(0.435, 0.668)	0.233
$\beta_{01}$	-0.20	<b>(-0.311, 0.040)</b>	0.351	(-0.259, -0.177)	0.082
$\beta_{11}$	-0.85	(-0.910, -0.774)	0.136	(-0.871, -0.838)	0.033
$\beta_{21}$	0.11	(0.087, 0.201)	0.114	(0.084, 0.128)	0.044
$\beta_{31}$	-0.16	(-0.188, -0.125)	0.063	(-0.172, -0.136)	0.036
$\beta_{02}$	-0.30	(-0.458, -0.192)	0.266	(-0.419, -0.282)	0.137
$\beta_{12}$	-0.53	(-0.587, -0.416)	0.171	(-0.561, -0.446)	0.115
$\beta_{22}$	0.15	(0.090, 0.266)	0.176	(0.139, 0.219)	0.08
$\beta_{32}$	0.30	(0.202, 0.322)	0.12	(0.298, 0.346)	0.048



## CHAPTER 4

### APPLICATION TO INSURANCE DATA

#### 4.1 Introduction: Modeling Insurance Claim Data

The prediction of the number of future claims is one of the major interests of insurance companies. It is challenging to model the insurance data due to highly right-skewed distribution and a large point mass at zero.

Poisson and negative binomial regression have been widely used for fitting count data. In the insurance area, for examples, Antonio et al. 2010 (56) used Poisson distribution to model the number of claims. David et al. 2015 (57) analyzed auto insurance claim data using Poisson and negative binomial models. Aitkin et al. 2005 and Renshaw 1994 (58; 59) fitted Poisson regression to data sets of U.K. motor claim data. Bartoszewicz 2005 (60) discussed Poisson and negative binomial and conclude that negative binomial regression with log link is more flexible comparing to the Poisson regression model. Lambert 1992 (32) proposed ZIP as a new technique for modeling sparse data with an application for modeling number of defects in manufacturing, since then ZIP has been applied to many sparse count data including insurance claim data and healthcare data. For example, Lee et al. 2002 (61) used ZIP for modeling young driver motor vehicle crashes. Yip et al. 2005 (62) suggested several zero-inflated distributions for insurance claim data. They show that zero-inflated models provide a better fit than the regular regression models. Mouatassim et al. 2012 (63) applied zero-inflated models to analyze private health insurance data. They show that ZIP regression performs better than Poisson model.

In this chapter, we aim to analyse the relationship between the number of claims and some covariates such as vehicle age, vehicle type, driver gender, etc . For example, questions like “Are bigger cars more likely to be damaged in an accident?” and “Do teenagers cause significantly more accidents than adults?” frequently arise and need an answer.

## 4.2 Model Selection

Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been widely used in the literature for model selection purpose (see, for example, (64), for a good review).

$$AIC = -2\loglik + 2d$$

For Logistic Regression Model AIC is given by:

$$AIC = \frac{-2}{N} \cdot \loglik + 2 \cdot \frac{d}{N}$$

The Bayesian information criterion (BIC) is given by:

$$BIC = -2 \cdot \loglik + (\log N) \cdot d$$

## 4.3 Parameters Estimation

We applied our zero-inflated and hurdle regression model the dataset “dataCar”. The parameters are estimated using our maximum likelihood estimation (MLE) method. Table X, and Table XI represent

TABLE VIII: POSSIBLE LINK FUNCTIONS FOR  $\phi$ 

Name	Link function: $\eta = g(\phi)$	$\phi = g^{-1}(\eta)$
logit	$\log(\phi/(1 - \phi))$	$\exp(\eta)/(1 + \exp(\eta))$
probit	$\Phi^{-1}(\phi)$	$\Phi(\eta)$
cloglog	$\log(-\log(1 - \phi))$	$1 - \exp(-\exp(\eta))$

the parameters estimates and their corresponding loglikelihood, AIC, and BIC with three link functions for the parameter  $\phi$ . Table VIII has the link functions and the corresponding inverse.

#### 4.4 Data set

We use the data set “dataCar” in R package “insuranceData”. The data consists of 67,856 one-year vehicle insurance policies issued in 2014-2015.

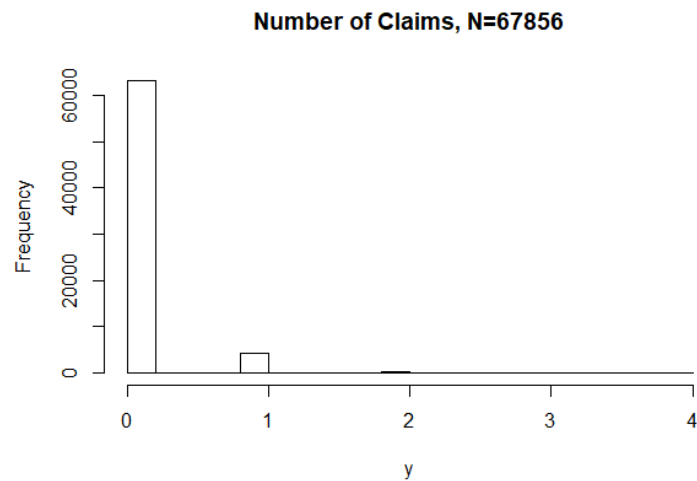


Figure 3: The percentage of total zero claims occurrences are 93.2%.

TABLE IX: NUMBER OF CLAIMS IN THE DATASET

Occurrence	Frequency	Percentage
0	63,232	<b>93.18%</b>
1	4,333	6.39%
2	271	0.40%
3	18	0.03%
4	2	0.00%
Total	67,856	100%

Table IX represents the response variable, which is number of claims. The table shows that the data is sparse; more than 93% of the response variable are zeros. Figure 3 shows that the data is highly right skewed. The mean of the data is 0.0728, and variance is 0.0773 which is greater than the mean.

In our application to Insurance Claim Data several zero-inflated and Hurdle regression models were built. Our Zero-Inflated Regression Models (ZIRM) not only cover currently available zero-inflated regression models, such as ZIP, ZINB with fixed  $r$ , ZIBB with constant prior parameters, but also include new regression models, including ZINB with flexible  $r$ , ZIBB with flexible prior parameters, and ZIBNB. We also build the corresponding Hurdle Regression Models for zero-altered responses. With the enriched model candidates, we perform model selection based on AIC and BIC criteria. We assign three different candidate link functions (logit, c-log-log, and probit) to the parameter  $\phi_i$ .

We conclude that ZINB with flexible  $r$  and probit link function for  $\phi$  has the smallest AIC comparing to some other regression models. In addition, ZINB with flexible  $r$  using probit link function for  $\phi$  has the smallest BIC. To the best of our knowledge, ZINB with fixed  $r$  is used in the literature for modeling sparse data. However, ZINB with flexible  $r$  is not included in the literature.

TABLE X: AIC USING DIFFERENT LINK FUNCTIONS FOR  $\phi$ .

Distribution	logit	c-log-log	probit
Hurdle Poisson(HP)	35078.2	53078.2	42449.1
Hurdle Negative Binomial(HNB)	34847.1	53142.1	42354.6
Hurdle Beta Binomial(HBB)	34938.2	53233.3	42445.8
Hurdle Beta Negative Binomial(HBNB)	35093.4	53357.1	4298.6
Zero-inflated Poisson(ZIP)	34742.1	34737.4	34743.6
Zero-inflated Negative Binomial, r is fixed(ZINB)	34722.7	34848.1	34721
Zero-inflated Negative Binomial(ZINB)	34745.6	34716.8	<b>34714.4</b>
Zero-inflated Beta Binomial(ZIBB)	34729.6	34744.1	34721.6
Zero-inflated Beta Binomial(ZIBB), n is fixed	34849	34898	34887.5
Zero-inflated Beta Binomial(ZIBB), fixed $\alpha, \beta$	35163.9	35008.1	34785.2
Zero-inflated Beta Negative Binomial(ZIBNB)	34935.1	34819.8	34878.1

TABLE XI: BIC USING DIFFERENT LINK FUNCTIONS FOR  $\phi$ .

Distribution	logit	c-log-log	probit
Hurdle Poisson(HP)	35151.2	53381.7	42512.9
Hurdle Negative Binomial(HNB)	34956.5	53251.6	42464.1
Hurdle Beta Binomial(HBB)	35084.2	53379.3	42591.8
Hurdle Beta Negative Binomial(HBNB)	35239.4	53408.5	42598.2
Zero-inflated Poisson(ZIP)	34815.1	34810.4	34816.3
Zero-inflated Negative Binomial, r is fixed(ZINB)	34804.8	34930.22	34802.89
Zero-inflated Negative Binomial(ZINB)	34855.1	34826.3	<b>34791.04</b>
Zero-inflated Beta Binomial(ZIBB)	34875.6	34938.4	34867.56
Zero-inflated Beta Binomial(ZIBB), n is fixed	34967.6	34954.7	34987.8
Zero-inflated Beta Binomial(ZIBB), fixed $\alpha, \beta$	35255.2	35141.5	34876.4
Zero-inflated Beta Negative Binomial(ZIBNB)	34887.1	34974.8	34908.6

## CHAPTER 5

### CONCLUSION

In this thesis, we develop a statistical procedure for identifying the most appropriate probabilistic models for discrete sparse data. Our procedure is based on the modified KS test for discrete distributions with unknown parameters. We developed a general approach for estimating the parameters for a large class of zero-inflated models and Hurdle models, such as ZIP, ZINB, ZIBB, ZIBNB, PH, NBH, BBH, and BNBH. We also proposed a general likelihood ratio test based on Neyman-Pearson lemma for choosing the best model from multiple candidate ones. Based on the a real dataset from a metagenomics experiment, we found out that zero-inflated Beta-Binomial, zero-inflated Beta Negative Binomial, and the corresponding Hurdle models (i.e. ZIBB, ZIBNB, BBH, BNBH) are more appropriate for modeling the sparse omics data comparing to the commonly used Poisson and Negative Binomial in the literature.

We develop a new regression model approach for modeling Hurdle models and zero-inflated models; called Hurdle Regression Model (HRM), and Zero-inflated Regression Models (ZIRM). Our approach assumes the existence of link functions  $g$  and  $h_1, \dots, h_b$  such that  $g(\phi_i) = \mathbf{G}_i^T \boldsymbol{\gamma}$  and  $h_j(\theta_{ij}) = \mathbf{B}_{ij}^T \boldsymbol{\beta}_j$ ,  $i = 1, \dots, n; j = 1, \dots, b$ , where  $\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_b$  are regression coefficients,  $\mathbf{G}_i = (r_1(\mathbf{x}_i), \dots, r_s(\mathbf{x}_i))^T \in \mathbb{R}^s$  and  $\mathbf{B}_{ij} = (q_{j1}(\mathbf{x}_i), \dots, q_{jt_j}(\mathbf{x}_i))^T \in \mathbb{R}^{t_j}$  are the corresponding predictors,  $r_i$ 's and  $q_{ji}$ 's are known functions.

Our ZIRM model is not only cover currently available zero-inflated regression models, such as ZIP, ZINB with fixed  $r$ , ZIBB with constant prior parameters, but also include new regression models, including ZINB with flexible  $r$ , ZIBB with flexible prior parameters, and ZIBNB. We also build the cor-

responding Hurdle Regression Models for zero-altered responses. With the enriched model candidates, we perform model selection based on AIC and BIC criteria. Our application to Insurance Claim Data shows that ZINB with flexible  $r$  is more appropriate than any others.

For general zero-inflated and Hurdle regression models, we derive and simply its general form of Fisher information matrix and then perform significance tests for variable selection. We compare the confidence intervals based on the Fisher information matrix with the ones built by bootstrapping. The results are consistent with each other. Compared with the bootstrapping solutions, the variable selection based on Fisher information matrix is apparently more efficient. Nevertheless, we suggest the use of bootstrapping confidence intervals when the sample size is moderate or small.

In Section 3.4, the definition of Fisher information matrix was described under the regularity conditions, which include exchangeability of the first and second differentiations and the integral sign. The Fisher information matrix under more general conditions is part of our future work.

## CITED LITERATURE

1. Metwally, A. A., Aldirawi, H., and Yang, J.: A review on probabilistic models used in microbiome studies. Communications in Information and Systems, 18(3):173–191, 2018. Copyright ©2018, International Press of Boston.
2. Metwally, A.: Computational Methods for Longitudinal Microbiome Analysis: Identification, Modeling, and Classification. Doctoral dissertation, 2018.
3. Vatanen, T., Kostic, A. D., D’Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T. D., Hämäläinen, A. M., Peet, A., Tillmann, V., Uibo, R., Mokurov, S., Dorshakova, N., Ilonen, J., Virtanen, S. M., Szabo, S. J., Porter, J. A., Lähdesmäki, H., Huttenhower, C., Gevers, D., Cullen, T. W., Knip, M., and Xavier, R. J.: Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. Cell, 165(4):842–853, 2016.
4. Pflughoeft, K. J. and Versalovic, J.: Human Microbiome in Health and Disease. Annual Review of Pathology: Mechanisms of Disease, 7(1):99–122, 2012.
5. Cho, I. and Blaser, M. J.: The human microbiome: At the interface of health and disease. Nature Reviews Genetics, 13(4):260–270, 2012.
6. Rani, A., Ranjan, R., McGee, H. S., Metwally, A., Hajjiri, Z., Brennan, D. C., Finn, P. W., and Perkins, D. L.: A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms. Scientific Reports, 6(September):1–13, 2016.
7. Schott, C., Weigt, S. S., Turturice, B. A., Metwally, A., Belperio, J., Finn, P. W., and Perkins, D. L.: Bronchiolitis obliterans syndrome susceptibility and the pulmonary microbiome. The Journal of Heart and Lung Transplantation, pages 1–10, 2018.
8. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-liggett, C., Knight, R., and Gordon, J. I.: The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature, 449(7164):804–810, 2007.
9. Reiman, D., Metwally, A., and Dai, Y.: Using convolutional neural networks to explore the microbiome. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), (November):4269–4272, 2017.



10. Ditzler, G., Polikar, R., and Rosen, G.: Multi-Layer and Recursive Neural Networks for Metagenomic Classification. IEEE Transactions on Nanobioscience, 14(6):608–616, 2015.
11. Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C., and Knight, R.: Human-associated microbial signatures: Examining their predictive value. Cell Host and Microbe, 10(4):292–296, 2011.
12. Metwally, A. A., Dai, Y., Finn, P. W., and Perkins, D. L.: WEVOTE: Weighted voting taxonomic identification method of microbial sequences. PLoS ONE, 11(9):1–18, 2016.
13. Wood, D. E. and Salzberg, S. L.: Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15(3), 2014.
14. Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N.: MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods, 12(10):902–903, 2015.
15. Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., Strauss, J. F., Jefferson, K. K., and Buck, G. A.: The truth about metagenomics: Quantifying and counteracting bias in 16S rRNA studies Ecological and evolutionary microbiology. BMC Microbiology, 15(1):1–14, 2015.
16. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J.: Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology, 8(NOV):1–6, 2017.
17. Paulson, J. N., Colin Stine, O., Bravo, H. C., and Pop, M.: Differential abundance analysis for microbial marker-gene surveys. Nature Methods, 10(12):1200–1202, 2013.
18. Love, M. I., Huber, W., and Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12):1–21, 2014.
19. Robinson, M. D., McCarthy, D. J., and Smyth, G. K.: edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, 2009.
20. Hu, M.-C., Pavlicova, M., and Nunes, E. V.: Zero-inflated and hurdle models of count data with extra zeros: examples from an hiv-risk reduction intervention trial. The American journal of drug and alcohol abuse, 37(5):367–375, 2011.
21. Cameron, A. C.: Regression analysis of count data.. Cambridge university press, 2013.

22. Hu, M. C., Pavlicova, M., and Nunes, E. V.: Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. American Journal of Drug and Alcohol Abuse, 37(5):367–375, 2011.
23. Yang, S., Harlow, L. L., Puggioni, G., and Redding, C. A.: A comparison of different methods of zero-inflated data analysis and an application in health surveys. Journal of Modern Applied Statistical Methods, 16(1):518–543, 2017.
24. McCullagh, P. and Nelder, J.: Generalized Linear Models. Chapman and Hall/CRC, 2 edition, 1989.
25. Xu, L., Paterson, A. D., Turpin, W., and Xu, W.: Assessment and selection of competing models for zero-inflated microbiome data. PloS one, 10(7), 2015.
26. Fletcher, D., MacKenzie, D., and Villouta, E.: Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. Environmental and ecological statistics, 12(1):45–54, 2005.
27. Welsh, A. H., Cunningham, R. B., Donnelly, C., and Lindenmayer, D. B.: Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological Modelling, 88(1-3):297–308, 1996.
28. Cohen Jr, A. C.: Estimation of the poisson parameter from truncated samples and from censored samples. Journal of the American Statistical Association, 49(265):158–168, 1954.
29. Heilbron, D.: Generalized linear models for altered zero probabilities and overdispersion in count data. Unpublished Technical report, University of California, San Francisco, Department of Epidemiology and Biostatistics, 1989.
30. Heilbron, D. C.: Zero-altered and other regression models for count data with added zeros. Biometrical Journal, 36(5):531–547, 1994.
31. Mullahy, J.: Specification and testing of some modified count data models. Journal of econometrics, 33(3):341–365, 1986.
32. Lambert, D.: Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics, 34(1):1–14, 1992.
33. Mamun, M. A. A.: Zero-inflated regression models for count data: an application to under-5 deaths. 2014.

34. Greene, W. H.: Accounting for excess zeros and sample selection in poisson and negative binomial regression models. 1994.
35. Majo, M. C. and van Soest, A.: The fixed-effects zero-inflated poisson model with an application to health care utilization. 2011.
36. Gilles, R. and Kim, S.: Distribution-free estimation of zero-inflated models with unobserved heterogeneity. Statistical methods in medical research, 26(3):1532–1542, 2017.
37. Sileshi, G., Hailu, G., and Nyadzi, G. I.: Traditional occupancy–abundance models are inadequate for zero-inflated ecological count data. Ecological Modelling, 220(15):1764–1775, 2009.
38. Chen, P., Liu, Q., and Sun, F.: Bicycle parking security and built environments. Transportation research part D: transport and environment, 62:169–178, 2018.
39. Yee, T.: Vgam: Vector generalized linear and additive models. 2017.
40. Aldirawi, H., Yang, J., and Metwally, A. A.: Identifying appropriate probabilistic models for sparse discrete omics data. In 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pages 1–4. IEEE, 2019. Copyright ©2019, IEEE.
41. Wang, L., Aldirawi, H., and Yang, J.: Identifying zero-inflated distributions with a new r package izid. Communications in Information and Systems, 20(1):23–44, 2020. Copyright ©2020, International Press of Boston.
42. Boucher, J.-P., Denuit, M., and Guillen, M.: Number of accidents or number of claims? an approach with zero-inflated poisson models for panel data. Journal of Risk and Insurance, 76(4):821–846, 2009.
43. Massey Jr, F. J.: The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association, 46(253):68–78, 1951.
44. Cameron, A. C. and Trivedi, P. K.: Regression analysis of count data, volume 53. Cambridge university press, 2013.
45. Pettitt, A. N. and Stephens, M. A.: The kolmogorov-smirnov goodness-of-fit statistic with discrete and grouped data. Technometrics, 19(2):205–210, 1977.
46. Dimitrova, D., Kaishev, V., and Tan, S.: Computing the kolmogorov-smirnov distribution when the underlying cdf is purely discrete, mixed or continuous. 2017.

47. Lilliefors, H. W.: On the kolmogorov-smirnov test for normality with mean and variance unknown. Journal of the American statistical Association, 62(318):399–402, 1967.
48. Lilliefors, H. W.: On the kolmogorov-smirnov test for the exponential distribution with mean unknown. Journal of the American Statistical Association, 64(325):387–389, 1969.
49. Parsons, F. and Wirsching, P.: A kolmogorov-smirnov goodness-of-fit test for the two-parameter weibull distribution when the parameters are estimated from the data. Microelectronics Reliability, 22(2):163–167, 1982.
50. Novack-Gottshall, P. and Wang, S. C.: KScorrect, 2018.
51. Ferguson, T.: A Course in Large Sample Theory. Chapman & Hall/CRC, 1996.
52. Manly, B. F.: Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall/CRC, 2006.
53. Neyman, J. and Pearson, E. S.: IX. on the problem of the most efficient tests of statistical hypotheses. Phil. Trans. R. Soc. Lond. A, 231(694-706):289–337, 1933.
54. Tipton, L., Müller, C. L., Kurtz, Z. D., Huang, L., Kleerup, E., Morris, A., Bonneau, R., and Ghedin, E.: Fungi stabilize connectivity in the lung and skin microbial ecosystems. Microbiome, 6(1):12, 2018.
55. Wu, C.-F. J. et al.: Jackknife, bootstrap and other resampling methods in regression analysis. the Annals of Statistics, 14(4):1261–1295, 1986.
56. Antonio, K., Frees, E., and Valdez, E.: A multilevel analysis of intercompany claim counts. ASTIN Bulletin: The Journal of the IAA, 40(1):151–177, 2010.
57. David, M. and Jemna, D.-V.: Modeling the frequency of auto insurance claims by means of poisson and negative binomial models. Annals of the Alexandru Ioan Cuza University-Economics, 62(2):151–168, 2015.
58. Aitkin, M., Francis, B., and Hinde, J.: Statistical Modelling in GLIM 4, volume 32. Oxford University Press, 2005.
59. Renshaw, A. E.: Modelling the claims process in the presence of covariates. ASTIN Bulletin: The Journal of the IAA, 24(2):265–285, 1994.

60. Bartoszewicz, B.: Modelling the claim count with poisson regression and negative binomial regression. In Innovations in Classification, Data Science, and Information Systems, pages 103–110. Springer, 2005.
61. Lee, A. H., Stevenson, M. R., Wang, K., and Yau, K. K.: Modeling young driver motor vehicle crashes: data with extra zeros. Accident Analysis & Prevention, 34(4):515–521, 2002.
62. Yip, K. and Yau, K.: On modeling claim frequency data in general insurance with extra zeros. Insurance: Mathematics and Economics, 36(2):153–163, 2005.
63. Mouatassim, Y. and Ezzahid, E. H.: Poisson regression and zero-inflated poisson regression: application to private health insurance data. European actuarial journal, 2(2):187–204, 2012.
64. Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2 edition, 2009.



Hani Aldirawi &lt;haldir2@uic.edu&gt;

---

**RE: Fw: CIS: Copyright Permission**

1 message

**Kristin Strobel** <kristin@intlpress.com>

Mon, Jul 27, 2020 at 7:47 AM

To: "haldir2@uic.edu" &lt;haldir2@uic.edu&gt;

Cc: Lixin Qin &lt;lixin@intlpress.com&gt;, Brian Bianchini &lt;ipb-mgmt@intlpress.com&gt;, "Lixin Qin (lixinuiuc@gmail.com)" &lt;lixinuiuc@gmail.com&gt;

Dear Hani,

Thank you for your inquiry. Please feel free to use the content of your papers published in CIS as part of your Ph.D. thesis. We ask only that International Press of Boston be properly credited as a footnote and in your bibliography.

Let us know if you should have any further questions.

Many thanks, and best regards,

Kristin Strobel, Editorial Assistant

Tel: (617) 623.3826

kristin@intlpress.com

International Press of Boston, Inc.

PO Box 502 | 387 Somerville Avenue | **Somerville, MA 02143**

Intlpress.com

**From:** lixin Qin [mailto:[lixinuiuc@gmail.com](mailto:lixinuiuc@gmail.com)]**Sent:** Sunday, July 26, 2020 11:36 AM**To:** Kristin Strobel**Cc:** Lixin Qin; Brian Bianchini**Subject:** Fwd: Fw: CIS: Copyright Permission

Hi Kristin,

Please take care of this permission request.


**RightsLink®**


Home



Help



Email Support



Sign in



Create Account



## Identifying Appropriate Probabilistic Models for Sparse Discrete Omics Data

Conference Proceedings:

2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)

Author: Hani [::Aldirawi::]; Jie Yang; Ahmed A. Metwally

Publisher: IEEE

Date: 19-22 May 2019

Copyright © 2019, IEEE

### Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis online.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

[BACK](#)
[CLOSE WINDOW](#)

<b>Contact Information</b>	Department of MSCS University of Illinois at Chicago 851 S. Morgan Street Chicago, IL 60607, USA	<i>Office:</i> MAGC 3.556 <i>Phone:</i> (708) 439-2269 <i>E-mail:</i> haldir2@uic.edu
<b>Education Background</b>	<b>University of Illinois at Chicago (UIC)</b> PhD candidate in Statistics <ul style="list-style-type: none"> <li>• <i>Dissertation:</i> Model Selection and Regression Analysis for Zero-Inflated Data</li> <li>• <i>Advisor:</i> Dr. Jie Yang</li> </ul>	<b>2015-2020</b>
	<b>University of Texas-Pan American (UTPA)</b> Master's degree in Applied Mathematics	<b>2014-2015</b>
	<b>Islamic University of Gaza (IUG)</b> , Palestine B.Sc. in Mathematics,	<b>2007-2011</b>
<b>Professional Experience</b>	Aug, 2015 - Present    University of Illinois at Chicago Jan, 2014 - July, 2015    University of Texas-Pan American Sep, 2013 - Dec, 2013    Ebn Othaymeen High School Jan, 2012 - Aug, 2013    Islamic University of Gaza	<b>Teaching Assistant</b> <b>Teaching Assistant</b> <b>Math Teacher</b> <b>Teaching Assistant</b>
<b>Awards and Honors</b>	May, 2020    Yeuk-Lam Yau-Leung Memorial Scholarship May, 2019    Statistics Department Travel Award May, 2019    UIC Statistics Research Award March, 2019    Graduate College Travel Award May, 2018    UIC Statistics Consulting Award April, 2018    Teaching Award Oct, 2017    MF Scholarship for Academic Achievements (\$1,000) Sep, 2015    UIC Merit Award for Academic Achievements (\$2,000) 2007-2011    IUG Excellence Awards due to my high GPA	
<b>Published Papers</b>	<ul style="list-style-type: none"> <li>• <b>Aldirawi, H.</b>, Wang, L., and Yang, J. (2020). <i>Identify zero-inflated distribution with a new R package iZID</i>. Communications in Information and Systems. Vol. 20, Number 1, 23-44, DOI: <a href="https://dx.doi.org/10.4310/CIS.2020.v20.n1.a2">https://dx.doi.org/10.4310/CIS.2020.v20.n1.a2</a></li> <li>• Carmean, C.M., Kirkley, A.G., Landeche, M., Ye, H., Chellan, B., <b>Aldirawi, H.</b>, Roberts, A.A., Parsons, P.J. and Sargis, R.M.(2020). <i>Arsenic Exposure Decreases Adiposity During HighFat Feeding</i>. Obesity, 28: 932-941. doi:10.1002/oby.22770 .</li> <li>• <b>Aldirawi, H.</b>, Yang, J., and Metwally, A. (2019). <i>Identifying Appropriate Probabilistic Models for Sparse Discrete Omics Data</i>. IEEE-BHI-2019, pp. 1-4.</li> <li>• Merhi, M., and <b>Aldirawi, H.</b> (2019). <i>Factors Impacting Seniors Usage of Technology</i>. MWAIS 2019 Proceedings. 11.</li> <li>• Ruiz, D., Haro, F., <b>Aldirawi, H.</b>, Dybala, M., Hara, M., Kirkley, A., Regnier, S., and Sargis, R. (2019). <i>Developmental Exposure to the Endocrine Disruptor Tolyfluanid Induces Sexually-Dimorphic Later-Life Metabolic Dysfunction</i>. Reproductive Toxicology, Vol. 89, 74-82.</li> <li>• McCarty, W. P., <b>Aldirawi, H.</b>, Dewald, S., and Palacios, M. (2019). <i>Burnout in Blue: An Analysis of the Extent and Primary Predictors of Burnout Among Law Enforcement Officers in the United States</i>. Police Quarterly, Vol.22, No.3, 278-304.</li> <li>• <b>Aldirawi, H.</b>, Metwally, A. A., and Yang, J. (2018). <i>A review on probabilistic models used in microbiome studies</i>. Communications in Information and Systems, Vol.18, No.3, 173-191.</li> </ul>	
<b>Working Papers</b>	<ul style="list-style-type: none"> <li>• <b>Aldirawi, H.</b>, Metwally, A., and Yang, J. (2020). Model Selection and Regression Analysis for Zero-inflated and Hurdle Data with Unknown Parameters.</li> <li>• Eldeirawi, K., and <b>Aldirawi, H.</b> (2020). Chronic conditions, lifestyle factors, and health screening practices among Arab Americans in Southwest Chicago.</li> </ul>	



<b>Seminar Talks</b>	<ul style="list-style-type: none"> <li>• California State University-San Bernardino, Department of Mathematics, Jan 2020. <i>Probabilistic Models and Regression Analysis for Spars Discrete Data.</i></li> <li>• Western New England University, Department of Mathematics, Dec 2019. <i>Probabilistic Models and Regression Analysis for Spars Discrete Data.</i></li> <li>• Penn State Behrend, Department of Mathematics, Dec 2019. <i>Probabilistic Models and Regression Analysis for Spars Discrete Data.</i></li> <li>• John Carroll University, Department of Mathematics and Computer Science, Dec 2019. <i>Probabilistic Models and Regression Analysis for Spars Discrete Data.</i></li> <li>• University of Wisconsin-Stout, Mathematics, Statistics and Computer Science Department. Nov 2019 <i>Probabilistic Models and Regression Analysis for Spars Discrete Data.</i></li> <li>• University of Illinois at Chicago, Statistics Research Seminar, March 2019. <i>Identifying Appropriate Probabilistic Models for Sparse Discrete Omics Data.</i></li> <li>• University of Illinois at Chicago, Graduate Students Seminar, Nov 2018. <i>A review on probabilistic models used in microbiome studies.</i></li> <li>• University of Illinois at Chicago, Graduate Students Seminar, Jan 2018. <i>Developmental Exposure to the Endocrine Disruptor Tolyfluanid Induces Sexually-Dimorphic Later-Life Metabolic Dysfunction.</i></li> <li>• University of Illinois at Chicago, Graduate Students Seminar, Oct 2017. <i>An Analysis of the Extent and Primary Predictors of Burnout Among Law Enforcement Officers in the United States.</i></li> </ul>	
<b>Conference talks</b>	<ul style="list-style-type: none"> <li>• Joint Statistical Meetings (JSM), Denver, July 2019. <i>Identifying Appropriate Probabilistic Models for Sparse Discrete Omics Data.</i></li> <li>• IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Chicago, May 2019. <i>Identifying Appropriate Probabilistic Models for Sparse Discrete Omics Data.</i></li> <li>• The 14th Annual Conference of the Midwest AIS, (MWAIS), Oshkosh, May 2019. <i>Factors Impacting Seniors Usage of Technology.</i></li> <li>• Eastern North American Region (ENAR) Meeting, Philadelphia, March 2019. <i>A review on probabilistic models used in microbiome studies.</i></li> <li>• ICSA Midwest Chapter meeting, Chicago, Oct 2018. <i>A review on probabilistic models used in microbiome studies.</i></li> <li>• The 6<sup>th</sup> Annual Pan American Collaboration for Ethnicity in the Professions Conference (P.A.C.E), Edinburg, March 2014. <i>Dietary Behaviors That Cause Diabetes and Hypertension.</i></li> </ul>	
<b>Professional Services</b>	<ul style="list-style-type: none"> <li>• Reviewer, IEEE- Biomedical and Health Informatics (BHI) Conference, 2019.</li> <li>• Reviewer, 14th Annual Conference of the Midwest AIS, 2019.</li> <li>• I am a co-chair of Statistics Graduate Student Committee at UIC. I help in running seminars, invite speakers for our weekly statistical seminars.</li> <li>• Mentoring several new Statistics PhD students during their first semester.</li> </ul>	
<b>Courses Taught at UIC</b>	Elements Statistical Methods (As a lecturer) Intermediate Algebra (As a lecturer) Calculus for Business Introduction to Statistics Precalculus Intermediate Algebra	Fall 2018, Spring (2019, 2020) Summer(2016, 2018) Fall(2016, 2017) Spring (2017, 2018) Spring 2016 Fall 2015
<b>Courses Taught at UTPA</b>	College Algebra (As a lecturer) Calculus I (As a lecturer) College Algebra	Spring 2014 Fall 2014 Spring 2013
<b>Courses Taught at IUG(As a TA)</b>	Calculus I & II Probability & Statistics Real Analysis	Fall (2012, 2013) Spring (2012, 2013) Summer 2013

**Technical Skills**

Statistical Software(R, SAS, Minitab)  
Python  
Mathematics Programs(Mathematica, Matlab)  
Office(Word, Excel, Powerpoint)  
SQL  
L<sup>A</sup>T<sub>E</sub>X  
Tableau

**References**

- *Dr. Jie Yang*(Associate Professor at UIC & Advisor)  
Email: jyang06@uic.edu  
Phone: 312-413-3748
- *Dr. Min Yang*(Professor at UIC)  
Email: myang2@uic.edu  
Phone: 312-996-8612
- *Dr. Jing Wang*(Associate Professor at UIC)  
Email: jiwang12@uic.edu  
PPhone: 312-996-4835
- *Dr. Jennifer Pajda-De La O*(Stat Instructor at UIC & Stat 361 Coordinator)  
Email: jpajda2@uic.edu