

**ANALYSIS OF ZERO-HEAVY DATA
USING A MIXTURE MODEL APPROACH**

By

Shin Cheng Wang

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

APPROVED

Eric P. Smith, Chairman

Jesse C. Arnold

Clint W. Coakley

Klaus H. Hinkelmann

Keying Ye

March 18, 1998

Blacksburg, Virginia

Key Words: *Ceriodaphnia Dubia*, Chronic toxicity testing, Generalized Estimating Equations, Inhibition Concentration, Longitudinal Data, Principal Component Analysis, Zero-inflated Poisson

Copyright 1998, Shin Cheng Wang

Analysis of Zero-Heavy Data Using A Mixture Model Approach

by

Shin Cheng Wang

Eric P. Smith, Chair

Department of Statistics
Virginia Polytechnic Institute and State University

(ABSTRACT)

The problem of high proportion of zeroes has long been an interest in data analysis and modeling, however, there are no unique solutions to this problem. The solution to the individual problem really depends on its particular situation and the design of the experiment. For example, different biological, chemical, or physical processes may follow different distributions and behave differently. Different mechanisms may generate the zeroes and require different modeling approaches. So it would be quite impossible and inflexible to come up with a unique or a general solution.

In this dissertation, I focus on cases where zeroes are produced by mechanisms that create distinct sub-populations of zeroes. The dissertation is motivated from problems of chronic toxicity testing which has a data set that contains a high proportion of zeroes. The analysis of chronic test data is complicated because there are two different sources of zeroes: mortality and non-reproduction in the data. So researchers have to separate zeroes from mortality and fecundity. The use of mixture model approach which combines the two mechanisms to model the data here is appropriate because it can incorporate the mortality kind of extra zeroes.

A zero inflated Poisson (ZIP) model is used for modeling the fecundity in *Ceriodaphnia dubia* toxicity test. A generalized estimating equation (GEE) based ZIP model is developed to handle longitudinal data with zeroes due to mortality. A joint estimate of inhibition concentration (IC_x) is also developed as potency estimation based on the mixture model approach. It is found that the ZIP model would perform better than the regular Poisson model if the mortality is high. This kind of toxicity testing also involves longitudinal data where the same subject is measured for a period of seven days. The GEE model allows the flexibility to incorporate the extra zeroes and a correlation structure among the repeated measures.

The problem of zero-heavy data also exists in environmental studies in which the growth or reproduction rates of multi-species are measured. This gives rise to multivariate data. Since the inter-relationships between different species are imbedded in the correlation structure, the study of the information in the correlation of the variables, which is often accessed through principal component analysis, is one of the major interests in multi-variate data. In the case where mortality influences the variables of interests, but mortality is not the subject of interests, the use of the mixture approach can be applied to recover the information of the correlation structure. In order to investigate the effect of zeroes on multi-variate data, simulation studies on principal component analysis are performed. A method that recovers the information of the correlation structure is also presented.

Dedicated to my wife, Julia

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Eric Smith for his patience and guidance. Despite his busy schedule and many responsibilities, he always made time for my research and to offer encouragement. His invaluable knowledge and support were indispensable for this work. I would also like to express my gratitude to Dr. Jesse Arnold, Dr. Clint Coakley, Dr. Klaus Hinkelmann, and Dr. Keying Ye for serving on my committee.

I am also indebted to my parents, Lai Foo Wang and Sue Chin Chung, my brother, Shinho, and my sister, Wen Hsien for their love, encouragement, and support throughout the years. I would also like to thank my best friend, Dr. Ian Yip for his support and faith in me.

Finally, to my lovely wife, Julia, I would like to thank her for her patience, encouragement and sacrifice. Without her support, this dissertation would not have come to existence.

TABLE OF CONTENTS

LIST OF TABLES		ix
LIST OF FIGURES		xii
CHAPTER ONE	INTRODUCTION	1
1.1	Zero-heavy data	1
1.2	True Zeroes, Sampling Zeroes	2
1.3	Zero-creating Mechanisms	2
CHAPTER TWO	BACKGROUND	5
2.1	Mixture Models For Zero Heavy Data	5
2.2	Zero Altered Model	7
	2.2.1 ANOVA Model	7
	2.2.2 Regression Model	10
2.3	Added Zero Model	13
	2.3.1 Zero-inflated Poisson Regression	14
2.4	Summary	16
2.5	Example: Multi-population Lachenbruch Model With Lognormal density	17
CHAPTER THREE	TOXICITY TESTING	23
3.1	Introduction	23
3.2	Current Potency Estimators	24
3.3	Mixture Model	26
3.4	Chronic Toxicity Testing with adjustment of Mortality Effects	27
3.5	E-M algorithm	29
3.6	Examples	33
3.7	Inhibition Concentration and the Confidence Intervals	42

3.8	Simulation Study	45
3.9	Results of Simulation Study	72
3.10	Conclusion	74
CHAPTER FOUR	REPEATED MEASURES IN TOXICITY TESTING	76
4.1	Introduction	76
4.2	Common Approaches to Longitudinal Data	77
4.3	Time-wise Approach	78
4.4	Derived Variables Approach	79
	4.4.1 The Mean Overall and Mean Ignoring Mortality	80
4.5	Generalized Estimating Equations (GEE) Approach	82
4.6	Example for GEE Approach	90
	4.6.1 Inhibition Concentration	95
4.7	MIM revisited	95
4.8	Conclusion	97
CHAPTER FIVE	PRINCIPAL COMPONENT ANALYSIS AND ZERO HEAVY DATA	99
5.1	Introduction	99
5.2	Principal Component Analysis and Factor Analysis	100
5.3	Principal Component Analysis for Zero-Heavy Data	102
5.4	Simulation Study	103
	5.4.1 Simulation I	103
	5.4.2 Results of Simulation I	106
	5.4.3 Simulation II	112
	5.4.4 Results of Simulation II	114
5.5	Conclusion	120
CHAPTER SIX	CONCLUSION	121
6.1	Summary	121
6.2	Further Research	122

BIBLIOGRAPHY		124
APPENDIX A	Koopmans Data	128
APPENDIX B	S-plus Program for ZIP model	129
APPENDIX C	S-plus Program for GEE-ZIP model	133
APPENDIX D	SAS macro for Principal Component Analysis	138
VITA		146

LIST OF TABLES

Table 2.1	Mixture models	7
Table 2.2	The parameters of the Poisson and truncated Poisson distribution	12
Table 2.3	Summary of the reviewed models	17
Table 2.4	Estimated parameters of Koopmans' data	20
Table 2.5	Results of the different hypotheses testing on different combination of parameters based on the likelihood ratio test for the Koopmans' data	22
Table 3.1	The means and variances for the zero mixed Poisson and Poisson distributions.	28
Table 3.2	Summary of Poisson models each with single toxicant	34
Table 3.3	Summary of ZIP regression models each with single toxicant	34
Table 3.4	Estimated coefficients for ZIP and Poisson models	37
Table 3.5	Estimated coefficients for the mortality part of ZIP models	38
Table 3.6.	Estimated coefficients for ZIP and Poison models	39
Table 3.7.	Estimated coefficients for ZIP and Poison models	40
Table 3.8.	Estimated coefficients for the mortality part of ZIP models	40
Table 3.9	Comparison of ZIP and Poisson model	41
Table 3.10	Explanation of the terms in the models	41
Table 3.11A	Estimates and confidence intervals for parameters on inhibition concentrations based on 2000 bootstrap samples -Mercury	44
Table 3.11B	Estimates and confidence intervals for parameters on inhibition concentrations based on 2000 bootstrap samples- Copper	44
Table 3.12	Simulation results	48
Table 3.13	Simulation results	49
Table 3.14	Simulation results	50
Table 3.15	Simulation results	51
Table 3.16	Simulation results	52
Table 3.17	Simulation results	53
Table 3.18	Simulation results	54
Table 3.19	Simulation results	55
Table 3.20	Simulation results	56

Table 3.21	Simulation results	57
Table 3.22	Simulation results	58
Table 3.23	Simulation results	59
Table 3.24	Simulation results	60
Table 3.25	Simulation results	61
Table 3.26	Simulation results	62
Table 3.27	Simulation results	63
Table 3.28	Simulation results	64
Table 3.29	Simulation results	65
Table 3.30	Simulation results	66
Table 3.31	Simulation results	67
Table 3.32	Simulation results	68
Table 3.33	Simulation results	69
Table 3.34	Simulation results	70
Table 3.35	Simulation results	71
Table 4.1	Sample data of <i>Ceriodaphnia</i> test	80
Table 4.2	Common correlation structures	85
Table 4.3	GEE-ZIP model for <i>Ceriodaphnia</i> data (Independent correlation structure)	91
Table 4.4	GEE-ZIP model for <i>Ceriodaphnia</i> data (Compound Symmetric correlation structure)	91
Table 4.5	GEE-ZIP model for <i>Ceriodaphnia</i> data (Autoregressive correlation structure)	92
Table 4.6	The correlation estimates for different models (symmetric)	94
Table 4.7	MIM and GEE-ZIP mean estimates and their variance	97
Table 5.1	Explanation of the notations for simulation I	103
Table 5.2	The three factors in the simulation study	106
Table 5.3	The combinations of the different factors in the study	106
Table 5.4	Eigenvalues for the simulated data (means of 100) in simulation I, n=100, p=5	107
Table 5.5	First 5 eigenvalues for the simulated data (means of 100) in simulation I, n=100, p=10	108
Table 5.6	Eigenvalues for the simulated data (means of 100) in simulation I, n=40, p=5	109
Table 5.7	First 5 eigenvalues for the simulated data (means of 100) in simulation I, n=40, p=10	110

Table 5.8	The percentage of variation explained by the first two eigenvalues in the simulation study I, $n=100$	111
Table 5.9	The percentage of variation explained by the first two eigenvalues in the simulation study I, $n=40$	111
Table 5.10	Explanation of the notations for simulation II	113
Table 5.11	Eigenvalues for the simulated data (means of 100) in simulation I, $n=100$, $p=5$	115
Table 5.12	First 5 eigenvalues for the simulated data (means of 100) in simulation II, $n=100$, $p=10$	116
Table 5.13	Eigenvalues for the simulated data (means of 100) in simulation II, $n=40$, $p=5$	117
Table 5.14	First 5 eigenvalues for the simulated data (means of 100) in simulation II, $n=40$, $p=10$	118
Table 5.15	The percentage of variation explained by the first two eigenvalues in the simulation study II, $n=100$	119
Table 5.16	The percentage of variation explained by the first two eigenvalues in the simulation study II, $n=40$	119

LIST OF FIGURES

Figure 2.1	Boxplots of Koopman Data	19
Figure 3.1	Copper model	35
Figure 3.2	Mercury model	35
Figure 3.3	The mortality probabilities at different values of γ_0 and γ_1 .	45
Figure 3.4	The mean response of the Poisson part of the ZIP model at different values of β_0 and β_1	45
Figure 3.5	Difference in log(likelihood) of ZIP and Poisson model at low mortality	72
Figure 3.6	Difference in log(likelihood) of ZIP and Poisson model at medium mortality	72
Figure 3.7	Difference in log(likelihood) of ZIP and Poisson model at high mortality	73
Figure 3.8	IC ₅₀ of the ZIP and Poisson model at low mortality	73
Figure 3.9	IC ₅₀ of the ZIP and Poisson model at medium mortality	73
Figure 3.10	IC ₅₀ of the ZIP and Poisson model at high mortality	73
Figure 4.1	IC levels of the four single toxicant models with time effect (Day = time after 3)	95
Figure 5.1	10x10 Correlation matrix (symmetric) with block diagonal elements. Relatively lower correlation at off diagonal elements	104
Figure 5.2	10x10 Correlation matrix (symmetric) with block diagonal elements. Relatively higher correlation at off diagonal elements	104
Figure 5.3	5x5 Correlation matrix (symmetric) with block diagonal elements. Relatively lower correlation at off diagonal elements	105
Figure 5.4	5x5 Correlation matrix (symmetric) with block diagonal elements. Relatively higher correlation at off diagonal elements	105