APPLICATION OF FINITE MIXTURE MODELS

FOR VEHICLE CRASH DATA ANALYSIS

A Dissertation

by

BYUNG JUNG PARK

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2010

Major Subject: Civil Engineering

APPLICATION OF FINITE MIXTURE MODELS

FOR VEHICLE CRASH DATA ANALYSIS

A Dissertation

by

BYUNG JUNG PARK

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Dominique Lord
Committee Members,	Yunlong Zhang
	Luca Quadrifoglio
	Jeffrey D. Hart
Head of Department,	John Niedzwecki

May 2010

Major Subject: Civil Engineering

ABSTRACT

Application of Finite Mixture Models for Vehicle Crash Data Analysis. (May 2010)
Byung Jung Park, B.E., Seoul National University;
M.S., Seoul National University
Chair of Advisory Committee: Dr. Dominique Lord

Developing sound or reliable statistical models for analyzing vehicle crashes is very important in highway safety studies. A difficulty arises when crash data exhibit over-dispersion. Over-dispersion caused by unobserved heterogeneity is a serious problem and has been addressed in a variety ways within the negative binomial (NB) modeling framework. However, the true factors that affect heterogeneity are often unknown to researchers, and failure to accommodate such heterogeneity in the model can undermine the validity of the empirical results.

Given the limitations of the NB regression model for addressing over-dispersion of crash data due to heterogeneity, this research examined an alternative model formulation that could be used for capturing heterogeneity through the use of finite mixture regression models. A Finite mixture of Poisson or NB regression models is especially useful when the count data were generated from a heterogeneous population. To evaluate these models, Poisson and NB mixture models were estimated using both simulated and empirical crash datasets, and the results were compared to those from a single NB regression model. For model parameter estimation, a Bayesian approach was adopted, since it provides much richer inference than the maximum likelihood approach.

Using simulated datasets, it was shown that the single NB model is biased if the underlying cause of heterogeneity is due to the existence of multiple counting processes. The implications could be poor prediction performance and poor interpretation. Using two empirical datasets, the results demonstrated that a two-component finite mixture of NB regression models (FMNB-2) was quite enough to characterize the uncertainty about the crash occurrence, and it provided more opportunities for interpretation of the dataset which are not available from the standard NB model. Based on the models from the empirical dataset (i.e., FMNB-2 and NB models), their relative performances were also examined in terms of hotspot identification and accident modification factors. Finally, using a simulation study, bias properties of the posterior summary statistics for dispersion parameters in FMNB-2 model were characterized, and the guidelines on the choice of priors and the summary statistics to use were presented for different sample sizes and sample-mean values.

To my families and lovely Youn-Mi

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and deep appreciation to my advisor, Dr. Dominique Lord, for his constant encouragement and guidance throughout the completion of this dissertation. His deep concern and continued advice, both scholastically and personally, have made this work possible. I also would like to extend my thanks to the committee members, Dr. Yunlong Zhang, Dr. Luca Quadrifoglio and Dr. Jeffrey D. Hart for their time, suggestions and critical reviews of this dissertation. Special thanks are given to Dr. Yunlong Zhang for his hearty support and for giving me many opportunities during my study to help achieve my career objective.

I'm greatly indebted to the Texas Transportation Institute (TTI) for providing funding for the entire period of my PhD study. The research experience at TTI has provided me with invaluable practical experience. Various hands-on experiences with the crash data and the modeling skills learned have been invaluable assets to carrying out my doctoral studies. Especially, I wish to express my deep gratitude to Dr. Kay Fitzpatrick for her continuous support for my research at TTI.

I also wish to thank my colleagues and friends, including Srinivas Geedipally, Fan Ye, and Pei-Fen Kuo, for their friendship, help, support and healthy discussions during my study.

Last but not least, I would like to express my deep gratitude to my parents, mother-in-law, and sisters for their encouragement and patience. Of all people, I am most grateful to my wife, Youn-Mi, who has been always there for me and looked at the bright side of our future.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Highway safety has been a major research topic in transportation studies since highway crashes account for more than 90% of all transportation-related fatalities and cause enormous socio-economic costs. Although much progress has been made in improving highway safety due to various research projects and safety programs, a large number of traffic collisions still occur. According to the *2008 Traffic Safety Fact Sheets* in the United States (NHTSA, 2009), approximately 5,811,000 police-reported traffic crashes in 2008 claimed the lives of nearly 37,261 road users and injured 2,346,000 people. This means that everyday more than 100 people are still killed on the U.S. highway network. The economic cost alone of motor vehicle crashes in 2000 was estimated about $230.6 billion (NHTSA, 2009).

A motor vehicle crash is generally a consequence of three major elements: driver errors, vehicle characteristics, and road environment. However, from a perspective of highway safety designers or engineers, it is more important to provide safer roadway environment to reduce the number of fatalities and injuries because a highway designed with explicit attention to safety can also mitigate the consequences of driver errors. Recognizing the importance of providing safer roadway environment, highway authorities have established the Highway Safety Improvement Program (HSIP) which is characterized by detecting hazardous locations (hotspot identification), diagnosing problems, and providing remedies. Especially, in regard to identifying hotspots, statistical models play an important role since crash frequency or rate at a specific location is a random variable whose underlying mean value is not known. Therefore, identifying locations based on

This dissertation follows the style of Accident Analysis and Prevention.

crash history is subject to uncertainty and the statistical techniques should be utilized for estimating the expected number of crashes at a location.

On the other hand, during the recent few years, increased emphasis has been placed on improving the explicit role of highway safety in making decisions on transportation planning, design, and operations. This can be achieved by quantifying the safety effects of geometric design elements for various transportation facilities, and incorporating the safety information in the planning and design stages of the project development process (Bonneson et al., 2007). To evaluate safety in a quantitative way, it is vital to identify the relationships between safety and various highway geometric design elements for a highway facility. These geometric design elements may include roadway cross-section (lane width and shoulder width), horizontal alignment (degree of curvature or radius, superelevation, and spiral transition curves), vertical alignment (grade, length of grade, and lengths of crest vertical curves), median width, roadside (clear zone width and sideslopes), etc (ASHTO, 2001). Statistical models can provide information about the impacts of these variables on safety. For example, the forthcoming *Highway Safety Manual (HSM)* uses the concept of accident modification factors derived from the statistical modeling results to predict the safety performance for various highway facilities before they are open to traffic.

As described above, statistical models play a key role in various highway safety analyses. Therefore, developing sound or reliable statistical models for analyzing motor vehicle crashes is very important. The primary goal of this research is to introduce a new type of statistical model for analyzing vehicle crashes as extensions to the traditional models. The remainder of this chapter consists of three sections. Section 1.1 provides the problem statement. In Section 1.2, specific objectives of this research are provided. The outline of the dissertation is presented in Section 1.3.

1.1 Problem Statement

From a statistical point of view, we treat highway crashes as random events by assuming that there is an underlying mean crash rate for each individual roadway segment or intersection. Although a number of statistical models have been used to estimate crash frequencies or rates at a specific location over a given interval of time, the primary assumption is that crash counts follow the Poisson probability law. What makes the analysis difficult in modeling crash data is that this kind of data mostly exhibits "over-dispersion". Over-dispersion results when the variability accounted for by the homogeneous Poisson process is not sufficient. In this case, we say that the data are over-dispersed. There are reasonable explanations of over-dispersion in crash data, which will be seen shortly in Chapter II.

To correct for the extra Poisson variation, highway safety analysts used negative binomial (NB) regression models since the NB models can effectively approximate the underlying crash process by introducing a probabilistic error term related to the mean of the Poisson variable. Within the NB regression modeling framework, many studies have focused on the structure of the dispersion parameter (ϕ, or its inverse $\alpha = 1/\phi$) of the NB distribution. Instead of using a fixed (or common) dispersion parameter, some researchers suggested to use a varying dispersion parameter in which the dispersion parameter is modeled as a function of the covariates (Heydecker and Wu, 2001; Hauer, 2001; Miaou and Lord, 2003). Despite the considerable efforts put in place to improve the performance of the NB models, several studies have documented important limitations associated with these models in a cross-sectional data analysis. First, as described above, the development of NB models using a fixed versus a varying dispersion parameter is still an on-going issue. Varying dispersion parameter models may be preferred because one can determine sources influencing over-dispersion (Hilbe, 2007). However, finding appropriate covariates that influence the over-dispersion can be problematic if it is partly caused by unobserved variables or conditions. If the fixed dispersion model is preferred in terms of parameter parsimony, it may not tell us about the nature of the over-dispersion in the data. It only confirms that evidence of over-dispersion has been found and that this has been taken into account in the NB model (Land et al., 1996). Second, a particular distribution (i.e., gamma distribution) assumed in a probabilistic error term related to the mean of the Poisson variable would be restrictive in terms of its ability to account for heterogeneity across observations. Furthermore, it is difficult to justify in practice because there is no a priori reason why the empirical frequency of crash data should be well approximated by that particular distribution (Lord et al., 2008). Third, NB models are usually estimated at the aggregate level for a sample, resulting in a common parameter vector and a dispersion parameter for all the cross-sections. This may mask the possibility of heterogeneity in the coefficients of the covariates across the sites. Fourth, some have reported that NB regression models have difficulties handling the heavily over-dispersed data with a very long-tail and relatively high mean value because a negligible probability is usually assigned to high counts (Guo and Trivedi, 2002). Last but not least, if the datasets are characterized by small sample sizes and low mean values, the performance of the NB models can be significantly affected in terms of parameter estimation (Lord, 2006) as well as goodness-of-fit (Maher and Summersgill, 1996; Wood, 2002; Park and Lord, 2008). Especially, many empirical crash data, in addition to over-dispersion, exhibit more zero observations that would be allowed for by a NB regression model.

Given the limitations of the NB regression model for addressing over-dispersion of crash data due to heterogeneity, this research examined an alternative model formulation that could be used for capturing heterogeneity through the use of finite mixture regression models. Modeling based on finite mixture distributions has a long history, and with the advancement of computing power and technology, it has continued to receive increasing attention in many areas, such as biometrics, genetics, medicine, and marketing (Frühwirth-Schnatter, 2006). The finite mixtures of Poisson regression models or NB regression models (abbreviated as FMP and FMNB, respectively, hereafter) are especially useful where count data were drawn from heterogeneous populations. In finite mixture models, it is assumed that the observations of a sample arise from more than two unobserved components with unknown proportions. There are many reasons to expect the existence of different subpopulations since the crash data are generally collected from various geographic, environmental and geometric design contexts over some fixed time periods. In such cases, it may be inappropriate to apply one aggregate NB regression model and the interpretation of the model could be misleading. Therefore, it would be reasonable to hypothesize that the individual crashes on highway entities (intersections, segments, etc.) are generated from a certain number (K) of hidden subgroups, or components that are unknown to the transportation safety analyst. The final outputs of FMNB-K regression models will be the number of components, component proportions, component-specific regression coefficients, and the degree of over-dispersion within each component.

1.2 Research Objectives

The primary goal of this research is to examine the application of finite mixture regression models (both for Poisson mixtures and NB mixtures) for analyzing motor vehicle crashes. To accomplish this goal, following objectives are planned to be addressed in this research.

1. Demonstrate the appropriateness of finite mixture model specification in describing a data generation process using simulated datasets. Three hypothetical examples will be set up to show that the standard NB regression model is biased if the underlying cause of data heterogeneity is due to the existence of multiple counting processes.

2. Apply mixture models to two real-world datasets (one for intersections and the other for roadway segments) to examine whether the suggested model would discern the underlying distinctions in the data if they exist. The results will be compared to those from the standard NB regression model in terms of goodness-of-fit, variance structure and parameter interpretation.

3. Apply the results from the finite mixture model to highway safety analyses and examine the relative performances over the standard NB regression model. Two application areas will be considered: hotspot identification and accident modification factors.

4. Characterize the bias properties of dispersion parameters of the FMNB-2 model through a simulation study and provide the guidelines on sample sizes and sample-mean values. The effect of using different prior distributions for the dispersion parameters will also be investigated.

5. Develop recommendations for implementing the mixture models in highway safety research and propose several extensions that merit further study in the future.

For model estimation, a Bayesian sampling approach will be adopted as a model estimation method since it provides much richer inference than the maximum likelihood approach. However, the maximum likelihood estimates will also be computed where appropriate and those will be compared with the Bayesian counterparts.

1.3 Outline of the Dissertation

The rest of this dissertation is organized as follows:

Chapter II overviews various crash count data models that have been approached for modeling highway safety. These models include crash count models for both over-dispersion and under-dispersion, and several emerging models. Based on the discussion about the shortcomings of some of those models, finite mixture models are introduced.

Chapter III provides basic essentials for a Bayesian estimation method. Then, it presents the methodology for analyzing crash data for both single count regression models and finite mixture models. Several issues about estimating finite mixture models with a

Bayesian approach are discussed including the label switch problem and the determination of optimal number of components.

Chapter IV examines the performance of the finite mixture models suing several simulated datasets. Three hypothetical examples are set up with different purposes. The main objective of this chapter is to show the poor prediction and interpretation of the NB regression model if the underlying cause of data heterogeneity is due to the existence of population heterogeneity.

Chapter V applies the finite mixture models to actual vehicle crash data and the results are compared with those from the NB model in various aspects such as goodness-of-fit, variance structure, and parameter interpretation. Analyses are carried out with two empirical crash datasets: one for intersection crash data and the other for segment crash dataset.

Chapter VI deals with the application side of the developed model in Chapter VI in terms of two important highway safety analyses: the identification of hotspots and the development of accident modification factors. The comparison of the results between the proposed model and the NB model is also summarized.

Chapter VII carries out a simulation study to examine the bias properties of the posterior summary statistics (i.e., posterior mean and median) of the dispersion parameters in the FMNB-2 model. Simulations are designed for various sample sizes under three sample-mean values, with two prior specifications for the dispersion parameters. Based on the simulation results, a brief guideline is provided on the choice of priors and the posterior summary statistic to use for different sample sizes and sample-mean values.

Chapter VIII summarizes the major results found in this research along with the general conclusions and future research.

CHAPTER II

BACKGROUND

Over the last few decades, there have been considerable efforts to develop the statistical models for crash data. All these models were directed to address three common properties observed in crash data: non-equal variance (heteroscedasticity), over-dispersion (variance > mean) and excess zeros. Especially, over-dispersion caused by unobserved heterogeneity in crash data is a serious problem and has been addressed in a variety ways. However, the true factors that affect heterogeneity are often unknown to researchers and failure to accommodate such heterogeneity in the model can undermine the validity of the empirical results. Hauer (2001) reported that over-dispersion observed in crash data can be described in terms of "*represented traits*" and "*unrepresented traits*"; the root cause of over-dispersion is that entities with the same represented traits have different means because of the unrepresented traits (measured or unmeasured) not included in the model. On the other hand, Lord et al. (2005) provided a more fundamental definition in which the over-dispersion arises from the actual nature of the crash process. This process dictates that the over-dispersion is the result of Bernoulli trials with unequal probability of independent events (known as Poisson trials) and all distributions, such as the Poisson-gamma (or negative binomial) or Poisson-lognormal, are used as approximation to capture the over-dispersion observed in crash data.

There are numerous ways in which the standard Poisson model may be modified for accommodating over-dispersion. The objective of this chapter is to overview various crash count models that have been approached for modeling highway safety. The chapter is divided into five sections. Section 2.1 provides crash count models for over-dispersion with the Poisson regression model as a starting model and moves onto many other variants which overcome the limitations of the Poisson model. For completeness, Section 2.2 overviews some crash count models which can accommodate under-dispersion although under-dispersion is rarely found in crash data. Section 2.3 provides a brief overview of several emerging models that have been recently proposed in highway safety literature. Section 2.4 provides the structure of finite mixture regression models proposed in this study, and briefly introduces several parameter estimation alternatives and available software. Section 2.5 summarizes the chapter.

2.1 Crash Count Models for Over-dispersion

2.1.1 Poisson regression model

Prior to using Poisson regression models, normal linear regression models with log-transformed scale of crash count data have been sometimes used to address the non-equal variance in crash data. As Miaou and Lum (1993) pointed out, however, underlying distributional assumption (i.e., normal distribution) of log-transformed regression models cannot adequately describe the discreteness and nonnegativity of accident occurrence. In addition, the fitted model provides the mean of the log of crash not the mean of crash itself, i.e., $E[\log(y_i)] \neq \log[E(y_i)]$. Instead, in the Poisson regression model, the conditional distribution is assumed to follow the Poisson distribution which is a probability distribution for non-negative integers. Joshua and Garber (1990) and Miaou et al. (1992) listed the advantages of the use of the Poisson model over these conventional models in terms both of theoretical justification and appropriateness in the model assumptions and of the improvement in the fit (Maher and Summersgill, 1995).

In the basic form of a Poisson model, the number of crashes per year, y_i for a particular site i is assumed to follow a Poisson distribution with the mean crashes per year, λ_i:

$$y_i \mid \lambda_i \sim Poisson\,(\lambda_i) \tag{2.1}$$

with a probability function and the corresponding mean and variance by

$$p(y_i \mid \lambda_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \tag{2.2}$$

$$E(y_i \mid \lambda_i) = Var(y_i \mid \lambda_i) = \lambda_i \tag{2.3}$$

In a regression setting, the expected number of crashes per year, λ_i is conditioned to the explanatory variables $\mathbf{x}_i$ (the traffic flows and geometric characteristics of the site) through a log link function: $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$. The vector $\boldsymbol{\beta}$ contains the parameters which have to be estimated by various estimation methods.

The Poisson regression model respects most of the properties of accident events by (1) providing a model for dealing with heteroscedasticity, and (2) by preserving the original nature (i.e., discreteness and nonnegativity) of crash counts without transforming them into an another scale. In spite of these advantages, a shortcoming of this model is that the variance is restraint to be equal to the mean. In crash data, the variance is usually greater than the mean because the vector of $\mathbf{x}_i$ usually does not explain completely the conditional mean because of omitted variables or randomness. For modeling crash count data with over-dispersion phenomenon the Poisson regression model is not suitable because it fails to capture the extra-variation which exceeds that which would be normally expected by the homogeneous Poisson process (McCullagh and Nelder, 1989). Under over-dispersion, the Poisson regression model still yields consistent parameter estimates, but their standard errors are inconsistent resulting in underestimation (Cameron and Trivedi, 1998). This leads to the invalidation of inference based on the estimated standard errors.

2.1.2 Quasi-Poisson regression

The quasi-Poisson regression model is an alternative way of dealing with over-dispersion in which the dispersion parameter is introduced to relate the mean and the variance from the Poisson regression model and is estimated from the data. This strategy leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion (Cameron and Trivedi, 1998).

A quasi-likelihood is a function of the data that behaves similarly to a likelihood function but can be specified even when the probability distribution of the data is unknown. In this approach, first, the analyst must define how the expected number of y_i depends on the explanatory variables $\mathbf{x}_i$, like $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$. Second, a dispersion parameter, or scale parameter ($\varphi > 0$) is introduced into the relationship between the variance and the mean (i.e., $Var(y_i) = \varphi E(y_i)$) to account for over-dispersion (Wedderburn, 1974; McCullagh and Nelder, 1989). When $\varphi = 1$ the ordinary Poisson model is obtained, and when $\varphi > 1$ we have the over-dispersed Poisson model. The introduction of the dispersion parameter gives a correction factor testing the regression parameter estimates under the Poisson model. That is, the regression parameter estimates resulting from the quasi-Poisson model are identical to those from the standard Poisson model, but their estimated covariance matrix is inflated by the dispersion parameter. McCullagh and Nelder (1989) suggested that the dispersion parameter φ be estimated as ratio of the scaled deviance (SD) or the Pearson Chi-Square (χ^2) to its associated degrees of freedom.

$$\hat{\varphi}_{SD} = \frac{1}{N-p}\sum_{i=1}^{N} 2\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i) \right] \text{, or} \tag{2.4}$$

$$\hat{\varphi}_{\chi^2} = \frac{1}{N-p}\sum_{i=1}^{N} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \tag{2.5}$$

Maher and Summersgill (1996) investigated the reliability of the three estimators of $\hat{\varphi}$ with a simulation study by changing the proportion of data points with low mean value. The estimators included: $SD/(N-p)$, $\chi^2/(N-p)$, and $SD/E(SD)$. They found that

$SD/(N-p)$ behaved very poorly as the proportion of data points with low mean value increases whereas $SD/E(SD)$ performed only a little better. The Pearson estimator $\chi^2/(N-p)$ was the best of the three. However, it also consistently underestimated the assumed dispersion parameter when the proportion of data points with low mean value was above 60% or so.

2.1.3 Negative binomial (NB) regression as a continuous mixture

Since the conventional Poisson model does not provide flexibility to accommodate frequently observed over-dispersion in crash data, several different mixed-Poisson distributions have been applied by assuming a particular distribution in the Poisson mean. In the NB regression model, the conditional mean of y_i is replaced with the random variable as follows:

$$\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta}\cdot\varepsilon_i) = \mu_i \exp(\varepsilon_i) \tag{2.6}$$

where, ε_i is a random error that is assumed to be uncorrelated with $\mathbf{x}_i$. The error ε_i can be thought either as the combined effects of unobserved variables that have been omitted from the model or as another source of pure randomness (Long, 1997). Therefore, even for all sites with the same covariate ($\mathbf{x}$), there is variation in λ_i due to unobserved heterogeneity introduced by ε_i. If ε_i has an arbitrary density $g(\varepsilon_i)$, then the probability function of y_i, $p(y_i)$, can be written as

$$p(y_i) = \int Pois(y_i \mid \lambda_i)\cdot g(\varepsilon_i)d\varepsilon_i \tag{2.7}$$

where $Pois(y_i \mid \lambda_i)$ denotes a Poisson distribution with mean λ_i.

Depending upon the parametric form imposed on $g(\varepsilon_i)$, various mixed-Poisson regression models can be derived (e.g. Poisson-gamma, Poisson-lognormal, Poisson-

Inverse Gaussian, etc.). The negative binomial (NB) regression model arises if one assumes that $g(\varepsilon_i)$, or equivalently the distribution of λ_i, follows a gamma distribution. Specifically, if $\exp(\varepsilon_i)$ follows a $Gamma\,(\phi,\phi)$, the marginal distribution of y_i is a $NB\,(\mu_i,\phi)$. The same $NB\,(\mu_i,\phi)$ can be derived by assuming that the distribution of λ_i follows a $Gamma\,(\phi,\phi/\mu_i)$. The probability mass function for $NB\,(\mu_i,\phi)$ is given by:

$$p(y_i)=\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i} \tag{2.8}$$

The derivation of Equation (2.8) from these two approaches is provided in Appendix A.

The Poisson-gamma distribution is the most common distribution used for modeling crash data because its marginal distribution has a closed form and this mixture results in a conjugate model (Hauer, 1997).[1] The interpretation and derivation of the negative binomial as a Poisson-gamma mixture is well described in Cameron & Trivedi (1998). Despite its reported limitations (see Lord (2006) and references herein), the NB regression model is still very popular, especially since all statistical software programs have built-in functions that can handle such models (Hilbe, 2007).

Unfortunately, the continuous parametric mixing distributions assumed in the Poisson mean rate may pose limitations in fitting the data, especially, with a small sample size or low sample mean value (Lord, 2006, Park and Lord, 2008). Furthermore, the choice of a particular distribution imposes a restrictive functional form between the mean and variance (e.g., quadratic relationship in the NB model), and is difficult to justify in practice because there is no a priori reason why the empirical frequency of crash data should be well approximated by that particular distribution, as discussed above.

[1] For those that use as mixing distribution the log-normal distribution or the inverse-Gaussian distribution, the marginal distribution cannot be expressed in a closed form. One may have to use numerical quadrature or simulated maximum likelihood to estimate the model (Cameron and Trivedi, 1998).

In addition, the continuous mixed Poisson models usually estimate a common parameter vector ($\boldsymbol{\beta}$) and inverse dispersion parameter (ϕ) for all the cross-sections (Ramawamy et al., 1994). In other words, they are estimated at the aggregate level with one standard probability distribution function, which can mask the possibility of heterogeneity in the coefficients of the covariates across the sites. Washington et al. (2003) noted that it could lead to inconsistent and biased parameter estimates when the coefficients actually vary across observations. Since the crash data are generally collected from various geographic, environmental and geometric design contexts, there are many reasons to expect the different effects of each variable on the crash occurrence. To capture unobserved heterogeneity in these parameters, one can first classify the data based on some criteria, and then apply several Poisson or negative binomial regression models at a disaggregate level. However, there may be some arbitrariness involved in the criteria dividing the groups. Gelman et al. (2004, p. 467) warned that this type of crude analysis completely ignores the uncertainty in the dividing indicators and thus can overestimate the differences between each model.

2.1.4 Extensions of NB regression Model

The traditional NB regression model can be extended in several ways depending how we parameterize the over-dispersion parameter ϕ. Up until early 2000s, most of researchers in highway safety have developed predictive models using a fixed or common dispersion parameter model (Hauer, 2001). In 2001, Heydecker and Wu (2001) suggested that ϕ could be modeled as a function of the covariates of the model (which can be defined as the varying dispersion parameter). Hauer (2001) argued that the inverse dispersion parameter should be modeled as a function of segment, $\phi_i = \delta L_i$, to correct for the unequal variance of the NB regression model. Since then, other researchers have investigated various structures of the dispersion parameter, both spatially and temporally (Lord and Park, 2008; Miaou and Lord, 2003; Miranda-Moreno et al., 2005; El-Basyouny and Sayed, 2006; Geedipally and Lord, 2008). Very recently, it was found that the structure of the dispersion parameter can greatly depend on how the mean function is modeled (Mitra and Washington, 2007). Models with a well-defined mean function may not have a structured variance. The varying dispersion parameter can be defined as follows:

$$y_i \mid \mu_i, \phi_i \sim NB\,(\mu_i, \phi_i) \text{ , or} \tag{2.9}$$

$$\phi_i = \delta \exp(\mathbf{z}_i \boldsymbol{\gamma}) \;\; (\delta > 0) \tag{2.10}$$

where $\mathbf{z}_i$ is a vector of observable covariates (not necessarily the same as $\mathbf{x}_i$) and $\boldsymbol{\gamma}$ is a vector of associated parameters.

Another extension is done by Greene (2008) which encompasses the two well-known variants of the negative binomial model: NB1 and NB2 models, originally termed by Cameron and Trivedi (1998). The model by Greene (2008) is termed as a NBP model. In the NBP model, ϕ_i is specified as a function of the conditional mean such that:

$$\phi_i = \delta \mu_i^{2-P} \;\; (\delta > 0) \tag{2.11}$$

Under this specification, the NB1 and NB2 models are special cases of $P=1$ and $P=2$. While the conditional mean for the NBP model is still μ_i, the variance structure is as follows:

$$Var(y_i) = \mu_i \left[1 + \frac{1}{\delta} \mu_i^{P-1} \right] \tag{2.12}$$

$$Var(y_i) = \mu_i \left[1 + \frac{1}{\delta} \right] \text{ for } P=1 \text{ (NB1)} \tag{2.13}$$

$$Var(y_i) = \mu_i \left[1 + \frac{\mu_i}{\delta} \right] \text{ for } P=2 \text{ (NB2)} \tag{2.14}$$

The NBP model is more flexible than the NB1 and NB2 models since the variance function is not restricted to a linear or quadratic form. Greene (2008) effectively applied this type of model to a health care study and used the maximum likelihood method for parameter estimation. Considering that no applications of this type of model have yet been found in highway safety literature, it is worthwhile to investigate the performance of this type of model with various crash data in the future.

On the other hand, Shankar et al. (1998) showed that when spatial and temporal effects are not explicitly included in the NB model, the random effect negative binomial model offered advantages. Miaou and Song (2005) showed that the inclusion of a spatial effect (induced by omitted variables) in the NB model could significantly improve the overall goodness-of-fit of the model.

2.1.5 Hurdle and zero-inflated regression models

Many empirical crash data, in addition to over-dispersion, exhibit more zero observations than would be allowed for by the Poisson or NB regression model. In this case the over-dispersion can also arise from the nature of the process generating the zeros. In highway safety literature, in order to accommodate the excess zeros, some researchers have applied the zero-inflated (or zero-altered) regression models (Shankar et al., 1997; Shankar et al., 2003; Lee and Mannering, 2002) and the hurdle regression models (Son et al., 2009).

The hurdle model relaxes the assumption that the zeros and the positive values come from the same data generating process. It partitions the data generating process into two parts. The first part models the probability that the zero value is observed, and the second part models the probability that positive values cross the zero hurdle (or threshold). In principle, the threshold need not be at zero; it could be any value (Cameron and Trivedi, 1997). The probability of zero count is determined by $p(y_i = 0) = p_1(0)$ and the probability of positive values, $p(y_i > 0)$ is determined by the truncated density $p_2(y_i)/(1 - p_2(0))$, which is multiplied by $1 - p_1(0)$ to ensure that probabilities sum to unity. Thus, the general form of a hurdle model can be formulated as follows:

$$
\begin{aligned}
p(y_i) = p_1(0), \qquad & \text{if } y_i = 0 \\
\frac{1-p_1(0)}{1-p_2(0)} p_2(y_i), \quad & \text{if } y_i \geq 1
\end{aligned}
\tag{2.15}
$$

The hurdle regression model for a Poisson or negative binomial can be obtained by specifying $p_1(\cdot)$ and $p_2(\cdot)$ to be a Poisson or negative binomial distribution. For example, in the Poisson hurdle regression model, the following are specified in Equation (2.15).

$$
\begin{aligned}
p_1(0) &= \exp(-\lambda_{1i}) \\
p_2(0) &= \exp(-\lambda_{2i}) \\
p_2(y_i) &= \frac{\exp(-\lambda_{2i})\lambda_{2i}^{y_i}}{y_i!}
\end{aligned}
$$

where, , $\lambda_{1i} = \exp(\mathbf{x}_i\boldsymbol{\beta}_1)$, and $\lambda_{2i} = \exp(\mathbf{x}_i\boldsymbol{\beta}_2)$.

The zero-inflated model is an extension of the hurdle model in which the zero outcomes can arise from one of two processes. The underlying assumption is that zero crash counts are generated by a dual-state process: a perfect state or an imperfect state with a certain mean value. Therefore, zeros may come from both a perfect state and from an imperfect state. For modeling the unobserved state, a binary process is assumed. For site i, if the binary process takes value 0 with probability w_i, then $y_i = 0$. If the binary process takes value 1 with probability $1-w_i$, then y_i takes count values $0, 1, 2, \cdots$ from a count distribution $p_2(\cdot)$. Thus, the probability density function is as follows:

$$p(y_i) = w_i + (1 - w_i) p_2(0), \quad \text{if } y_i = 0 \qquad (2.16)$$
$$(1 - w_i) p_2(y_i), \quad \text{if } y_i \geq 1$$

The probability w_i can either be a constant or often be parameterized as a certain function of the vector of covariates $\mathbf{z}_i$. In order to ensure $0 < w_i < 1$, w_i is determined by a logit or probit model. The zero-inflated regressions of the Poisson (ZIP) or negative binomial (ZINB) can be obtained by specifying $p_2(\cdot)$ to be the Poisson or negative binomial distributions. For example, in the zero-inflated Poisson regression model with a logistic function of w_i, the following are specified in Equation (2.16).

$$p_2(0) = \exp(-\lambda_i)$$
$$p_2(y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$
$$w_i = \frac{\exp(\mathbf{z}_i\boldsymbol{\gamma})}{1 + \exp(\mathbf{z}_i\boldsymbol{\gamma})}$$

where, the vector $\boldsymbol{\gamma}$ contains the parameters associated with the observable covariates $\mathbf{z}_i$. The mean for both ZIP and ZINB models is the same as follows (Long, 1997):

$$E(y_i \,|\, \mathbf{x}_i, \mathbf{z}_i) = \lambda_i (1 - w_i) \qquad (2.17)$$

The variances of the ZIP and the ZINB model are as follows, respectively (Long, 1997):

$$Var(y_i \,|\, \mathbf{x}_i, \mathbf{z}_i) = \lambda_i (1 - w_i)(1 + \lambda_i w_i) \qquad (2.18)$$
$$Var(y_i \,|\, \mathbf{x}_i, \mathbf{z}_i) = \lambda_i (1 - w_i)(1 + \lambda_i (w_i + \phi^{-1})) \qquad (2.19)$$

For both cases, $Var(y_i \,|\, \mathbf{x}_i, \mathbf{z}_i)$ is always greater than $E(y_i \,|\, \mathbf{x}_i, \mathbf{z}_i)$ unless w_i is zero in the ZIP model.

Although above-described models have provided an improved fit to data as compared to other count models, they have undergone some criticism in modeling vehicle crash data (Lord et al. 2005, 2007). The underlying assumption may be unrealistic when considering the probabilistic structure of data generating process for vehicle crashes. In the ZIP or ZINB model, it is assumed that there is a group of sites that never experience crashes. This is unrealistic since a roadway segment or an intersection always has a likelihood of having crashes unless there is no traffic on it. It is also worth noting the researches in other areas such as environment and ecology. For example, Warton (2005) argued that many of the uses of excess zero models are probably unnecessary and the negative binomial probability model by itself is sufficient to handle most occurrences of zero-inflation in environmental and ecological data.

As will be seen shortly in Section 2.4, the ZIP and the ZINB model are the special cases of the FMP-2 and the FMNB-2 model where the strict dual-state assumption is relaxed.

2.2 Crash Count Models for Under-dispersion

Models for under-dispersion have generally been neglected in highway safety analysis since crash data rarely exhibit under-dispersion. However, under-dispersed data have been sporadically encountered by highway safety researchers and they have tried to deal with under-dispersion by introducing some special models, yet relatively unknown. Among them are the gamma probability model, the generalized Poisson model, and the Conway-Maxwell-Poisson (COM-Poisson) model.

2.2.1 Gamma probability model

Oh et al. (2006) investigated the performance of the gamma probability model to deal with slight under-dispersion observed in railroad crossing related vehicle crashes in Korea. They concluded that the gamma probability model was the most appropriate statistical model for their dataset among all the models they considered. The gamma probability model for the count data is given by:

$$p(y_i = k) = Gamma\ (\alpha k, \lambda_i) - Gamma\ (\alpha k + \alpha, \lambda_i) \tag{2.20}$$

where, $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$. Similar to the hurdle model, the probability function consists of two-parts: the probability that the zero value is observed and the probability that positive values are observed.

$$\begin{aligned} &Gamma\ (\alpha k, \lambda_i) = 1, && \text{if } k = 0 \\ &Gamma\ (\alpha k, \lambda_i) = \frac{1}{\Gamma(\alpha k)}\int_0^{\lambda_i} u^{\alpha k-1}e^{-u}du\ , && \text{if } k > 1 \end{aligned} \tag{2.21}$$

where α is a dispersion parameter. Depending on the value α, the gamma probability model can be used for analyzing under-dispersed and over-dispersed data: $\alpha > 1$ represents under-dispersion; $\alpha < 1$ represents over-dispersion. When $\alpha = 1$, the gamma probability model reduces to a Poisson model.

The conditional mean function is given by:

$$E(y_i \mid \mathbf{x}_i) = \sum_{k=1}^{\infty} kGamma\ (\alpha k, \lambda_i) \tag{2.22}$$

and the cumulative distribution function is:

$$\begin{aligned} F(T \mid \alpha, \lambda_i) &= \int_0^T \frac{\lambda_i^{\alpha k}}{\Gamma(\alpha k)} u^{\alpha k-1} e^{-\lambda_i u} du \\ &= \frac{1}{\Gamma(\alpha k)}\int_0^{\lambda_i T} u^{\alpha k-1}e^{-u}du \\ &= Gamma\ (\alpha k, \lambda_i T) \end{aligned} \tag{2.23}$$

Although this model can provide the better goodness-of-fit, it begs the same question as in the hurdle model in which the data generation process is split into two processes (zero vs. count).

2.2.2 Generalized Poisson regression model

Although less widely used in highway safety area, the generalized Poisson regression (GPR) model proposed by Famoye (1993) has been used in modeling various data sets that exhibit either over-dispersion or under-dispersion (e.g. see Wang and Famoye, 1997 for under-dispersion). The application to accident data can be found in Famoye et al. (2004). The probability density function is given by

$$p(y_i) = \left(\frac{\lambda_i}{1+\alpha\lambda_i} \right)^{y_i} \frac{(1+\alpha y_i)^{y_i - 1}}{y_i!} \exp\left(-\frac{\mu_i(1+\alpha y_i)}{1+\alpha\lambda_i} \right) \tag{2.24}$$

with mean $E(y_i) = \lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$, and variance $Var(y_i) = \lambda_i(1+\alpha\lambda_i)^2$. The GPR is a natural extension of the Poisson regression model. If $\alpha = 0$, the model reduces to the Poisson regression model. If $\alpha > 0$, the model represents count data with over-dispersion. If $\alpha < 0$, the model represents count data with under-dispersion.

The dispersion parameter α can be estimated along with the regression parameters using the maximum likelihood method (Famoye, 1993). When $\alpha > 0$, Equation (2.24) always sums to 1. However, if $\alpha < 0$, Equation (2.24) gets truncated and it may not sum to 1. Furthermore, when the iterative programming is implemented for parameter estimation, the program should check that when $\alpha < 0$ (under-dispersion), α mush satisfy both $1+\alpha\lambda_i > 0$ and $1+\alpha y_i > 0$ by placing appropriate restrictions. Ismail and Jemain (2007) provided the fitting procedure by showing how the Integrative Weighted Least Square (IWLS) method similar to the negative binomial regression model can be applied to obtain the maximum likelihood estimates for the GPR model.

2.2.3 Conway-Maxwell-Poisson regression model

The Conway-Maxwell-Poisson (COM-Poisson) distribution, originally proposed by Conway and Maxwell (1962), is a two-parameter extension of Poisson, Bernoulli, and geometric distributions. This is an improved alternative by allowing both over- and under-dispersion. The probabilistic and statistical properties of the distribution were derived and summarized by Shmueli et al. (2005). The probability density function of the COM-Poisson distribution is given by

$$p(y_i) = \frac{1}{Z(\lambda,\nu)} \frac{\lambda^{y_i}}{(y_i!)^{\nu}} \tag{2.25}$$

$$Z(\lambda,\nu) = \sum_{j=0}^{\infty} \frac{\lambda^{j}}{(j!)^{\nu}} \tag{2.26}$$

where λ is a centering parameter that is related directly to the mean of the observations and ν is a shape parameter. Depending on the value of ν, the model represents for under-dispersed data ($\nu > 1$), over-dispersed data ($\nu < 1$), and equi-dispersed data ($\nu = 1$), respectively.

Since $Z(\lambda,\nu)$ is not a closed form, the COM-Poisson distribution does not have closed-form expressions for its moments in terms of λ and ν. Shmueli et al. (2005) used an asymptotic expression for Z and derived the mean and variance as follows:

$$E(y_i) \approx \lambda^{1/\nu} + \frac{1}{2\nu} - \frac{1}{2} \tag{2.27}$$

$$Var(y_i) \approx \frac{1}{\nu} \lambda^{1/\nu} \tag{2.28}$$

The basic COM-Poisson distribution above has been extended to a generalized linear model (GLM) framework. Guikema and Coffelt (2008) developed a dual-link GLM based on this distribution, and Lord et al. (2008) applied this model to evaluating vehicle crash data. In this new form, λ is replaced with $\mu = \lambda^{1/\nu}$ representing a clear centering parameter. The pdf, mean and variance are redefined as follows:

$$p(y_i) = \frac{1}{S(\mu,\nu)}\left(\frac{\mu^{y_i}}{y_i!}\right)^{\nu} \tag{2.29}$$

$$S(\lambda,\nu) = \sum_{j=0}^{\infty}\left(\frac{\lambda^j}{j!}\right)^{\nu} \tag{2.30}$$

$$E(y_i) \approx \mu + \frac{1}{2\nu} - \frac{1}{2} \tag{2.31}$$

$$Var(y_i) \approx \mu/\nu \tag{2.32}$$

Under this formulation, the asymptotic approximations of the mean and the variance are especially accurate once $\mu > 10$. The interpretations of ν are the same as before.

In a dual-link COM-Poisson GLM framework, both the mean and the variance depend on covariates as follows, but if a single-link model is desired the second link given by Equation (2.34) can be removed allowing a single ν to be estimated directly.

$$\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta}) \tag{2.33}$$

$$\nu_i = \exp(\mathbf{z}_i\boldsymbol{\gamma}) \tag{2.34}$$

where, $\mathbf{z}_i$ is a vector of covariates associated influencing the variance and $\boldsymbol{\gamma}$ is a vector of corresponding parameters to be estimated. For the parameter estimation, Sellers and Shmueli (2008) developed the code for maximum likelihood estimation, and Guikema and Coffelt (2008) and Lord et al. (2008) used a full Bayesian estimation approach. The model characteristics and the performance with application to vehicle crash data have been investigated and the results are well documented in Geedipally (2008).

2.3 Other Models

In addition to the models overviewed in the previous sections, several advances in the development of statistical models for analyzing vehicle crashes have been made recently. These include (i) the application of the multivariate Poisson-lognormal to model crash count data at different levels of severity (Tunaru, 2002; Park and Lord, 2007); (ii) the application of Beta-binomial model based on the fact that crash data are the product of Bernoulli trials with unequal probability of events (Lord et al., 2005; Tong and Lord, 2007); (iii) the application of neural Bayesian network models and support vector machine models for crash predictions (Xie et al., 2007; Li et al, 2008); (iv) the use of a two-state Markov switching model to analyze crash frequencies by assuming that there are two unobserved states of roadway safety over time (Malyshkina et al., 2009); (v) the application of random-parameters count models which provides a fuller understanding of the factors determining crash frequencies (Anastasopoulos and Mannering, 2009).

The random-parameters models deserve additional comment. In random-parameters models it is assumed that some of all model parameters vary across observations while most of the traditional models constrain the coefficients to be fixed. A normal error term in the coefficients is usually employed to allow them to vary (i.e., $\beta_i = \beta + \delta_i$, where δ_i is a normally distributed term with mean 0 and variance σ^2). Using simulated maximum likelihood estimation, Greene (2007) has developed estimation procedures for incorporating random parameters in Poisson and NB regression models. The finite mixture model which will be described in the following section is different from the random-parameters models in that the parameter heterogeneity is approximated by a finite number of support points and their probability masses without making a distributional assumption on the regression coefficients or mixing variable. The finite mixture model allows the data to determine the true relationships by choosing a finite number of unobserved latent components.

2.4 Finite Mixture Model

The finite mixture model allows for extremely flexible modeling of heterogeneous data because it incorporates a combination of discrete and continuous representation of population heterogeneity. Its flexibility and advantages have been extensively recognized in many different modeling environments (e.g. Ramaswamy et al., 1994; Deb & Trivedi, 1997, Wang et al., 1998; Guo & Trivedi, 2002; Karlis and Rahmouni, 2007, to name a few). These models have also been applied before in the traffic safety context: for instance, see the work of Viallefont et al. (2002). For a comprehensive list of the applications and numerical derivations of finite mixture models, readers are referred to Titterington, et al (1985), McLachlan & Peel (2000) and Frühwirth-Schnatter (2006). Especially the last reference deals with finite mixture models from a Bayesian viewpoint.

2.4.1 Model structure

The random vector $\mathbf{y} = (y_1, y_2, \cdots, y_N)'$ is said to arise from a finite mixture distribution, if the probability density function $p(\mathbf{y})$ of this distribution has the following form:

$$p(\mathbf{y} \mid \boldsymbol{\Theta}) = w_1 f_1(\mathbf{y} \mid \boldsymbol{\theta}_1) + w_2 f_2(\mathbf{y} \mid \boldsymbol{\theta}_2) + \cdots + w_K f_K(\mathbf{y} \mid \boldsymbol{\theta}_K) \qquad (2.35)$$

where, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_k)', \mathbf{w})$ denotes the vector of all parameters, and $\mathbf{w} = (w_1, w_2, \cdots, w_K)'$ is called a weight distribution whose elements are restricted to be positive and sum to unity ($w_k > 0$ and $\sum w_k = 1$). A single density $f_k(\cdot \mid \boldsymbol{\theta}_k)$ is referred to as the component distribution for component k $(k = 1, 2, \cdots, K)$, and K is the number of components. In most applications, it is assumed that all component distributions arise from the same parametric distribution family, $f(\cdot \mid \boldsymbol{\theta}_k)$. In our case, it is a Poisson or a NB distribution.

The raw moments of a finite mixture distribution are quite easily derived (Frühwirth-Schnatter, 2006). The mean and the variance are given, respectively, by

$$\mu = E(\mathbf{y} \mid \mathbf{\Theta}) = \sum_{k=1}^{K} \mu_k w_k \tag{2.36}$$

$$\sigma^2 = Var(\mathbf{y} \mid \mathbf{\Theta}) = \sum_{k=1}^{K} (\mu_k^2 + \sigma_k^2) w_k - \mu^2 \tag{2.37}$$

provided that the component moments $\mu_k = E(\mathbf{y} \mid \boldsymbol{\theta}_k)$ and $\sigma_k^2 = Var(\mathbf{y} \mid \boldsymbol{\theta}_k)$ exist.

Under this formulation, the heterogeneity in the data can be accounted for in two ways. First, it accounts for the population heterogeneity by choosing a finite number of unobserved latent components, each of which may be regarded as a sub-population. This is a discrete representation of heterogeneity in the data since the mean event rate is approximated by a finite number of support points. In this respect, the finite mixture model assumes that there is more than one component in the data set. If such a distinct difference is not observed from modeling using Equation (2.35), in other words, if the probability density function does not take the stated mixture density, the resulting parameter estimates would be very unstable and inaccurate. In such cases, it is possible to choose the traditional regression model (Poisson or NB regression model) that does not account for the heterogeneity due to the existence of different sub-populations. This can be done by setting K=1 in Equation (2.35). Second, depending on the choice of the component distribution, $f(\cdot \mid \boldsymbol{\theta})$, it can also accommodate heterogeneity within each component. For example, for FMP-K and FMNB-K regression models, the heterogeneity within each component is accounted for by including the explanatory variables in the mean event rate function. Using the NB distribution as a component distribution would explain additional over-dispersion within component not captured by those explanatory variables. Thus, the formulation is flexible enough to allow for both between-component and within-component variations. It should be noted that the finite mixture approach does not require any distributional assumptions for the mixing variable.

2.4.2 FMP-K and FMNB-K regression models

The general set-ups for FMP-K and FMNB-K regression models can be extended from the Equation (2.35) and their means and variances are obtained from Equations (2.36) and (2.37). The FMP-K regression model assumes that the marginal distribution of y_i follows a mixture of Poisson distributions,

$$p\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right) = \sum_{k=1}^{K} w_k Pois(\mu_{i,k}) = \sum_{k=1}^{K} w_k \left(\frac{e^{-\mu_{k,i}} (\mu_{i,k})^{y_i}}{y_i!} \right) \tag{2.38}$$

$$E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \mu_{i,k} w_k \tag{2.39}$$

$$Var(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = E(y_i \mid \boldsymbol{\Theta}) + \left(\sum_{k=1}^{K} w_k \mu_{i,k}^2 - E(y_i \mid \boldsymbol{\Theta})^2 \right) \tag{2.40}$$

where $\mu_{i,k} = \exp(\mathbf{x}_i \boldsymbol{\beta}_k)$ and $\boldsymbol{\Theta} = \{(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K)', \mathbf{w}\}$. It can be readily seen that unless the entire component's means are the same $(\mu_{i,1} = \cdots = \mu_{i,K})$, the variance is always greater than the mean.

For the FMNB-K regression model, it is assumed that the marginal distribution of y_i follows a mixture of negative binomial distributions,

$$p\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right) = \sum_{k=1}^{K} w_k NB(\mu_{i,k}, \phi_k) = \sum_{k=1}^{K} w_k \left[\frac{\Gamma(y_i + \phi_k)}{\Gamma(y_i + 1)\Gamma(\phi_k)} \left(\frac{\mu_{i,k}}{\mu_{i,k} + \phi_k} \right)^{y_i} \left(\frac{\phi_k}{\mu_{i,k} + \phi_k} \right)^{\phi_k} \right] \tag{2.41}$$

$$E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \mu_{i,k} w_k \tag{2.42}$$

$$Var(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) + \left(\sum_{k=1}^{K} w_k \mu_{i,k}^2 (1 + 1/\phi_k) - E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta})^2 \right) \tag{2.43}$$

where $\mu_{k,i} = \exp(\mathbf{x}_i \boldsymbol{\beta}_k)$ and $\boldsymbol{\Theta} = \{(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K)', (\phi_1, \cdots, \phi_K)', \mathbf{w}\}$. In this case, even if all the component's means are the same, the variance of y_i is always greater than the mean. When ϕ_k in each component goes to infinity, the FMNB-K model is reduced to the FMP-K model. Thus, the FMNB-K models allow for additional over-dispersion within components not captured by the explanatory variables. If additional heterogeneity is present within components, the Poisson mixture model is misspecified. An implication of such additional heterogeneity is that the standard errors are underestimated (Cameron and Trivedi, 1998).

The FMP-K and FMB-K models can be equivalently formulated in a hierarchical manner using a latent variable z_i representing the allocation of each observation y_i to one of the components. The mixture model can thus be written as:

$$\begin{gathered} p\left(y_i \mid \lambda_{i,k}, z_i = k\right) = Pois(\lambda_i) \\ p\left(z_i = k\right) = w_k \end{gathered} \tag{2.44}$$

For the FMP-K model, $\lambda_{i,k}$ is replaced with $\mu_{i,k}$, and for the FMNB-K model, $\lambda_{i,k}$ is replaced with $\mu_{i,k} \exp(\varepsilon_{i,k})$ with $\varepsilon_{i,k} \sim Gamma\,(\phi_k, \phi_k)$. Note that under this formulation, the posterior probability that the site i belongs to a certain component k is expressed as:

$$\begin{aligned} p\left(z_i = k \mid y_i\right) &= \frac{p\left(y_i \mid \lambda_{i,k}, z_i = k\right) p(z_i = k)}{p(y_i)} \\ &= \frac{p\left(y_i \mid \lambda_{i,k}, z_i = k\right) w_k}{\sum_{k=1}^{K} p\left(y_i \mid \lambda_{i,k}, z_i = k\right) w_k} \end{aligned} \tag{2.45}$$

Therefore, each site can be classified into one of the K components according to the posterior probability of component membership, $p\left(z_i = k \mid y_i\right)$. It is common that each site is assigned to the component with maximum posterior probability. In this case some information loss is involved because of disregarding the ‘fuzziness’ of the classifications (Wedel et al., 1993).

It should be noted that the weight distribution ($\mathbf{w}$) used in both FMP-K and FMNB-K was treated as a constant variable. The constant weight model can be extended to a more generalized model by parameterizing the weight distribution as a function of covariates (Wang, et al., 1998; Frühwirth-Schnatter and Kaufmann, 2006; Grün and Leisch, 2007). This parameterization allows each observation to have a different weight that is dependent on the covariates, similar to the application of the varying dispersion parameter for the standard Negative Binomial model (see, e.g., Miaou and Lord, 2003; Lord and Park, 2008). Unfortunately, the use of varying weight factors was beyond the scope of this study, since the estimation process can be very complex and choosing a suitable link function might be an additional problem (i.e., there are various link functions that can be used to define the varying weights). The majority of applications in the literature have used fixed weights. Furthermore, the varying weight model may not always provide the best modeling result (Frühwirth-Schnatter, 2006).

It is noteworthy that the finite mixture of regression models as defined in Equation (2.38) or (2.41) embrace the zero-inflated Poisson (ZIP) or zero-inflated negative binomial (ZINB) regression models as a special case (Cameron and Trivedi, 1998); see Lord et al. (2005, 2007) for a discussion about their use in highway safety. This can be obtained by setting K=2 and $\mu_{1,i} = 0$ for all i. However, the generalized two-component mixture model does not make this somewhat strict dual-state process assumption and allows mixing with respect to both zeros and positives. The group separation is characterized by low mean with low variance and high mean with high variance. Recently, Malyshkina et al. (2008) demonstrated a superior statistical fit of two-state Markov switching negative binomial models using time series crash data in Indiana interstate highway segments. Therefore, the FMP or FMNB models are expected to improve the goodness-of-fit relative to the conventional one-component NB model even when the sample mean is very low although this still needs to be verified in the future.

2.4.3 Parameter estimation and available software

Frühwirth-Schnatter (2006) lists four parameter estimation methods for the finite mixture model: method of moments; maximum likelihood-based methods; Bayesian method; and distance-based methods. Since the maximum likelihood-based methods are most widely used among others, it is briefly reviewed here and the Bayesian method which was adopted in this study will be fully described in the following Chapter III.

The maximum likelihood (ML) estimates of $\hat{\mathbf{\Theta}}$ in Equation (2.35) are obtained either by directly maximizing the mixture likelihood $L(\mathbf{\Theta} \mid \mathbf{y})$ with respect to $\mathbf{\Theta}$ using some methods such as Newton's method or a gradient method, or by maximizing the likelihood function using an iterative scheme such as the Expectation Maximization (EM) algorithm. The mixture likelihood function in the direct maximization method takes the form:

$$\ln L(\mathbf{\Theta} \mid \mathbf{y}) = \prod_{i=1}^{N} \ln\left[p(\mathbf{y} \mid \mathbf{\Theta})\right] = \prod_{i=1}^{N}\left(\sum_{k=1}^{K} \ln(w_k f(y_i \mid \mathbf{\theta}_k))\right) \tag{2.46}$$

The advantage of the direct maximization method is that the convergence is very quick if it is achieved. However, the convergence greatly depends on the choice of the initial values for the model parameters. It can be implemented using the NLMIXED procedure in SAS (SAS Institute Inc., 2002) by directly defining the mixture likelihood and initial values. The procedure uses a dual quasi-Newton algorithm as a default algorithm. According to our experience it was difficult to get stable estimates as the number of components increased more than two.

An alternative most commonly applied method to find ML estimates is the EM algorithm introduced by Dempster et al. (1977). It is known to be much slower to converge compared to the direct maximization (Brännäs and Rosenqvist, 1994). The EM algorithm is implemented based on the complete-data likelihood function in which the mixture likelihood $\ln L(\mathbf{\Theta} \,|\, \mathbf{y})$ in Equation (2.46) is augmented with latent random variables as follows:

$$\ln L(\mathbf{\Theta} \,|\, \mathbf{y}, \mathbf{Z}) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \ln(w_k f(y_i \,|\, \mathbf{\theta}_k)) \tag{2.47}$$

where $z_{i,k}$ is a binary value (0 or 1) of allocation of each site. Starting from $\hat{\mathbf{\Theta}}^{(0)}$, the EM algorithm iterates between an E-step and an M-step. In the E-step for $r \geq 1$, the following estimate of $z_{i,k}^{(r)}$ is obtained by:

$$\text{E-step:} \qquad \hat{z}_{i,k}^{(r)} = \frac{\hat{w}_k^{(r-1)} f(y_i \,|\, \hat{\mathbf{\theta}}_k^{(r-1)})}{\sum_{k=1}^{K} \hat{w}_k^{(r-1)} f(y_i \,|\, \hat{\mathbf{\theta}}_k^{(r-1)})} \tag{2.48}$$

and based on the estimate $\hat{z}_{i,k}^{(r)}$, all unknown parameters $\hat{\mathbf{\Theta}}^{(r)} = (\hat{\mathbf{\theta}}_1, \hat{\mathbf{\theta}}_2, \cdots, \hat{\mathbf{\theta}}_k)', \hat{\mathbf{w}})$ are obtained in the M-step by maximizing the following:

$$\text{M-step:} \qquad \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{z}_{i,k}^{(r)} \ln(w_k f(y_i \,|\, \mathbf{\theta}_k)) \tag{2.49}$$

Currently the FlexMix package in R (R development Core Team, 2006) provides a general framework for finite mixtures of regression models using the EM algorithm (Grün and Leisch, 2007). It provides the finite mixtures of generalized linear models including mixtures of normal, binomial, gamma, and Poisson regression models, but it does not include support for mixtures of negative binomial regression models.

A Bayesian parameter estimation method relies on the posterior distribution which is a product of the mixture likelihood and the prior. However, since there is no natural conjugate prior is available for the mixture likelihood in the form of Equation (2.46), the resulting posterior distribution does not belong to any tractable distribution family. For this reason, Bayesian estimation of even simple mixture problems proved to be a challenge. Markov chain Monte Carlo (MCMC) techniques and their application to Bayesian estimation of finite mixture models have greatly improved the situation. Like the EM algorithm, Bayesian estimation of finite mixture models is based on an incomplete data problem by introducing the allocations as missing data. A detailed account of data augmentation and Gibbs sampling methods is given in Chapter III. Available software for Bayesian estimation of finite mixture models is limited. Currently a MATLAB software package called *bayesf* (version 2.0) [2] allows performing Bayesian inference for some of the finite mixture and Markov switching models discussed in Frühwirth-Schnatter (2006). Because of lack of software at the time of writing this dissertation, we coded the algorithm for finite mixtures of Poisson and negative binomial regression models using the R program (R Development Core Team, 2006). Several example codes used in this research can be found in Appendix D.

2.5 Chapter Summary

In this chapter various crash count models have been overviewed in terms of both over-dispersion and under-dispersion. Since over-dispersion is more frequent in crash data, the modeling approaches for accommodating over-dispersion were the main focus of this chapter. These approaches can be broadly grouped into three.

The first approach is to use the quasi-Poisson regression in which the mean regression function and the variance function from the Poisson GLM are used but the dispersion parameter is left unrestricted. The dispersion parameter is not assumed to be fixed at one

[2] Available at http://www.ifas.jku.at/e2571/e2626/e2632/index_ger.html (Accessed in Sept. 2009).

but is estimated from the data. The second approach for dealing with over-dispersion is a parametric mixture model in which a continuous parametric form of the mixing distribution is assumed for the Poisson mean rate λ_i. For instance, the NB model can be derived as a continuous mixture of Poisson-gamma distribution. Another approach is a finite mixture model which arises by assuming a discrete random variable for the mixing distribution which overcomes the specific distributional assumptions for the Poisson mean rate. Furthermore, it includes the ZIP and the ZINB regression models as special cases by relaxing the dual-state data generation process.

With the rapid development in statistical modeling for vehicle crash data, producing a good statistical fit to data per se may be no longer a challenge as noted by Miaou and Lord (2003). The bottom line is that we need to develop a logical model which conforms to the crash data generation process. Recall that, in finite mixture models, it is assumed that the observations of a sample arise from more than two unobserved components with unknown proportions. The estimation of a single aggregate regression model across all observations in a sample may be inadequate if the observations arise from a number of unknown components in which the regression coefficients or dispersion parameters differ. In this sense, the assumption underlying the finite mixture model seems to be logical and acceptable. The following Chapter III will provide the methodology for estimating the finite mixture model within a Bayesian framework.

CHAPTER III

METHODOLOGY

This chapter provides the methodology on how to analyze count data within a Bayesian framework. In finite mixture regression models, the component distribution in the mixture can be virtually any count distributions, but the attention will be primarily given to Poisson, negative binomial, and hierarchical Poisson models since these have been the most commonly adopted models in highway safety analyses. The Bayesian approach applied to these single count regression models will be used as a building block for the subsequent analysis of finite mixture regression models.

This chapter consists of seven sections. Section 3.1 provides the advantages of using the Bayesian method along with some practical difficulties with the maximum likelihood method. Section 3.2 reviews the Bayesian concept and describes the fundamental difference from the Frequentist analysis. Since the empirical Bayes (EB) method is widely used in the highway safety analysis, its concept and difference from the full Bayesian method are also provided. Section 3.3 provides the Bayesian method for single (or standard) count data regression models – that is, Poisson, negative binomial, and hierarchical Poisson regression models. In Section 3.4, the methodology for finite mixture regression models is provided by introducing a data augmentation and Gibbs sampling method. When the Bayesian sampling approach is adopted for finite mixture models, a label switching problem can be an issue. Section 3.5 briefly discusses this issue and illustrates how this problem will be addressed throughout this study. Given the difficulty in determining the correct number of components in finite mixture models, Section 3.6 provides alternative methods that this study has adopted. Finally, Section 3.7 summarizes the chapter.

3.1 Maximum Likelihood Method vs. Bayesian Method

Estimating a finite mixture model demands extensive computational work. Traditionally, the expectation-maximization (EM) algorithm has been most commonly applied based on the work of Dempster et al. (1977). They realized that a finite mixture model can always be expressed in terms of an incomplete data problem by introducing the allocations as missing data. While this algorithm is based on a maximum likelihood estimation method and is relatively easy to implement, it has several known drawbacks (Mclachlan and Peel, 2000, Frühwirth-Schnatter, 2006). According to Frühwirth-Schnatter (2006), the practical difficulties with the maximum likelihood estimation of finite mixture models are as follows:

- It is difficult to find a global maximum of the likelihood numerically: the EM algorithm tends to lead to a local maximum and thus a grid of many different starting points is needed for finding a global maximum.
- The algorithm can fail to converge particularly when the sample size is small or the components in a dataset are not well separated.
- Like in any incomplete data problems, it is not straightforward to obtain the standard errors of maximum likelihood estimates of a finite mixture model. Various methods have been suggested how to obtain approximate standard errors from the EM algorithm.
- The sample size has to be very large because the maximum likelihood method is based on the asymptotic theory: the regularity conditions are often violated in cases of small datasets, mixtures with small components weights, and over-fitting mixtures with too many components.

In contrast, a Bayesian approach can provide a smoothing effect on the mixture likelihood function by introducing proper priors and reduce the risk of obtaining spurious modes in cases where the EM algorithm leads to degenerate solutions. A

Bayesian approach is less likely to get stuck in a local optimum of the mixture likelihood function because of a more flexible way of searching the parameter space.[3] Since a full posterior distribution is available, it can also provide much richer inference than the maximum likelihood approach in that it can address the issue of parameter uncertainty via the full posterior distribution. If we adopt a sampling-based method for a parameter estimation method, any summary statistics, such as a posterior mean, median or mode, can be computed for a sample from a posterior distribution and they can be used to describe the posterior distribution. Furthermore, a Bayesian approach does not rely on the asymptotic normality, and yields valid inference in cases where regularity conditions are violated. For a full discussion of comparing various estimation methods for finite mixture models, see Frühwirth-Schnatter (2006, pp. 49-56).

In this respect, this study adopted the Bayesian sampling approach. Following the work of Diebolt and Robert (1994) (data augmentation and Gibbs sampling), Bayesian mixture models can be applied routinely when the number of components is assumed to be known. According to Richardson and Green (1997), Bayesian method is the only sensible way if the number of mixture components is allowed to vary. However, this study does not intend to address the mixture models with varying numbers of components because it is not only computationally very intensive but also it is still an on-going issue in the statistical community: there are some issues with regard to prior selection for the number of component (K) and the sensitivity of its posterior distribution (Mclachlan and Peel, 2000; Aitkin, 2001; Jasra et al., 2005). Instead, to determine the appropriate number of components in the mixture, a series of models with increasing numbers of components are fitted and then the most plausible model is selected by various model selection criteria, such as a Bayes factor via marginal likelihoods or information criteria.

[3] However, this flexibility may cause a so-called "label switching" problem. This issue will be reviewed in Section 3.5.

3.2 Introduction to Bayesian Method

The fundamental difference between Frequentists and Bayesians is the way they view the unknown parameter θ. Frequentists regard θ as a fixed value and estimate the probability distribution of the data (y), $p(y \mid \theta)$, being the result of a sampling event from an unknown but fixed parameter space of θ (Gelman et al., 2004). They estimate values $\hat{\theta}$ through maximization of a likelihood function. The curvature of the likelihood function provides the accuracy of these maximum likelihood estimates. Thus, in the Frequentist method, the uncertainty about parameter estimates is quantified by investigating how such estimates would vary one to the next in a repeated sampling from the same population. In contrast, Bayesians regard the unknown parameter θ as a random variable and are interested in the probability distribution of a model parameter (Gelman et al., 2004). This probability distribution is called a posterior distribution, denoted as $\pi(\theta \mid y, \eta)$ which is a product of a likelihood function $p(y \mid \theta)$ and a prior distribution $\pi(\theta \mid \eta)$, in which η is a low-dimensional hyper-parameter. The hyper-parameters can be assumed either to be known or to be drawn from some second-stage prior. To Bayesians, the prior distribution on θ is important. The uncertainty about parameter estimates is quantified by determining how much prior opinion about parameter values change given the observed data. The posterior distribution takes on the following form (Carlin and Louis, 2000):

$$\pi(\theta \mid y, \eta) = \frac{p(y \mid \theta)\pi(\theta \mid \eta)}{m(y \mid \eta)} \tag{3.1}$$

where $m(y) = \int_{\Theta} p(y \mid \theta)\pi(\theta \mid \eta) d\theta$ does not depend on θ. It is a normalizing constant of integration, also known as a marginal likelihood. The normalizing constant is often difficult to evaluate unless the inside part of the integral $p(y \mid \theta)\pi(\theta \mid \eta)$ is a kernel of a familiar distribution. Therefore, the standard approach is to omit the normalizing constant and write the posterior distribution as follows:

$$\pi(\theta \mid y, \eta) \propto p(y \mid \theta)\pi(\theta \mid \eta) \tag{3.2}$$

As will be seen later in this chapter, modern simulation-based methods do not require an evaluation of the normalizing constant at all.

The algorithms employed are Markov Chain Monte Carlo (MCMC) sampling techniques introduced by Geman and Geman (1984), Tanner and Wong (1987), and Gelfand and Smith (1990). The MCMC sampling methods have their roots in the Metropolis-Hastings (MH) algorithm (Metropolis et al. 1953, Hastings 1970). Before the advent of MCMC sampling methods, Bayesians used various numerical approximation methods (for example, quadrature methods, Taylor series expansions)[4] to summarize a posterior distribution since the posterior distribution can be a rather high dimensional form. The numerical approximation methods relied upon normality assumptions or asymptotic arguments which undermined a key benefit of having a complete posterior distribution.

A sampling-based method (e.g. MCMC) is an alternative to these approximation methods. The basic idea of a sampling method is that we generate a large number of samples from a posterior distribution, $\pi(\theta)$, and then use discrete formulas applied to these samples to summarize the posterior distribution by approximating the integrals necessary to calculate the posterior mean and variance. For example, we can estimate the posterior mean and variance of θ by

$$\hat{E}(\theta) = \int \theta \cdot \pi(\theta)\, d\theta \approx \frac{1}{n}\sum_{i=1}^{n} \theta_i \tag{3.3}$$

$$V\hat{a}r(\theta) = \int \left[\theta - \hat{E}(\theta)\right]^2 \cdot \pi(\theta)\, d\theta \approx \frac{1}{n}\sum_{i=1}^{n} \left[\theta_i - \hat{E}(\theta)\right]^2 \tag{3.4}$$

[4] These methods often require extensive knowledge of advanced numerical methods that non-statisticians generally do not have. This also limited the usefulness of the Bayesian approach.

where θ_i is a draw from the posterior distribution $\pi(\theta)$ and n is a very large number. Various quantiles can also be computed to create a Bayesian credible interval based on the large number of sampled draws. For general introductions to MCMC methods, see Gelman et al., (2004) and Gill (2002).

On the other hand, the empirical Bayes (EB) method has been widely used in the highway safety literature to better estimate the long-term mean of a site. The EB estimates can be used for hotspot identifications and countermeasure analyses via before-and-after study (Hauer, 1997). It proved to be the most consistent and reliable method in identifying hotspots (Cheng and Washington, 2005; Elvik, 2007). The EB approach uses a hierarchy idea. The final stage parameters within a hierarchical model (for example, η above) are estimated from using the observed data, and then the usual Bayesian method is proceeded to estimate the unknown parameters as if the priors were known. This is where the name "empirical Bayes" originated since we are using the observed data to estimate the hyper-parameter η (Carlin and Louis, 2000). If η was known, the posterior inference for θ would be carried out using Equation 3.1. However, since η is unknown, the marginal distribution of all data $m(y \mid \eta)$ is used to compute an estimate $\hat{\eta}$ through the maximum likelihood estimation (MLE) or method of moments (MOM) methods. Therefore, in EB analysis, inference about the parameters is based on the estimated posterior distribution $\pi(\theta \mid y, \hat{\eta})$. In contrast to the full Bayesian method, a significant computational simplification is achieved by replacing the integration in Equation 3.1 by maximization. However, Bayesian "purists" do not like the EB approach since it uses the data twice: that is, the data are first used to estimate the parameters in the hyper-prior distribution and, once these values are determined, the observed crash count is used for making inference about the posterior estimation. (Carlin and Louis, 2000). Moreover, the EB approach does not explicitly account for the uncertainty of associations of covariates and safety since the point estimates of all covariate effects will be assumed to be true without any uncertainty (Miaou and Lord, 2003). Table 3.1

illustrates the above-described differences between the Frequentist, full Bayesian, and EB methods.

Table 3.1 Fundamental differences between Frequentist, Bayesian, and EB methods

Method	Parameter assumption	Functions for inference	Summary statistics
Frequentist	Fixed (θ)	$p(y\mid\theta)$: Likelihood function Estimate $\hat{\theta}$ by MLE or MOM	- $\hat{\theta}_{MLE}$ or $\hat{\theta}_{MOM}$ - Standard error - Confidence interval
Full-Bayesian	Random (θ,η)	$\pi(\theta\mid y,\eta)=\dfrac{p(y\mid\theta)\pi(\theta\mid\eta)}{m(y\mid\eta)}$ Use hyper-prior on η (informative or non-informative)	- Posterior mean, median or mode - Quantiles - Credible interval
Empirical-Bayesian	Fixed (θ,η)	$\pi(\theta\mid y,\hat{\eta})=\dfrac{p(y\mid\theta)\pi(\theta\mid\hat{\eta})}{m(y\mid\hat{\eta})}$ Estimate $\hat{\eta}$ by MLE or MOM	

3.3 Bayesian Analysis of Count Data

3.3.1 Poisson regression model

Consider regression count data ($y_i, \mathbf{x}_i), i=1,\cdots,n$, where y_i is crash frequency and $\mathbf{x}_i$ is a vector of observed explanatory variables which includes one in the first column. The dimension of $\mathbf{x}_i$ is $n\times(p+1)$, where p denotes the number of covariates. In the framework of generalized linear models (GLM), a link function is employed to connect the mean number of crashes with related covariates. Given the linear predictor $\eta_i=\mathbf{x}_i\boldsymbol{\beta}$ with unknown regression parameters, $\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots\beta_p)'$ and using a log-linear link function (i.e., $\log\mu_i=\eta_i$), the Poisson model assumes conditionally independent observations:

$$y_i \mid \mu_i \sim Poisson\,(\mu_i) \tag{3.5}$$

with probability function given by

$$p(y_i \mid \mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \tag{3.6}$$

The mean and variance are $E(y_i \mid \mu_i) = \mu_i$ and $Var(y_i \mid \mu_i) = \mu_i$, respectively.

In a Bayesian approach, priors have to be assigned to all unknown parameters $\boldsymbol{\beta}$. Prior for $\boldsymbol{\beta}$ in the linear predictor η_i is usually assumed to follow the $MVN_{p+1}(\mathbf{b}_0, \mathbf{B}_0)$ distribution, where MVN_{p+1} denotes the multivariate normal distribution with $p+1$dimension, and $\mathbf{b}_0$ and $\mathbf{B}_0$ are prior parameters.[5]

$$\boldsymbol{\beta} \sim MVN_{p+1}(\mathbf{b}_0, \mathbf{B}_0) \tag{3.7}$$

$$\pi(\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0)\right] \tag{3.8}$$

For a non-informative prior specification, we usually assume $\mathbf{b}_0 = (0, \cdots, 0)'$ and a large variance such as $\mathbf{B}_0 = 100\mathbf{I}_{p+1}$, where $\mathbf{I}_{p+1}$ denotes the ($p+1$)-dimensional identity matrix.

Then, the Bayesian inference is based on the posterior distribution of all unknown parameters. The posterior distribution is defined by

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta}) \tag{3.9}$$

[5] An alternative choice is to place independent normal priors on each of the regression parameters, e.g. $\beta_j \sim Norm\,(0, 100)$, $j = 0, 1, \cdots, p$.

where $p(\mathbf{y} \mid \boldsymbol{\beta})$ is the likelihood of the Poisson model, $\pi(\boldsymbol{\beta})$ is defined in Equation (3.8). If the posterior distribution $\pi(\boldsymbol{\beta} \mid \mathbf{y})$ belongs to a kernel of any known parametric distribution, the evaluation of unknown parameters are relatively easy. Otherwise, an MCMC (Markov Chain Monte Carlo) sampling technique can be employed to evaluate the exact posterior distribution via Metropolis-Hastings (MH) algorithms (see Appendix B). In this case, the posterior density $\pi(\boldsymbol{\beta} \mid \mathbf{y})$ is proportional to

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}) \propto \prod_{i=1}^{n} \left[\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \right] \cdot \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0) \right] \tag{3.10}$$

and it is not in a standard form. Thus, a Random-Walk Metropolis algorithm using a normal proposal density may be implemented. This is a special case of MH algorithms. The proposal function is symmetric and the usual choice for the update is the normal distribution (Rossi et al., 2005). The new values for $\boldsymbol{\beta}$ is drawn in an iteration step from the following relationship:

$$\boldsymbol{\beta}_{new} \mid \boldsymbol{\beta}_{old} \sim Norm\,(\boldsymbol{\beta}_{old},\ s_\beta \cdot \Sigma_\beta) \tag{3.11}$$

where, Σ_β are the variances for the increments of $\boldsymbol{\beta}_{old}$ and s_β is a variance inflation factor or a scaling parameter. The efficient work of the Random-Walk Metropolis algorithm greatly depends on the choice of Σ_β and s_β. Chib et al. (1998) suggested that the values for Σ_β can be taken from the asymptotic covariance matrix of $\boldsymbol{\beta}$ which can be obtained during the maximum likelihood estimation procedure. Another factor that should be considered is the scaling parameter s_β. The Random-Walk Metropolis algorithm must be tuned by choosing an appropriate value of this scaling parameter in order to induce good mixing behavior of chains. The statistical literature (Winkelmann, 2008; Rossi et al., 2005) recommends that we do the trial runs first during a burn-in period and then adjust the value to obtain acceptance rates between 40% and 60%.[6]

Finally, the probability of move is determined by the following rule:

$$\alpha(\boldsymbol{\beta}_{old}, \boldsymbol{\beta}_{new} \mid y) = \min\left\{\frac{\pi(\boldsymbol{\beta}_{new} \mid \mathbf{y})}{\pi(\boldsymbol{\beta}_{old} \mid \mathbf{y})}, 1\right\} \tag{3.12}$$

Note that since the proposal density is symmetric in $(\boldsymbol{\beta}_{old}, \boldsymbol{\beta}_{new})$ and hence it cancelled out in the ratio. If $\pi(\boldsymbol{\beta}_{new} \mid \mathbf{y}) > \pi(\boldsymbol{\beta}_{old} \mid \mathbf{y})$, the chain moves to $\boldsymbol{\beta}_{new}$ with probability 1. Otherwise, it moves with probability $0 < \alpha(\boldsymbol{\beta}_{old}, \boldsymbol{\beta}_{new} \mid y) < 1$. If rejected, the chain does not move and keep the old values for the next evaluation. This step ensures that the accepted candidates come from the distribution of interest – that is, the posterior distribution of $\boldsymbol{\beta}$.

3.3.2 Negative binomial regression model

Given the same log-link function in the conditional mean (i.e., $\log \mu_i = \eta_i$) and a fixed dispersion parameter $\phi > 0$, the negative model assumes conditionally independent observations:

$$y_i \mid \mu_i, \phi \sim NB(\mu_i, \phi) \tag{3.13}$$

with the probability function given by

$$p(y_i \mid \mu_i, \phi) = \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i} \left(\frac{\phi}{\mu_i + \phi}\right)^{\phi} \tag{3.14}$$

[6] On the other hand, according to Roberts (1996), about 25% and 45% acceptance rates are generally accepted for Metropolis algorithms.

The mean and the variance are as follows:

$$E(y_i \mid \mu_i, \phi) = \mu_i \tag{3.15}$$

$$Var(y_i \mid \mu_i, \phi) = \mu_i + \frac{\mu_i^2}{\phi} \tag{3.16}$$

The prior distribution for $\boldsymbol{\beta}$ is defined as the same in Poisson regression models (i.e., non-informative multivariate normal distribution). For the dispersion parameter ϕ, we assume a Gamma prior to ensure that ϕ is positive:

$$\phi \sim Gamma\ (a, b) \tag{3.17}$$

$$p(\phi) = \frac{b^a}{\Gamma(a)} \phi^{a-1} \exp(-b\phi) \tag{3.18}$$

It has a mean $E(\phi) = a / b$ and a variance $Var(\phi) = a / b^2$. The prior parameters a and b can be chosen such that the Gamma distribution has a non-informative prior; for example, $a = 0.01$ and $b = 0.01$. The prior parameters can be defined in a further stage of the hierarchy by introducing a hyper-prior. The typical flat Gamma hyper-prior would be to set $a = 1$, and $b \sim Gamma\ (1, 0.01)$. It should be noted that other prior specifications on the NB dispersion parameter are also possible as alternatives to the Gamma distribution. Christiansen and Morris (1997) suggested $p(\phi) \propto a_0 / (a_0 + \phi)^2$, where $a_0 > 0$ as a prior guess for the median of ϕ. Small values of a_0 provide less information. Miranda-Moreno et al. (2009) compared the performance this distribution with other prior distribution. Frühwirth-Schnatter et al. (2009) assumed $p(\phi) \propto 2d\phi / (d + \phi)^3$, having a median of $d(1 + \sqrt{2})$. They used this prior setting for applying the FMNB-K models on fabric fault data (Aitkin, 1996).

After specifying priors on each parameter, the Bayesian inference is based on the posterior distribution of all unknown parameters. That is, the posterior distribution is defined by

$$\pi(\boldsymbol{\beta}, \phi \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}, \phi)\pi(\boldsymbol{\beta}, \phi) \tag{3.19}$$

where $p(\mathbf{y} \mid \boldsymbol{\beta}, \phi)$ is the likelihood of the NB model. If we assume an independent relationship between $\boldsymbol{\beta}$ and ϕ, and then we can consider $\pi(\boldsymbol{\beta}, \phi) = \pi(\boldsymbol{\beta})\pi(\phi)$ and apply separate priors for each as defined in Equations (3.8) and (3.18). In this case, since the posterior distribution $\pi(\boldsymbol{\beta}, \phi \mid \mathbf{y})$ does not belong to the kernel of any known parametric distribution, we explain here the MH algorithm within the Gibbs sampling method. It is also known as a "hybrid" or "Metropolis within Gibbs" method (Rossi et al., 2005). This algorithm uses the Gibbs sampling technique as an outer loop and within each loop the MH algorithm is implemented to draw samples from the full conditionals of parameters given the remaining parameters and the data.[7]

The full conditionals of each parameter for the NB regression model are as follows:

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid \theta_{-\boldsymbol{\beta}}) &\propto p(\mathbf{y} \mid \boldsymbol{\beta}, \phi)\pi(\boldsymbol{\beta}) \\ &\propto \prod_{i=1}^{n}\left[\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\right]\cdot\exp[-0.5\cdot(\boldsymbol{\beta}-\mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta}-\mathbf{b}_0)]\end{aligned} \tag{3.20}$$

$$\begin{aligned}\pi(\phi \mid \theta_{-\phi}) &\propto p(\mathbf{y} \mid \boldsymbol{\beta}, \phi)\pi(\phi) \\ &\propto \prod_{i=1}^{n}\left[\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\right]\cdot\phi^{a-1}\exp(-b\phi)\end{aligned} \tag{3.21}$$

where, θ_{-x} denotes all parameters in the posterior other than x parameter.

[7] The full conditionals mean the conditional distribution of each variable given all else. For example, if we have a joint pdf $p(\theta_1, \theta_2, \cdots, \theta_k)$ of unknown parameters $\theta_1, \theta_2, \cdots, \theta_k$, Gibbs sampling simulates each parameter sequentially from the following k full conditional distributions instead of simulating directly from the joint distribution: $p(\theta_1 \mid \theta_2, \theta_3, \cdots, \theta_k)$, $p(\theta_2 \mid \theta_1, \theta_3, \cdots, \theta_k)$, $\cdots$, $p(\theta_k \mid \theta_1, \theta_2, \cdots, \theta_{k-1})$.

Under this setup, the expressions above do not belong to any standard distribution family, so they have no analytical closed form. Similar to the Poisson regression model, we can use the Random-Walk Metropolis algorithm using a normal proposal density. The new values for $\boldsymbol{\beta}$ and ϕ are drawn in an iteration step from the following relationship:

$$\boldsymbol{\beta}_{new} \mid \boldsymbol{\beta}_{old} \sim Norm\,(\boldsymbol{\beta}_{old},\ s_{\beta} \cdot \Sigma_{\beta}) \tag{3.22}$$

$$\log(\phi_{new}) \mid \log(\phi_{old}) \sim Norm\,(\log(\phi_{old}),\ s_{\phi} \cdot V_{\phi}) \tag{3.23}$$

where, Σ_{β} and V_{ϕ} are the variances for the increments of $\boldsymbol{\beta}_{old}$ and $\log(\phi_{old})$, and s_{β} and s_{ϕ} are the scaling parameters. Taking the logarithm on ϕ is for ensuring that the drawn value for ϕ is positive. For an efficient sampling the values for Σ_{β} and V_{ϕ} can be taken from the asymptotic covariance matrix of $\boldsymbol{\beta}$ and ϕ from the maximum likelihood estimation procedure (Chib et al., 1998).

Finally, the probability of move for $\boldsymbol{\beta}$ and ϕ is determined by the following rule:

$$\alpha(\boldsymbol{\beta}_{old}, \boldsymbol{\beta}_{new} \mid y) = \min\left\{ \frac{\pi(\boldsymbol{\beta}_{new} \mid \theta_{-\boldsymbol{\beta}})}{\pi(\boldsymbol{\beta}_{old} \mid \theta_{-\boldsymbol{\beta}})}, 1 \right\} \tag{3.24}$$

$$\alpha(\phi_{old}, \phi_{new} \mid y) = \min\left\{ \frac{\pi(\phi_{new} \mid \theta_{\theta_{-\phi}})}{\pi(\phi_{old} \mid \theta_{\theta_{-\phi}})}, 1 \right\} \tag{3.25}$$

Again, the scaling parameters s_{β} and s_{ϕ} must be tuned in order to induce good mixing behaviors of each chain and to satisfy the acceptance rate range at the same time.

Above-described procedure can be extended to the Bayesian analysis of the varying dispersion parameter model, in which ϕ_i can be a function of a vector of site attributes $\mathbf{z}_i$. The vector $\mathbf{z}_i$ may or may not include covariates of the vector $\mathbf{x}_i$. In order to prevent a zero or negative values, we can parameterize ϕ_i using an exponential link function, so that

$$\phi_i = \exp(\mathbf{z}_i \boldsymbol{\gamma}) \tag{3.26}$$

where $\boldsymbol{\gamma} = (\gamma_0, \cdots, \gamma_m)'$ is a vector of parameters. Then, analogous to the vector of regression parameters, the non-informative multivariate normal distribution may be specified for $\boldsymbol{\gamma}$.

3.3.3 Hierarchical Poisson regression models

In hierarchical Poisson regression models, the treatment of over-dispersion is made more explicit by introducing the random effects into the Poisson mean (λ_i). Depending on the parametric distribution imposed on the Poisson mean, various mixed Poisson regression models can be derived (e.g. Poisson-Gamma, Poisson-Lognormal, Poisson-Inverse Gaussian, etc.). In this subsection, we present the cases for Poisson-Gamma and Poisson-Lognormal within the Bayesian framework.

The standard two-stage hierarchical Poisson model can be expressed as follows (Carlin and Louis, 2000):

(Likelihood) $\quad y_i \mid \lambda_i \sim Poisson\ (\lambda_i) \qquad (3.27)$

(First-stage) $\quad \lambda_i \mid \eta \sim \pi_\lambda(\eta) \qquad (3.28)$

(Second-stage) $\quad \eta \sim \pi_\eta(\cdot) \qquad (3.29)$

where $\pi_\lambda(\eta)$ is the prior distribution imposed on the Poisson mean λ_i with a prior parameter η, and $\pi_\eta(\cdot)$ is the hyper-prior on η with known hyper-parameters (a, b, for example). The structure of this two-stage Bayesian hierarchical model is depicted in Figure 3.1. It can be readily seen that uncertainty in the parameters is introduced in hierarchical Poisson model by considering a full hyper-prior framework. The posterior distribution can be written as:

$$\pi(\lambda_i \mid y_i) \propto p(y_i \mid \lambda_i)\pi_\lambda(\lambda_i \mid \eta)\pi_\eta(\eta \mid a, b) \tag{3.30}$$

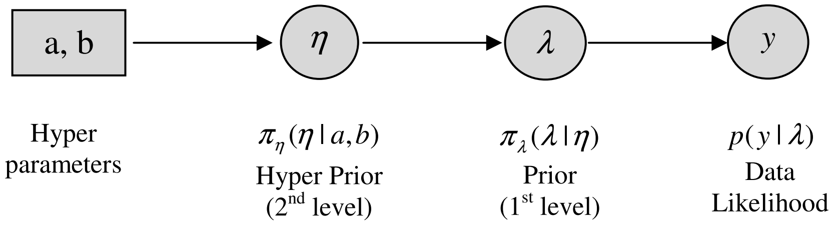


Figure 3.1 Graphical representation of a two-stage Bayesian hierarchical model

We will present two hierarchical Poisson models which have been widely used in highway safety analysis: Poisson-Gamma and Poisson-Lognormal regression models.

Poisson-Gamma Regression Model. The Poisson-Gamma regression model assumes that π_λ is the gamma distribution with two parameters: shape and scale parameters. In Equations (3.28) and (3.29), if we specify $\lambda_i = v_i \mu_i$ (where $\mu_i = e^{\mathbf{x}_i \boldsymbol{\beta}}$) where $v_i \sim Gamma\,(\phi, \phi)$ in the first stage and $\phi \sim Gamma\,(a, b)$ in the second stage, these result in exactly the NB regression model that we described in the previous subsection. This is because integrating out the v_i parameters from $p(y_i \mid \mu_i, v_i)$ results in the following marginal distribution of y_i without depending on v_i, and this leads to the NB regression model (Cameron and Trivedi, 1998; also see Appendix A for derivation):

$$\begin{aligned} m(y_i \mid \mu_i, \phi) &= \int_0^\infty p(y_i \mid \mu_i, v_i)\pi(v_i \mid \phi)dv_i \\ &= \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i} \left(\frac{\phi}{\mu_i + \phi}\right)^{\phi} \end{aligned} \tag{3.31}$$

For the Poisson-Gamma regression model, the posterior distribution is defined by

$$\pi(\boldsymbol{\beta}, \nu_i, \phi \mid y_i) = p(y_i \mid \boldsymbol{\beta}, \nu_i)\pi(\boldsymbol{\beta})\pi(\nu_i \mid \phi)\pi(\phi) \tag{3.32}$$

where the Poisson likelihood $p(y_i \mid \boldsymbol{\beta}, \nu_i)$ and $\pi(\boldsymbol{\beta})$ were previously defined in *Subsection 3.3.1* (see Equation (3.10)). Once we know the full conditionals for each parameter, we can implement the Gibbs sampling by drawing samples of each parameter sequentially. The full conditionals for each parameter are easy to derive and given as follows:

$$\begin{aligned} \pi(\nu_i \mid \theta_{-\nu_i}) &\propto p(y_i \mid \boldsymbol{\beta}, \nu_i)\pi(\nu_i \mid \phi) \\ &\propto \mathit{Gamma}\ (y_i + \phi,\ \mu_i + \phi) \end{aligned} \tag{3.33}$$

$$\pi(\boldsymbol{\beta} \mid \theta_{-\boldsymbol{\beta}}) \propto p(y_i \mid \boldsymbol{\beta}, \nu_i)\pi(\boldsymbol{\beta}) \text{ (given in Equation (3.10))}$$

$$\begin{aligned} \pi(\phi \mid \theta_{-\phi}) &\propto \pi(\nu_i \mid \phi)\pi(\phi \mid a, b) \\ &\propto \frac{\phi^{n\phi+a-1}}{\Gamma(\phi)^n}\left(\prod_{i=1}^{n} \nu_i\right)^{\phi-1} \exp\left\{-\phi\left(b + \sum_{i=1}^{n} \nu_i\right)\right\} \end{aligned} \tag{3.34}$$

The expression for $\boldsymbol{\beta}$ and ϕ have no analytical closed forms, so the MH algorithm with a Random walk proposal can be implemented. Instead of programming the code, the specialized software WinBUGS can also be utilized in which these algorithms are already available (Lunn et al., 2000). The key advantage of WinBUGS is that it derives conditional distributions automatically which significantly simplifies the Gibbs sampling for a variety of models. In WinBUGS, the analysts are only required to write down the probability and specify the priors, which significantly reduce the start-up costs of performing the Bayesian analysis.

Since the posterior of ν_i is a Gamma distribution, the posterior mean of λ_i, $E(\lambda_i \mid y_i)$ given that y_i crashes are observed at site *i*, can be derived:

$$E(\lambda_i \mid y_i) = \mu_i E(\nu_i) \tag{3.35}$$

$$= \mu_i \frac{y_i + \phi}{\mu_i + \phi} \tag{3.36}$$

$$= w_i \mu_i + (1 - w_i) y_i \text{ (where } w_i = \phi / (\phi + \mu_i)\text{)} \tag{3.37}$$

The simple expression of $E(\lambda_i \mid y_i)$ is due to the conjugacy between the gamma and the Poisson distributions. Under this form the posterior mean of crashes at site *i* is a weighted average of conditional mean of crashes (μ_i) and the observed number of actual crashes (y_i). It should be noted that in a full Bayesian approach all the uncertainty about unknown parameters ($\boldsymbol{\beta}$ and ϕ) was taken into consideration by specifying prior distributions on them. In the EB approach, however, $\boldsymbol{\beta}$ and ϕ are estimated from the marginal distribution of y_i (i.e., the NB distribution in Equation (3.31)) using the MLE or MOM methods. Therefore, it is obvious that the EB approach does not allow for any uncertainty in the model parameters and assumes that the mean and the variance of crash frequency at the individual site are estimated without errors, which may be not true in practice (Miaou and Lord, 2003; Lord and Miranda-Moreno, 2008).

Poisson-Lognormal Regression Model. This model has been used effectively for modeling accident frequency especially when the outliers are present, since its tails are known to be asymptotically heavier than those of the Gamma distribution (Kim et al., 2002; Lord and Miranda-Moreno, 2008; El-Basyouny and Sayed, 2009). The only difference between this model and the Poisson-Gamma model lies in the assumption of the distribution of the random effects. This model assumes a log-normal distribution for the i.i.d random effects ν_i in the Poisson mean λ_i. This can be rewritten by adding an additive random effect that is assumed to be normally distributed in the linear link function:

$$y_i \sim Poisson\ (\lambda_i) \tag{3.38}$$

$$\lambda_i = \nu_i \mu_i = \exp(\mathbf{x}_i \boldsymbol{\beta} + \tilde{\nu}_i) \tag{3.39}$$

$$\tilde{\nu}_i \sim Norm\,(0, \tau^2) \tag{3.40}$$

where τ^2 is a hyper-parameter that captures the variance of the random effect. The expression above is equivalent to writing that $\nu_i \sim Lognorm\,(0, \tau^2)$. Under this setup, it is easy to verify that $E(\nu_i) = \exp(\tau^2 / 2)$ and $Var(\nu_i) = \exp(\tau^2)\{\exp(\tau^2) - 1\}$.[8] It became evident that τ^2 indicates the magnitude of the over-dispersion in the data. The lognormal random effects can be comparable to the Gamma random effects by specifying $\tilde{\nu}_i \sim Norm\,(-0.5\tau^2, \tau^2)$ and $\tau^2 = \log(1 + 1/\phi)$, which leads to $E(\nu_i) = 1$ and $Var(\nu_i) = 1/\phi$. In contrast to the Poisson-Gamma regression model, the marginal distribution of the Poisson-Lognormal regression model does not have a closed form (Winkelmann, 2008). According to Hinde (1982), the maximum likelihood estimates can be obtained using a combination of numerical integration, the EM algorithm and iteratively re-weighted least squares. Modern computing power enabled the direct computation by Gauss-Hermite quadrature, and the maximum likelihood estimation of the Poisson-Lognormal models is as fast as estimation of the Poisson-Gamma models (Winkelmann, 2008).

For a hierarchical Bayesian analysis, hyper-prior for variance τ^2 is introduced in a further stage of the hierarchy. A common choice is the Inverse-Gamma distribution which is a conjugate distribution to the normal distribution.[9]

$$\tau^2 \sim IG\,(a, b) \tag{3.41}$$

[8] If $X \sim Lognorm(\mu, \sigma^2)$, then $\log(X) \sim Norm\,(\mu, \sigma^2)$ with $E(X) = \exp(\mu + \sigma^2 / 2)$ and $Var(X) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$.

[9] A random variable X follows an Inverse-Gamma distribution $X \sim IG\,(a, b)$, if 1/X has a Gamma distribution: $1/X \sim Gamma\,(a, b)$. The density is proportional to $(1/X)^{a+1} \exp(-b/X)$.

The hyper-parameters a and b are usually selected as highly dispersed to reflect no prior information on τ^2. Given all prior specifications, the posterior distribution for the Poisson-Lognormal model can be defined as:

$$\pi(\boldsymbol{\beta}, \nu_i, \tau^2 \mid y_i) = p(y_i \mid \boldsymbol{\beta}, \nu_i)\pi(\boldsymbol{\beta})\pi(\nu_i \mid \phi)\pi(\tau^2) \tag{3.42}$$

where the Poisson likelihood $p(y_i \mid \boldsymbol{\beta}, \nu_i)$ and $\pi(\boldsymbol{\beta})$ were previously defined in *Subsection 3.3.1* (see Equation (3.10)). Now we can derive the full conditionals for each parameter from which we draw samples sequentially.

$$\begin{aligned} \pi(\nu_i \mid \theta_{-\nu_i}) &\propto p(y_i \mid \boldsymbol{\beta}, \nu_i)\pi(\nu_i \mid \tau^2) \\ &\propto \nu_i^{\,y_1 - 1} \exp\left\{ -\nu_i \mu_i - \frac{1}{2\tau^2}(\log \nu_i)^2 \right\} \end{aligned} \tag{3.43}$$

$$\pi(\boldsymbol{\beta} \mid \theta_{-\boldsymbol{\beta}}) \propto p(y_i \mid \boldsymbol{\beta}, \nu_i)\pi(\boldsymbol{\beta}) \text{ (given in Equation (3.10))}$$

$$\begin{aligned} \pi(\tau^2 \mid \theta_{-\tau^2}) &\propto \pi(\nu_i \mid \tau^2)\pi(\tau^2 \mid a, b) \\ &\propto IG\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n} (\log \nu_i)^2 \right) \end{aligned} \tag{3.44}$$

In this case, the expression for $\boldsymbol{\beta}$ and ν_i have no analytical closed forms, so the Random-Walk Metropolis algorithm with a normal proposal can be implemented. Again, the specialized software WinBUGS can also be utilized in which the algorithm above is already available.

One particular advantage of the Poisson-Lognormal model is that it is readily extended to the multivariate case (Winkelmann, 2008). In highway safety studies, the application of multivariate Poisson-Lognormal models using Bayesian methods includes Park and Lord (2007) and Ma et al. (2008). They modeled crash counts by severity data based on the multivariate Poisson-Lognormal model by considering the correlations that may exist among different severity levels.

3.4 Estimation of Finite Mixture Regression Models

Having described how we perform the MCMC updates for parameters within a single regression model, we now move onto the estimation of finite mixture of these models. Like the EM algorithm, the Bayesian estimation using MCMC methods for a finite mixture model is based on the incomplete data problem where the allocations are assumed as missing data (Dempster, et al., 1977). In particular, the Gibbs sampling (Gelfand and Smith, 1990) coupled with data augmentation method (Tanner and Wong, 1987) is used to obtain a large number of random variates from the posterior distribution. Tanner and Wong (1987) introduced data augmentation as a method for dealing with missing data or unknown parameter values by augmenting known information with candidate values.

In this section, we will present how to draw samples from the mixture posterior distribution using MCMC techniques coupled with data augmentation. For an illustrative purpose, a negative binomial regression model will be used as a component model – that is, the FMNB-K regression model. The algorithm can be easily modified for other component models.

3.4.1 Complete-data likelihood and posterior distribution

To facilitate the parameter estimation, the data are augmented with a latent random variable $\mathbf{z}_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,K})'$. Here, $\mathbf{z}_i$ can be regarded as an unobserved categorical variable which indicates the component membership of site i. Each element $z_{k,i}$ is defined as follows:

$$z_{i,k} = \begin{cases} 1 & \text{if } y_i \text{ is drawn from the } k^{\text{th}} \text{ component} \\ 0 & \text{otherwise} \end{cases} \tag{3.45}$$

For a particular site i, the unobserved random variables, $\mathbf{z}_i$, are assumed to be independently and identically multinomial distributed with probabilities $\mathbf{w}$ (Diebolt and Robert, 1994), such that:

$$p(\mathbf{z}_i \mid \mathbf{w}) = \prod_{k=1}^{K} w_k^{z_{i,k}} \tag{3.46}$$

where $p(\mathbf{z}_i \mid \mathbf{w})$ is the likelihood of observing the component membership vector $\mathbf{z}_i$ for each site i, given the component proportions $\mathbf{w}$. The complete-data likelihood for site i is then defined as follows:

$$p(y_i, \mathbf{z}_i \mid \mathbf{x}_i, \mathbf{\Theta}) = p(y_i \mid \mathbf{z}_i, \mathbf{x}_i, \mathbf{\Theta})\, p(\mathbf{z}_i \mid \mathbf{x}_i, \mathbf{\Theta}) = \prod_{k=1}^{K} \left[p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_k, \phi_k)\right]^{z_{i,k}} w_k^{z_{i,k}} \tag{3.47}$$

where $p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_k, \phi_k)$ is the likelihood function of the NB model for the k$^{\text{th}}$ component. Therefore, the complete-data likelihood function over all sites N now becomes:

$$\begin{aligned} p(\mathbf{y}, \mathbf{Z} \mid \mathbf{\Theta}, \mathbf{X}) &= \prod_{i=1}^{N} \prod_{k=1}^{K} \left[p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_k, \phi_k)\right]^{z_{i,k}} w_k^{z_{i,k}} \\ &= \prod_{i=1}^{N} \left(\prod_{k=1}^{K} \left[p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_k, \phi_k)\right]^{z_{i,k}} \right) \left(\prod_{k=1}^{K} w_k^{n_k} \right) \end{aligned} \tag{3.48}$$

where $n_k = \sum_{i=1}^{N} z_{i,k}$ denotes the number of observations allocated to component k. The vector of all parameters $\mathbf{\Theta}$ for the FMNB-K model was defined in Equations (2.41-2.43). In practice the component indicators $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_N)$ are unknown but are considered as missing data to be sampled at each iteration during the course of MCMC runs.

By the Bayes' theorem, the complete-data posterior distribution $\pi(\mathbf{Z}, \mathbf{\Theta} | \mathbf{y}, \mathbf{X})$ is proportional to the complete-data likelihood defined in Equation (3.48) times the prior distribution $\pi(\mathbf{\Theta})$ for the parameters, and is given by,

$$\pi(\mathbf{Z}, \mathbf{\Theta}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}, \mathbf{Z} | \mathbf{\Theta}, \mathbf{X}) \pi(\mathbf{\Theta}) \tag{3.49}$$

where the matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)'$ denotes the covariates across all sites.

3.4.2 Prior distributions

The choice of suitable prior distributions and their prior parameters should be done carefully because the priors, especially for the weight distribution, may have significant effects on the posterior distribution (Frühwirth-Schnatter, 2006). For the FMNB model, the unknown parameter is $\mathbf{\Theta} = \{(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K), (\phi_1, \cdots, \phi_K), \mathbf{w}\}$. It is assumed that the parameters $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K$, $\phi_1, \cdots, \phi_K$, and $\mathbf{w}$ are, *a priori*, mutually independent:

$$\pi(\mathbf{\Theta}) = \pi(\boldsymbol{\beta}_1) \cdots \pi(\boldsymbol{\beta}_K) \pi(\phi_1) \cdots \pi(\phi_K) \pi(\mathbf{w}) \tag{3.50}$$

For the finite mixture models, the standard conjugate prior for the weight distribution $\mathbf{w}$ is the $\mathit{Dirichlet}\,(e_1, \cdots, e_K)$ distribution on the simplex $\{(w_1, \cdots, w_K) : w_1 + \cdots + w_K = 1\}$, and the prior parameters are assumed to be the same (i.e. $e_k = e_0$) such that:

$$\pi(\mathbf{w}) = \mathit{Dirichlet}\ (e_0, \cdots, e_0) \propto \prod_{k=1}^{K} {w_k}^{e_0 - 1} \tag{3.51}$$

Note that the Dirichlet distribution is a multivariate generalization of the beta distribution. When $e_0 = 1$, the Dirichlet distribution is the uniform one. This is the usual choice when no information is available for the weights.

The prior distributions for the regression coefficient $\boldsymbol{\beta}_k$ and the dispersion parameter ϕ_k in each component were already defined in Equations (3.8) and (3.18), respectively. They are reproduced here for completeness.

$$\pi(\boldsymbol{\beta}_k) = MVN_{P+1}(\mathbf{b}_0, \mathbf{B}_0) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta}_k - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta}_k - \mathbf{b}_0)\right] \tag{3.52}$$

$$\pi(\phi_k) = Gamma\ (a, b) \propto \phi_k^{a-1} \cdot \exp(-b\phi_k) \tag{3.53}$$

3.4.3 Full conditional distributions

Based on the augmented data likelihood (Equation (3.48)) and the given priors (Equations from (3.51) to (3.53)), the following full conditional distribution for each parameter can be obtained:

$$\pi(\boldsymbol{\beta}_k \mid \boldsymbol{\Theta}_{-\boldsymbol{\beta}_k}, \mathbf{Z}, \mathbf{X}, \mathbf{y}) \propto \prod_{i=1}^{N}\left(\prod_{k=1}^{K}\left[p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_k, \phi_k)\right]^{z_{i,k}}\right) \cdot \exp\left[-\frac{1}{2}(\boldsymbol{\beta}_k - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta}_k - \mathbf{b}_0)\right] \tag{3.54}$$

$$\pi(\phi_k \mid \boldsymbol{\Theta}_{-\phi_k}, \mathbf{Z}, \mathbf{X}, \mathbf{y}) \propto \prod_{i=1}^{N}\left(\prod_{k=1}^{K}\left[p(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta})\right]^{z_{i,k}}\right) \cdot \phi_k^{a-1} \cdot \exp(-b\phi_k) \tag{3.55}$$

$$\pi(\mathbf{w} \mid \mathbf{Z}) = Dirichlet\ (e_0 + n_1, \cdots, e_0 + n_K) \propto \prod_{k=1}^{K} w_k^{e_0 + n_k - 1} \tag{3.56}$$

where $\boldsymbol{\Theta}_{-x}$ denotes all parameters in the posterior other than x.

Then, conditional on knowing all component parameters, $\boldsymbol{\Theta}$, the component indicator vector $\mathbf{z}_i$ allocates the site i into the k$^{\text{th}}$ component by Bayes' rule. The probability of each element of $\mathbf{z}_i$ is calculated by the following equation (Frühwirth-Schnatter, 2006).

$$\Pr(z_{i,k} \mid \boldsymbol{\Theta}, \mathbf{x}_i, y_i) = \frac{p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_k, \phi_k) \cdot w_k}{\sum_{j=1}^{K} p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_j, \phi_j) \cdot w_j} \propto p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_k, \phi_k) \cdot w_k \tag{3.57}$$

Given the probabilities for all components, the conditional distribution of $\mathbf{z}_i$ is a multinomial distribution, satisfying $\sum_{k=1}^{K} z_{i,k} = 1$. That is,

$$\pi(\mathbf{z}_i \mid \mathbf{\Theta}, \mathbf{x}_i, y_i) = Multinom\left(1, \left[\Pr(z_{i,1} \mid \mathbf{\Theta}, \mathbf{x}_i, y_i), \cdots, \Pr(z_{i,K} \mid \mathbf{\Theta}, \mathbf{x}_i, y_i)\right]\right) \quad (3.58)$$

3.4.4 Gibbs sampling algorithm

Now it is straightforward to draw samples from the posterior distribution using the Gibbs sampling method. Conditional on $\mathbf{Z}$ which classifies each observation into a component, the component parameters are drawn sequentially from Equations (3.54) to (3.56). On the other hand, conditional on knowing the component parameters, each component indicator vector $\mathbf{z}_i$ is drawn from the multinomial distribution as defined in Equation (3.58). Since the conditional distributions for $\boldsymbol{\beta}_k$ and ϕ_k do not belong to any standard distribution family, the Random-Walk Metropolis algorithm is utilized within the larger Gibbs sampler as described in *Subsection 3.3.2*.

The MCMC estimation procedure using the Random-Walk algorithm within Gibbs sampling, therefore, can be summarized as follows:

> Start with initial allocations $\mathbf{Z}^{(0)}$, and initial values $\boldsymbol{\beta}_k^{(0)}$ and $\phi_k^{(0)}$. Repeat the following steps for $r = 1, \cdots, R_0, \cdots, R_0 + R$.
>
> *Step 1*: Conditional on the allocations $\mathbf{Z}^{(r-1)}$,
>
> *1-1*. Draw $\mathbf{w}^{(r)}$ from Equation (3.58).
>
> *1-2*. Draw $\boldsymbol{\beta}_1^{(r)}, \cdots, \boldsymbol{\beta}_K^{(r)}$ from Equation (3.54) independently for all k. If accepted by the Random-Walk algorithm, then $\boldsymbol{\beta}_k^{(r)} =$ sampled values, otherwise $\boldsymbol{\beta}_k^{(r)} = \boldsymbol{\beta}_k^{(r-1)}$.

1-3. Draw $\phi_1^{(r)},\cdots,\phi_K^{(r)}$ from Equation (3.55) independently for all k. If accepted by the Random-Walk algorithm, then $\phi_k^{(r)}$= sampled value, otherwise $\phi_k^{(r)}=\phi_k^{(r-1)}$.

1-4. Store the values of all parameters:

$$\mathbf{\Theta}^{(r)}=\{(\boldsymbol{\beta}_1^{(r)},\cdots,\boldsymbol{\beta}_K^{(r)}),(\phi_1^{(r)},\cdots,\phi_K^{(r)}),\mathbf{w}^{(r)}\}.$$

Step 2: Conditional on knowing $\mathbf{\Theta}^{(r)}$,

2-1. Draw $\mathbf{z}_i$ for each observation y_i from Equation (3.58) and store all allocations as $\mathbf{Z}^{(r)}$.

2-2. Increase r by one, and return to *Step 1*.

Step 3: Discard the first R_0 draws as a burn-in period.

After equilibrium is reached at the R_0 th iteration, sampled values are averaged to provide the consistent estimates of the parameters:

$$\hat{E}[h(\theta_k)]=\frac{\sum_{r=R_0+1}^{R}h(\theta_k)^{(r)}}{R} \tag{3.59}$$

where θ_k denotes any interest parameter in the model. For the initial allocations, $\mathbf{Z}^{(0)}$ can be generated from the multinomial distribution with the same weight (i.e. $w_k=1/K$). For $\boldsymbol{\beta}_k^{(0)}$ and $\phi_k^{(0)}$, the maximum likelihood estimates for a single regression model can used. In this research, the Software R (R Development Core Team, 2006) was used for coding the algorithm above. The codes were tested with simulated datasets for FMP-2 and FMNB-2 models. The example codes are provided in Appendix D and are used in Chapter IV.

3.5 Label Switching Problem

In finite mixture regression models, there is a generic identification problem generally known as the “label-switching problem” in the literature of finite mixture models (Redner and Walker, 1984; Celeux et al., 2000; Stephens 2000b; Früwirth-Schnatter, 2001). This is caused by the invariance of the mixture likelihood function under a permutation of the component labels in $\mathbf{\Theta}$. Since the likelihood is the same under relabeling the components of a mixture model, it effectively has K! modes. This problem is viewed as a form of model non-identification. The effect of label switching is very important when the solution is being searched by an iterative method and there is the possibility that the component labels may be switched on different iterations.

Label switching does not create difficulty in maximum likelihood estimation via the EM algorithm, because the goal is to find one of the equivalent modes of the likelihood function (Früwirth-Schnatter, 2006) and solutions converging to different permutations of a single mode are easily identified (Chung, et al., 2004). Although the EM algorithm is free from the label switching problem, its solution can be stuck in the local maximum since the global maximum will actually occur at K! different locations on the mixture likelihood surface.

In the context of Bayesian estimation, label switching is a serious issue because the parameters are estimated by averaging the MCMC output during the simulation run. One solution to this problem may be to put different priors on the component parameters which make the marginal posterior distributions for the parameters be different for each mixture component. However, in practice, it would be difficult to obtain such prior information that allows one to discriminate between the components of a mixture model belonging to the same parametric family. When exchangeable priors are placed on the component parameters, component labels may switch during the simulation run because the Markov chain may visit different one of K! modes. Therefore, without correcting it, it is meaningless to draw inference directly from MCMC output using ergodic averaging

(Jasra et al., 2005). Many techniques have been proposed for overcoming this problem: imposing identifiability constraints on the parameters (Richardson and Green, 1997); clustering methods (Stephens 2000a; Celeux et al., 2000); random permutation sampling (Früwirth-Schnatter, 2001).

Imposing identifiability constrains is a rather simpler method which applies an appropriate constraint to the posterior draws during the MCMC simulation. The constraints usually take the form of ordering the components in terms of their weights ($w_1 > \cdots > w_K$) or their regression parameters (e.g. $\beta_{1,1} > \cdots > \beta_{1,K}$). Whenever a draw does not satisfy the constraint, the component labels are permuted such that the constraint is fulfilled. While this ordering can be easily incorporated into the MCMC simulation, it can be done post-simulations – that is, we can run the MCMC simulation from the unconstrained posterior distribution and then impose an identifiability constraint. This approach has an advantage in that it does not cause any adverse effect on simulation (Jasra et al., 2005). However, according to Celeux (1998) and Celeux et al (2000), this approach does not always work. They suggested that the simulations should be run without any constraints on the parameters and then, at the end of the simulations, a cluster-like method can be applied to change the component labels of the simulation values for $\mathbf{\Theta}$.

Früwirth-Schnatter (2001), acknowledging the difficulty in finding suitable identifiability constraints because of unbalanced labeling switching in the Gibbs sampler, proposed the random permutation sampling method. In this approach, a random permutation of the component labels is performed after each draw to ensure that the samples explore the whole unconstrained parameter space and jump between the various labeling subspaces in a balanced fashion. For example, for K=2, there are only two permutations. That is, with probability 0.5 the draws remain unswitched, whereas with probability 0.5 the labels are interchanged. The MCMC output from this random permutation sampler is explored to find suitable identifiability constraints. Once a suitable identifiability constraint is found, then the MCMC simulation is run again by imposing that constraint on the parameters, in which the draws are permuted if the identifiability constraint is violated. Geweke (2007) also used this concept and showed that widely used MCMC algorithms with data augmentation reliably recover the entire posterior distribution.

In this study, we basically adopted the idea of imposing the identifiability constraints. The ordering constraints are placed on the component regression parameters (e.g. $\beta_{1,1} > \cdots > \beta_{1,K}$) or weight parameters ($w_1 > \cdots > w_K$). In order to find out an appropriate constraint, the simulations are first run without any constraints. If there is a sign of label switching, different order constraints are tested by trial and error to look for the best constraint. It is often obvious from the visual inspection of the MCMC trace plots to tell which constraint is most appropriate. Once a suitable ordering constraint is determined, then the simulation is run again by reordering the MCMC output by means of inequality constraints.

3.6 Determination of Number of Components

When applying mixture regression models to real data, the actual number of components (K) is unknown and must be inferred from the data. The determination of the correct number of components is one of the well-known difficulties in finite mixture models, and it is still a major contemporary issue in mixture modeling.

Within the Bayesian framework, it has been approached in two ways: one is to assume that K is an unknown variable and it is estimated within the modeling process; the other is to fit a series of models with increasing numbers of components, and then select the most plausible model by various model selection criteria. The first approach may appear to be more appealing. The methods in this category include Dirichlet process mixtures (Escobar and West, 1995), distributional distances (Mengersen and Robert, 1996), reversible jump MCMC (Richardson and Green, 1997), and Birth-and-Death MCMC (Stephens, 2000a). However, these methods are not only computationally intensive, but also they have some issues with regards to prior selection for the unknown parameter K and the sensitivity of its posterior distribution (Jasra et al., 2005; Aitkin, 2001). The discussion of all these issues is beyond the scope of this. Instead, we chose the second approach, which is relatively easy to implement and thus widely used.

To determine the best model and the number of components, various model selection criteria were examined: Information-based criteria (AIC, BIC, and DIC) and Bayes factor via marginal likelihoods. The Akaike Information Criterion, or AIC is defined as $-2LL+2p$, where LL is a log-likelihood value and p is the number of parameters in the model. It penalizes the models by the number of parameters included. Smaller values represent better overall fits. The Bayesian information criterion, or BIC is uses a penalty term of $p \cdot \log(n)$, where n is the total number of observation. The BIC is more conservative than the AIC by requiring a greater improvement in fit before it will accept a more complex model. As a rule of thumb, an AIC or BIC difference greater than 10 indicates very strong evidence in favor of the model with lower values (Kass and Raftery, 1995; Burnham and Aderson, 2004). The Deviance information criterion, or DIC is defined as $\hat{D}+2(\overline{D}-\hat{D})$, where $\overline{D}$ is the average of the deviance ($-2LL$) over the posterior distribution, and $\hat{D}$ is the deviance calculated at the posterior mean parameters. As with AIC and BIC, DIC uses $p_D = \overline{D}-\hat{D}$ (effective number of parameters) as a penalty term on the goodness of fit. Differences in DIC from 5-10 indicate that one model is clearly better (Spiegelhalter et al., 2002).

Formal Bayesian model assessment is based on the Bayes factor, B_{12}, for comparing model M_1 to model M_2 after observing the data (Lewis and Raftery, 1997). The Bayes factor is the ratio of the marginal likelihoods of the two models being compared if we assume that the prior probabilities for the two models are equal

($B_{12} = p(\mathbf{y} \mid M_1) / p(\mathbf{y} \mid M_2)$). However, in practice, computing Bayes factors for a particular set of models can be demanding because it requires either complicated multidimensional integrals or some kind of stochastic sampling from the prior distribution. For calculating the marginal likelihood, we adopted the method developed by Lewis and Raftery (1997), who suggested using the posterior simulation output for the computation of the marginal likelihoods (so-called Laplace-Metropolis estimator). The approximation of the marginal likelihood is carried out on the natural logarithmic scale such as:

$$\log\{p(\mathbf{y} \mid M_j)\} \approx \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\{\mid \mathbf{H}^* \mid\} + \log\{f(\mathbf{y} \mid \mathbf{\Theta}^*)\} + \log\{\pi(\mathbf{\Theta}^*)\} \quad (3.60)$$

where d is the number of parameters, $\log\{f(\mathbf{y} \mid \mathbf{\Theta}^*)\}$ is the log-likelihood of data at $\mathbf{\Theta}^*$, and $\log\{\pi(\mathbf{\Theta}^*)\}$ is the log-likelihood of prior distribution at $\mathbf{\Theta}^*$. One way of estimating $\mathbf{\Theta}^*$ is to find the value of $\mathbf{\Theta}$ at which $\log\{f(\mathbf{y} \mid \mathbf{\Theta}^*)\} + \log\{\pi(\mathbf{\Theta}^*)\}$ achieves its maximum from the posterior simulation output. $\mid \mathbf{H}^* \mid$ is the determinant of the variance-covariance matrix estimated from the Hessian at the posterior mode, and it is asymptotically equal to the posterior variance-covariance matrix. This can be estimated from the sample variance-covariance matrix of the posterior simulation output. Assuming that the prior probabilities for the competing models are equal, B_{12} is expressed as follows:

$$\log(B_{12}) = \log\{p(\mathbf{y} \mid M_1)\} - \log\{p(\mathbf{y} \mid M_2)\} \quad (3.61)$$

According to Kass and Raftery (1995), the values of $\log(B_{12})$ between 1 and 3 are positive evidence and the values between 3 and 5 are strong evidence in support of model 1 (see Table 3.2).

The advantage of using information criteria is that they are easy to calculate, and do not depend on prior information except for DIC. However, from a Bayesian perspective, since the posterior model probability should be the tool for model comparison, the information criteria do not have such a formal Bayesian justification in their use for model comparison (Koop, 2003). In this study, the log of marginal likelihood of a model, which is an essential ingredient in calculating the Bayes factor in Equation (3.61), was used as a primary criterion for model selection, whereas the information criteria were used as secondary criteria. Higher value of the log of marginal likelihood of a model is indicative of an improved model.

Table 3.2 Model selection guidelines (Kass and Raftery, 1995)

$2\log\,(B_{12})$	(B_{12})	Evidence against model 2 (M_2)
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

3.7 Chapter Summary

In this chapter, we have provided the fundamental methodology on how to analyze count data within the Bayesian framework for both single count regression models and finite mixture regression models. Prior to describing these count models in details, basic essentials for the Bayesian analysis have also been provided. The fundamental difference between Frequentists and Bayesians was the way they view the unknown parameter θ. The different perspectives resulted in different ways of quantifying the uncertainty about parameter estimates. Frequentists regard θ as a fixed value and the uncertainty about parameter estimates is quantified by the repeated sampling scheme, while Bayesians regard θ as a random variable and the uncertainty about parameter estimates is quantified by determining how much prior opinion about parameter values change in light of the observed data.

There are numerous single count regression models, but in this chapter we have focused on the Bayesian methodology for Poisson, negative binomial, and hierarchical Poisson regression models in that they have been extensively used in highway safety analyses. The degree of complexity for estimating parameters was different from model to model, but the general procedure was similar for all models as follows: i) specify a likelihood function for the data; ii) specify a prior distribution for the model parameters; iii) derive the posterior distribution for the model parameters; iv) simulate the parameter samples from the posterior distribution; v) summarize the parameter samples using basic descriptive statistics. In sampling from a posterior distribution, the key ingredient was the Gibbs sampling technique which draws samples sequentially from full conditional posterior distributions of parameters given the remaining parameters and the data. When the full conditional posterior distribution of a parameter did not belong to the standard distribution, the Random-Walk Metropolis algorithm with a normal proposal could be used.

While any single count regression model can be a component model for a finite mixture regression model, the estimation method was illustrated with the FMNB-K model since the algorithm can be easily adapted for other component models. We showed how the Gibbs sampler coupled with data augmentation could be used to draw samples from the mixture posterior distribution. The algorithm consisted of three steps: first, the data are augmented with a latent random variables, $\mathbf{z}_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,K})'$ which indicates the component membership of site *i*; second, conditional on $\mathbf{z}_i$, the component parameters are drawn sequentially from the full conditional posterior distribution; third, conditional on knowing the component parameters, each component indicator vector $\mathbf{z}_i$ is drawn from a multinomial distribution, satisfying $\sum_{k=1}^{K} z_{i,k} = 1$.

While the Bayesian approach has fewer problems with local optima in the likelihood function and empty components than the likelihood-based approach (EM algorithm) used in the Frequentist approach, it is susceptible to a label switching problem. It is caused by the invariance of a finite mixture model to relabeling the components. Although more rigorous methods can be adopted for correcting it, this study chose a rather ad-hoc approach by imposing a suitable identifiability constraint on the parameters where appropriate.

Finally, to determine the optimal number of components, a series of models is fitted with the fixed number of components and then the best model is selected based on the model selection criteria. For model selection criteria, while information-based criteria are shortcut methods, they are often interpreted as approximations when the Bayesian approach is used and hence do not have a formal Bayesian justification. In this chapter, we showed how the Bayes factor can be obtained between two competing models by calculating the log of marginal likelihood values and it was used as a primary criterion throughout this study along with other information criteria as secondary criteria. The next two chapters apply the methodology described here to simulated datasets (Chapter IV) and empirical crash datasets (Chapter V).

CHAPTER IV

HYPOTHETICAL EXAMPLES

In this chapter, we will examine the performance of finite mixture models with several simulated datasets. It should be noted that this is not a Monte Carlo study; we only consider the results of finite mixture models for a single simulated sample. In Chapter VII, however, we will carry out a Monte Carlo simulation study by generating many samples to investigate the potential bias and variability in the parameter estimates for various combinations of sample sizes and sample mean values. The objectives of this chapter are, first, to examine the appropriateness of the mixture model specification in describing the count data generation process which exhibits over-dispersion and, second, to investigate how the finite mixture models can effectively capture the sub-populations, thereby, explain the population heterogeneity existing in the data. Working with the numerical examples is effective in illustrating the theoretical aspects of the finite mixture models in that we can generate and analyze a random sample with known characteristics.

Three examples are presented in this chapter. The first example described in Section 4.1 is used to illustrate the mechanism how finite mixture regression models can provide good numerical approximations when the underlying mixing distribution is continuous. The effects of sample mean, sample size and the degree of dispersion on the number of components are also examined. The second and third examples shown in Section 4.2 and 4.3 are used to illustrate the appropriateness of the mixture model specifications when the data were actually generated from a two-component Poisson (FMP-2) or NB distribution (FMNB-2). We will show from these examples that how effectively the finite mixture regression models can capture the sub-populations, and thereby emphasize the disadvantage of using single aggregate NB regression models in such situations.

4.1 Example 1

The objective of this example is to examine how well a K-component Poisson mixture regression model (FMP-K) can approximate (or replicate) the data which were originally generated by a continuously mixed Poisson distribution. For this purpose, first, various datasets will be generated by a Poisson-Gamma (NB) distribution. The NB model assumes that the unobserved heterogeneity in the Poisson mean follows a continuous gamma distribution. Each generated dataset will be fitted with both the NB regression model and the FMP-K models, and the results will be compared. It should be noted that, in FMP-K models, no distributional assumption is made on the mixing distribution but a few finite number of mass (or support) points and their respective proportions approximate the continuous gamma distribution.

4.1.1 Data generation method

For generating NB random variates, first, we introduced two covariates, $\mathbf{x}_i = (1, x_{i1}, x_{i2})^T$, in the link function, which were randomly generated from the standard normal distribution. The Poisson mean was then constructed from the two covariates by assuming a log-linear relationship using known (assigned) regression coefficients $\boldsymbol{\beta}_i = (\beta_0, \beta_1, \beta_2)^T$, which results in $\mu_i = \exp(\mathrm{x}_i^T \cdot \mathrm{\beta})$. Specifically, rather than simulating the data directly using the probability density function of the NB distribution, the random variables ($\mathbf{y}$) were simulated in the following step-wise fashion:

- Step 1: Set sample size, $\boldsymbol{\beta}$, and ϕ to the required values (see table on page 78).
- Step 2: Generate two covariates (x_{i1}, x_{i2}) from the $N(0,1)$ distribution.
- Step 3: Generate the error term $v_i (= e^{\varepsilon_i})$ from the $Gamma\ (\phi, \phi)$ distribution.
- Step 4: Set the Poisson mean, $\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) \cdot v_i$.
- Step 5: Generate the count variable y_i from the $Poisson\ (\lambda_i)$ distribution.
- Step 6: Repeat Steps 3 through 5 N times and save all the generated values.

4.1.2 FMP-K model estimation

Suppose that the Poisson mean parameter λ_i has a random intercept term, and the random term enters the conditional mean function multiplicatively, that is,

$$\begin{aligned} \lambda_i &= \exp(\mathbf{x}_i \cdot \boldsymbol{\beta} + \varepsilon_i) && (4.1) \\ &= \exp(\beta_0 + \mathbf{x}_i^* \cdot \boldsymbol{\beta}^* + \varepsilon_i) \\ &= e^{\mathbf{x}_i^* \cdot \boldsymbol{\beta}^*} \cdot e^{(\beta_0 + \varepsilon_i)} \\ &= \mu_i^* \cdot \delta_i && (4.2) \end{aligned}$$

where $\mathbf{x}_i^* = (x_{i1}, x_{i2})$, $\boldsymbol{\beta}_i^* = (\beta_1, \beta_2)'$, and $\delta_i = e^{(\beta_0 + \varepsilon_i)}$ is interpreted as a random intercept. Note that $\mu_i^* = \exp(\mathbf{x}_i^* \cdot \boldsymbol{\beta}^*)$ contains no intercept term.

The unconditional probability of y_i have the following form:

$$\begin{aligned} \Pr(y_i) &= \int_0^\infty \Pr(y_i \mid \mu_i^*, \delta_i) \cdot g(\delta_i) d\delta_i \\ &= \int_0^\infty e^{\delta\mu_i^*} (\delta\mu_i^*)^{y_i} / y_i! \cdot g(\delta_i) d\delta_i \end{aligned} \qquad (4.3)$$

where $\Pr(y_i \mid \mu_i^*, \delta_i)$ is the conditional Poisson density for given δ_i, and $g(\cdot)$ is a mixing distribution. Equation (4.3) indicates that $\Pr(y_i \mid \mu_i^*, \delta_i)$ is averaged by the probability of each value of δ_i. If we assume that δ_i has only two mass points m_1 and m_2, the counterpart to Equation (4.3) is expressed as a two-component mixture:

$$\Pr(y_i) = \Pr(y_i \mid \mu_i^*, \delta_i = m_1) \cdot \Pr(\delta_i = m_1) + \Pr(y_i \mid \mu_i^*, \delta_i = m_2) \cdot \Pr(\delta_i = m_2) \qquad (4.4)$$

This implies that Equation (4.3) can be obtained by weighting $\Pr(y_i \mid \mu_i^*, \delta_i)$ by $\Pr(\delta_i)$ and then adding over all values of δ_i. It has been shown that a continuous mixing distribution function $g(\cdot)$ can be consistently estimated with a finite number (K) of mass points and their weights (Brännäs and Rosenqvist, 1994):

$$\Pr(y_i) = \sum_{k=1}^{K} \Pr(y_i \mid \hat{\delta}_k) \cdot \hat{w}_k \tag{4.5}$$

A maximum likelihood estimator based on the equation above yields a consistent estimator $\hat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}^*$, where $\hat{\delta}_k$ is an estimated mass point and $\hat{w}_k$ is the associated estimated probability.

In estimating the parameters of FMP-K models in this exercise, the regression parameters for the covariates (slope parameters) were constrained to be fixed across all K components assuming that heterogeneity arises from differences in the intercepts only. This constraint is reasonable since the original data will be generated from a single set of regression parameters from an NB model. The slope-constrained, or random intercept, modeling approach has been taken by several authors to account for unobserved heterogeneity in count data (Simar, 1976; Laird, 1978; Heckman and Singer, 1984; Brännäs and Rosenqvist, 1994). Especially, Heckman and Singer (1984) have shown that the coefficients of the covariates estimated from this approach are consistent, and asymptotically normal. In Example 2 in the following section, however, we relax this constraint and the regression coefficients of the covariates as well as the intercept are allowed to vary across the sample. The constrained FMP-K model will be termed as a CFMP-K model hereafter.

For parameter estimation, although it can be done within the Bayesian framework, we adopted the maximum likelihood method for this exercise. Using the existing software package (i.e. R package “FlexMix”) had an advantage to reduce the amount of computing time. The FlexMix provides a general framework for the finite mixture of regression models using the EM algorithm which is available as an extension package for the statistical software R (Grün and Leisch, 2007).

4.1.3 Results

To illustrate the theoretical aspects and compare the results with the estimated NB model and CFMP-K models, three datasets were generated based on the moderate mean value scenario for $\phi = 0.5$, 2.0, and 5.0, respectively (see table on page 78). Sample size, $N = 300$, was used for this exercise. To determine the number of components, K was sequentially increased from 2 until either AIC or BIC value reached its minimum value.

Table 4.1 shows the true values and estimated results for $\phi = 0.5$ (high-dispersed data). K=4 was determined as the best for this dataset. Although the NB estimates of slope parameters look much closer to the true values than the CFMP-4 model, it is believed that repeating the sampling would produce the estimates clustered around the true values. When we compared the AIC and BIC values, the approximation by the four components seems to be quite satisfactory.

Table 4.1 True values and estimation results ($\phi = 0.5$)

	True Values	NB Model	CFMP-2		CFMP-3			CFMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.0	0.9040 (0.1006)	2.0709 (0.0630)	-0.3880 (0.1078)	2.3897 (0.0749)	0.8458 (0.1080)	-3.0245 (0.6929)	2.4893 (0.1119)	1.5949 (0.2794)	0.5257 (0.2167)	-3.5507 (0.8821)
β_1	0.5	0.4529 (0.0945)	0.4444 (0.0369)		0.3194 (0.0441)			0.3461 (0.0681)			
β_2	-0.5	-0.5038 (0.1019)	-0.5298 (0.0512)		-0.5956 (0.0501)			-0.6120 (0.0566)			
ϕ	0.5	0.403 (0.047)	-		-			-			
w	-	-	0.249	0.751	0.140	0.431	0.428	0.104	0.150	0.353	0.393
-2LL	-	1184.2	1340.6		1194.3			1172.3			
AIC	-	1192.2	1350.6		1208.3			1190.3			
BIC	-	1207.1	1369.1		1234.3			1223.6			

NOTE: Sample mean=3.01; Sample variance=40.79; ( ) indicates the standard error of the estimate.

Figure 4.1 illustrates how the approximation approach operates. The fitted mixing density of $\hat{\delta}_{CFMP-4}$ is compared to the gamma density based on the true NB (δ_{true}) and estimated NB ($\hat{\delta}_{NB}$) models. Since the error term, e^{ε} follows the $gamma\,(0.5, 0.5)$

distribution in the true NB model, the random intercept, $\delta_{true}(=e^{1.0+\varepsilon})$ follows the $gamma\,(0.5, 0.5/e^{1})$ distribution. In the same manner, the $\hat{\delta}_{NB}$ follows the $gamma\,(0.403, 0.403/e^{0.904})$ distribution in the estimated NB model. As shown in the figure, CFMP-4 model effectively approximates the continuous δ_{true} distribution with four numbers of mass points ($e^{-3.5507}, e^{0.5257}, e^{1.5949}$, and $e^{2.4893}$) and their respective probabilities (0.393, 0.353, 0.150, and 0.104).

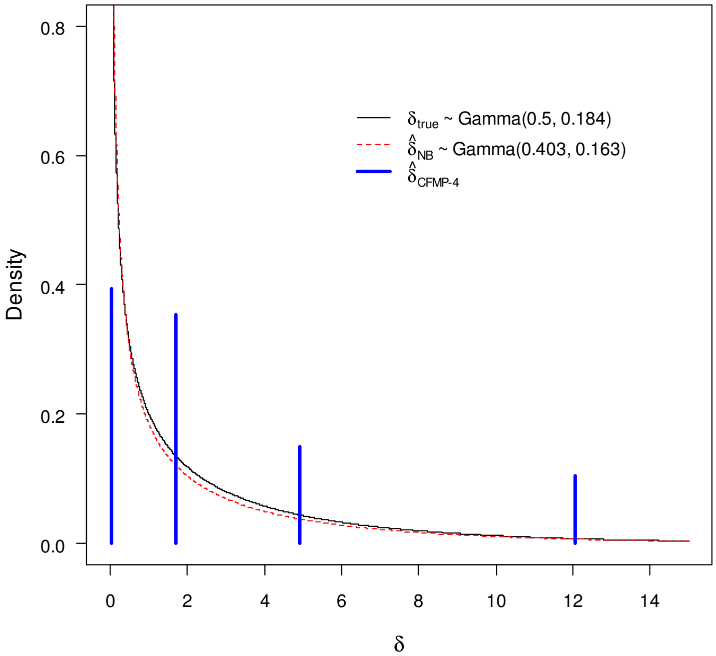


Figure 4.1 Density functions for random intercept, δ $(\phi = 0.5)$

On the other hand, the plot in Figure 4.2 visualizes the goodness-of-fit comparison between the generated (or observed) frequencies and the predicted frequencies from each model. To make the comparison more informative the histogram was truncated at 39 counts, while the maximum count was 62. It clearly shows that the CFMP-4 model provides almost as good a result as the NB model. The explanation on how to calculate the predicted frequencies is in order. The predicted probabilities can be computed for each observation for each count m that is of interest (0 to 62 in this case). Then the mean predicted probability for each count m can be used to summarize the predictions of the model (Long, 1997):

$$\overline{\Pr}(y=m)=\frac{1}{N}\sum_{i=1}^{N}\Pr(y_i=m\mid \mathbf{x}_i) \tag{4.6}$$

Then, the predicted frequencies for each count m can be obtained by $N\times\overline{\Pr}(y=m)$ and can be compared to the observed frequencies of the sample at each count. This method was utilized throughout this study.

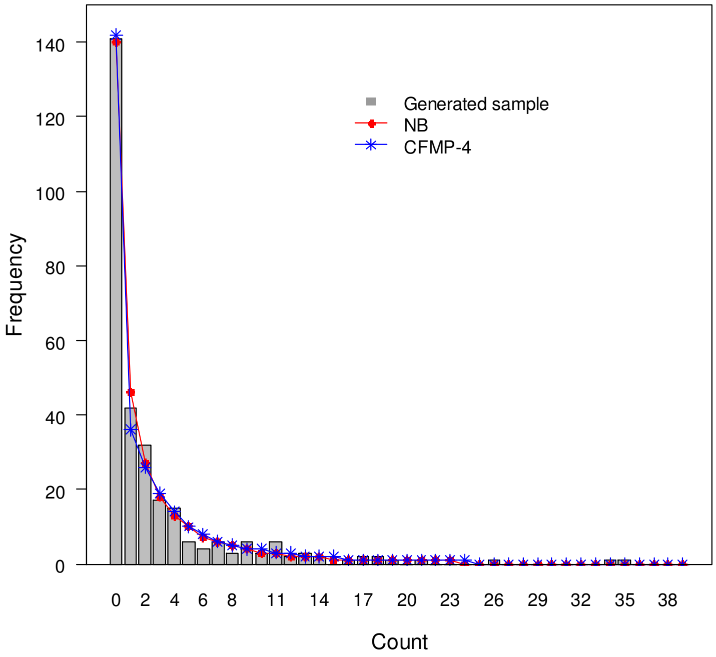


Figure 4.2 Goodness-of-fit comparison $(\phi=0.5)$

Table 4.2 shows the results for $\phi=2.0$ (moderate-dispersed data). In this case, either K=2 or K=3 produced as good a result as the NB model. This is in contrast to the result in the first exercise where more number of components was required. It seems that as the data are less dispersed, smaller number of components is required, and vice versa (This will be confirmed later by the third exercise). Figure 4.3 compares the estimated mixing density of $\hat{\delta}_{CFMP-2}$ and $\hat{\delta}_{CFMP-3}$ to the gamma density based on the true NB (δ_{true}) and the estimated NB $(\hat{\delta}_{NB})$ models. Note that the difference between $\hat{\delta}_{NB}$ and δ_{true} becomes noticeable as the true dispersion parameter become larger.[10] The goodness-of-fits are compared in Figure 4.4. The predicted frequency plots by CFMP-2 and CFMP-3 do not perform very well at very low counts. However, the overall approximation looks very satisfactory.

Table 4.2 True values and estimation results $(\phi = 2.0)$

	True Values	NB Model	CFMP-2		CFMP-3		
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3
β_0	1.0	0.9208 (0.0554)	1.6602 (0.0969)	0.4765 (0.0901)	1.8000 (0.1048)	0.8618 (0.1448)	-0.0439* (0.2769)
β_1	0.5	0.3950 (0.0507)	0.3771 (0.0423)		0.3676 (0.0450)		
β_2	-0.5	-0.5719 (0.0552)	-0.5818 (0.0433)		-0.5628 (0.0465)		
ϕ	2.0	2.318 (0.384)	-		-		
w	-	-	0.251	0.749	0.329	0.509	0.162
-2LL	-	1219.2	1221.5		1213.7		
AIC	-	1227.2	1231.5		1227.7		
BIC	-	1242.0	1250.0		1253.6		

NOTE: Sample mean=3.10; Sample variance=17.14; *indicates the coefficient which is not significant at 5% significance level; ( ) indicates the standard error of the estimate.

Table 4.3 shows the results for $\phi = 5.0$ (low-dispersed data). In this case, it is clear from the BIC value that only two components are quite enough for the approximation. This confirms the speculation in the second example that as the data become less dispersed, smaller number of components is required.

[10] Using a simulation study, Park and Lord (2008) showed that the bias for the maximum likelihood estimate of the dispersion parameter (ϕ) is a function of true dispersion parameter (if it is known) as well as sample size and sample mean value. The bias is largely due to the size of the true dispersion parameter, and it becomes larger as the true dispersion parameter increases.

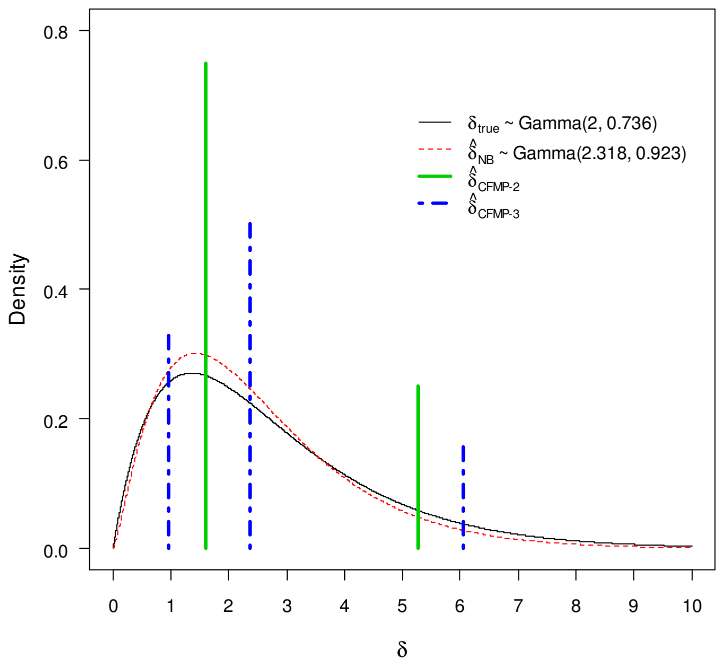


Figure 4.3 Density functions for random intercept, δ ($\phi = 2.0$)

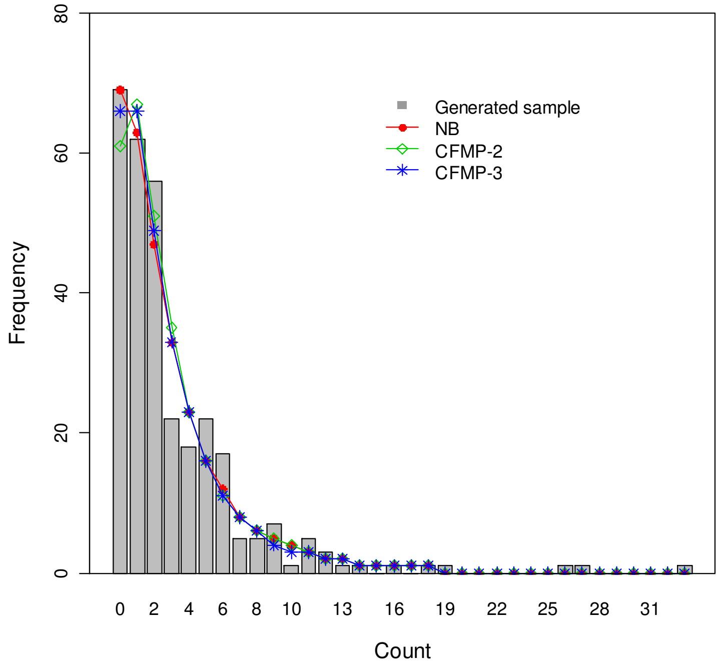


Figure 4.4 Goodness-of-fit comparison ($\phi = 2.0$)

Table 4.3 True values and estimation results $(\phi = 5.0)$

	True Values	NB Model	CFMP-2		CFMP-3		
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3
β_0	1.0	0.9321 (0.0456)	1.2800 (0.0884)	0.4721 (0.1307)	1.2797 (0.0883)	0.4714* (0.3545)	0.4714* (0.3511)
β_1	0.5	0.4826 (0.0401)	0.4862 (0.0392)		0.4861 (0.0392)		
β_2	-0.5	-0.4205 (0.0425)	-0.4347 (0.0415)		-0.4347 (0.0415)		
ϕ	5.0	6.540 (1.750)	-		-		
w	-	-	0.468	0.532	0.459	0.271	0.270
-2LL	-	1169.6	1166.0		1165.9		
AIC	-	1177.6	1176.0		1179.9		
BIC	-	1192.4	1194.5		1205.9		

NOTE: Sample mean=3.10; Sample variance=11.50; *indicates the coefficient which is not significant at 5% significance level; ( ) indicates the standard error of the estimate.

Figure 4.5 compares the estimated mixing density of to the gamma density based on the true NB (δ_{true}) and estimated NB $(\hat{\delta}_{NB})$ models. Note that the CFMP-2 model effectively approximates the continuous δ_{true} distribution with only two numbers of mass points $(e^{0.4765}, e^{1.6602})$ and their respective probabilities (0.749, 0.251). It is also worth noting that the difference between $\hat{\delta}_{NB}$ and δ_{true} is much larger than the more dispersed dataset cases. This agrees with the findings in Park and Lord (2008) who noticed that the bias of the maximum likelihood estimate $\hat{\phi}_{MLE}$ value in NB distribution becomes larger as the true ϕ value increases. The goodness-of-fits are compared in Figure 4.6. As evidenced by the AIC and BIC values, the goodness-of-fit is almost the same between the NB model and the CMFP-2 model.

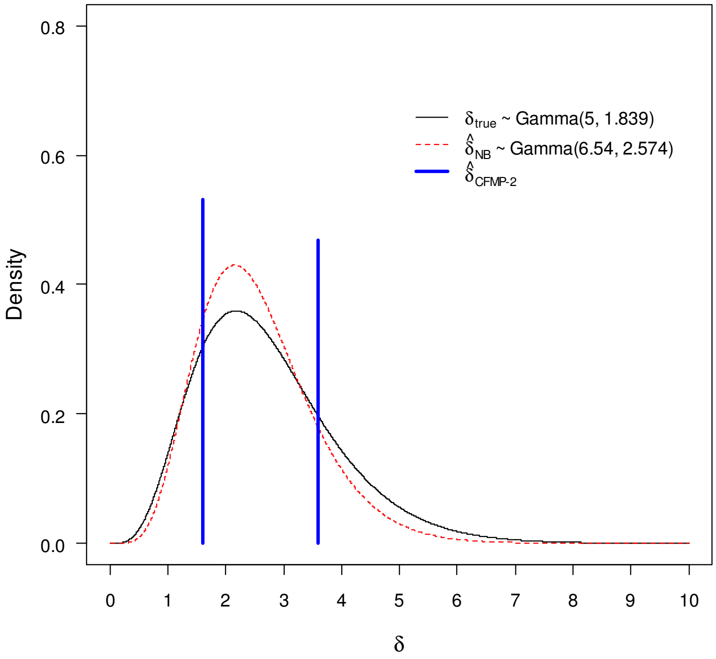


Figure 4.5 Density functions for random intercept, δ $(\phi = 5.0)$

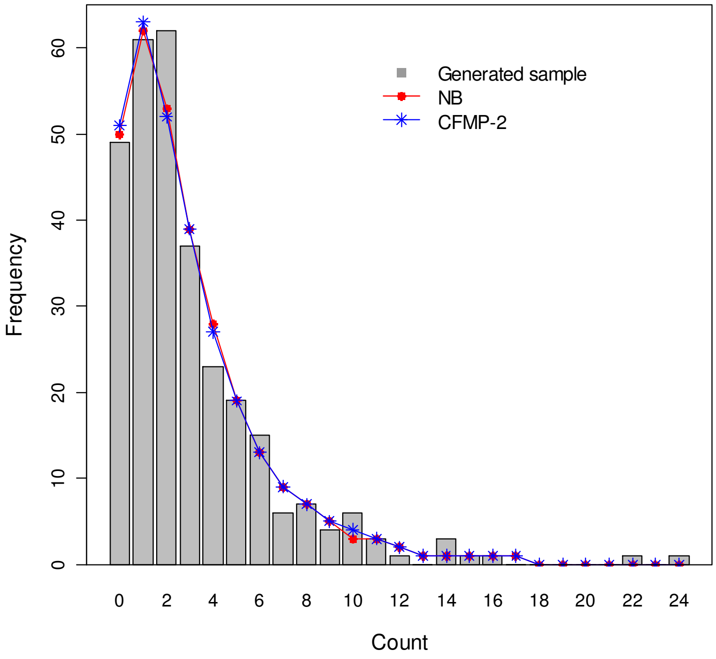


Figure 4.6 Goodness-of-fit comparison $(\phi = 5.0)$

4.1.4 Effects of sample size and sample mean

In the previous exercises, the potential effect of the dispersion parameter on the number of components was explained. It was evident that assuming that the dataset is within the similar sample mean and sample size, the less dispersed the data are, the smaller number of components is required.

In order to examine the effects of the sample mean value and the sample size on the number of components, we generated the datasets under the various combinations of sample size, sample mean value, and the dispersion parameter. In Step 1 described in Subsection *4.1.1*, the regression coefficients were controlled to produce high mean $(\bar{\mathbf{y}} > 5)$, moderate mean $(1 < \bar{\mathbf{y}} < 5)$, and low mean $(\bar{\mathbf{y}} < 1)$, respectively. For ϕ values, 0.5 for high-dispersed, 2.0 for moderate-dispersed, and 5.0 for low-dispersed were used. For sample sizes, 50 (small), 100 (moderate), and 500 and 1000 (large) were used. Therefore, a total of 36 datasets (3×3×4) were generated. Table 4.4 shows the values used for data generation.

Table 4.4 Values used for generating NB random variates

	Small mean $(\bar{\mathbf{y}} < 1)$	Moderate mean $(1 < \bar{\mathbf{y}} < 5)$	High mean $(\bar{\mathbf{y}} > 5)$
β_0 β_1 β_2	-0.5 0.5 -0.5	1.0 0.5 -0.5	1.7 0.5 -0.5
ϕ	0.5 (High-dispersed) 2.0 (Moderate-dispersed) 5.0 (Low-dispersed)		
N	50 (Small sample size) 100 (Moderate sample size) 500, 1000 (Large sample size)		

Table 4.5 shows the number of components in CFMP models required to adequately approximate the NB distribution under the various combinations of sample size, sample mean value, and the dispersion parameter. As before, the number of components was determined in terms of AIC and BIC values. All the simulation results are attached in the Appendix C. The results are summarized here as follows:

- In all example scenarios, the NB regression models could be effectively approximated by a few mixtures of Poisson regression models (K=2~5).
- While the AIC values of CFMP-K models were sometimes superior to NB models, the CFMP-K models did not perform better than the NB models based on the BIC value.
- As sample mean value increases, there is a trend that more components are required. This trend was more pronounced in the more dispersed datasets.
- As the sample size increases, there is a trend that more components are required. This trend was more evident in the higher mean value cases.
- As the ϕ value increases (less dispersed), smaller components were needed. This was true regardless of sample mean values and sample sizes.

Table 4.5 Number of components in CFMP models required to approximate NB models

	Small mean $(\bar{\mathbf{y}} < 1)$			Moderate mean $(1 < \bar{\mathbf{y}} < 5)$			High mean $(\bar{\mathbf{y}} > 5)$		
ϕ	0.5	2.0	5.0	0.5	2.0	5.0	0.5	2.0	5.0
$N = 50$	2	2	2	3	2	2	3	2	2
$N = 100$	2	2	2	3	2	2	4	3	2
$N = 500$	3	2	2	4	3	2	5	4	3
$N = 1000$	3	2	2	4	4	2	4	5	4

4.2 Example 2

The objective of this example is to illustrate how the interpretation from the single aggregate NB model can be misleading when the data were actually generated by the two-component finite mixture of Poisson regression models (FMP-2). This hypothetical example is meant to show how poor the prediction capability of the standard NB model will be because of the model misspecification. Since the standard NB model estimates a single set of regression coefficients, the interpretation of its coefficients may be wrong if the population is heterogeneous with respect to the impact of explanatory variables. In this example, therefore, the data are to be generated by the FMP-2 whose regression coefficients of the covariates as well as the intercept are allowed to vary across the two components.

4.2.1 Data generation method

For generating FMP-2 random variates, similar to Example 1, we introduced two covariates, $\mathbf{x}_i = (1, x_{i1}, x_{i2})'$, in the link function, which were randomly generated from the standard normal distribution. The Poisson means for each component $\mu_{i,1}$ and $\mu_{i,2}$ were then constructed from the independent variables by assuming a log-linear relationship using known (or assigned) regression coefficients $\boldsymbol{\beta}_1 = (\beta_{0,1}, \beta_{1,1}, \beta_{2,1})'$ and $\boldsymbol{\beta}_2 = (\beta_{0,2}, \beta_{1,2}, \beta_{2,2})'$. This results in each component mean $\mu_{i,1} = \exp(\mathbf{x}_i \cdot \boldsymbol{\beta}_1)$ and $\mu_{i,2} = \exp(\mathbf{x}_i \cdot \boldsymbol{\beta}_2)$, respectively. Based on these two components' means, the FMP-2 random variate for site i was generated by introducing a mixing proportion, w. Thus, with probability w, the random variate for the site i is generated from the $Poisson\,(\mu_{i,1})$ distribution whereas with probability $1-w$, it is generated from the $Poisson\,(\mu_{i,2})$ distribution. The data generation procedures can be summarized as follows:

- Step 1: Set N (sample size), $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and w to the required values.

- Step 2: Generate two covariates (x_{i1}, x_{i2}) from the $N\,(0,1)$ distribution.
- Step 3: Generate the binary value (0 or 1) from the $Binom\,(1, w)$ distribution.
- Step 4: Save the generated binary value as z_i for site *i*.
- Step 5: Generate the count variable y_i from the following relationship.

$$z_i \cdot Poisson\,(\mu_{i,1}) + (1 - z_i) \cdot Poisson\,(\mu_{i,2}) \qquad (4.7)$$

- Step 6: Repeat Steps 3 through 5 N times and save all the generated values.

In this manner, a dataset was generated from a FMP-2 distribution with the sample size of $N = 500$ for this example. Then, it was fitted with the NB and FMP-2 models, respectively. In this case, the NB regression model is a misspecification. The assumed values for parameters are shown in Table 4.6 and the histograms of the generated count data are shown in Figure 4.7. The data appear to be highly dispersed and resemble empirical crash frequency plots which are likely to be encountered by highway safety analysts. The sample mean and sample variance were 3.14 and 27.54, respectively.

4.2.2 Parameter estimation method

The Bayesian estimation method was implemented for this example. The method was described in details in Chapter III. For prior distributions for each model parameter, non-informative prior specifications were used: i.e., as prior for the weight distribution w_k, the $Dirichlet\,(1,1)$ was used; as prior for the regression coefficient $\boldsymbol{\beta}_k$, the $MVN_3(\mathbf{b}_0, \mathbf{B}_0)$ distribution was used, in which $\mathbf{b}_0 = (0,0,0)'$ and $\mathbf{B}_0 = 100\mathbf{I}_3$, where $\mathbf{I}_3$ denotes the 3-dimensional identity matrix. For the NB model, the prior for the regression coefficients was again $MVN_3(\mathbf{b}_0, \mathbf{B}_0)$, and the non-informative $\Gamma(0.01, 0.01)$ prior was used for the dispersion parameter. The codes for data generation and estimation with the FMP-2 model are provided in Appendix D.

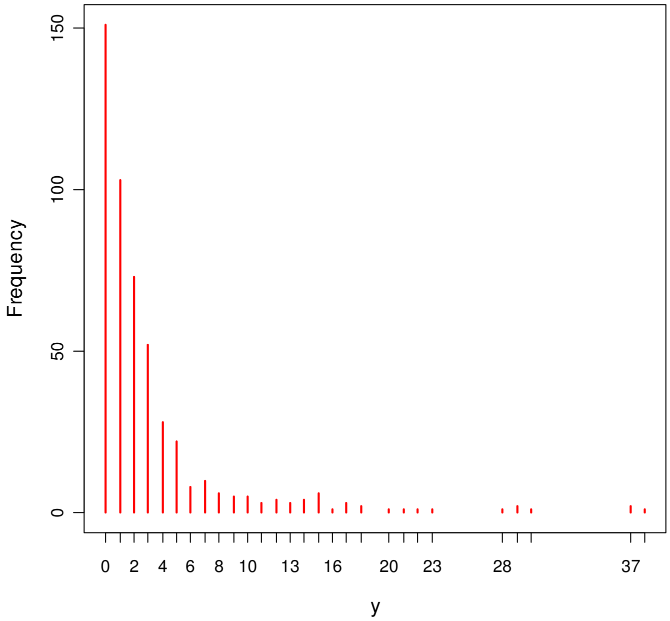


Figure 4.7 Histograms of generated counts from FMP-2

4.2.3 Results

For a preliminary NB model assessment, we first fitted the data with the NB distribution using the maximum likelihood method and checked the quality of the fit between the observed values y_i and the fitted values $\hat{\mu}_i$ with the Pearson X^2 statistic. For a well-fitting, or adequate, model the value of X^2 should come from a χ^2 distribution with ($N-p$) degrees of freedom (McCullagh and Nelder, 1989), where N is the number of observations and p is the number of parameters which have been estimated. Therefore, the ad-hoc assessment is if $X^2/(N-p)$ is close to 1, we conclude that the model's goodness-of-fit is satisfactory. For this dataset, the Pearson X^2 statistic was 1.09. It seems that the NB model produced a very satisfactory goodness-of-fit and addressed the over-dispersion. However, the NB regression model is a misspecification and any inference or prediction from this model can be misleading because it totally ignores the existence of different coefficients.

For the Bayesian estimation of the FMP-2 model, a total of 5,000 MCMC iterations were used without thinning (i.e., keeping every 1st samples), and half the iterations were discarded (burn-in period). From the remaining 2,500 samples, the posterior means and standard deviations were calculated.

Table 4.6 displays the estimated parameters and computed values of model selection criteria for each model. For this dataset, as shown in the table, the coefficients estimated from the FMP-2 model are close to the true values and, as expected, all model selection criteria supported the choice of FMP-2 model. The NB model, by nature, could not explain the heterogeneous impact of the covariates.

Table 4.6 True values and estimation results (FMP-2)

Model Parameters	True Values		NB Regression	FMP-2	
	Comp 1	Comp 2		Comp 1	Comp 2
β_0	2.0	0.0	1.1401 (0.0648)[a]	1.9825 (0.0493)	0.0173* (0.0622)
β_1	-0.5	0.5	-0.1107* (0.0600)	-0.5741 (0.0350)	0.5198 (0.0469)
β_2	0.5	-0.5	0.0861* (0.0553)	0.4389 (0.0354)	-0.5118 (0.0444)
ϕ	-	-	0.6040 (0.049)	-	-
w	0.2	0.8	1	0.213 (0.023)	0.787 (0.023)
-2LL	The smaller the better		2242.3	1918.4	
AIC	〃		2250.3	1932.4	
BIC	〃		2267.1	1961.9	
DIC	〃		2250.4	1932.0	
Log(ML)	The larger the better		-1142.8	-995.5	

NOTE: [a] indicates the standard deviation of the coefficient; * indicates the coefficient whose 95% credible interval includes zero.

Figure 4.8 compares the goodness-of-fits between the NB and FMP-2 models. The NB model is showing a very poor predictive capability, especially at the portion of the smaller numbers of counts. It results from the fact that the NB model could not consider the population heterogeneity by completely ignoring the discrete nature of the data generation process.

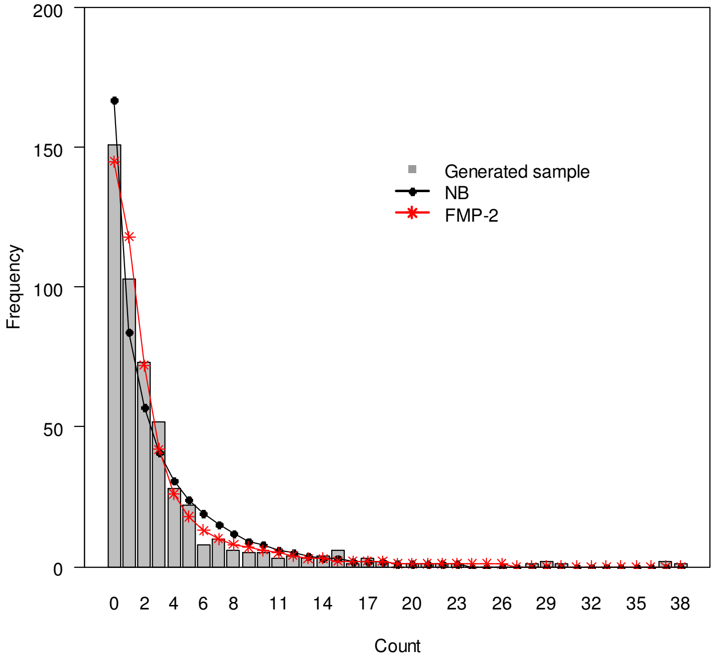


Figure 4.8 Goodness-of-fit comparison between NB and FMP-2

Convergence was checked by monitoring the trace plots of the samples, autocorrelations and marginal posterior distributions of the model parameters. The MCMC trace plots (Figure 4.9) indicate that the chains appear to have reached stationary distributions and the chains have good mixing and are dense.

The autocorrelation plots (Figure 4.10) indicate that although there is small evidence of autocorrelations between samples, increasing MCMC iterations (for example, 50,000 iterations) or thinning value did not change the resulting coefficient values very much.

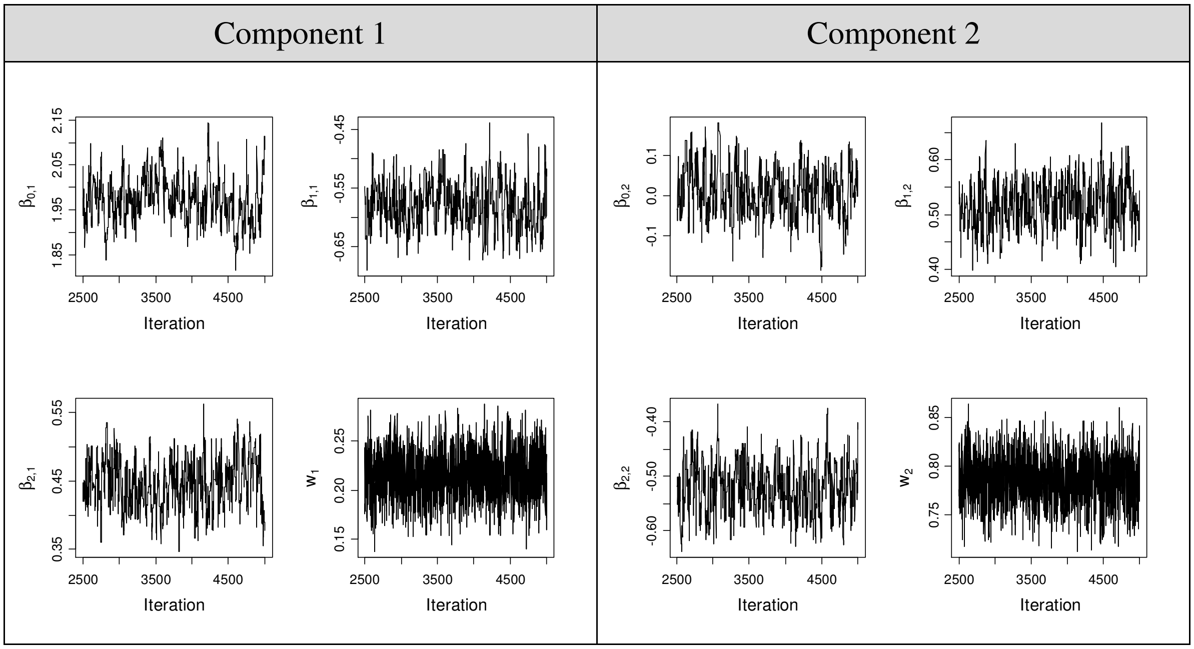


Figure 4.9 MCMC trace plots (FMP-2)

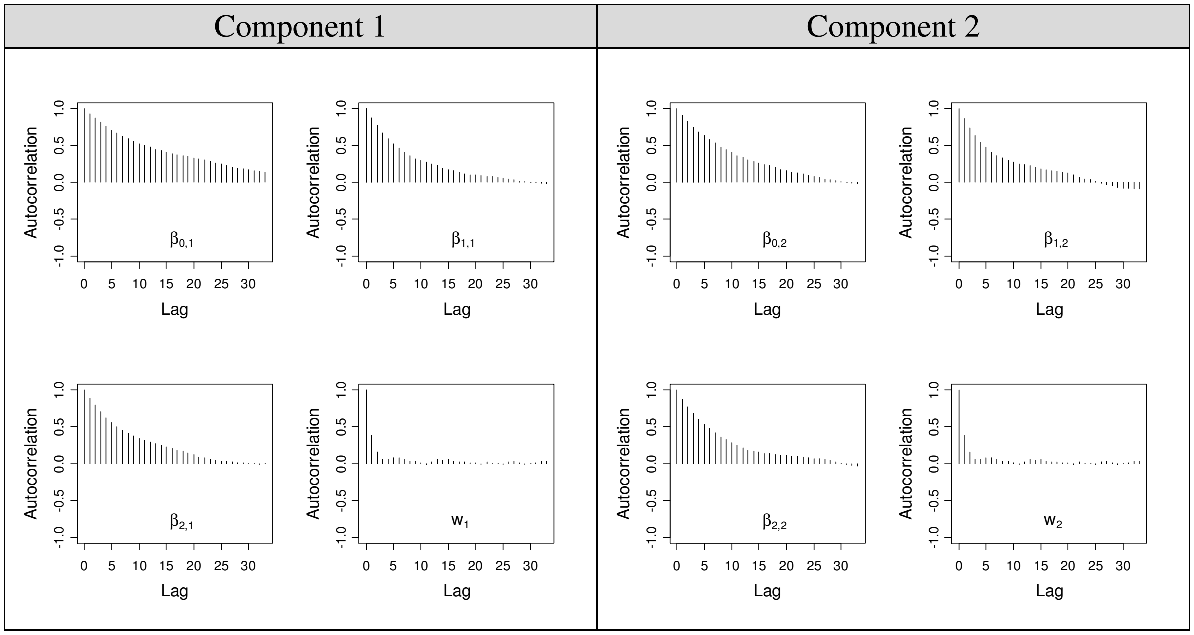


Figure 4.10 Autocorrelation plots (FMP-2)

Finally, the kernel density plots (Figure 4.11) show the uni-modal shape of marginal posterior distributions for each model parameter, which is very close to a normal distribution. If label switches have occurred, there must be jumps in the trace plots or multimodal density plots.

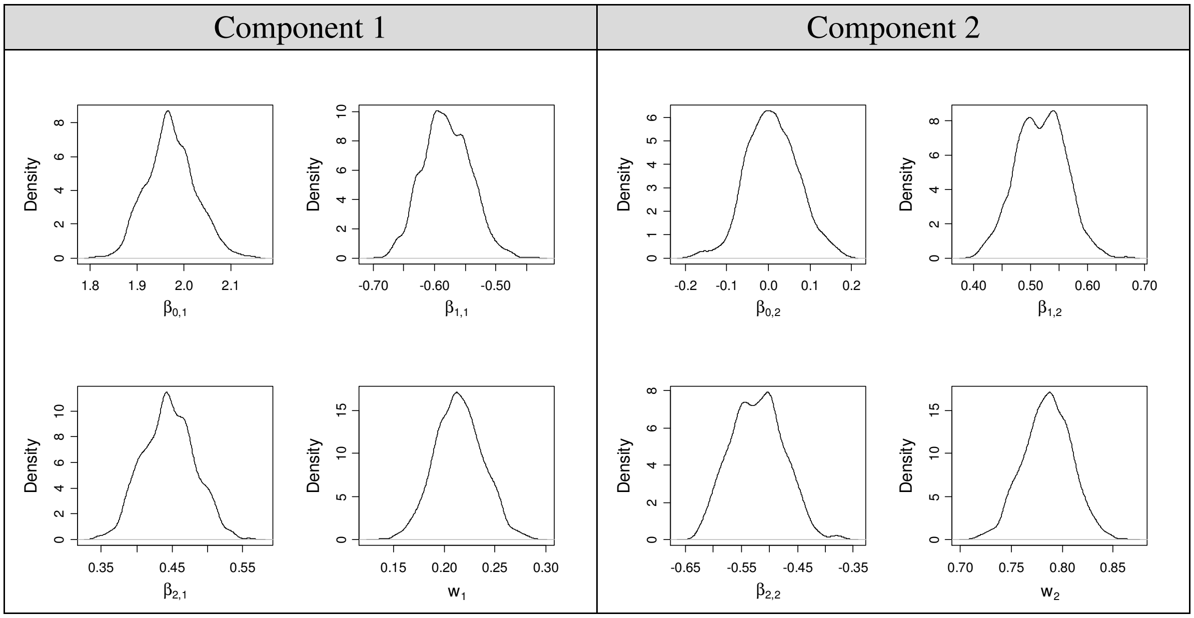


Figure 4.11 Marginal posterior distributions (FMP-2)

4.3 Example 3

The objective of this example is to examine how well the two-component finite mixture of negative binomial regression models (FMNB-2) can replicate the data as compared to the standard NB model, when the data were originally generated from a FMNB-2 distribution. As in Example 2, this example is meant to show how poor the prediction capability of the standard NB model will be because of the model misspecification. The results of this example also will support the idea of using the FMNB-2 model when the data are suspected to belong to different sub-groups and each sub-component exhibits over-dispersion. Note that the FMP-2 model can accommodate the population heterogeneity, but cannot handle the over-dispersion within sub-groups.

4.3.1 Data generation method

Generating FMNB-2 random variates is very similar to the generation of FMP-2 random variates, except that the negative binomial distribution is used in each component. The component means $\mu_{i,1}$ and $\mu_{i,2}$ are constructed in the same manner as in Example 2. Thus, with probability w, the binary value (z_i) is generated from the $Binomial\ (1, w)$ distribution for each site *i*, and then the FMNB-2 random variates are generated from $z_i \cdot NB\ (\mu_{i,1}, \phi_1) + (1 - z_i) \cdot NB\ (\mu_{i,2}, \phi_2)$. The data generation procedures are summarized as follows:

- Step 1: Set N (sample size), $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \phi_1, \phi_2$ and w to the required values
- Step 2: Generate two covariates (x_{i1}, x_{i2}) from the $N\ (0,1)$ distribution.
- Step 3: Generate the binary value (0 or 1) from the $Binomial\ (1, w)$ distribution.
- Step 4: Save the generated binary value as z_i for site *i*.
- Step 5: Generate the count variable y_i from the following relationship.

$$z_i \cdot NB\ (\mu_{i,1}, \phi_1) + (1 - z_i) \cdot NB\ (\mu_{i,2}, \phi_2) \tag{4.8}$$

- Step 6: Repeat Steps 3 through 5 N times and save all the generated values.

In this example, a dataset was generated from a FMNB-2 distribution with the sample size of $N = 500$, and then it was fitted with three models: NB, FMNB-2 and FMP-2 models. In this case, both NB and FMP-2 models are misspecifications. The assumed values used for data generation are shown in Table 4.7, and Figure 4.12 shows the histograms of the generated count data. The data appear to be highly dispersed and resemble empirical crash data. The sample mean and sample variance were 2.84 and 28.85, respectively.

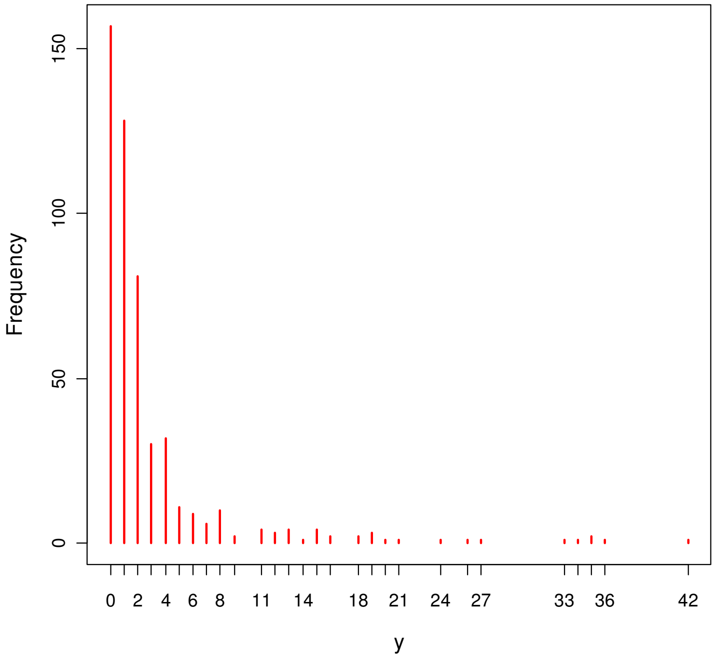


Figure 4.12 Histogram of generated counts from FMNB-2

4.3.2 Parameter estimation method

The Bayesian estimation method was implemented for this example. As priors for the weight distribution and the regression coefficients, the specifications are the same as in Example 2. As prior for the dispersion parameter ϕ_k , the $\Gamma\,(0.01, 0.01)$ prior was used which represents no prior information on this parameter. For the NB model, the prior specifications for the parameters are the same as in Example 2. The codes for data generation and estimation with the FMNB-2 model are provided in Appendix D.

4.3.3 Results

As in Example 2, the initial check of the goodness-of-fit of the NB model was checked with the $X^2/(N-p)$ statistic. For this dataset, it was around 1.19 which indicates that the observations are slightly over-dispersed with respect to the NB model. However, all the coefficients turned out to be significant at 5% significance level. Note that the negative binomial can also be over-dispersed. Hilbe (2007) suggests that if the Pearson X^2 statistic is greater than 1.25 for moderate sized models and 1.05 for large numbers of observations, a correction for over-dispersion may be warranted. It is obvious in this case that the correction should be made and the source of over-dispersion is the population heterogeneity. Thus, the finite mixture model is a good option among others.

For the Bayesian estimation of the FMNB-2 model, a total of 5,000 MCMC iterations were used, keeping every 1^{st} samples (thinning), and half the iterations were discarded (burn-in period). From the remaining 2,500 samples, the posterior means and standard deviations were calculated.

Table 4.7 displays the posterior means of parameters and computed values of model selection criteria for each model. For this dataset, as shown in the table, the coefficients estimated from the FMNB-2 model are very close to the true values and all model selection criteria support the FMNB-2 model. The NB model, by nature, could not explain the heterogeneous impact of the covariates. Apparently, the FMP-2 depicts the true regression parameters quite well. However, because of the model misspecification, it could not account for the additional heterogeneity present within components. Such heterogeneity resulted in the underestimation of the standard deviation in the FMP-2 model. As shown in the table, the standard deviations for each parameter of FMP-2 model are consistently lower than those of FMNB-2. This is typical of the standard Poisson regression model when the over-dispersion was not accounted for.

On the other hand, the posterior mean of ϕ_2 for the smaller-mean component is by no means close to the true value, and its standard deviation is also very large. As already observed in Figure 4.14, this is because the posterior distribution of ϕ_2 included implausibly large values during the sampling process, which rendered its posterior to be skewed with a long right tail. One probable reason for this is the use of a non-informative gamma prior with an extremely large variance. The posterior mean from such skewed distribution is biased and its use as a posterior summary statistic is not a good option. The posterior medians for ϕ_1 and ϕ_2 were 3.893 and 12.715, respectively. In contrast, the maximum likelihood estimates for ϕ_1 and ϕ_2 were 5.103 and 32.676, respectively. Van Dongen (2006) noted that if the posterior is skewed, the mean or median of the posterior will not necessarily be close to the maximum likelihood estimate even if a non-informative prior is used. This prompts the suggestion that different prior specifications should be considered on the dispersion parameter so that its posterior distribution is less skewed and the posterior mean or median is less biased.

Table 4.7 True values and estimation results (FMNB-2)

Model Parameters	True Values		NB Model	FMNB-2		FMP-2	
	Comp 1	Comp 2		Comp 1	Comp 2	Comp 1	Comp 2
$\beta_{0,k}$	2.0	0.0	1.0221 (0.0634)	1.8333 (0.1133)[a]	-0.0292* (0.0674)	2.0692 (0.0550)	-0.0277* (0.0576)
$\beta_{1,k}$	-0.5	0.5	-0.1587 (0.0627)	-0.5195 (0.0828)	0.4641 (0.0582)	-0.3690 (0.0382)	0.4344 (0.0471)
$\beta_{2,k}$	0.5	-0.5	0.1409 (0.0614)	0.6351 (0.0907)	-0.5034 (0.0632)	0.5284 (0.0451)	-0.4729 (0.0490)
ϕ_k	5	10	0.575 (0.047)	4.152 (1.634)	19.492 (19.174)	-	-
w_k	0.2	0.8	-	0.218 (0.030)	0.782 (0.030)	0.179 (0.022)	0.821 (0.022)
Model Comparison Criteria							
-2LL	The smaller the better		2135.6	1891.6		1943.8	
AIC	〃		2143.6	1909.6		1957.8	
BIC	〃		2160.5	1947.6		1987.3	
DIC	〃		2143.7	1911.8		1957.7	
Log(ML)	The larger the better		-1089.3	-987.2		-1007.7	

NOTE: *indicates the coefficient whose 95% credible interval includes zero; ( ) indicates the standard deviation of the posterior mean.

Figure 4.13 compares the goodness-of-fit of the three models. It is not surprising to see that the predicted frequency of the standard NB model does not fit the data very well, especially at the smaller numbers of counts. We can clearly see the significant improvement made by the FMNB-2 model which incorporated the population heterogeneity in the model. The goodness-of-fit of the FMP-2 model looks as good as the one provided by the FMNB-2 model.

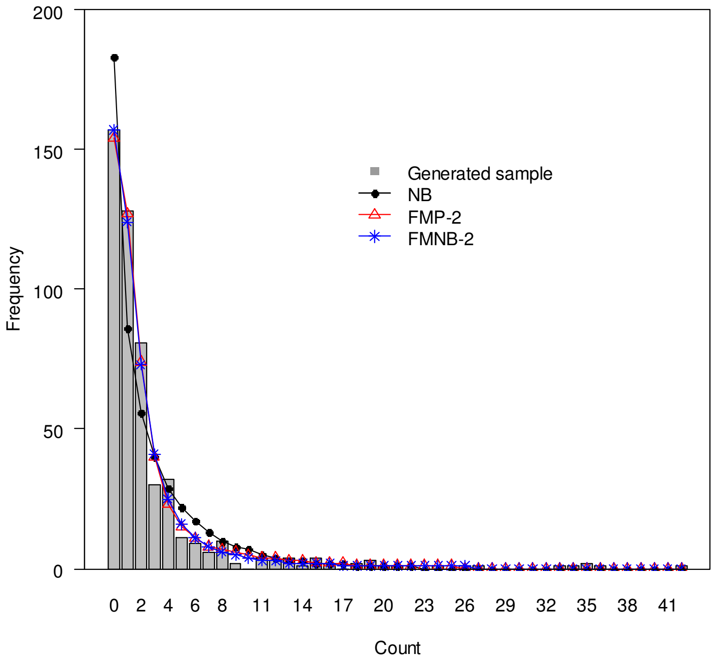


Figure 4.13 Goodness-of-fit comparisons between NB, FMP-2, and FMNB-2

The three convergence diagnostic graphs are shown in Figures 4.14-4.16. The MCMC trace plots indicate that the chains appear to have reached stationary distributions, and the chains have good mixing and are dense. The autocorrelation plots indicate low autocorrelation and efficient sampling. The marginal posterior distributions of regression parameters are very close to a normal distribution whereas those for dispersion parameters are positively skewed with a relatively long right tail because of the occasional very large samples. The probable reason for this may be attributed to three factors; i.e., the sample size, sample mean, and the use of a non-informative prior with a large variance for the dispersion parameters. In Chapter VII, we will investigate the bias and variability associated with the estimates of dispersion parameters using both non-informative prior and weakly-informative prior for various combinations of sample sizes and sample-mean values.

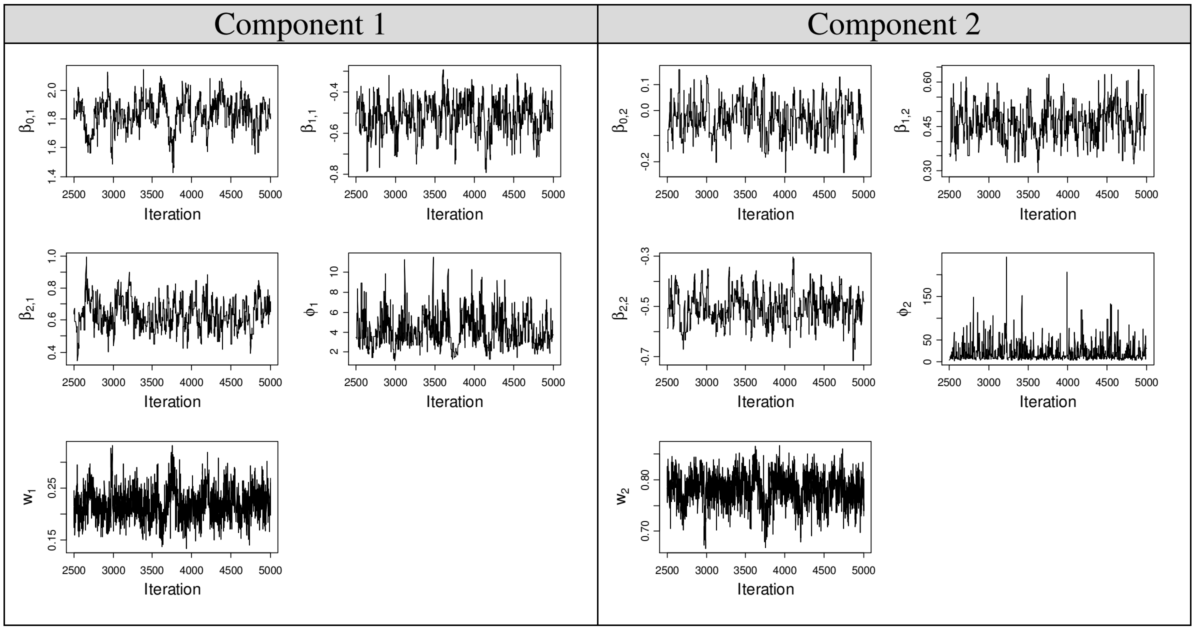


Figure 4.14 MCMC trace plots (FMNB-2)

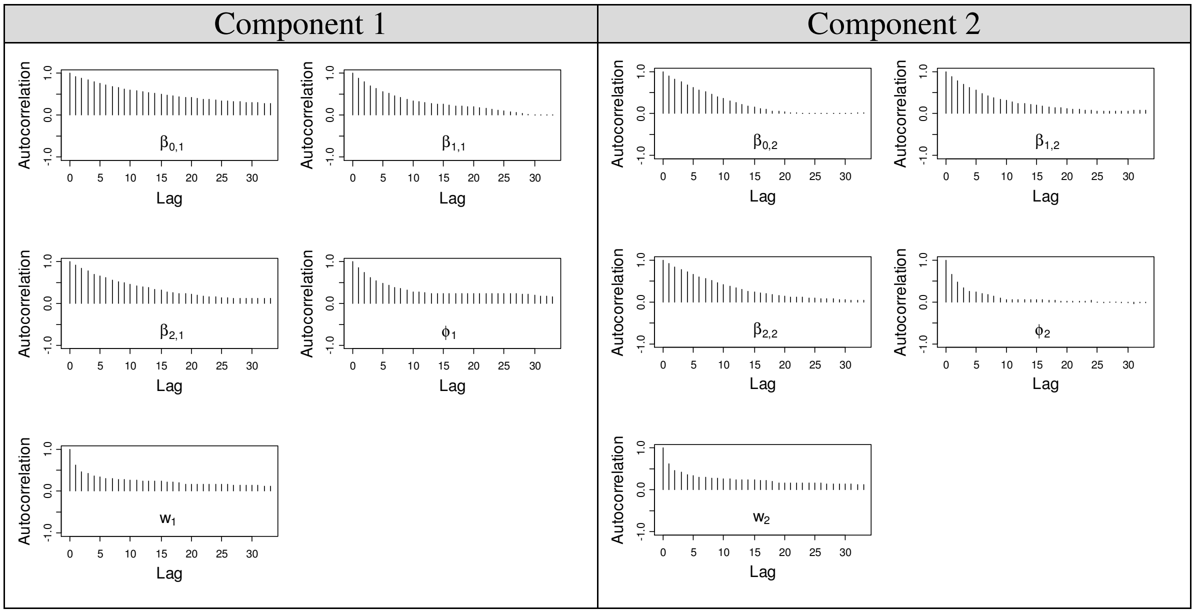


Figure 4.15 Autocorrelation plots (FMNB-2)

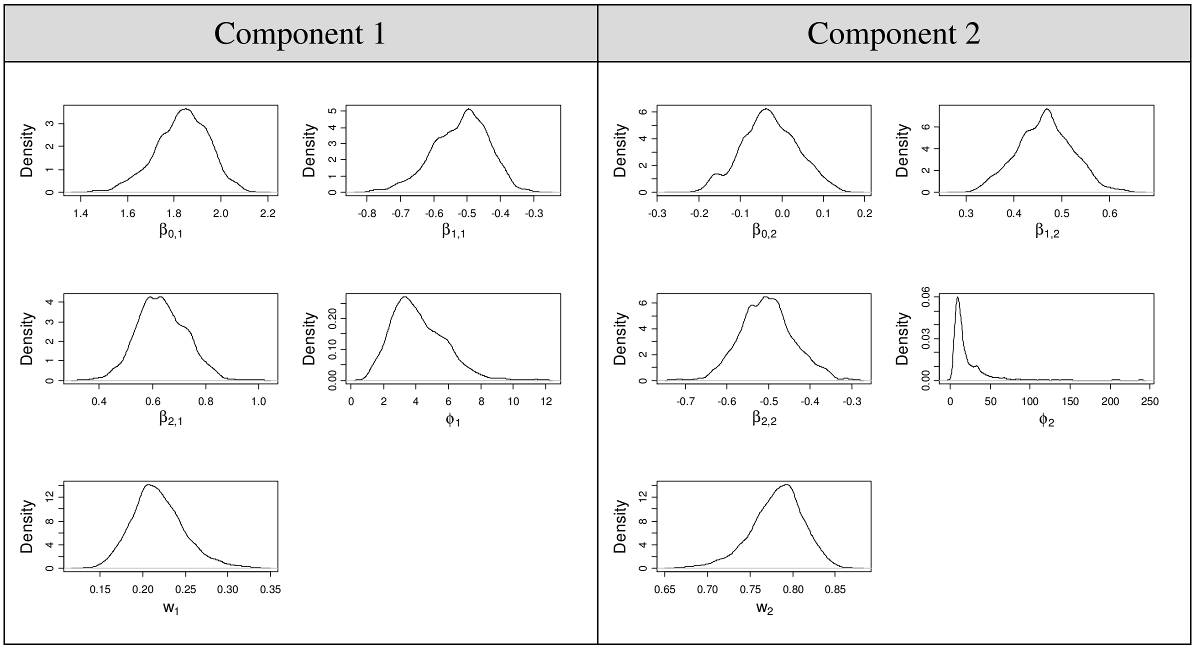


Figure 4.16 Marginal posterior distributions (FMNB-2)

4.4 Chapter Summary

In this chapter, we have shown with Example 1 that the continuous mixture of Poisson/Gamma (NB) model could be effectively approximated with the finite mixture of Poisson regression models with a few numbers of support points and their respective weights without making a distributional assumption on the mixing variable. The necessary number of components was as a function of sample size, sample mean value, and the degree of dispersion. It has been also demonstrated with two examples that the use of standard NB regression models was disadvantageous because it was incapable of addressing the existence of population heterogeneity when the data were actually generated in the form of the finite mixture distribution. The implications could be poor prediction performance and poor interpretation of parameters. The satisfactory goodness-of-fit produced by the NB models could mask the possibility of the different effects of covariates on the count frequency. The FMP-2 model could not handle the extra-variation within components which may be often the case in vehicle crash data. In both cases, the interpretation of the model could be misleading. Instead, FMNB-2 model was shown to be a good candidate model by effectively capturing the population heterogeneity by estimating separate sets of regression parameters and dispersion parameters for sub-components. In the following chapter, we will apply the finite mixture models with more than two components to real crash data.

CHAPTER V

APPLICATION TO EMPIRICAL CRASH DATA

In the previous chapter, we have shown using hypothetical examples that the standard NB regression model is not a viable option if the source of the over-dispersion is due to the population heterogeneity, and the finite mixture models were good alternatives to address this unobserved heterogeneity. At this moment, it is worth a while to recall the underlying assumption in the finite mixture regression models – that is, it assumes that there are a finite number of unobservable categories of observations and the heterogeneity arises from different values of regression coefficients caused by missing variables. In fact, there are many reasons to expect the existence of different subpopulations in vehicle crash data since the crash data are generally collected from various geographic, environmental and geometric design contexts over some fixed time periods. In such cases, group membership of the individual roadway segments (or intersections) is usually unknown or latent, and hence population heterogeneity is unobserved.

The objective of this chapter is to apply the finite mixture regression models to actual vehicle crash data and to demonstrate the effectiveness in discerning the underlying distinctions in the data if they exist. The results of these models will be compared with those produced from the standard NB regression model in terms of various aspects.

Two datasets are considered for application: intersection crash data (Section 5.1) and segment crash data (Section 5.2). For intersection crash data, Park and Lord (2009) applied the finite mixture models to the signalized intersection crash data in Toronto and showed the potential advantage of the FMNB-2 model in addressing the unobserved heterogeneity as well as providing useful information on features of the population. We will utilize the same dataset in this study, but with a different mean functional form for component models. As we will see shortly, many different mean functional forms have been suggested for traffic flow-only models.

For segment crash data, this study will utilize the rural multilane segment crash data for divided highways in California and Texas, which were analyzed during the NCHRP 17-29 project (Lord et al., 2009). While the intersection dataset contains only traffic flow variables, the segment dataset has more covariates (such as median width, shoulder width) as well as traffic flow. Therefore, we can examine whether or not the mixture models would work better even in a more fully-specified model.

5.1 Intersection Crash Data Analysis

This section presents the dataset and the analysis results for intersection crash data.

5.1.1 Data description

To test the applicability of the finite mixture models for intersection crash data, data collected for 1995 at urban 4-legged signalized intersections in Toronto, Canada were used. Even though the data are a little outdated, there are two main reasons for using this dataset. First, the data have been extensively used for various study purposes and have been found to be of relatively good quality (Lord, 2000; Persaud et al., 2002; Miaou and Lord, 2003; Lord et al., 2008). Second, more importantly, despite many factors that may have influenced crash occurrences around and within intersections, the dataset contains only traffic flows for major and minor approaches. There are many evidences in this dataset to support the idea that the un-modeled heterogeneity could have come from the existence of the several different sub-populations. For example, the data were collected across different business environments (e.g. shopping centers, schools, office compounds, etc.). The data contain a mix of fixed and actuated traffic signals with permissive, semi-protected, and protected left turns. It also includes divided and undivided approaches with different speed limits and different number of approaching lanes. Therefore, once the mixture model is estimated with several sub-components, one can go back to the data and see if there are common traits among the different observations that have separated the dataset (if those variables are available).

The summary statistics for the data are provided in Table 5.1. It contains 868 intersections, which have a total of 10,030 reported crashes. Individual intersections experienced crashes from 0 to 54 crashes which resulted in the sample mean equals 11.56 (crashes/intersection) and the variance around 100 (crashes/intersection)^2. Since the type of crashes includes both injury and non-injury crashes, the sample mean value exhibit pretty high. The observed crash frequency plot is shown in Figure 5.1. Entering traffic volumes vary widely from intersections to intersections: from about 5469 to 72,178 vehicles/day for major approaches and from 53 to about 42,644 vehicles/day for minor approaches. The intersection crashes defined here include both intersection and intersection-related crashes as reported by the police that are located within about 15 m (50 ft) from the center of the intersection. For more detailed descriptions of the dataset, the readers are referred to Lord (2000).

Table 5.1 Summary statistics for intersection dataset

Variable	Maximum	Minimum	Average	Standard Deviation
Major-Approach AADT (F_1), (veh/day)	72,178	5,469	28,045	10,660
Minor-Approach AADT (F_2), (veh/day)	42,644	53	11,010	8,599
Crashes	54	0	11.56	10.02

5.1.2 Mean functional form for component model

Despite many factors that may influence crash occurrences around and within intersections, may transportation safety analysts have often favored using traffic flow-only models over models with covariates, even though the former models may be affected by the omitted variables bias (Hauer, 1997; Persaud et al., 2001). They are often preferred over models that include several covariates because they can be easily recalibrated when they are developed in one jurisdiction and applied to another (Persaud et al., 2002; Lord and Bonneson, 2005). As initially discussed by Miaou and Lord (2003) and later confirmed by Mitra and Washington (2007), the un-modeled heterogeneity across sites might be structured spatially in some way, especially when a limited number of covariates are used in the model. The finite mixture regression models assume that part of the heterogeneity come from the existence of the several different sub-populations because of the omitted variables and/or the interaction between the observed and the omitted variables.

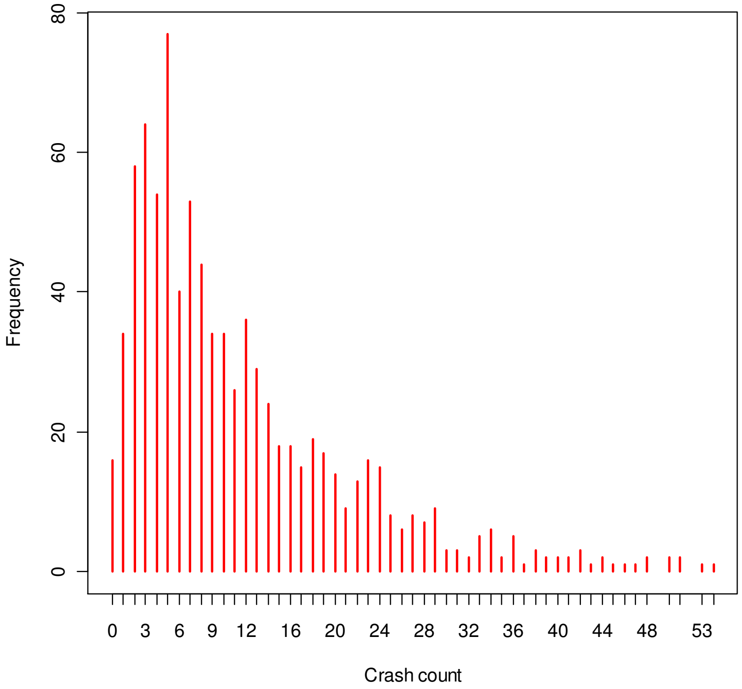


Figure 5.1 Observed crash frequency plot, intersection dataset

Within traffic flow-only models for intersections, Miaou and Lord (2003) listed commonly-used five functional forms based on the previous studies and proposed an additional one:

Form 1: $\mu_i = \beta_0 (F_{1i} + F_{2i})^{\beta_1}$

Form 2: $\mu_i = \beta_0 F_{1i}^{\beta_1} F_{2i}^{\beta_2}$

Form 3: $\mu_i = \beta_0 (F_{1i} F_{2i})^{\beta_1}$

Form 4: $\mu_i = \beta_0 (F_{1i} + F_{2i})^{\beta_1} (F_{2i} / F_{1i})^{\beta_2}$

Form 5: $\mu_i = \beta_0 F_{1i}^{\beta_1} F_{2i}^{\beta_2} \exp(\beta_3 F_{2i})$

Form 6: $\mu_i = F_{1i} \exp(\beta_0 + \beta_1 F_{2i}) + F_{2i} \exp(\beta_0^* + \beta_2 F_{1i})$

Among these functional forms, Form 2 is the most popular one which has been favored most by transportation safety modelers for modeling crash data at intersections. It should be noted, however, it does not appropriately fit the data near the boundary conditions since vehicles on the major approach can still be involved in crashes with vehicles on the same approach even when F_2 is zero.[11] To overcome the boundary value limitation, Miaou and Lord (2003) proposed an alternative form (Form 6) which represents two different risk levels for vehicles entering the two approaches. On the other hand, the functional form 4 was used in Lord and Park (2008) for modeling the three-legged rural intersections in California and was found as the best fitted model amongst others although they didn’t estimate the functional form 6. Note that the functional form 4 also suffers from the same boundary limitations.

Within the finite mixture modeling approach, Park and Lord (2009) used the functional form 2 for a component model and identified the existence of two distinct sub-

[11] “Boundary conditions” refer to (1) when the flow at main approach is close to zero, and (2) when the flow at minor approach is close to zero, or equivalently the flow ratio F_2/F_1 approaches zero (Miaou and Lord, 2003).

populations, each having different degrees of over-dispersion and regression coefficients for the major and minor approach flows. In this study, we apply finite mixture models to the data with a different functional form – that is, the functional form 4 – to examine whether or not the finite mixture models will still reveal the existence of different regression coefficients and degrees of over-dispersion between components. Under this specification the mean functional form for each component is as follows:

$$\mu_{i,k} = \beta_{0,k}(F_{1i} + F_{2i})^{\beta_{1,k}}(F_{2i} / F_{1i})^{\beta_{2,k}} \tag{5.1}$$

where,

$\mu_{i,k} =$ k -th component's estimated number of crashes for intersection i

F_{1i}= entering flows in veh/day from the major approaches at intersection i

F_{2i}= entering flows in veh/day from the minor approaches at intersection i

$\boldsymbol{\beta}_k = (\beta_{0,k}, \beta_{1,k}, \beta_{2,k})'$ = estimated regression coefficients for component k

5.1.3 Model estimation

For the initial check of the standard NB model, it was estimated by the maximum likelihood method. The results are shown in Table 5.2. The Pearson Chi-Square statistic (indicated as Value/DF in the SAS output) is very close to 1 and all model coefficients turn out to be very significant. The estimated dispersion parameter is 7.097 (=0.1409^{-1}). While it is evident that the standard NB model works very well and the model need not be corrected for the over-dispersion, it does not tell us the source of over-dispersion. The purpose of applying the finite mixture models for this dataset is, therefore, to examine the possible existence of different sub-populations and identify the best mixture model which can separate such sub-populations without deteriorating the goodness-of-fit.

We then fitted the data with increasing number of components for FMP-K and FMNB-K models, respectively, until the log of marginal likelihood reached its maximum. Table

5.3 shows the computed values of log-marginal likelihoods along with other model selection criteria. The log-likelihood value (LL) was evaluated at the posterior means of model parameters and the AIC and BIC values were computed based on these log-likelihood values and the number of parameters.

Table 5.2 SAS output for NB model, intersection dataset

Criteria for Assessing Goodness of Fit			
Criterion	DF	Value	Value/DF
Deviance	865	935.5058	1.0815
Scaled Deviance	865	935.5058	1.0815
Pearson Chi-Square	865	875.6736	1.0123
Scaled Pearson X^2	865	875.6736	1.0123
Log Likelihood		17031.3689	
Full Log Likelihood		-2531.1901	

Maximum Likelihood Parameter Estimates					
Parameter	Estimate	Standard Error	Confidence Interval	Wald Chi-Square	Pr>ChiSq
Intercept	-11.0518	0.4927	(-12.0174, -10.0862)	503.23	< .0001
F_1	1.3082	0.0459	(1.2182, 1.3982)	812.13	< .0001
F_2	0.3873	0.0214	(0.3455, 0.4292)	329.01	< .0001
Dispersion	0.1409	0.0122	(0.1169, 0.1649)		

Table 5.3 Model selection criteria, intersection dataset

Models	No. of Parameters	LL	AIC	BIC	DIC	$\log[p(\mathbf{y} \mid M_K)]$
Poisson	3	-2793.9	5593.9	5608.2	5594.2	-2813.9
FMP-2	7	-2558.0	5130.0	5163.3	5130.2	-2596.8
FMP-3	11	-2529.0	5080.0	5132.5	5079.7	-2582.3
FMP-4	15	-2518.8	5067.6	5139.1	5068.0	-2584.8
NB	4	-2531.2	5070.4	5089.5	5070.0	-2556.1
FMNB-2	9	-2527.3	5072.6	5115.5	5061.7	-2560.2
FMNB-3	14	NA	NA	NA	NA	-2571.3

For FMP-K models, while the number of components were identifiable up to K=4 without exhibiting the label switching problem, the largest value of log-marginal likelihood was achieved at K=3. Nevertheless, this value is still much smaller than that of the standard NB model. For FMNB-K models, on the other hand, the models up to K=3 were fitted, but because of unbalanced label switching and difficulty in correcting it even with the identifiability constraints, the FMNB-3 model was not identifiable. Under this situation the computing of the posterior means for each parameter is invalid, which made the log-likelihood and the information criteria unavailable. The marginal likelihood, however, can be computed since it is insensitive to label switching. The largest marginal likelihood was achieved with the FMNB-2 model, and it is comparable to that of the standard NB model.

The results for each Poisson mixture model (from K=2 to K=3) are given in Table 5.4.

Table 5.4 Modeling results for FMP-K models (Bayesian method), intersection dataset

FMP-K models			w_k	$Ln(\beta_{0,k})$	$\beta_{1,k}$	$\beta_{2,k}$
Poisson	Post. mean (Std. dev.)		1.0	-11.0030 (0.3047) [a]	1.3034 (0.028)	0.3844 (0.0144)
FMP-2	Comp. 1	Post. mean (Std. dev.)	0.504 (0.039)	-12.0801 (0.7323)	1.3708 (0.0669)	0.4497 (0.0314)
	Comp. 2	Post. mean (Std. dev.)	0.496 (0.039)	-10.1577 (0.4658)	1.2460 (0.0425)	0.3328 (0.0233)
FMP-3	Comp. 1	Post. mean (Std. dev.)	0.555 (0.044)	-11.7397 (0.6811)	1.3570 (0.0618)	0.4233 (0.0311)
	Comp. 2	Post. mean (Std. dev.)	0.347 (0.050)	-9.9319 (0.5884)	1.2310 (0.0539)	0.3109 (0.0285)
	Comp. 3	Post. mean (Std. dev.)	0.098 (0.030)	-6.3917* (3.4172)	0.7851 (0.3212)	0.5225 (0.1325)

NOTE: MCMC iterations=50,000; Burn-in iterations=25,000; N=868 intersections; Non-informative priors for the regression parameters and the weight distribution were assumed; [a] indicates the standard deviation of the posterior mean; *indicates the coefficient whose 95% credible interval includes zero.

For the sake of comparison, the results from the maximum likelihood estimation method (via EM algorithm) are also provided in Table 5.5. Those were obtained by using the FlexMix package in R (R development Core Team, 2006). As can be seen from the two tables, the posterior means are very close to the maximum likelihood estimates (MLEs) due to the relatively large sample size (N=868) and the use of non-informative priors for regression parameters and weight distribution. Also note that the asymptotic standard errors for MLEs are very similar to the standard deviations for posterior means since the marginal posterior distributions of regression parameters are very close to a normal distribution.

Table 5.5 Modeling results for FMP-K models (Frequentist method), intersection dataset

FMP-K models			w_k	$Ln(\beta_{0,k})$	$\beta_{1,k}$	$\beta_{2,k}$
Poisson	MLE (Std. err.)		1.0	-11.0270 (0.3014) [a]	1.3055 (0.0277)	0.3828 (0.0140)
FMP-2	Comp. 1	MLE (Std. err.)	0.502 (NA)	-12.1396 (0.7230)	1.3766 (0.0661)	0.4467 (0.0308)
	Comp. 2	MLE (Std. err.)	0.498 (NA)	-10.1925 (0.4687)	1.2493 (0.0428)	0.3312 (0.0228)
FMP-3	Comp. 1	MLE (Std. err.)	0.558 (NA)	-11.7032 (0.6821)	1.3539 (0.0619)	0.4186 (0.0298)
	Comp. 2	MLE (Std. err.)	0.342 (NA)	-9.9255 (0.5826)	1.2303 (0.0533)	0.3084 (0.0280)
	Comp. 3	MLE (Std. err.)	0.100 (NA)	-6.6417* (3.4271)	0.8093 (0.3222)	0.5047 (0.1200)

NOTE: Each mixture model was estimated after 20 different initializations and choosing the one with the maximum likelihood; [a] indicates the asymptotic standard error of the MLE; *indicates the coefficient which is not significant at 5% significance level.

For the FMNB-2 model, three cases were estimated: two cases by the Bayesian estimation method with two priors and one case by the maximum likelihood method. The results are shown in Table 5.6. Due to the non-informative priors for the regression parameters, their posterior means are similar to the corresponding maximum likelihood estimates (MLEs). The MLEs were obtained by the nonlinear modeling procedure (NLMIXED) in SAS. For the Bayesian approach, two prior specifications for the dispersion parameter were compared: a non-informative gamma prior ($\phi_k \sim \Gamma(0.01, 0.01)$) and a weakly-informative prior ($\phi_k \sim \Gamma(0.5, 0.1)$). While the justification for using the weakly-informative gamma prior will be described in details in Chapter VII (Simulation Study), it is intended to reduce the implausibly large values of

ϕ_k in the posterior samples, and thereby improve the behavior of the posterior distribution of ϕ_k. As indicated in the table, the use of $\phi_k \sim \Gamma(0.5, 0.1)$ has reduced the posterior means for each component's dispersion parameter and stabilized the estimates greatly by reducing their standard deviations. However, the posterior medians were estimated rather similar to each other, indicating that the posterior medians were less influenced by the choice of priors. This may also imply that the posterior median is a more consistent estimator than the posterior mean. This speculation will be confirmed by simulation study in Chapter VII. On the other hand, the maximum likelihood estimate of the dispersion parameter for component 2 was estimated much larger than that from the Bayesian method. This tendency was also observed in Example 3 in Chapter IV.

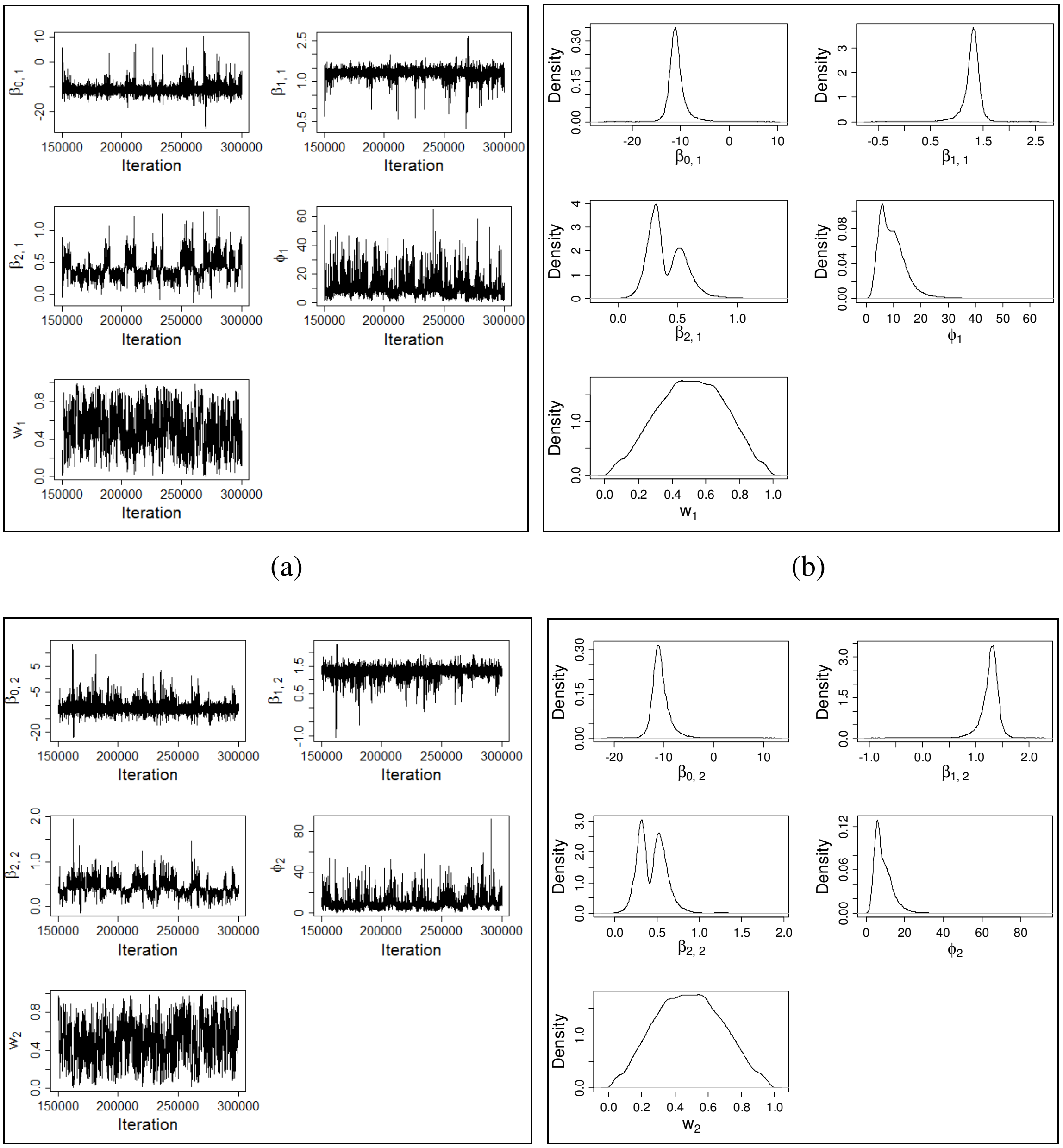
Table 5.6 Modeling results for FMNB-2 models, intersection dataset

Model Parameters	NB	FMNB-2 $\phi_k \sim \Gamma(0.01, 0.01)$		FMNB-2 $\phi_k \sim \Gamma(0.5, 0.1)$		FMNB-2 (MLE)	
		Comp. 1	Comp. 2	Comp. 1	Comp. 2	Comp. 1	Comp. 2
$\hat{\beta}_{0,k}$	-11.0154 (0.4727)	-10. 1445 (1.9007) [a]	-11.3797 (1.1972)	-9.9416 (2.3049)	-11.3708 (1.2611)	-10.9214 (1.0554) [b]	-11.1492 (1.1766)
$\hat{\beta}_{1,k}$	1.3047 (0.0441)	1.2208 (0.1802)	1.3387 (0.1121)	1.2013 (0.2173)	1.3376 (0.1189)	1.2913 (0.0966)	1.3246 (0.1053)
$\hat{\beta}_{2,k}$	0.3873 (0.0206)	0.5451 (0.0961)	0.2831 (0.0639)	0.5566 (0.1176)	0.2914 (0.0622)	0.4935 (0.0658)	0.2577 (0.0767)
$\hat{\phi}_{k,\,mean}$	7.056 (0.612)	7.991 (7.589)	13.176 (8.472)	6.918 (3.489)	11.641 (4.799)	6.825 (1.453)	19.090 (11.019)
$\hat{\phi}_{k,\,median}$	7.021 (0.612)	6.401 (7.589)	11.785 (8.472)	6.206 (3.489)	11.108 (4.799)		
$\hat{w}_k$	-	0.461 (0.181)	0.539 (0.181)	0.433 (0.187)	0.567 (0.187)	0.622 (0.231)	0.378 (0.231)
Model Comparison Criteria							
-2LL	5062.3	5054.6		5052.5		5048.3	
AIC	5070.4	5072.6		5070.5		5066.3	
BIC	5089.5	5115.5		5113.4		5109.2	
DIC	5070.0	5061.7		5064.4		-	
Log(ML)	-2556.1	-2560.2		-2557.0		-	

NOTE: MCMC iterations=300,000; Burn-in iterations=150,000; Non-informative priors for the regression parameters were assumed; Priors for the dispersion parameters and the weight distribution were described in the text; [a] indicates the standard deviation of the posterior mean; [b] indicates the asymptotic standard error of the MLE.

There is a noticeable difference in the estimates of weight distribution, $\hat{w}_k$, between the two methods (i.e. Bayesian vs. MLE). The usual prior choice for the weight distribution so far has been the non-informative $Dirichlet\,(1, 1)$ distribution. However, when this prior was used for this dataset, one of the components with an empty observation was often produced after some period of MCMC iterations. It was also noticed that this had an undesirable effect on the estimation of $\boldsymbol{\beta}_k$ and ϕ_k. One of the advantages of a Bayesian approach is we can avoid solutions with empty components by using proper priors. Because of non-regular characteristics of the likelihood if component k is not observed, Frühwirth-Schnatter (2006) suggests using $e_0 > 1$ in order to pull the posterior of $\mathbf{w}$ away from the boundary of the parameter space. Following the suggestion, $e_0 = 4$ was used for this dataset, which is considered a mildly informative prior on the weight distribution. This may have caused the difference between the two methods.

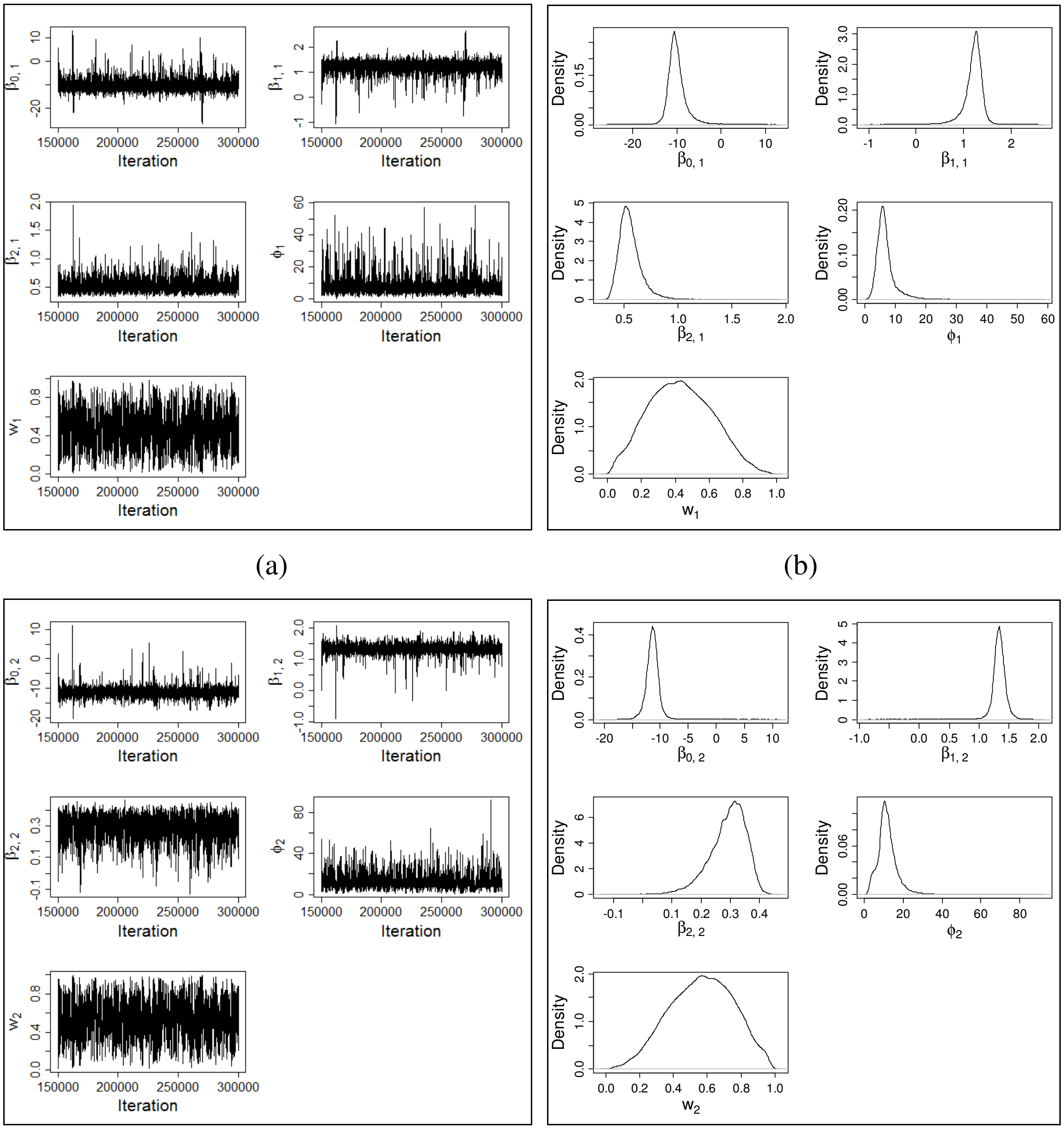
Regarding the label switching issue, the MCMC samples for the FMNB-2 model exhibited the frequent label switches between components during the course of MCMC iterations. Figure 5.2 shows the example of the MCMC trace plots and marginal posterior densities of model parameters when no identifiability constraint was imposed. It is clear from the figure that we should not take the average of the posterior samples for a parameter estimate before correcting the label switching problem. It is also evident that a sensible identifiability constraint can be imposed on $\beta_{2,\,k}$. Figure 5.3 provides the corrected MCMC trace plots and marginal posterior densities after imposing the identifiability constraint (i.e. $\beta_{2,\,1} > \beta_{2,\,2}$). While the label switching problem has been corrected, the MCMC trajectories for the regression parameters are showing occasional distant excursions from the normal course, which rendered their marginal posterior densities skewed to the left or right. It was observed that these digressions occurred when no or very small number of observations were assigned to one of the components (i.e. $w_k \approx 0$). The MCMC trajectories for w_k indicate that one of the components with an empty observation was often produced during the course of MCMC iterations. This may imply that the data are ill-separated and the component centroids are not sufficiently separated for this particular dataset.



(a) (b)

(c) (d)

Figure 5.2 Unconstrained MCMC trace plots and marginal posterior densities: (a) and (b) for component 1, (c) and (d) for component 2. This is an example for the FMNB-2 model with $\phi_k \sim \Gamma(0.5, 0.1)$



(a)

(b)

(c)

(d)

Figure 5.3 Corrected MCMC trace plots and marginal posterior distributions: (a) and (b) for component 1, (c) and (d) for component 2. This is an example for the FMNB-2 model with $\phi_k \sim \Gamma(0.5, 0.1)$

5.1.4 Discussion of the results

In a strict sense, no finite mixture model could be selected as the best model in terms of model selection criteria as compared to the standard NB model for this application dataset. NB model itself produced a very satisfactory goodness-of-fit. This is why those criteria favored the simpler model. However, the difference between NB model and FMNB-2 model is considered small, and the discrepancy of predicted frequencies between two models is negligible as shown in Figure 5.4 (The result is for the FMNB model with $\phi_k \sim \Gamma(0.5, 0.1)$).

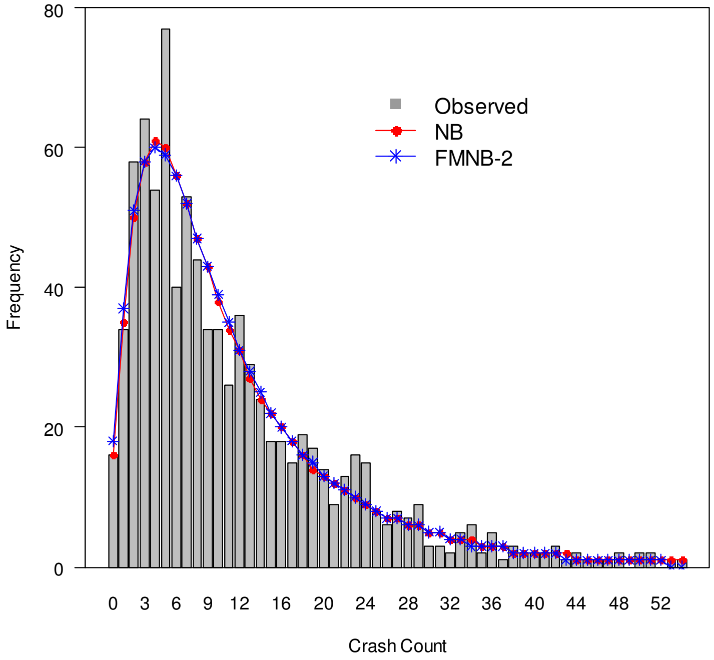


Figure 5.4 Goodness-of-fit comparison between NB and FMNB-2

Furthermore, the FMNB-2 model can provide more opportunities for interpretation of the dataset not available from the standard NB model. First, it seems that the data were generated from two sub-populations with each population having different regression coefficients and degrees of dispersion although the separation was not so distinct in this case. Different regression parameters resulted in different sample averages of fitted means for each component: i.e., $\bar{\mu}_1 = 9.78$ (crashes/year) for Component 1 and $\bar{\mu}_2 = 12.49$ (crashes/year) for Component 2.[12] This indicates that Component 1 is associated with smaller-mean value observations and Component 2 with higher-mean value observations. The over-dispersion parameter in the NB model (7.056) has been split into two values: i.e., 6.918 for Component 1 and 11.641 for Component 1. This indicates that the observations in Component 1 are more dispersed than those in Component 2.

The different estimates of the over-dispersion parameters for Component 1 and Component 2 in the FMNB-2 model have resulted in a different effect on the variance. To visualize such effects, Figure 5.5 shows the variance function of the two models with respect to the corresponding mean values. The variance function for the NB model follows the quadratic function of the mean values with a constant over-dispersion parameter which is estimated from Equation (3.16), whereas that of the FMNB-2 model does not follow a simple curve but can be estimated from Equation (2.43). The figure shows that the variance function of the FMNB-2 model is very close to a quadratic function with a few observations deviant from the general trend. Although the difference in variance between the two models is small in this case, the FMNB-2 model characterizes the uncertainty more accurately in the number of an intersection's crashes.

Second, after assigning each intersection into the component with the highest posterior probability using Equation (3.57), we can identify the population heterogeneity more clearly due to the different effects of covariates on crash frequency. We can see from Figure 5.6 that there is a clear different effect of total traffic flow (F_1+F_2) on Component 1 and Component 2. The total traffic flow has much greater impact on the crash frequency for Component 2 than for Component 1. Figure 5.7 also shows the different

[12] $\bar{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N}\mu_{i,1} = \frac{1}{N}\sum_{i=1}^{N}\exp(\mathbf{x}_i \cdot \hat{\boldsymbol{\beta}}_1)$, $\bar{\mu}_2 = \frac{1}{N}\sum_{i=1}^{N}\mu_{i,2} = \frac{1}{N}\sum_{i=1}^{N}\exp(\mathbf{x}_i \cdot \hat{\boldsymbol{\beta}}_2)$

effects of traffic flow ratio (F_2/F_1) on crash frequency on each component. This type of information is usually difficult to obtain from the standard NB regression models unless we divide the data into components based on a pre-specified criterion and estimate the parameters separately. If additional descriptive data are available, it enables us to go back to the data and relate those variables with the separation of the data.

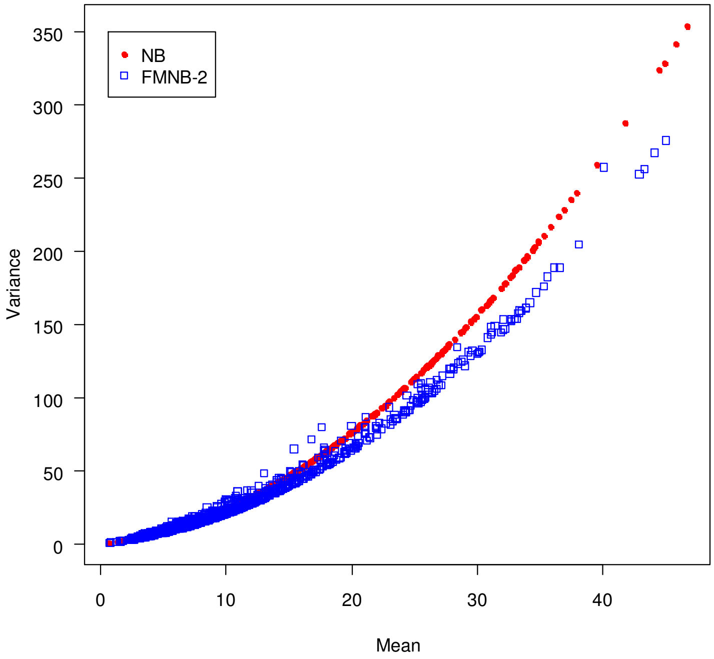


Figure 5.5 Mean-variance relationships for NB and FMNB-2 models

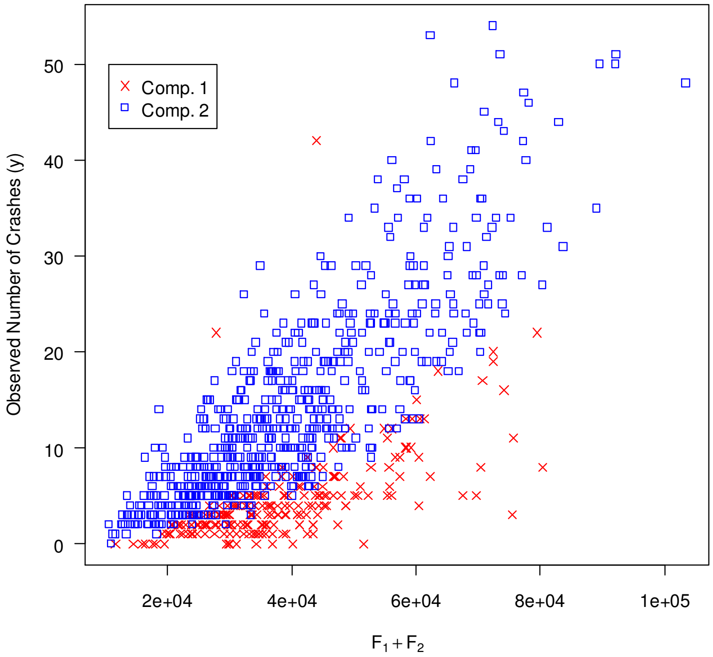


Figure 5.6 Effects of total traffic flow (F_1+F_2) on crash frequency

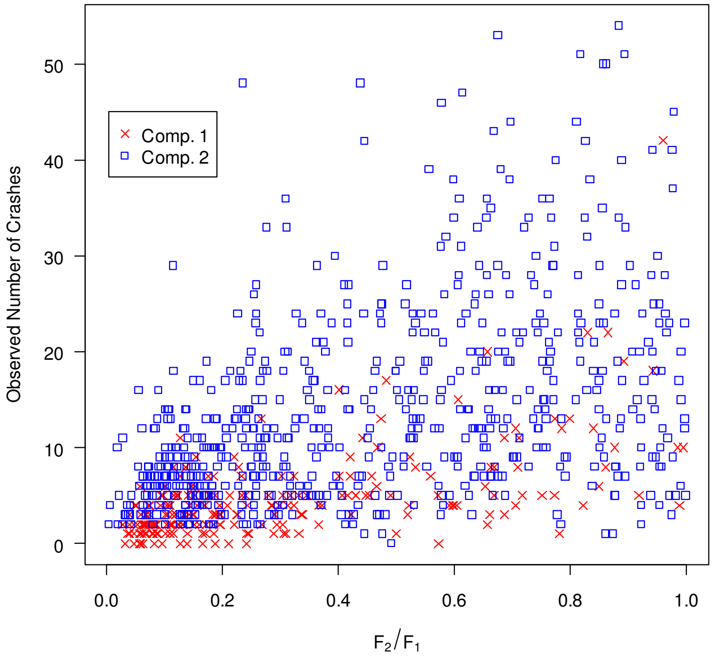


Figure 5.7 Effects of traffic flow ratio (F_2/F_1) on crash frequency

5.2 Segment Crash Data Analysis

This section presents the dataset and the analysis results for segment crash data.

5.2.1 Data description

For segment crash data this study utilized the rural multilane segment data for divided highways in California and Texas. The data were analyzed during the NCHRP 17-29 project (Lord et al., 2009). The California data were originally obtained from the FHWA's HSIS maintained by the University of North Carolina, and the Texas data were from the Department of Public Safety (DPS) and the Texas Department of Transportation (TxDOT). The dataset contained a total of 2,587 roadway segments with 12-ft lane width only in order to estimate the NB regression models with baseline conditions, and used for developing accident modification factors for divided rural multilane highways. The same dataset was used in this study to test the applicability of the finite mixture regression models. It is suitable for application because the data were geographical combined. Since we know the area to which each roadway segment belongs, we could build the separate models for California and Texas. However, this assumes, *a priori*, that the area is a main source of heterogeneity. However, we do not know whether similar groupings of the sample can be made on the basis of individual-level segments, or whether statistical criteria of model fit would suggest that fewer or more groupings are optimal. Therefore, a general modeling strategy using the finite mixture models is more appropriate when drawing inferences from such heterogeneous count data.

Table 5.7 shows the summary statistics of the input data for modeling. Unlike the intersection crash data modeling, more covariates such as segment length, median width and shoulder width along with traffic flow were used when developing the appropriate models. Figure 5.8 shows the histograms of the observed crash frequency.

Table 5.7 Summary of statistics for segment dataset

Variable	Maximum	Minimum	Average	Standard Deviation
Average AADT (*F*), (veh/day)	89,264	158	13,799	11,281
Segment length (*L*), (mile)	11.21	0.1	0.82	1.05
Median width[a] (*MW*), (feet)	240	1	47.07	29.41
Right-shoulder width[b] (*RSW*), (feet)	19	0	7.68	1.98
Injury crashes[c]	148	0	3.17	6.30

NOTE: [a] Median width includes the left shoulder widths; [b] Average right-shoulder width (both sides); [c] Injury crashes include only KAB crashes for five to ten years (K=fatal, A=incapacitating injury, and B=non-incapacitating injury).

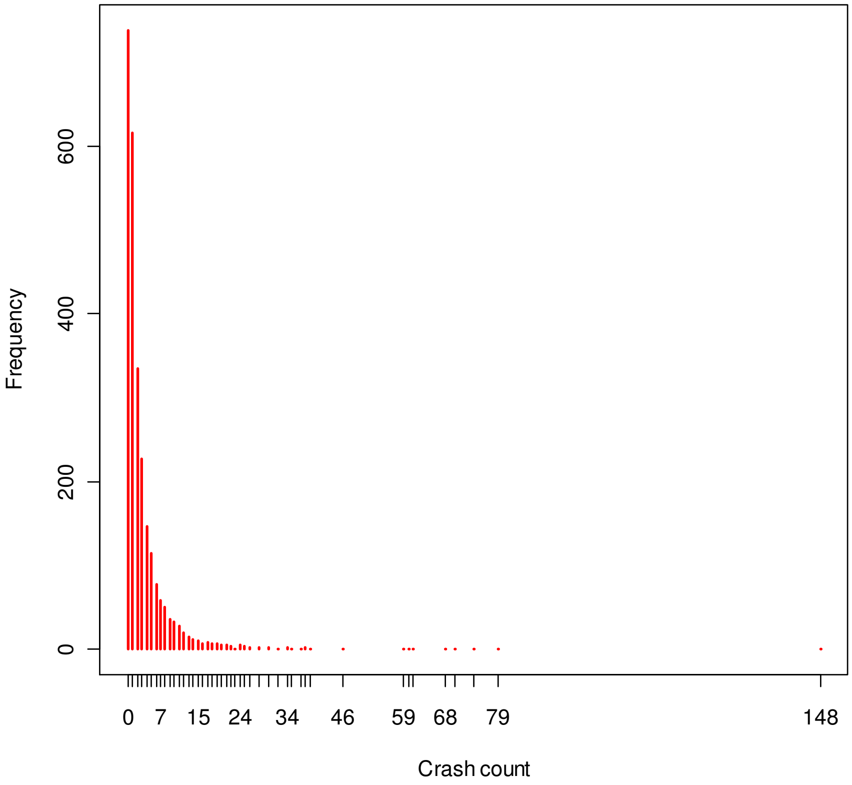


Figure 5.8 Observed crash frequency plot, Segment dataset

5.2.2 Mean functional form for component model

For the component-wise mean functional form for the finite mixture models, the following was used.

$$\mu_{i,k} = t_i \, L_i F_i^{\alpha_k} \exp(\beta_{0,k} + \beta_{1,k} MW_i + \beta_{2,k} RSW_i) \qquad (5.2)$$

where, $\mu_{i,k}$ is the k^{th} component's estimated number of injury crashes per year for segment i, t_i is the number of years, and $\{\alpha_k, \beta_{0,k}, \beta_{1,k}, \beta_{2,k}\}$ are the parameters to be estimated for component *k*. In the present analysis, three measures (traffic flow, median width, and right-shoulder width) were used as independent variables and two variables (segment length and number of years) were used as offset variables. This type of functional form is very common among highway safety analysts (for example, see Lord and Bonneson, 2007; Bonneson et al., 2007). The important characteristics about this functional form are: first, the segment length (*L*) is used as an exposure along with the number of years (t_i) indicating that it is directly proportional to the segment crash frequency; second, rather than using the flow (*F*) as an direct exposure, flow to a power function is used. The exponent α determines the manner in which the segment crash frequency depends on F. For example, if $\alpha < 1$, the number of crashes increases at a decreasing rate as the traffic volume increases.

5.2.3 Model estimation

For the initial check of the goodness-of-fit of the NB model, Table 5.8 shows the results of the maximum likelihood estimation. While the regression parameters are significant at 5% level except for the median width, the Pearson Chi-Square statistic (indicated as Value/DF in the SAS output) appears to be rather high (1.212). According to Hilbe (2007), a correction for over-dispersion may be necessary for this dataset. The over-dispersion with respect to the NB model indicates that there remains a factor in the variance of segment crashes that the NB regression model does not capture. The indicator variable for state (1 for California and 0 for Texas) indicates that the roadway segments in California experience slightly less crashes than those in Texas, assuming everything else the same. However, adding or removing the indicator variable did not change the Pearson X^2 statistic a lot, signifying that there are other factors that influence the variability of the crash occurrence. This supports the application of the finite mixture models and, in this case the finite mixture models are expected to increase the goodness-of-fit as well.

Table 5.8 SAS output for NB model, segment dataset

Criteria for Assessing Goodness of Fit			
Criterion	DF	Value	Value/DF
Deviance	2582	2717.9322	1.0526
Scaled Deviance	2582	2717.9322	1.0526
Pearson Chi-Square	2582	3129.4008	1.2120
Scaled Pearson X^2	2582	3129.4008	1.2120
Log Likelihood		6747.3098	
Full Log Likelihood		-4707.1381	

Maximum Likelihood Parameter Estimates					
Parameter	Estimate	Standard Error	Confidence Interval	Wald Chi-Square	Pr>ChiSq
Intercept	-7.9608	0.2793	(-8.5083, -7.4133)	812.12	< .0001
F	0.8584	0.0258	(0.8078, 0.9090)	1104.62	< .0001
MW	-0.0009	0.0006	(-0.0022, 0.0003)	2.13	0.1444
RSW	-0.0611	0.0099	(-0.0806, -0.0416)	37.85	< .0001
State (Calif)	-0.1795	0.0417	(-0.2611, -0.0978)	18.54	< .0001
State (Texas)	0.0000	0.0000	(0.0000, 0.0000)	-	-
Dispersion	0.3000	0.0201	(0.2613, 0.3403)		

We then fitted the data with increasing number of components for FPM-K and FMNB-K models, respectively, until the log of marginal likelihood reached its maximum. Table 5.9 shows the summary of the computed model selection criteria. Again, the log-likelihood value (LL) was evaluated at the posterior means of model parameters and the AIC and BIC values were computed based on the obtained log-likelihood values and the corresponding number of model parameters. For FMP-K models, the largest value of log-marginal likelihood was achieved at K=4. However, this value is still smaller than that of the standard NB model, which indicates that the FPM-3 model does not perform better than the NB model.

Table 5.9 Model selection criteria, segment dataset

Models	No. of Parameters	LL	AIC	BIC	DIC	$\log(p(\mathbf{y} \mid M_K)$
Poisson	4	-5321.3	10650.5	10673.9	10650.9	-5352.5
FMP-2	9	-4788.5	9595.1	9647.8	9595.6	-4850.1
FMP-3	14	-4697.7	9423.4	9505.5	9423.7	-4784.0
FMP-4	19	-4654.9	9347.8	9459.2	9346.2	-4762.7
NB	5	-4716.3	9442.7	9472.0	9442.9	-4752.2
FMNB-2	11	-4650.3	9322.7	9387.1	9323.2	-4708.3
CFMNB-2	**9**	**-4652.4**	**9322.8**	**9375.6**	**9323.0**	**-4691.7**

For FMNB-K models, the models up to K=3 were fitted, but because of unbalanced label switching and difficulty in correcting it even with the identifiability constraints, it was not considered for model comparison. Instead, two models were estimated and compared. One is a regular two-component mixture model (FMNB-2) and the other is a constrained FMNB-2 model (termed as a CFMNB-2) in which some of the model parameters were constrained to be zero in one component model because their 95% credible interval included zero. The detailed description of each model will be provided later. As shown in Table 5.9, the largest log of marginal likelihood value was achieved with the CFMNB-2 model. The value is significantly larger than that of the standard NB model indicating that the CFMNB-2 model provides a superior goodness-of-fit for this dataset as compared to the NB model.

In what follows, detailed modeling results for FMP-K and FMNB-K models are presented and the results are compared with those from the maximum likelihood method.

Tables 5.10 and 5.11 provide the modeling results for FMP-K models from two approaches. As can be seen from the two tables, for models up to K=3, the posterior means for regression parameters and weight parameters are very close to the maximum likelihood estimates (MLEs) and their standards deviations are very similar to the asymptotic standard errors for the MLEs since the marginal posterior distributions of parameters were very close to a normal distribution. For the FMP-4 model, however, while the regression parameter estimates from the two approaches are similar for components and 1 and 2, those for components 3 and 4 are not very close to each other, especially for the intercept parameter (β_0). This is because, as the number of components increases, components with very small observations are produced and hence their parameter estimates become unstable because of a possible over-fitting.

Table 5.10 Modeling results for FMP-K models (Bayesian method), segment dataset

Poisson Mixture			w	β_0	β_1	β_2	β_3
Standard Poisson	Post. mean (Std. dev.)		1.0	-9.0398 (0.1560) [a]	0.9701 (0.0153)	-0.0021 (0.0004)	-0.0677 (0.0062)
FMP-2	Comp. 1	Post. mean (Std. dev.)	0.867 (0.015)	-8.7890 (0.2162)	0.8815 (0.0209)	0.0004* (0.0006)	-0.0284 (0.0081)
	Comp. 2	Post. mean (Std. dev.)	0.133 (0.015)	-7.6871 (0.4006)	0.9943 (0.0368)	-0.0101 (0.0012)	-0.1277 (0.0098)
FMP-3	Comp. 1	Post. mean (Std. dev.)	0.551 (0.043)	-8.3557 (0.2671)	0.8518 (0.0244)	-0.0004* (0.0008)	-0.0010* (0.0133)
	Comp. 2	Post. mean (Std. dev.)	0.384 (0.044)	-8.0981 (0.6800)	0.7688 (0.0706)	0.0019* (0.0016)	-0.0495 (0.0196)
	Comp. 3	Post. mean (Std. dev.)	0.065 (0.013)	-6.4943 (0.7395)	0.8911 (0.0675)	-0.0148 (0.0032)	-0.1116 (0.0145)
FMP-4	Comp. 1	Post. mean (Std. dev.)	0.489 (0.053)	-8.3343 (0.3342)	0.8444 (0.0308)	0.0002* (0.0008)	0.0064* (0.0136)
	Comp. 2	Post. mean (Std. dev.)	0.391 (0.051)	-8.2260 (0.8095)	0.7720 (0.0819)	0.0039 (0.0013)	-0.0390* (0.0213)
	Comp. 3	Post. mean (Std. dev.)	0.080 (0.023)	-7.0957 (3.4960)	0.9259 (0.3021)	-0.0299 (0.0075)	-0.1384 (0.0818)
	Comp. 4	Post. mean (Std. dev.)	0.040 (0.011)	-7.0907 (0.9341)	1.0101 (0.0856)	-0.0132 (0.0040)	-0.1794 (0.0248)

NOTE: MCMC iterations: 50,000; Burn-in iterations=25,000; N=2,587 segments; Non-informative priors for the regression parameters and the weight distribution were assumed; [a] indicates the standard deviation of the coefficient; *indicates the coefficient whose 95% credible interval includes zero.

Table 5.11 Modeling results for FMP-K models (Frequentist method), segment dataset

Poisson Mixture			w	β_0	β_1	β_2	β_3
Standard Poisson	MLE (Std. err.)		1.0	-9.0494 (0.1535) [a]	0.9705 (0.0149)	-0.0021 (0.0004)	-0.0669 (0.0058)
FMP-2	Comp. 1	MLE (Std. err.)	0.867 (NA)	-8.7971 (0.2146)	0.8824 (0.0209)	0.0004* (0.0006)	-0.0284 (0.0080)
	Comp. 2	MLE (Std. err.)	0.133 (NA)	-7.7102 (0.3942)	0.9964 (0.0363)	-0.0100 (0.0012)	-0.1278 (0.0098)
FMP-3	Comp. 1	MLE (Std. err.)	0.554 (NA)	-8.3050 (0.2599)	0.8471 (0.0236)	-0.0004* (0.0008)	-0.0023* (0.0119)
	Comp. 2	MLE (Std. err.)	0.380 (NA)	-8.0190 (0.6495)	0.7606 (0.0651)	0.0019* (0.0015)	-0.0499 (0.0195)
	Comp. 3	MLE (Std. err.)	0.066 (NA)	-6.8516 (0.6668)	0.9222 (0.0613)	-0.0126 (0.0022)	-0.1139 (0.0120)
FMP-4	Comp. 1	MLE (Std. err.)	0.495 (NA)	-8.2983 (0.2839)	0.8374 (0.0262)	-0.0000* (0.0008)	0.0043* (0.0129)
	Comp. 2	MLE (Std. err.)	0.380 (NA)	-7.8411 (0.6648)	0.7339 (0.0672)	0.0039 (0.0013)	-0.0490 (0.0200)
	Comp. 3	MLE (Std. err.)	0.074 (NA)	-5.8281 (0.8641)	0.7582 (0.0782)	-0.0078 (0.0025)	-0.0678 (0.0150)
	Comp. 4	MLE (Std. err.)	0.051 (NA)	-8.6109 (1.0101)	1.1391 (0.0913)	-0.0348 (0.0045)	-0.1617 (0.0201)

NOTE: Each mixture model was estimated after 20 different initializations and choosing the one with the maximum likelihood; [a] indicates the asymptotic standard error of the coefficient; *indicates the coefficient which is not significant at 5% significance level.

As can be seen from the two tables, for models up to K=3, the posterior means for regression parameters and weight parameters are very close to the maximum likelihood estimates (MLEs) and their standards deviations are very similar to the asymptotic standard errors for the MLEs since the marginal posterior distributions of parameters were very close to a normal distribution. For the FMP-4 model, however, while the regression parameter estimates from the two approaches are similar for components and 1 and 2, those for components 3 and 4 are not very close to each other, especially for the intercept parameter (β_0). This is because, as the number of components increases, components with very small observations are produced and hence their parameter estimates become unstable because of a possible over-fitting.

For the FMNB-K models, the FMNB-2 model was initially fitted, but it was found that the estimates for median width and shoulder width in one component (specifically, a smaller-mean component) were considered not much different from zero because their 95% credible interval included zero. Therefore, an alternative model (CFMNB-2) was estimated by constraining those parameters to be zero. Tables 5.12 and 5.13 show the modeling results for each model based on the two approaches (i.e., Bayesian vs. Frequentist).

For both models, non-informative priors for the regression parameters and the weight distribution were assumed. For this dataset, the use of a non-informative prior $\mathit{Dirichlet}\,(1,1)$ for the weight distribution did not cause any problems. For the dispersion parameter, two prior specifications were compared as before: a non-informative gamma prior ($\phi_k \sim \Gamma(0.01, 0.01)$) and a weakly-informative prior ($\phi_k \sim \Gamma(0.5, 0.1)$). As shown in both tables, the use of different priors did not have much influence on the posterior means of dispersion parameters although the use of a weakly-informative prior has increased the posterior mean for component 1 slightly upward and reduced the posterior mean for component 2 slightly downward. The same trend is true for the posterior medians, but the disparity from using different priors appear to be minor. This indicates that, because of the large sample size (N=2,587), the posterior means or medians are not much influenced by the choice of priors on the dispersion parameter.

For the CFMNB-2 model, the estimated coefficients are all significant at 5% level, and the population is considered still heterogeneous in that two subpopulations require different component parameters to adequately capture their characteristics. While the changes in the parameter estimates from the FMNB-2 model are considered small, the improvement in the log of marginal likelihood appears to be large. According to the guidelines by Kass and Raftery (1995), the evidence for choosing the CFMNB-2 model over the FMNB-2 model is very strong (see, Table 3.2 in Chapter III).

Table 5.12 Modeling results for FMNB-2 models, segment dataset

Model Parameters	NB	FMNB-2 ϕ_k ~ $\Gamma(0.01, 0.01)$		FMNB-2 ϕ_k ~ $\Gamma(0.5, 0.1)$		FMNB-2 (MLE)	
		Comp 1	Comp 2	Comp 1	Comp 2	Comp 1	Comp 2
$\hat{\beta}_{0,k}$	-8.5574 (0.2397)	-8.5239 (0.3007) [a]	-6.9009 (1.2048) [a]	-8.5272 (0.2862)	-6.8581 (1.2664)	-8.5530 (0.2884) [b]	-7.1243 (1.0500) [b]
$\hat{\beta}_{1,k}$	0.9015 (0.0234)	0.8386 (0.0300)	0.9125 (0.1108)	0.8387 (0.0286)	0.9078 (0.1151)	0.8378 (0.0286)	0.9238 (0.0961)
$\hat{\beta}_{2,k}$	-0.0015 (0.0006)	0.0012* (0.0007)	-0.0186 (0.0055)	0.0013* (0.0008)	-0.0191 (0.0067)	0.0015 (0.0007)	-0.0163 (0.0042)
$\hat{\beta}_{3,k}$	-0.0455 (0.0094)	0.0015* (0.0114)	-0.1548 (0.04137)	0.0014* (0.0113)	-0.1509 (0.0412)	0.0037* (0.0112)	-0.1504 (0.0364)
$\hat{\phi}_{k,\,mean}$	3.225 (0.222)	6.877 (1.167)	2.030 (0.737)	6.7945 (1.143)	2.149 (0.803)	7.223 (1.251)	1.9810 (0.529)
$\hat{\phi}_{k,\,median}$	3.206 (0.222)	6.729 (1.167)	1.881 (0.737)	6.646 (1.143)	1.979 (0.803)		
$\hat{w}_k$	-	0.856 (0.038)	0.144 (0.038)	0.857 (0.040)	0.143 (0.040)	0.836 (0.041)	0.164 (0.041)
Model Comparison Criteria							
-2LL	9432.7	9300.6		9300.7		9300.1	
AIC	9442.7	9322.6		9322.7		9322.1	
BIC	9472.0	9387.0		9387.1		9386.5	
DIC	9442.9	9323.6		9323.2		-	
Log(ML)	-4752.2	-4714.8		-4708.3		-	

NOTE: MCMC iterations=100,000; Burn-in iterations=50,000; Non-informative priors for the regression parameters and the weight distribution were assumed; [a] indicates the standard deviation of the coefficient; [b] indicates the asymptotic standard error of the coefficient; * indicates the coefficient whose 95% credible interval or confidence interval includes zero.

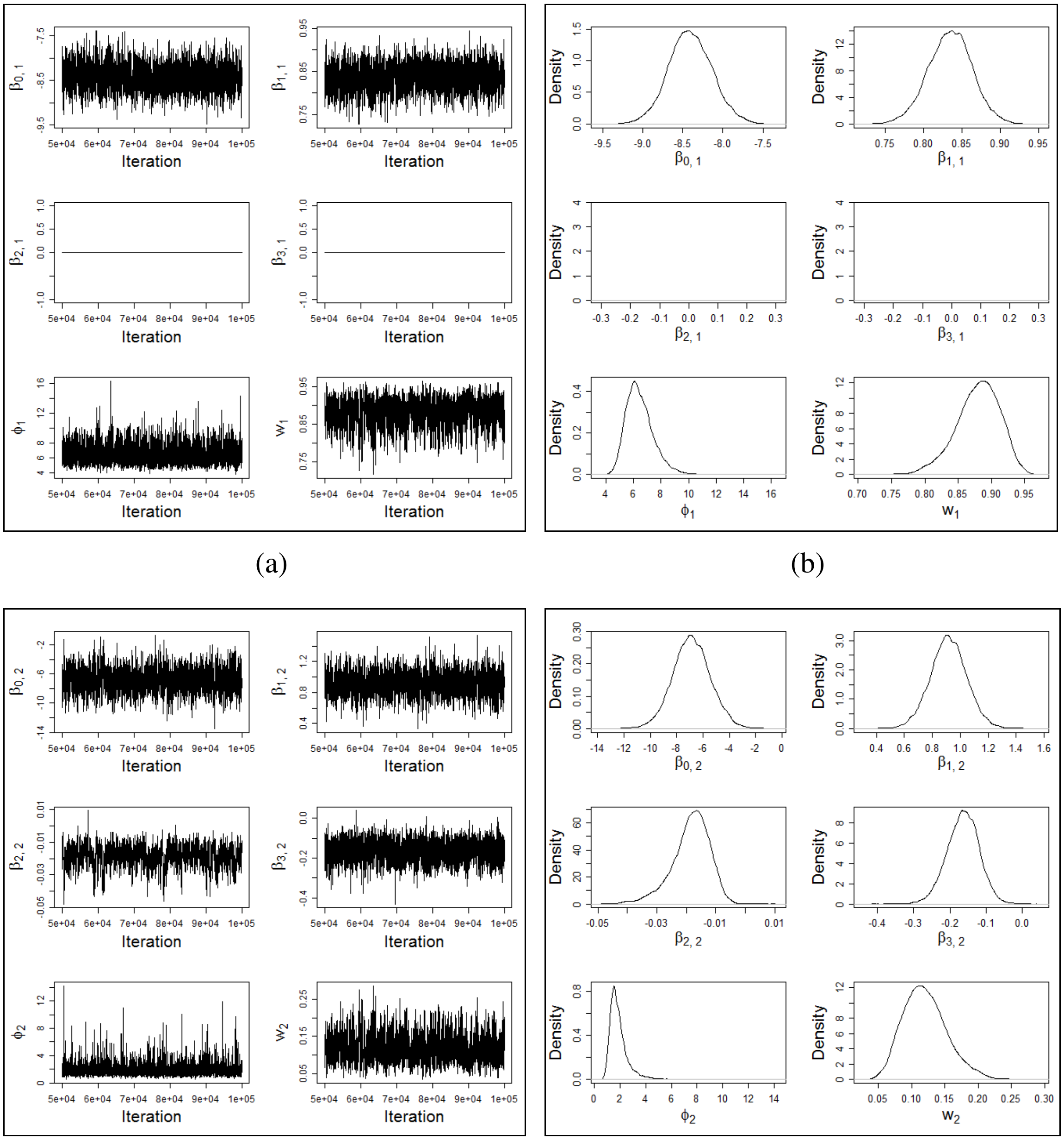
When the results from the Bayesian method were compared with those from the maximum likelihood method, the parameter estimates were overall very similar to each other except the dispersion parameter for a smaller-mean component (Component 1). The maximum likelihood estimate of the dispersion parameter for one component tends to be greater than the corresponding posterior mean from the Bayesian method. This tendency was also true for the intersection crash data in the previous section and for Example 3 in Chapter IV. Therefore, it is evident from the analyses of three datasets (one artificial dataset and two empirical datasets) that, for a FMNB-2 model, the maximum likelihood estimates of the dispersion parameters may be biased by over-estimating the true values. This does not necessarily mean that the Bayesian statistics are unbiased. In this respect, the bias properties of Bayesian statistics (posterior mean or median) will be investigated by a simulation study in Chapter VI.

Table 5.13 Modeling results for CFMNB-2 models, segment dataset

Model Parameters	NB	CFMNB-2 $\phi_k \sim \Gamma(0.01, 0.01)$		CFMNB-2 $\phi_k \sim \Gamma(0.5, 0.1)$		CFMNB-2 (MLE)	
		Comp 1	Comp 2	Comp 1	Comp 2	Comp 1	Comp 2
$\hat{\beta}_{0,k}$	-8.5574 (0.2397)	-8.4226 (0.2584) [a]	-6.8420 (1.3718)	-8.4073 (0.2712)	-6.8646 (1.4437)	-8.4027 (0.2669) [b]	-7.1796 (1.2334)
$\hat{\beta}_{1,k}$	0.9015 (0.0234)	0.8361 (0.0270)	0.9142 (0.1261)	0.8344 (1.4437)	0.9168 (0.1327)	0.8333 (0.0279)	0.9369 (0.1129)
$\hat{\beta}_{2,k}$	-0.0015 (0.0006)	-	-0.0186 (0.0063)	-	-0.0184 (0.0065)	-	-0.0158 (0.0050)
$\hat{\beta}_{3,k}$	-0.0455 (0.0094)	-	-0.1633 (0.0449)	-	-0.1643 (0.0447)	-	-0.1593 (0.0401)
$\hat{\phi}_{k,\,mean}$	3.225 (0.222)	6.483 (0.988)	1.829 (0.750)	6.448 (1.006)	1.893 (0.789)	6.666 (1.051)	1.762 (0.523)
$\hat{\phi}_{k,\,median}$	3.206 (0.222)	6.367 (0.988)	1.658 (0.750)	6.316 (1.006)	1.717 (0.789)		
$\hat{w}_k$	-	0.880 (0.031)	0.120 (0.031)	0.880 (0.033)	0.120 (0.033)	0.866 (0.034)	0.134 (0.034)
Model Comparison Criteria							
-2LL	9432.7	9304.8		9304.8		9304.5	
AIC	9442.7	9322.8		9322.8		9322.5	
BIC	9472.0	9375.6		9375.6		9375.2	
DIC	9442.9	9322.9		9323.0		-	
Log(ML)	-4752.2	-4695.0		-4691.7		-	

NOTE: MCMC iterations=100,000; Burn-in iterations=50,000; Non-informative priors for the regression parameters were assumed; Priors for the dispersion parameters and the weight distribution were described in the text; [a] indicates the standard deviation of the coefficient; [b] indicates the asymptotic standard error of the coefficient.

In order to prevent the label switching problem during the MCMC sampling process, after testing with various constraints on the model parameters, the order constraint on weight parameters (i.e., $w_1 > w_2$) was found to be most appropriate for this dataset. Figure 5.9 shows the example of the MCMC trace plots and marginal posterior densities of the parameters for the CFMNB-2 model with $\phi_k \sim \Gamma(0.5, 0.1)$.



(a) (b)

(c) (d)

Figure 5.9 Constrained MCMC trace plots and marginal posterior density plots: (a) and (b) for component 1, (c) and (d) for component 2. This is an example for the CFMNB-2 model with $\phi_k \sim \Gamma(0.5, 0.1)$.

The MCMC trajectories exhibit good mixing behaviors without particular jumps in the trace plots. The marginal posterior density plots also show the uni-modal shape for each model parameter, which is very close to a normal distribution. For the dispersion parameters (ϕ_1 and ϕ_2), however, it was observed that the occasional very large samples skewed the marginal posterior distribution to the right with a long tail.

5.2.4 Discussion of the results

In the previous subsection, it has been shown that the CFMNB-2 model is the best model for this dataset based on the model selection criteria. The goodness-of-fits of negative binomial (NB) regression and CFMNB-2 models can be visualized, as in Figure 5.10, by comparing the observed and predicted frequencies of each crash count outcome by taking the probability distribution into consideration. For aiding a better visual comparison, the crash count was truncated at 39 counts, but the maximum count was 148 (see Figure 5.8). As compared to the NB model, we can see a moderate improvement at crash counts 0 and 1 after considering the population heterogeneity in CFMNB-2 model (with $\phi_k \sim \Gamma(0.5, 0.1)$).

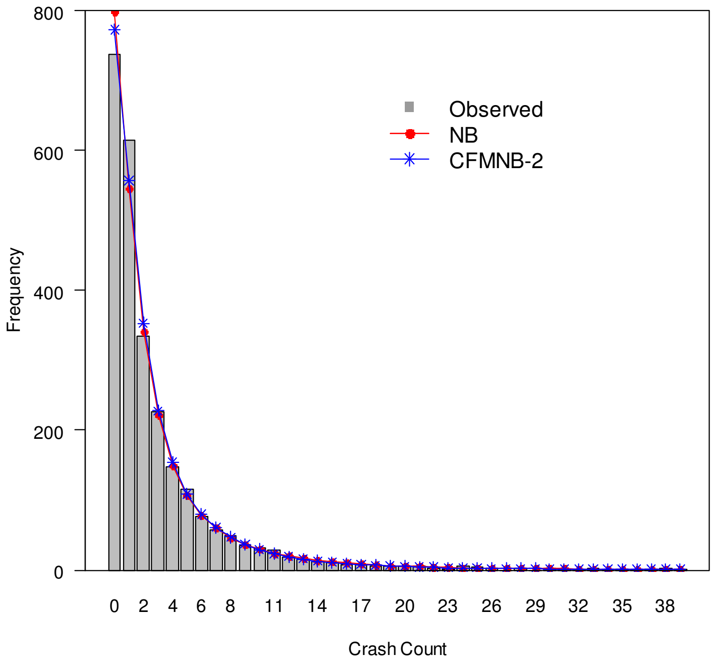


Figure 5.10 Goodness-of-fit comparison between NB and CFMNB-2

While the improvement in goodness-of-fit is moderate, the CFMNB-2 model can provide further information about the population. The population consists of two distinct sub-populations whose regression parameters and degrees of dispersion are different each other. With the coefficients estimated in Table 5.13, the sample averages of the estimated means for Component 1 and Component 2 were computed as $\bar{\mu}_1 = 2.96$ (crashes/year) and $\bar{\mu}_2 = 4.39$ (crashes/year), respectively. This indicates that Component 1 is associated with smaller-mean value observations and Component 2 with higher-mean value observations. The over-dispersion parameter in the NB model (3.225) has been split into two values: i.e., 6.448 for Component 1 and 1.893 for Component 2. This indicates that the higher-mean value observations (Component 2) are more dispersed than the smaller-mean value observations (Component 1).

Figure 5.11 shows the variance functions of the NB and CFMNB-2 models with respect to the corresponding mean values.

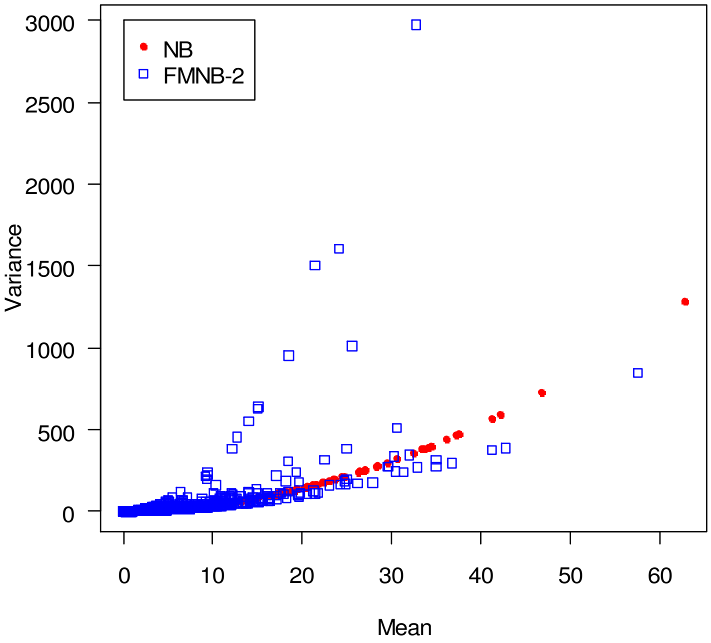


Figure 5.11 Mean-variance relationships for NB and CFMNB-2 models

In this case, it is evident from the figure that many data points for the CFMNB-2 model do not follow the theoretical quadratic function and some of them are much larger than the variances assumed for the NB model. A larger variance does not mean that the CFMNB-2 model is less accurate than the NB model, but it means that the former characterizes the uncertainty more accurately about the crash occurrence at a particular roadway segment.

Figures 5.12, 5.13 and 5.14 show the relationship between each model covariate and crash frequency by component. Again, the posterior probability given in Equation (3.57) was used to segment data by assigning each observation to the component with maximum posterior probability. Note that some information loss was involved in this method because of disregarding the 'fuzziness' of the classifications (Wedel et al., 1993). This resulted in a very small number of observations (sites) assigned to Component 2 which accounts for only about 2%. Nevertheless, the use of a finite mixture model enables us to identify difference in effects of covariates on crash occurrence across observations. For example, as already shown in the parameter estimation results (see Table 5.13), the observations in Component 1 are not likely to be much influenced by median width and right-shoulder width, whereas the observations assigned to Component 2 are significantly affected by those variables. The negative effects by those variables in Component 2 are much higher than the effects from the NB model which considers the average effect of a covariate across all observations in the sample by ignoring the potential presence of population heterogeneity. The difference in the effects of covariates on crash occurrence between the two models will result in different shapes of accident modification factors, which will be investigated in the next chapter.

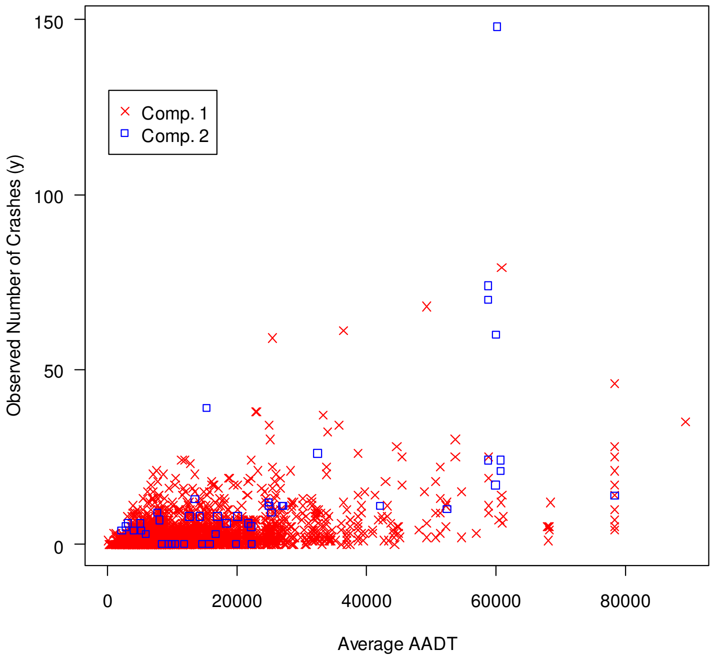


Figure 5.12 Relationship between average AADT and crash frequency by component

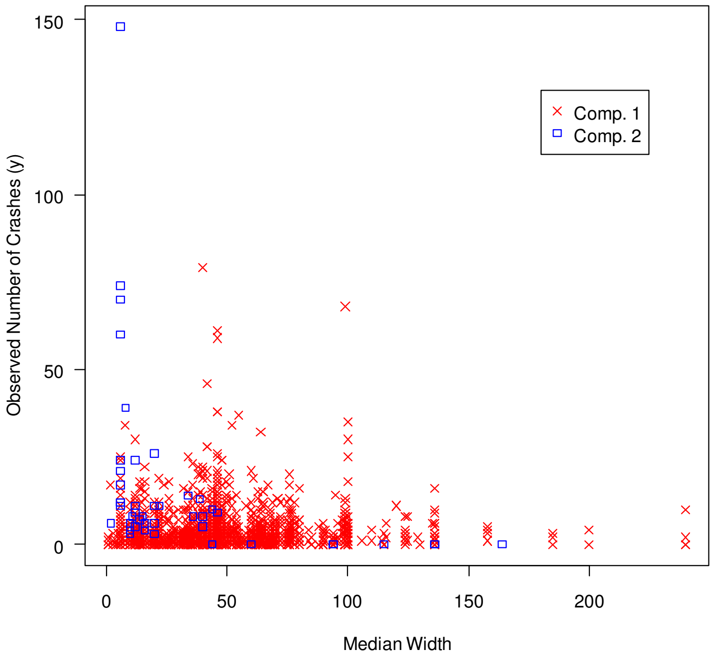


Figure 5.13 Relationship between median width and crash frequency by component

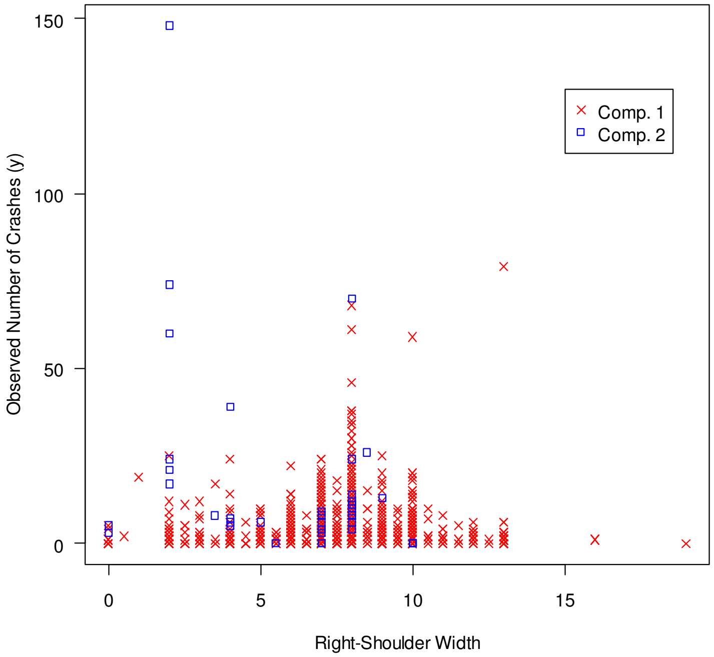


Figure 5.14 Relationship between right-shoulder width and crash frequency by component

5.3 Chapter Summary

In this chapter we have applied the finite mixture regression models to actual vehicle crash data (i.e., intersection crash data and segment crash data) and demonstrated the effectiveness in discerning the underlying distinctions in the data. Both application datasets exhibited the possible presence of several sub-populations in the sample. The FMP-K models generally produced more components than the FMNB-K models, but any of the FMP-K models could not be selected as the best model in terms of model selection criteria. For FMNB-K models, two components seemed to be quite enough to describe the population heterogeneity.

For the intersection crash data, no finite mixture model could be selected as the best model in a strict sense. However, the difference in the log of marginal likelihood value between the NB model and the FMNB-2 model was small, and the discrepancy of predicted frequencies between the two models was negligible. The choice of the FMNB-2 model could provide more opportunities for interpretation of the dataset not available from the standard NB model. It was also noticed that the use of a non-informative $Dirichlet\,(1,1)$ distribution as a prior for the weight distribution caused a problem by producing an empty component too often during the MCMC iterations. This could be indicative of a poor separation for this dataset.

For the segment crash data, the data separation was more distinct and the FMNB-2 (and CFMNB-2) model actually improved the goodness-of-fit as compared to the NB model. When the variance functions of the NB and CFMNB-2 models were compared, it was evident that many data points for the CFNM-2 model did not follow the quadratic function and some of them were much larger than the variances assumed for the NB model. This clearly illustrated that the CFMNB-2 model characterizes the uncertainty more accurately about the crash occurrence at a roadway segment without being restrained to a particular variance function. In the following chapter we apply this model to highway safety analyses such as hotspot identification and the accident modification factor development.

CHAPTER VI

APPLICATION TO HIGHWAY SAFETY ANALYSES

The main theme of the previous chapter was developing the most appropriate statistical model for a given dataset. For example, the CFMNB-2 model was chosen as the best model among many alternative models for the segment crash data. The developed statistical model is usually used as the basis for various traffic safety analyses such as the hotspot identification or development accident modification factors (AMFs). This chapter focuses on the application side of the finite mixture model in evaluating highway safety. Given the superior performance of the finite mixture model for a particular dataset, there is a need to investigate whether this type of model will result in important differences in various highway safety analyses as compared to the standard NB regression model. Among many usages of the statistical models in evaluating highway safety, this section will focus on two application areas: (i) the identification of hotspots or hazardous sites; and (ii) the development of accident modification factors via the coefficients of a model.

Thus, the main objective of this chapter is to examine the relative performance of the two alternative models (i.e., FMNB-2 model and NB model) in terms of their applications to the afore-mentioned two areas. Section 6.1 deals with the hotspot identification and compares the two models with both empirical and simulated data. In Section 6.2, the accident medication factor equations are derived from the two models and their respective characteristics are discussed. Section 6.3 summarizes the chapter.

6.1 Hotspot Identification

This section briefly introduces the hotspot identification in highway safety study and compares the results from the FMNB-2 and NB models.

6.1.1 Introduction

While a hotspot, also referred to as a blackspot (Maher and Mountain, 1988; Elvik, 2007), site with promise (Hauer, 1996; Hauer et al., 2002), or hazardous location, can be generally defined as a location (roadway segment, intersection or interchange) with high crash risk, it has been defined in many different ways depending on how to measure the crash risk at a particular location. For example, Hakkert and Mahalel (1978) proposed that a hotspot be defined as a site that has a crash frequency which is significantly higher than expected at some prescribed level of significance. McGuigan (1981) proposed the use of potential for accident reduction, as the difference between the observed and expected number of crashes at a site given exposure. Recently, Elvik (2008) proposed a theoretical definition of a hotspot as being any location that has a higher expected number of accidents than other similar locations as a result of local risk factors.

A naïve approach to identifying hotspots is to rank locations based on their observed accident frequencies. However, because of a rare and random nature of accident occurrence, this approach tends to be very sensitive to random variations. Miaou and Song (2005) illustrated the limitation of using naïve or raw crash-risk approach in ranking through simple simulations. To better address the random fluctuation of crash occurrence, researchers have used the statistical modeling-based approaches that apply random effect or Bayesian methods and compared their relative performances in identifying hotspots (Miranda-Moreno et al., 2005; El-Basyouny and Sayed, 2006). While many alternative statistical models and ranking criteria are available in the literature for identifying hotspots, a main difficulty arises from the inability to differentiate between sites that are truly high risk and sites that happen to have experienced random fluctuations in crashes during a period of observation (Cheng and Washington, 2005). In this respect, recently some researchers have adopted the epidemiological criteria such as "sensitivity" or "specificity"[13] to compare different statistical models or ranking criteria for identifying hotspots (Cheng and Washington, 2005; Miranda-Moreno, 2006; Elvik, 2008). These criteria can provide information about "false positives" (identifying a safe site as a hotspot) and "false negatives" (identifying a hotspot as a safe site). These criteria along with others will be used later in this section to compare the relative performance of the FMNB-2 and NB models in identifying hotspots. Among many ranking criteria, the following conditional mean of crash frequency assumed for both models will be used:

$$\hat{\mu}_i^{NB} = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}) \text{ (for NB model)} \tag{6.1}$$

$$\hat{\mu}_i^{FMNB-2} = \hat{w}_1\exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}_1) + \hat{w}_2\exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}_2) \text{ (for FMNB-2 model)} \tag{6.2}$$

6.1.2 Comparison by empirical crash data

First, in order to compare the difference in ranking orders between the NB and FMNB-2 models, we calculated $\hat{\mu}_i^{NB}$ and $\hat{\mu}_i^{FMNB-2}$ based on the parameter estimation results in Table 5.13. In this case, $\hat{\mu}_i^{FMNB-2}$ was calculated for the CFMNB-2 model with $\phi_k \sim \Gamma(0.5, 0.1)$. Figure 6.1 illustrates the relationship between the hotspot identification lists ranked by the two models. For comparison, 500 sites were selected from the top of a list sorted according to $\hat{\mu}_i^{FMNB-2}$ values. Smaller values in the ranking order imply more hazardous roadway segments (i.e., higher values in terms of $\hat{\mu}_i^{FMNB-2}$) and vice versa.

The figure shows that there is a strong positive association between the two rankings although the discrepancy in rankings becomes a little larger as the ranking order

[13] $\text{Sensitivity} = \dfrac{\text{number of detected hotspots}}{\text{number of true hotspots}}$, $\text{Specificity} = \dfrac{\text{number of detected non-hotspots}}{\text{number of true non-hotspots}}$

increases. We then compared the ranking orders from the NB model with those from the CFMNB-2 model with differing numbers of total hotspots denoted as m in the figure. The value m represents the total number of hazardous sites selected from the top of the list sorted by the CFMNB-2 model. When 100 sites ($m = 100$) were selected as hotspots from the CFMNB-2 model, six sites of them were not included as hotspots by the NB model. Likewise, 9 sites for $m = 200$, 11 sites for $m = 300$, 10 sites for $m = 400$, and 11 sites for $m = 500$ were excluded from a hotspot list by the NB model, whereas the CFMNB-2 model included them as hotspots.

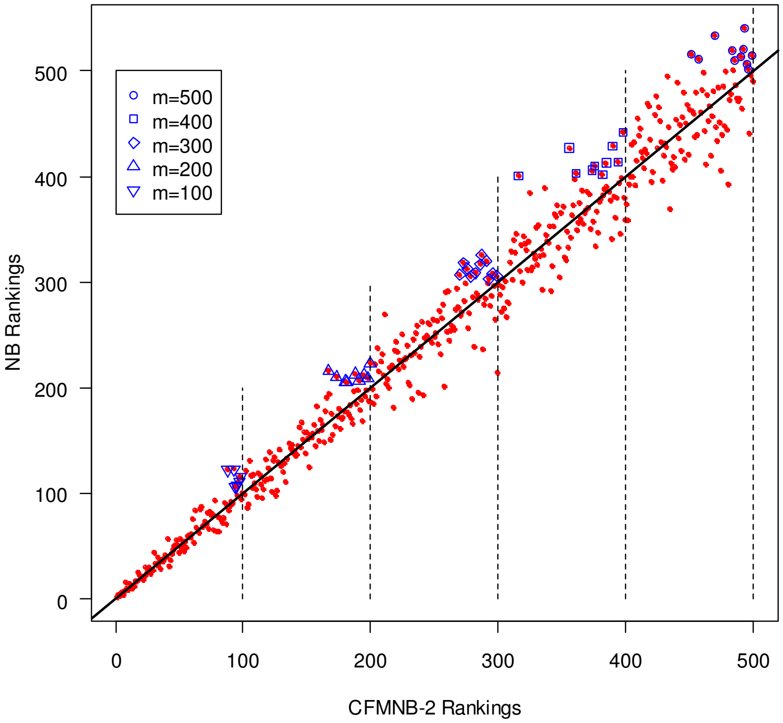


Figure 6.1 Comparison of rankings between NB and CFMNB-2 models

The percentage deviation can be computed as follows to compare two ranking orders for the number of sites that are different in two lists of hotspots (Miranda-Moreno et al., 2005).

$$\% \text{ deviation} = 100 \times (1 - s/m) \qquad (6.3)$$

where s is the number of hotspots that are common in the two lists and m is defined previously. Table 6.1 shows the computation results. As m increases the difference tends to decrease. Although the difference is not large, the ranking results from the CFMNB-2 model may be more reliable than the NB model because of a better model specification.

Table 6.1 Percent deviation between NB and CFMNB-2 models

m	100	200	300	400	500
s	94	191	289	390	489
% deviation	6.0%	4.5%	3.7%	2.5%	2.2%

6.1.3 Comparison by simulation

While the previous analysis showed the possible difference in ranking orders between the two models, it is difficult to count false positives or negatives from a respective model since we do not know, *a priori*, which sites are truly hazardous or safe. For this reason, some researchers prefer the simulation approach to using empirical crash data in assessing the relative performance of different models or various ranking criteria (Cheng and Washington, 2005; Miranda-Moreno, 2006). The simulation-based approach was, therefore, adopted in this study to compare the performance of the two models in the identification of hotspots. In simulation, the true hotspot is defined as a site whose expected conditional mean is greater than a pre-specified threshold value. Once the true hotspots are identified from the true crash frequency distribution, those are then compared with the detected hotspots determined by the two alternative models. In what follows, the performance evaluation criteria and simulation design will be described in details.

Table 6.2 shows the possible outcomes when n sites are classified according to a given hotspot identification method (Miranda-Moreno, 2006). V and R correspond to the Type I and Type II errors, respectively. When a threshold-based strategy is used for selecting hotspot lists, the threshold value should be selected so that both Type I and Type II errors are minimized. However, these two errors conflict each other: the lower the Type I error is, the higher the Type II error, and vice versa. No specific guideline exists for identifying hotspots (Elvik, 2008). In a real application, however, the optimal threshold value should be carefully selected based on the objective that we want to achieve in order to reduce the costs induced by false positives or false negatives. In this study we carried out the sensitivity analysis by using different threshold values to examine their impacts on each performance criteria.

Table 6.2 Possible outcomes of classification (Miranda-Moreno et al., 2006)

	Number of sites “detected” as non-hotspots	Number of sites “detected” as hotspots	
Number of “true” non-hotspot	U	V	n_0
Number of “true” hotspot	R	S	n_1
	$n-D$	D	n

where: n • Total number of sites in the set under analysis

n_0 • Number of “true” non-hotspots

n_1 • Number of “true” hotspots

U • Number of sites correctly classified as non-hotspots

V • Number of false positives or Type I errors

R • Number of false negatives or Type II errors

S • Number of sites correctly classified as hotspots

D • Number of sites detected hotspots as hotspots

In order to evaluate the relative performance of the two models in detecting the true hotspots, the following five measures were used as performance criteria. (Miranda-Moreno, 2006):

- False Discovery Rate (FDR): the ratio of false positives (Type I errors) among all the detected hotspots by a model. A model with a smaller value is considered to be a better model.

$$FDR = \frac{V}{D} \tag{6.4}$$

- False Negative Rate (FNR): the ratio of false negatives (Type II errors) among all the detected non-hotspots by a model. A model with a smaller value is preferred.

$$FNR = \frac{R}{n - D} \tag{6.5}$$

- Sensitivity (SENS): the ratio of correctly detected hotspots by a model among the true hotspots. A model with a larger value is preferred.

$$SENS = \frac{S}{n_1} \tag{6.6}$$

- Specificity (SPEC): the ratio of correctly detected non-hotspots by a model among the true non-hotspots. A model with a larger value is preferred.

$$SPEC = \frac{U}{n_0} \tag{6.7}$$

- Risk (RISK): the ratio of total number of false positives and false negatives among all the sites under analysis. A model with a smaller value is preferred.

$$RISK = \frac{V + R}{n} \tag{6.8}$$

In simulation design, we utilized the same covariates and model parameters used for Example 3 in Chapter IV. That is, the crash frequency at each site was assumed to follow the FMNB-2 distribution with known parameters. The simulation was carried out based on the following steps:

- *Step 1*: The "true" crash mean at site *i* is generated using the following conditional mean functional form: $\mu_i^{true} = w_1 \exp(\mathbf{x}_i \boldsymbol{\beta}_1) + w_2 \exp(\mathbf{x}_i \boldsymbol{\beta}_2)$. The covariates ($\mathbf{x}_i$) and the parameters ($w_1, w_2, \boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$) are defined as the same in Example 3 in Chapter IV (see Table 4.7). The data are generated for 500 sites.

- *Step 2*: Specify a threshold value k . In this study, four alternative threshold values are considered: sample mean, 80th-percentile, 85th-percentile, and 90th-percentile crash in the sample. Under each specified *k* value, the following selection rule is applied for each site:
  - If $\mu_i^{true} > k$, set $h_i = 1$ and site i is defined as a "true" hotspot
  - Otherwise, set $h_i = 0$ and site i is defined as a "non-true" hotspot

  Then, summing h_i over n sites results in the total "true" number of hotspots:

$$n_1 = \sum_{i=1}^{n} h_i$$

- *Step 3*: For each site, simulate crash frequency based on the method described in Subsection *4.3.1*.

- *Step 4*: Based on the simulated crash frequency, the model parameters are estimated for the NB and FMNB-2 models, respectively. The parameter estimation method follows the one described in Subsection *4.3.2*. This step results in $\hat{\mu}_i^{NB}$ and $\hat{\mu}_i^{FMNB-2}$ as defined in Equations in (6.1) and (6.2).

- *Step 5*: Once $\hat{\mu}_i^{NB}$ and $\hat{\mu}_i^{FMNB-2}$ are obtained for each site, the following selection rule is applied for identifying the "detected" hotspots:
  - If $\hat{\mu}_i > k$, set $d_i = 1$ and site i is defined as a "detected" hotspot
  - Otherwise, set $d_i = 0$ and site i is defined as a "non-detected" hotspot

  Summing d_i over n sites results in the total "detected" number of hotspots (*D*):

$$D = \sum_{i=1}^{n} d_i$$

- *Step 6*: At the end of each simulation replication, the five performance criteria (FDR, FNR, SENS, SPEC and RISK) are computed, which were defined in Equations (6.4) to (6.8).

The simulation was replicated 100 times by repeating Steps 3 to 6. The average of the 100 replications was used as final results.

6.1.4 Results

Table 6.3 shows the results of five performance criteria for the two models. Note that the results were obtained from using the sample mean value as a threshold value (k). The average values of five performance criteria for the FMNB-2 model are all superior to those for the NB model. This was expected because the data were generated from the FMNB-2 model. Nevertheless, this simulation can demonstrate what the consequences will be if a mis-specified model is used for the hotspot identification.

The results show that the false discovery rate for the NB model (0.578) is considerably larger than the corresponding value for the FMNB-2 model (0.128) and the sensitivity rate for the NB model (0.691) is much smaller than that for the FMNB-2 model (0.888). This means that, on average, 57.8% of the hotspots detected by the NB model are actually non-hotspots. This false positive rate is very high and may be unacceptable in practice. Furthermore, the sensitivity value, 0.691 for the NB model means that, on average, the NB model was able to detect only 69.1% among all the true hotspots. This power to detect the true hotspots is also considered very low.

Table 6.3 Results of performance criteria measures

Criteria	FMNB-2 model			NB model		
	Average	Min.	Max.	Average	Min.	Max.
FDR (Smaller is better)	**0.128**	0.000	0.471	**0.578**	0.323	0.801
FNR (Smaller is better)	**0.060**	0.000	0.271	**0.204**	0.058	0.571
SENS (Larger is better)	**0.888**	0.624	1.000	**0.691**	0.245	0.795
SPEC (Larger is better)	**0.946**	0.811	1.000	**0.584**	0.188	0.767
RISK (Smaller is better)	**0.080**	0.018	0.192	**0.385**	0.246	0.682

On the other hand, another simulation was carried out with different threshold values for the FMNB-2 model. This simulation was meant to examine the effects of different threshold values on the performance criteria. The results from this simulation can provide an insight into how the threshold value can be selected in practice. Table 6.4 shows the simulation results from using four alternative threshold values which include the sample mean, 80th-percentile, 85th-percentile, and 90th-percentile crash in the sample. The effects of those values on the performance criteria are visualized in Figure 6.2. As higher threshold values are used, the total error rate indicated by RISK is decreasing in general. However, the FDR and SPEC criteria are exhibiting an increasing trend and the FNR and SENS tend to decrease. Using a higher threshold value reduces the number of target hotspots for treatment. The results demonstrate that if we increase the threshold value with the hope of reducing target hotspots, we are more likely to have the increased number of false positives (identifying non-hotspots as hotspots) and less power in detecting true hotspots. At the same time, we can reduce false negatives (identifying hotspots as non-hotspots) and increase the specificity. This is a conflict result.

The costs associated with false positives and false negatives may be different depending on the type of crashes and the improvement expenses. For example, false positives incur unnecessary improvement costs by improving actually safe sites without any safety benefits. Similarly, false negatives will cause crash-related costs by leaving true hotspots untreated. Therefore, a decision on the threshold value can be made by considering the trade-off between these two costs so that it can minimize the unnecessary costs.

Table 6.4 Simulation results of four different threshold values for FMNB-2 model

Criteria	Alternative threshold values, *k*			
	Sample mean	80th percentile	85th percentile	90th percentile
FDR	0.128	0.138	0.169	0.236
FNR	0.060	0.017	0.012	0.004
SENS	0.888	0.898	0.842	0.785
SPEC	0.946	0.977	0.986	0.994
RISK	0.080	0.035	0.024	0.010

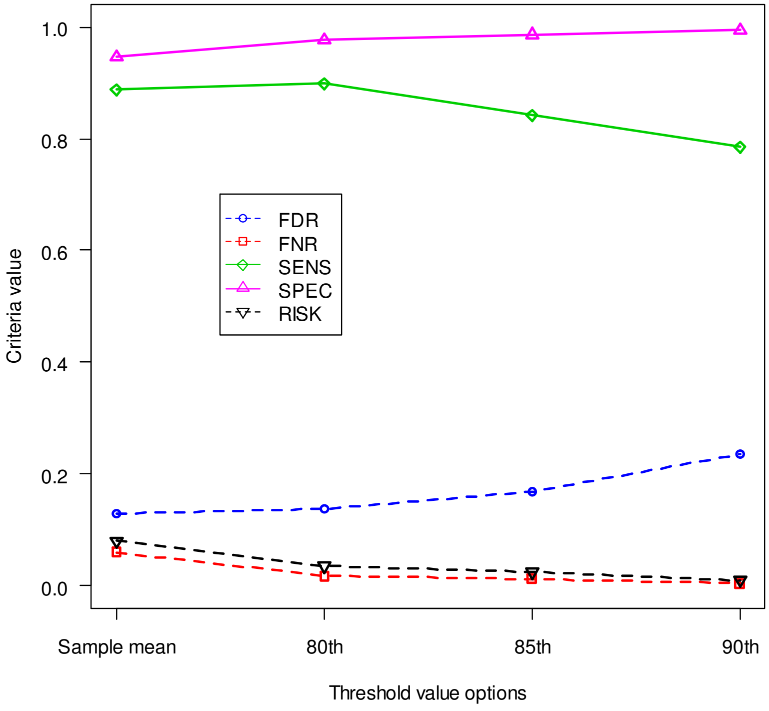


Figure 6.2 Effects of different threshold values on performance criteria

6.2 Accident Modification Factors

This section introduces the concept of the accident modification factor (AMF) and discusses the results from the two models (NB model and FMNB-2 model).

6.2.1 Introduction

The statistical model developed for vehicle crash data can also be used to developing accident modification factors of interest highway geometric variables such as shoulder width or median width. An AMF represents the change in safety when a particular geometric design element changes in size with respect to the base (or typical) condition. An AMF greater than 1.0 represents the situation where the design change is associated with more crashes while an AMF less than 1.0 indicates fewer crashes. The development and use of AMFs in highway safety has gained a lot of popularity because of the recent efforts to quantify and incorporate safety proactively into design process. The *Highway Safety Manual* (HSM)[14], which is envisioned to become a nationwide predictive tool, is currently under development and utilizes the AMF concept to evaluate the safety performance for various highway facilities (Fitzpatrick et al, 2008).

AMFs have been developed by various techniques which include the before-and-after study, cross sectional study, use of expert panels, and regression-based models (Bonneson and Lord, 2005; Li et al., 2009). In this study, we are particularly interested in deriving AMFs from regression-based models. The AMFs are estimated directly from the coefficients of the model. This approach for AMF development explicitly assumes that each AMF is independent, since the model parameters are assumed independent. In practice, however, AMFs may not be completely independent since changes in geometric design characteristics on highways are not done independently (e.g., lane and shoulder width may be changed simultaneously) and the combination of these changes

[14] The first edition of the HSM is expected for public release shortly, and it will contain safety prediction methodologies for rural two-lane highways, rural multilane highways, and urban and suburban arterials. Additional information is available on the HSM website (http://www.highwaysafetymanual.org/).

can influence crash risk. Nevertheless, experience in deriving AMFs in this manner indicates that the assumptions are reasonable and, with thoughtful model development, the resulting AMFs can yield useful information about the first-order effect of a given variable on safety. Others who have used this approach for developing AMFs include Fitzpatrick et al. (2008), Lord and Bonneson (2007) and Washington et al. (2005). On the other hand, Bonneson et al. (2007) and Gross et al. (2009) have argued that the interaction between design features should be included in the development of AMFs. In line with this effort, Li et al. (2009) tried to incorporate the interactions by using general additive models. Addressing this issue may be beyond the scope of this study. The objective of this section is to compare the relative performance of the two models (i.e., NB and FMNB-2) in terms of the difference in determining AMFs as a result of different model coefficients.

Once we obtained the AMFs for various highway geometric design elements, they are applied multiplicatively for adjusting crash frequency estimated from a baseline model. The baseline model represents the calibrated statistical model using data that meet specific base conditions, such as 12-ft lane width and 8-ft shoulder width for divided rural multilane highway segments. Therefore, the finally predicted number of crashes is computed as follows:

$$\mu_{final} = \mu_{baseline} \times AMF_1 \times \cdots \times AMF_n \tag{6.9}$$

where, μ_{final} = final predicted number of crashes per unit of time

$\mu_{baselinel}$ = baseline predicted number of crashes per unit of time

$AMF_1 \times \cdots \times AMF_n$ =accident modification factors (assumed to be independent)

6.2.2 Deriving AMFs from NB model

In additive models, such as a linear regression with $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}$, the coefficient $\hat{\beta}_j$ for a covariate x_j is readily interpreted as the effect of a one-unit change in x_j on the conditional mean.[15] That is, a unit increase in x_j is associated with a $\hat{\beta}_j$ increase in $\hat{\mu}_i$. In multiplicative models, such as the Poisson or negative binomial regression models, the conditional mean functional form is usually expressed as a log-linear form: $\ln(\hat{\mu}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}$. In such a case, the difference between two conditional means ($\Delta\hat{\mu}_i$) induced by a one-unit change in x_j is no longer constant across sites and depends on the values of the covariates. A more convenient way to examine the effect of a covariate is to take the ratio of the two conditional means as below instead of focusing on their difference.

$$\frac{\hat{\mu}_i(x_j+1)}{\hat{\mu}_i(x_j)} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_j(x_{ji}+1) + \cdots + \hat{\beta}_p x_{pi})}{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_j x_{ji} + \cdots + \hat{\beta}_p x_{pi})} = \exp(\hat{\beta}_j) \qquad (6.10)$$

The ratio in Equation (6.10) is now constant across all sites without depending on the values of any covariates. Hence, the effect of a covariate is interpreted as follows: a one-unit increase in x_j is associated with a factor of $\exp(\hat{\beta}_j)$ increase in $\hat{\mu}_i$ (Long, 1997).

In developing the AMF for a covariate x_j, however, we are not interested in the safety effect of a covariate x_j by changing a one-unit, but interested in the safety effect of x_j when it changes from its base condition value. In this way, the AMF for x_j can be derived in a continuous functional form with respect to x_j . If we condition the

[15] $\Delta\hat{\mu}_i = \hat{\mu}_i(x_j+1) - \hat{\mu}_i(x_j) = \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_j(x_{ji}+1) + \cdots + \hat{\beta}_p x_{pi}\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_j x_{ji} + \cdots + \hat{\beta}_p x_{pi}\right) = \hat{\beta}_j$

denominator in Equation (6.10), $\mu_i(x_j)$ on the base condition value for x_j, i.e., $\hat{\mu}_i(x_j = x_j^{base})$, the accident modification function for x_j is derived as follows:

$$AMF_{x_j} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_p x_p)}{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j^{base} + \cdots + \hat{\beta}_p x_p)} = \exp\left[\hat{\beta}_j (x_j - x_j^{base})\right] \qquad (6.11)$$

Without loss of generality the subscript i was removed from Equation (6.11) since the AMF_{x_j} is identical for all sites. The AMF_{x_j} represents the change in the expected crash frequency when the variable x_j changes from its base condition value, and it follows an exponential function with $AMF_{x_j} = 1$ when $x_j = x_j^{base}$. If $\hat{\beta}_j > 0$, the AMF_{x_j} is an strictly increasing function, and if $\hat{\beta}_j < 0$, it is an strictly decreasing function. This relationship is depicted in Figure 6.3.

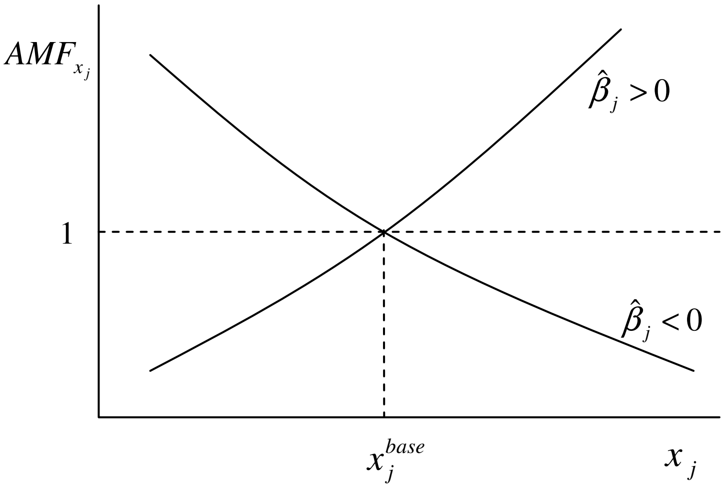


Figure 6.3 Concept of AMF

6.2.3 Deriving AMFs from FMNB-2 model

In the FMNB-2 model, coefficient interpretation is not as straightforward as in the NB model since the relationship between the conditional mean and the covariates is a mix of additive and multiplicative forms as expressed in Equation (6.2). The effect of an individual covariate on the conditional mean is determined by two sets of interactions between parameters and covariates. For instance, analogous to Equation (6.10), if we examine the effect of a one-unit change in covariate x_j by taking the ratio of the two conditional means, the ratio is expressed as follows:

$$\frac{\hat{\mu}_i(x_j+1)}{\hat{\mu}_i(x_j)} = \frac{\hat{w}_1 \exp[\hat{\beta}_{0,1} + \hat{\beta}_{j,1}(x_{ji}+1) + \sum_{k=1,\,k\neq j}^{p} \hat{\beta}_{k,1}x_{ki}] + \hat{w}_2 \exp[\hat{\beta}_{0,2} + \hat{\beta}_{j,2}(x_{ji}+1) + \sum_{k=1,\,k\neq j}^{p} \hat{\beta}_{k,2}x_{ki}]}{w_1 \exp(\hat{\beta}_{0,1} + \sum_{k=1}^{p} \hat{\beta}_{k,1}x_{ki}) + w_2 \exp(\hat{\beta}_{0,2} + \sum_{k=1}^{p} \hat{\beta}_{k,2}x_{ki})}$$

$$= \frac{\hat{w}_1\hat{\mu}_{i,1}}{\hat{w}_1\hat{\mu}_{i,1} + \hat{w}_2\hat{\mu}_{i,2}} \exp(\hat{\beta}_{j,1}) + \frac{\hat{w}_2\hat{\mu}_{i,2}}{\hat{w}_1\hat{\mu}_{i,1} + \hat{w}_2\hat{\mu}_{i,2}} \exp(\hat{\beta}_{j,2}) \tag{6.12}$$

where, $\hat{\mu}_{i,1} = \exp(\hat{\beta}_{0,1} + \sum_{k=1}^{p} \hat{\beta}_{k,1}x_{ki})$ and $\hat{\mu}_{i,2} = \exp(\hat{\beta}_{0,2} + \sum_{k=1}^{p} \hat{\beta}_{k,2}x_{ki})$. As can be seen in Equation (6.12) the difficulty arises because the conditional mean ratio varies across all sites, and also depends on the coefficients of the other covariates. Two options can be considered when we want to report a single value for the effect of a one-unit change in x_j. One option is first to calculate the Equation (6.12) for all sites and then to take the average value. Another option is to evaluate the Equation (6.12) at selected values of the covariates (e.g., sample average).

In developing the AMF of x_j, the same problem arises. Analogous to Equation (6.11), the AMF of x_j in the FMNB-2 model is expressed as follows:

$$AMF_{x_j} = \frac{\hat{w}_1 \exp(\hat{\beta}_{0,1} + \hat{\beta}_{j,1}x_{ji} + \sum_{k=1,\,k\neq j}^{p} \hat{\beta}_{k,1}x_{ki}) + \hat{w}_2 \exp(\hat{\beta}_{0,2} + \hat{\beta}_{j,2}x_{ji} + \sum_{k=1,\,k\neq j}^{p} \hat{\beta}_{k,2}x_{ki})}{\hat{w}_1 \exp(\hat{\beta}_{0,1} + \hat{\beta}_{j,1}x_j^{base} + \sum_{k=1,\,k\neq j}^{p} \hat{\beta}_{k,1}x_{ki}) + \hat{w}_2 \exp(\hat{\beta}_{0,2} + \hat{\beta}_{j,2}x_j^{base} + \sum_{k=1,\,k\neq j}^{p} \hat{\beta}_{k,2}x_{ki})} \tag{6.13}$$

In this case, the AMF_{x_j} differs across sites by depending on the values of covariates. In order to obtain a single continuous function of the AMF_{x_j} with respect to x_j like the one in Figure 6.3, we need to fix each covariate (except for the interest covariate x_j) at a selected value. For this purpose, we used the sample average of each covariate. This is, x_{ki} values for site i is replaced with $\bar{x}_k$ which is $(1/N)\sum_{i=1}^{N} x_{ki}$. This leads the Equation (6.13) to the following form:

$$AMF_{x_j} = \frac{\hat{w}_1 \exp(\hat{\beta}_{0,1} + \hat{\beta}_{j,1} x_j + \sum_{k=1,\, k \neq j}^{p} \hat{\beta}_{k,1} \bar{x}_k) + \hat{w}_2 \exp(\hat{\beta}_{0,2} + \hat{\beta}_{j,2} x_j + \sum_{k=1,\, k \neq j}^{p} \hat{\beta}_{k,2} \bar{x}_k)}{\hat{w}_1 \exp(\hat{\beta}_{0,1} + \hat{\beta}_{j,1} x_j^{base} + \sum_{k=1,\, k \neq j}^{p} \hat{\beta}_{k,1} \bar{x}_k) + \hat{w}_2 \exp(\hat{\beta}_{0,2} + \hat{\beta}_{j,2} x_j^{base} + \sum_{k=1,\, k \neq j}^{p} \hat{\beta}_{k,2} \bar{x}_k)} \qquad (6.14)$$

6.2.4 AMFs comparison

Based on the AMF functions provided in Equations (6.11) and (6.13), a comparison was carried out between the two models with the parameter estimation results for the segment crash data in Chapter V. The AMF function for the FMNB-2 model was calculated for the CFMNB-2 model with $\phi_k \sim \Gamma(0.5, 0.1)$. For the base condition for each variable, the values recommended in NCHRP Project 17-29 (Lord et al., 2009) were basically adopted: i.e., 30 ft for median width including left shoulder widths and 8 ft for right-shoulder width. Since the median width used for modeling in this study also included the left shoulder width for both sides (see Table 5.7 in Chapter V), we can used 30 ft as a base condition for the median width. The summary statistics of the variables are reproduced in Table 6.5 along with the respective base condition value. Note that the base condition value for the right-shoulder width is very close to the average of the sample data, while the base condition value for the median width is much smaller than the sample average.

Table 6.5 Summary statistics of variables and their base condition values

Variable	Min.	Max.	Average	Standard Deviation	Base Condition
Median width (feet)	1	240	47.07	29.41	30
Right-shoulder width, (feet)	0	19	7.68	1.98	8

The resultant AMF functions of median width and right-shoulder width are presented in Figures 6.4 and 6.5, respectively. While the AMF functions of the NB model are very close to a straight line for both variables, those of the CFMNB-2 model take on a more marked curve-shape. The difference in the shape mainly results from the fact that the NB model takes the average effect of a covariate across all sites, whereas the CFMNB-2 model takes into account the differential responsiveness of crash frequency to the covariate. As already noticed in the parameter estimation results (see Table 5.13 in Chapter V), the observations assigned to Component 1 (smaller-mean component) were not influenced by median width and right-should width, whereas the observations in Component 2 (higher-mean component) were significantly affected by these variables. This effect was reflected in the shape of AMF derived from the CFMNB-2 model.

Another good property about the shape of AMF in the CFMNB-2 model is that the safety effect of a covariate eventually levels off as the covariate increases significantly from the base condition. For example, it can be seen from Figure 6.4 that the safety effect of median width stabilizes after around 150 ft. The same tendency is noticed for right-shoulder width after around 16 ft. These trends are not observable in the NB model. This is partly supported by a few researchers who have noted that design elements, such as shoulder or lane width could follow a U-shaped relationship with safety (Hauer, 2000; Xie et al., 2007; Li et al., 2008). In a U-shaped relationship, narrow and wide widths experience more crashes. McLean (1996) explained the U-shaped relationship between safety and should width by suggesting that very wide shoulders can often be used as an additional lane, which may lead to an increase in accidents rates (Hauer, 2000).

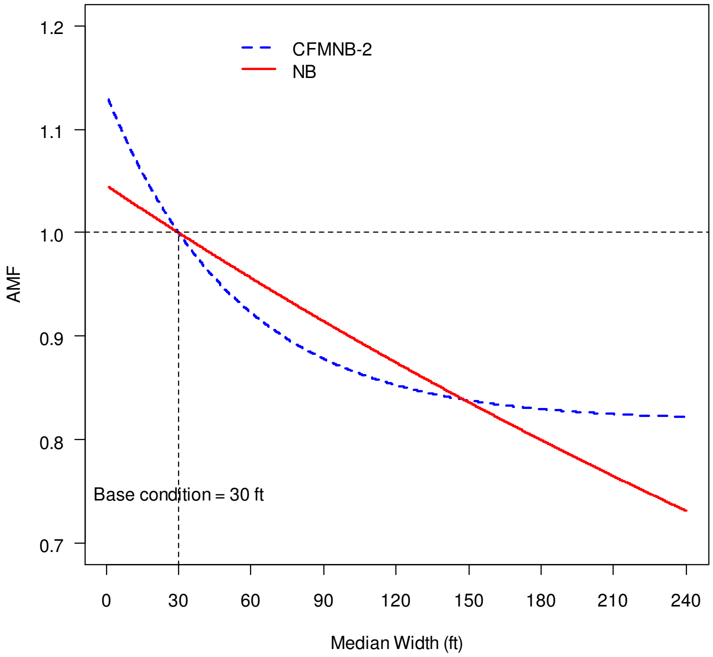


Figure 6.4 AMF function of median width (CFMNB-2 vs. NB)

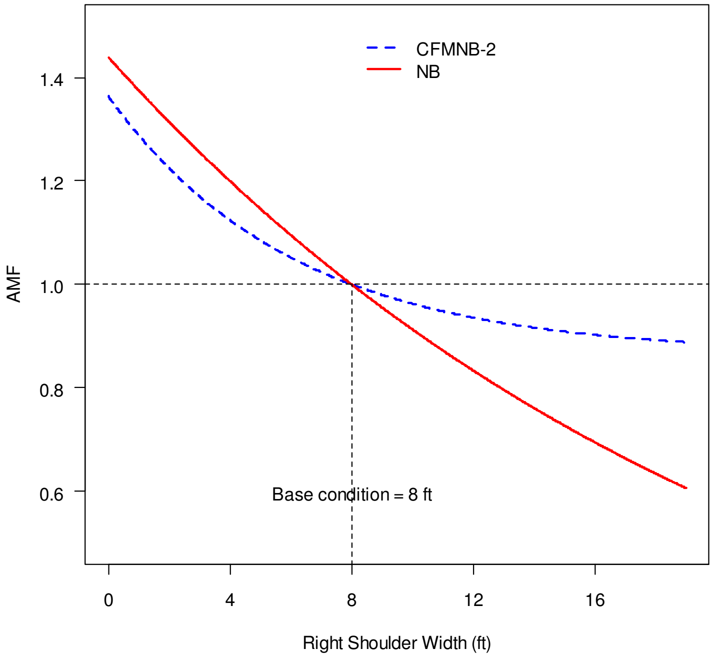


Figure 6.5 AMF function of right-shoulder width (CFMNB-2 vs. NB)

The AMF curves from the current CFMNB-2 model did not exhibit a complete U-shaped relationship within the sample boundary, but when the coefficients from the FMNB-2 model (see Table 5.12 in Chapter V) was used, the AMF curve for median width revealed a U-shaped relationship (Figure 6.6). Although this relationship can be debatable, the bottom line is that the AMF functions derived from the FMNB-2 model is more flexible and leave much more possibilities about the true effect of a design element on crash occurrence.

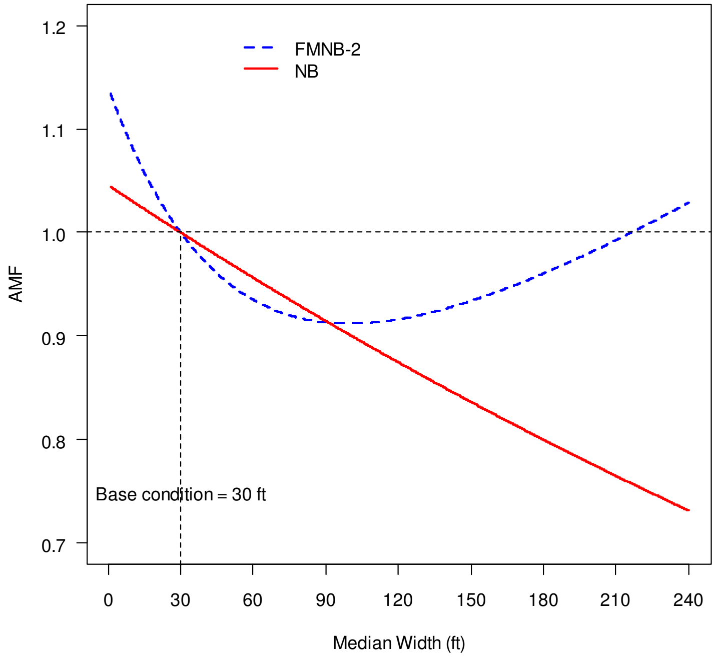


Figure 6.6 AMF function of median width (FMNB-2 vs. NB)

On the other hand, the AMF curve for right-shoulder width from the FMNB-2 model (Figure 6.7) did not exhibit a U-shaped curve and remained almost unchanged from the CFMNB-2 model. This is because the coefficient of the right-shoulder width in component 2 was much larger than that in component 1 (i.e., -0.1509 vs. 0.0014). The small value, even with the large weight, exercised little influence on the calculation of the AMF curve.

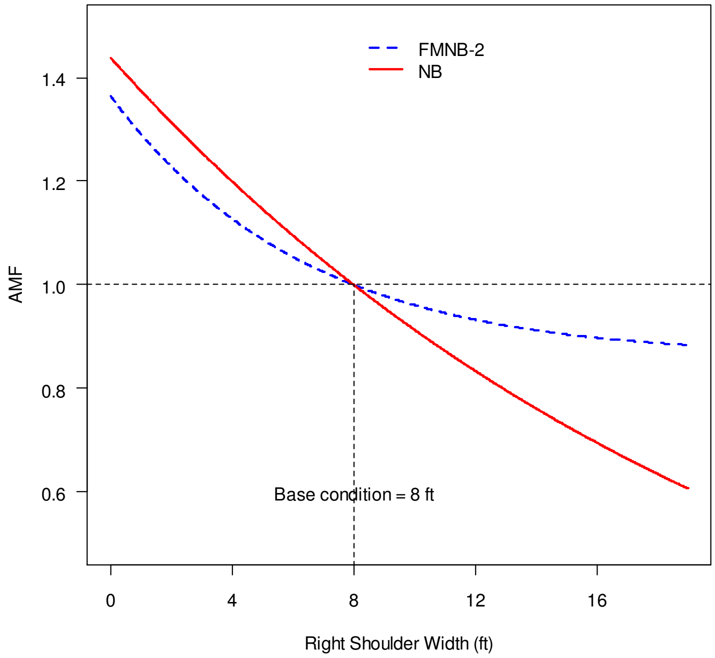


Figure 6.7 AMF function of right-shoulder width (FMNB-2 vs. NB)

6.3 Chapter Summary

Given the superior performance of the finite mixture model for a particular dataset, this chapter has focused on the application side of this type of model in terms of two important highway safety analyses: the identification of hotspots and development of accident modification factors.

With the modeling results for the segment crash data in Chapter V, the hotspot rankings were compared between the NB and CFMNB-2 models. The difference was measured by the percentage deviation in ranking orders between the two models, but the difference seemed to be small for this dataset. However, the ranking results from the CFMNB-2 model may be more reliable than the NB model because of a better model specification, Depending on the dataset under consideration there is a possibility that the difference can be more pronounced. A simulation study was also carried out to demonstrate what the consequences will be if a mis-specified model is used for the hotspot identification. The consequences turned out to be significant by producing a high number of false positives and negatives. This will lead to a waste of federal, state and local government resources by investing them into the improvement of wrong sites. In order to gain an insight into the selection of an optimal threshold value for identifying hotspots, another simulation was designed with four different threshold values: sample mean, 80th-percentile, 85th-percentile, and 90th-percentile crash in the sample. The use of a higher threshold value can reduce the number of target hotspots for treatment, but the results of simulation indicated that there is a conflict between false discovery rate (increasing) and false negative rate (decreasing), and also between sensitivity (decreasing) and specificity (increasing). Since the costs associated with false positives and false negatives are different, a decision on the optimal threshold value can be made by considering the trade-off between these two costs so that it can minimize the unnecessary costs.

The accident modification factor (AMF) function for the FMNB-2 model has been derived in this chapter. Its form was not as simple as in the NB model since the conditional mean takes on the mix of additive and multiplicative terms. However, the AMF function from the FMNB-2 model has an advantage over the one from the NB model since it can consider the interactions between parameters and covariates, and hence can better account for the differential responsiveness of crash frequency with respect to a certain covariate. The AMF curves for median width and right-shoulder width were derived based on the CFMNB-2 model results and they were compared with those from the NB model. The AMF shapes produced by the CFMNB-2 model had a better property in that the safety effect of a covariate eventually levels off as the covariate increases significantly from the base condition. On the other hand, when the FMNB-2 model – which is inferior to the CFMNB-2 model, but superior to the NB model – was used, the AMF curve for median width showed a U-shaped relationship. This is arguable, but the AMFs from the FMNB-2 model may open up interesting new prospects for finding the true effect of a design element on crash occurrence.

CHAPTER VII

SIMULATION STUDY

In Chapter IV, we have shown the usefulness of finite mixture of regression models for accommodating over-dispersion with a single random sample – that is, the results were evaluated with only one-time simulated dataset under a specific condition. If we adopt a Bayesian method for parameter estimation and summarize the posterior distribution with a single point estimate,[1] it is important to obtain consistent posterior summary values for model parameters under a repeated random sampling. In order to appreciate the potential bias and variability in posterior summary values, we need to run the simulation as many times as possible under the same condition. This point is illustrated in this chapter by means of a Monte Carlo simulation. As we already observed, the posterior mean of the dispersion parameters in the FMNB-2 model was biased upwards, and hence the objective of the simulation is primarily to investigate the bias associated with the posterior summary values of dispersion parameters. While the posterior mean is often favored as a summary statistic because it minimizes the posterior expected mean squared loss, the posterior median is also a useful summary value especially in a skewed distribution. To this end, the simulation is carried out under various sample sizes and sample-mean categories and then we investigate the biases associated with the posterior mean and median values. In addition, since the prior specification for the dispersion parameter has a potential influence on the posterior mean and median values, the results from non-informative and informative prior specifications are compared in terms of the magnitude of the bias introduced by various sample sizes and sample-mean values.

[1] However, remember that the full posterior distribution of a parameter provides much richer information than a single posterior summary value (i.e. posterior mean, median or mode), as it incorporates all information, and all uncertainty about the parameter.

This chapter consists of four sections. Section 7.1 briefly reviews previous work on the bias properties of the dispersion parameter within a standard NB modeling framework. Section 7.2 describes the simulation design for the FMNB-2 model in details, in which sample sizes, sample mean values, and two prior specifications for the dispersion parameters are defined. Section 7.3 presents the simulation results for three different sample-mean value scenarios. Finally, Section 7.4 summarizes the results and recommends a brief guideline about the choice of priors and the posterior summary statistics to use for different sample sizes and sample-mean values.

7.1 Previous Work

Within the standard NB modeling framework, many researchers have examined the biasness of the various estimators of the NB dispersion parameter under different scenarios. All the researchers used the simulation method. The key studies are summarized in Table 7.1. Note that the majority of the studies are based on the Frequentist method. Although earlier researches mainly focused on the impact of a small sample size on the performance of the maximum likelihood estimator (MLE) and compared the results with other estimators, more recent studies have examined more extreme cases in terms of sample mean values and true dispersion parameters (ϕ). For example, Lord (2006) examined the effects of a very low sample mean ($\mu < 1$) combined with a small samples size on the estimation of the dispersion parameter. The sample-mean values used for developing NB models for vehicle crash data are often below 1.0 (crashes/unit of time). In another example, Lloyd-Smith (2007) explored the bias, precision, and confidence interval coverage of the MLE of the dispersion parameter when the data are highly over-dispersed ($\phi < 1$). Highly dispersed data are commonly found in epidemiological studies.

It is worth noting that although researchers working on this topic are from a wide variety of fields their findings have much in common. First, when the dataset is characterized by a small sample size, the MLE of ϕ is less accurate than other estimators, such as the method of moment or the quasi-likelihood method. Second, small sample sizes tend to overestimate the true dispersion parameter under all conditions. Third, the bias for the MLE gets larger as sample mean decreases and the true ϕ increases (if known). It should be noted that unless a sufficiently large sample size is used, the bias for the MLE seems to be inevitable.

Table 7.1 Summary of previous studies for biasness of dispersion parameter estimate

Authors (year)	Estimators	Results
Clark and Perry (1989)	MME[a], MQLE[b]	Both become biased when $\mu \leq 3$, $n < 20$. Bias becomes worse when $\phi \rightarrow \infty$.
Piegorsch (1990)	MME, MQLE, MLE[c]	MLE is less accurate than MQLE and MME when n is small. α was allowed to have negative values.
Dean (1994)	MME, MLE	MLE produces a biased estimate as n decreases and ϕ increases. The bias influences coefficients of NB model.
Toft et al. (2006)	MLE	Estimator is unstable (even for $\mu = 10$ and $n = 100$) as $\alpha \rightarrow 0$ (i.e. $\phi \rightarrow \infty$).
Lord (2006)	MME, WRE[d], MLE	All three estimators are biased and skewed when n and μ are small. MLE method overestimates true dispersion parameter.
Lloyd-Smith (2007)	MLE	MLE becomes more biased and less precise for higher ϕ. MLEs are not biased downward by any of the factors considered.
Zhang et al. (2007)	MME, MLE, BMLE[f]	BMLE is more accurate and stable, and the improvements are more pronounced with small sample sizes and low sample means.
Lord and Miranda-Moreno (2008)	HBME[g]	An appropriate non-vague prior minimizes the bias in the posterior mean. Poisson-lognormal models are recommended over Poisson-gamma models when assuming vague priors for a low sample-mean sample.

NOTE: [a] Method of moment estimator; [b] Maximum Quasi-likelihood estimator,
[c] Maximum likelihood estimator; [d] Weighted regression estimator,
[f] Bootstrapped Maximum likelihood estimator; [g] Hierarchical Bayes Method estimator

In highway safety studies, it is not unusual that the analysts have to develop models under the limited sample size because of the prohibitive costs involved in collecting the crash data (Lord, 2000; Oh et al., 2003). At the same time, many of the analysis units (highway segments or intersections) tend to have a zero crash because of a rare and random nature of a crash occurrence. This may result in a low sample mean value. The sample mean values used for developing NB regression models are often below 1.0 (crashes/unit of time). In this context, Lord (2006) examined the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter, and recommended that no Poisson-gamma (NB) models be estimated for a sample size below 100, even when the sample mean is equal to 5 in highway safety analysis. To reduce the bias in the estimation of the dispersion parameter, Zhang et al. (2007) proposed to use the bootstrapped maximum likelihood estimator because it produced the most accurate and stable estimates for the data with low sample mean and small sample sizes. On the other hand, recently Bonneson et al. (2007), and Park and Lord (2008) tried to adjust the bias induced by the MLE method by deriving a reasonable relationship (a quadratic function) between estimated and true values of the dispersion parameter.

As a follow-up research to Lord (2006), Lord and Miranda-Moreno (2008) examined how low sample mean values and small sample sizes affect the posterior mean of the dispersion parameter of Poisson-gamma models estimated using the hierarchical Bayes method. Especially, they looked into the role of prior specifications for dispersion parameter. They found that the posterior mean of the dispersion parameter was seriously affected by a small sample size and low sample mean values as the MLE approach. This was more pronounced when a non-informative prior was used. However, the bias started at a lower sample mean value and smaller sample size than that found for the MLE, which suggests that a Bayesian method is more robust than an MLE method in such case.

In summary, judging from the previous work no estimators seem to be free from the bias in the NB dispersion parameter caused by small sample sizes and low sample mean values. Therefore, from an application-oriented point of view, it is important to know what sample size or sample mean is required at minimum in order to guarantee unbiased or bias-reduced estimates of model parameters. Within the finite mixture models, however, the necessary sample size may depend on the data at hand – that is, the sample size need not be large for well-separated data, but it can be huge for a poorly-separated case. Therefore, rather than searching for the minimum sample size, it would be better to focus on the bias and variability properties of an estimator we choose. A Bayesian approach can provide an asset in this respect since we can simply choose the best posterior summary statistic that minimizes the bias. This can also obviate an additional correction process.

7.2 Simulation Design

We first designed the simulation scenarios for generating FMNB-2 random variates. The regression parameters ($\boldsymbol{\beta}_k$), mixing proportions (w_k), and dispersion parameters (ϕ_k) were controlled in order to generate three sample mean categories: high mean $(\bar{\mathbf{y}} > 5)$, moderate mean $(1 < \bar{\mathbf{y}} < 5)$, and low mean $(\bar{\mathbf{y}} < 1)$. For each sample-mean value category, fixed values of ϕ and w were used. To allow for a high level of heterogeneity the higher-mean component (Component 1) was combined with a lower ϕ value, and the smaller-mean component (Component 2) was combined with a higher ϕ value. Note that to appreciate the sensitivities of different ϕ and w values, it is necessary to test with more combinations of their values. However, since the number of combinations can be prohibitive, the simulation was limited to the one provided in Table 7.2.

Table 7.2 True values used for generating FMNB-2 random variates for simulation

Parameters	High mean $(\bar{\mathbf{y}} > 5)$		Moderate mean $(1 < \bar{\mathbf{y}} < 5)$		Small mean $(\bar{\mathbf{y}} < 1)$	
	Comp. 1	Comp. 2	Comp. 1	Comp. 2	Comp. 1	Comp. 2
$\beta_{0,k}$ $\beta_{1,k}$ $\beta_{2,k}$	2.5 -0.5 0.5	1.0 0.5 -0.5	2.0 -0.5 0.5	0.0 0.5 -0.5	0.5 -0.5 0.5	-1.0 0.5 -0.5
ϕ_k	5	10	5	10	5	10
w_k	0.4	0.6	0.2	0.8	0.2	0.8
Sample size N	100 ~ 1,000 (one-hundred step) and 2,000		100 ~ 1,000 (one-hundred step) and 2,000		500 ~ 3,500 (five-hundred step) and 5,000	

Second, two prior specifications for the dispersion parameter are compared. The first one is the non-informative gamma prior: $\phi_k \sim \Gamma(0.01, 0.01)$. This is by far the most common prior distribution within the standard NB models (e.g., Miaou et al., 2003). This prior has a spike near zero with a mean=1 and a large variance which is 100. However, the automatic assignment of such flat or wide priors can be problematic in some cases (Van Dongen, 2006; Lord and Miranda-Moreno, 2008), so the prior specification should be done with great care. The extremely large variance can create problems especially when the sample size is small or sample mean is low. Recently, less vague priors have been proposed to use in analyzing vehicle crash data (Washington and Oh, 2006; Miranda-Moreno et al., 2008). Nevertheless, in the FMNB-2 model, it is difficult to assign informative priors on each component model because such information is rarely available. Therefore, as a comparative purpose, we introduced the weakly-informative gamma prior: $\phi_k \sim \Gamma(0.5, 0.1)$, and investigated its performance in terms of the bias. This prior also has a mode near zero, but it has a mean=5 and the much reduced variance which is 50. This prior was suggested in this paper with the hope of reducing the implausibly large values of ϕ_k in the posterior samples, thereby improving the behavior of the posterior distribution of ϕ_k. In addition, it is unlikely that this prior will prevent the chain from exploring the plausible space of the dispersion parameter since the prior variance is still large. Within the standard NB model, Lord and Miranda-Moreno (2008) found that the priors with very small mean and variance (for example, $\phi_k \sim \Gamma(0.1, 1.0)$) often generated extremely small values for the dispersion parameter, which resulted in a significant underestimation of the true value. Figure 7.1shows three different gamma distributions.

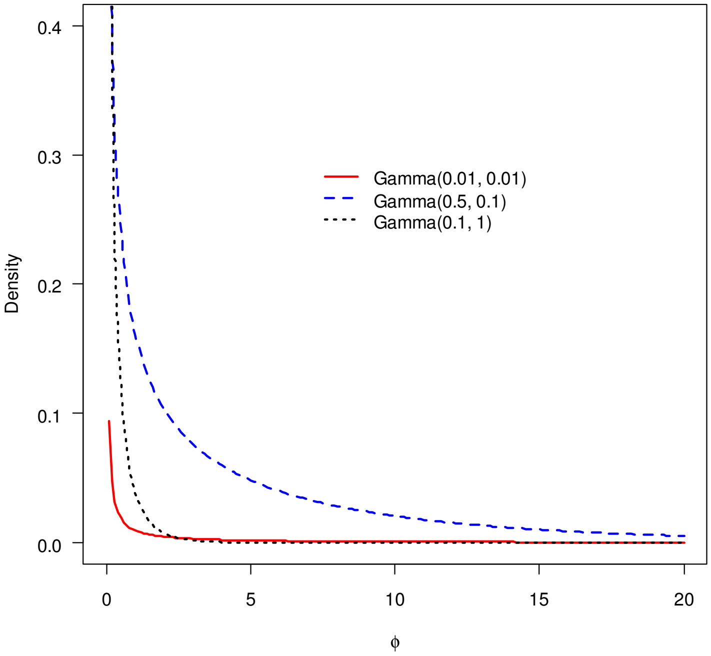


Figure 7.1 Shapes of three different gamma distributions

Based on the above-described simulation scenarios, the FMNB-2 random variable generation process was replicated 100 times for each category, and then for each of the datasets, Bayesian estimation was carried out using 2,500 draws after a burn-in of 2,500 draws. The prior specifications for the weight distribution and the regression parameters are the same as in Example 2. Due to the partial manual manipulations needed to adjust the acceptance rates in the Random Walk Algorithm and the amount of computing time, the numbers of replications and MCMC iterations were limited to 100 times and 5,000 runs, respectively. Another special consideration was also taken during the simulation to prevent the label switching which is caused by the invariance of a finite mixture distribution to relabeling the components. According to Frühwirth-Schnatter (2006), finding identifiability constraints is not trivial in the finite mixture regression models. After trying with several datasets from each mean value category, the order constraint on the weight parameters (i.e., $w_1 < w_2$) was found to be appropriate for the moderate and the small mean value scenarios. For the high mean scenario, the order constraint on the intercepts (i.e., $\beta_{0,1} > \beta_{0,2}$) was found to be most appropriate.

At the end of each replication, the posterior summary statistics such as posterior mean, median, standard deviation for each parameter estimate were computed. This provides the bias information $[E(\hat{\phi}_r) - \phi_{true}]$ in the parameter estimation, where *r* is the number of replications. The mean squared error, $\text{MSE} = Bias^2 + Var(\hat{\phi}_r)$, is another appropriate measure to check the quality of an estimator since it comprises both bias and variability.

7.3 Simulation Results

This section presents the simulation results for the three sample-mean value scenarios: high-mean, moderate-mean, and small-mean value scenarios.

7.3.1 High-mean value scenario

Table 7.3 shows the ranges of generated sample means and variances for each sample size. The minimum and maximum values of the generated sample means indicate that every dataset generated during 100 replications falls into the high-mean value category $(\bar{\mathbf{y}} > 5)$. Based on the values from Tables 7.4 and 7.5, the bias trends by sample size for the high-mean value scenario are visualized in Figure 7.2. The upper figures are for the non-informative prior and the lower ones are for the weakly-informative prior. If there is no bias, all the points would rest on the true values indicated as dotted lines. As shown in the figure, the biases for the regression parameters and weight parameters are considered to be very small regardless of the different prior choices on ϕ_k .

Table 7.3 Ranges of generated sample means and variances (High-mean)

Sample size		100	200	300	400	500	600	700	800	900	1000	2000
$\overline{\mathbf{y}}$	Min.	5.84	6.59	7.07	7.05	6.83	7.36	7.37	7.43	7.18	7.38	7.72
	Max.	10.98	11.53	10.69	9.46	9.91	9.18	9.91	9.38	9.42	9.24	9.36
var($\mathbf{y}$)	Min.	30.86	66.77	68.56	69.36	75.54	93.81	101.40	96.37	106.10	105.20	106.50
	Max.	233.00	283.10	362.60	171.50	271.50	208.10	550.40	226.50	244.70	204.10	224.40

For the dispersion parameters, the bias is negligible for the higher-mean component (component 1) unless the sample size is too small (about $N = 300$) for both priors. The bias is more significant in the smaller-mean component (component 2). This is particularly true if we choose the posterior mean as a summary statistic with the non-informative prior. For the non-informative prior case, there is an upward-bias trend for both posterior mean and median in component 2. It is evident that the posterior median has much better bias properties than the posterior mean. On the other hand, for the weakly-informative prior case, there is a slightly upward-bias for the posterior mean, but the bias appears to be small. For the posterior median, there is a downward-bias in component 2 when the sample size is less than 1,000, but as the sample size increases this trend disappears and the bias becomes negligible. We can see that even though the prior information is weak, it introduces a bias in posterior median for the smaller mean value component when the sample size is small or moderate. On the other hand, as shown in the mean squared errors (MSE) for each case in Tables 7.6 and 7.7, because of the reduced variance in the weakly-informative prior case, its posterior mean and median perform better than those for the non-informative prior case. In sum, for the high mean value scenario, if we use the non-informative prior, the choice of posterior mean should be avoided in terms of bias and MSE. The bias risks for other cases seem to be minimal. However, as the sample size becomes larger (more than $N = 1{,}000$), the posterior median with a weakly-informative prior seems to be preferred.

Table 7.4 Averages and standard deviations (High-mean, $\phi \sim \Gamma(0.01, 0.01)$)

Sample size	100	200	300	400	500	600	700	800	900	1000	2000
$E(\hat{\beta}_{0,1})$	2.312	2.431	2.413	2.474	2.493	2.505	2.498	2.501	2.488	2.501	2.496
$sd(\hat{\beta}_{0,1})$	0.476	0.317	0.664	0.222	0.049	0.043	0.040	0.041	0.033	0.036	0.026
$E(\hat{\beta}_{1,1})$	-0.386	-0.456	-0.479	-0.478	-0.499	-0.501	-0.503	-0.498	-0.497	-0.499	-0.500
$sd(\hat{\beta}_{1,1})$	0.353	0.223	0.173	0.147	0.050	0.040	0.041	0.037	0.032	0.031	0.025
$E(\hat{\beta}_{2,1})$	0.380	0.448	0.490	0.480	0.502	0.502	0.501	0.498	0.503	0.502	0.498
$sd(\hat{\beta}_{2,1})$	0.330	0.213	0.110	0.161	0.043	0.041	0.034	0.033	0.036	0.032	0.023
$E(\hat{\phi}_{1,\, mean})$	6.743	5.696	5.098	5.302	5.178	5.143	5.025	5.030	4.885	5.059	4.972
$sd(\hat{\phi}_{1,\, mean})$	4.754	3.688	1.517	1.533	1.404	0.953	0.844	0.822	0.619	0.509	0.474
$E(\hat{\phi}_{1,\, median})$	5.344	5.178	4.897	5.122	5.070	5.051	4.957	4.969	4.836	5.011	4.950
$sd(\hat{\phi}_{1,\, median})$	3.319	2.620	1.422	1.316	1.314	0.919	0.830	0.816	0.606	0.500	0.472
$E(\hat{w}_1)$	0.404	0.408	0.400	0.403	0.400	0.402	0.404	0.405	0.405	0.401	0.402
$sd(\hat{w}_1)$	0.053	0.045	0.057	0.032	0.029	0.027	0.023	0.025	0.021	0.020	0.016
$E(\hat{\beta}_{0,2})$	1.184	1.068	1.035	1.037	1.003	1.000	1.013	0.997	0.998	0.998	1.001
$sd(\hat{\beta}_{0,2})$	0.472	0.300	0.202	0.214	0.054	0.045	0.046	0.039	0.041	0.036	0.026
$E(\hat{\beta}_{1,2})$	0.370	0.465	0.478	0.482	0.486	0.506	0.489	0.499	0.501	0.497	0.501
$sd(\hat{\beta}_{1,2})$	0.320	0.218	0.133	0.148	0.053	0.041	0.035	0.037	0.038	0.032	0.021
$E(\hat{\beta}_{2,2})$	-0.404	-0.450	-0.483	-0.468	-0.507	-0.498	-0.495	-0.496	-0.500	-0.500	-0.500
$sd(\hat{\beta}_{2,2})$	0.339	0.208	0.134	0.143	0.042	0.040	0.034	0.039	0.030	0.031	0.022
$E(\hat{\phi}_{2,\, mean})$	11.328	15.389	12.474	13.920	12.000	12.553	11.783	11.299	11.850	12.035	10.875
$sd(\hat{\phi}_{2,\, mean})$	6.345	8.833	6.978	7.287	5.616	6.000	4.700	5.353	5.634	5.100	2.612
$E(\hat{\phi}_{2,\, median})$	7.431	10.597	9.580	10.895	9.920	10.426	10.111	9.936	10.499	10.707	10.374
$sd(\hat{\phi}_{2,\, median})$	3.763	5.819	4.498	4.764	3.888	4.121	3.446	4.023	4.240	3.805	2.306
$E(\hat{w}_2)$	0.596	0.592	0.600	0.597	0.600	0.598	0.596	0.595	0.595	0.599	0.598
$sd(\hat{w}_2)$	0.053	0.045	0.057	0.032	0.029	0.027	0.023	0.025	0.021	0.020	0.016

Table 7.5 Averages and standard deviations (High-mean, $\phi \sim \Gamma(0.5, 0.1)$)

Sample size	100	200	300	400	500	600	700	800	900	1000	2000
$E(\hat{\beta}_{0,1})$	2.312	2.431	2.413	2.474	2.493	2.505	2.498	2.501	2.488	2.501	2.496
$sd(\hat{\beta}_{0,1})$	0.476	0.317	0.664	0.222	0.049	0.043	0.040	0.041	0.033	0.036	0.026
$E(\hat{\beta}_{1,1})$	-0.386	-0.456	-0.479	-0.478	-0.499	-0.501	-0.503	-0.498	-0.497	-0.499	-0.500
$sd(\hat{\beta}_{1,1})$	0.353	0.223	0.173	0.147	0.050	0.040	0.041	0.037	0.032	0.031	0.025
$E(\hat{\beta}_{2,1})$	0.380	0.448	0.490	0.480	0.502	0.502	0.501	0.498	0.503	0.502	0.498
$sd(\hat{\beta}_{2,1})$	0.330	0.213	0.110	0.161	0.043	0.041	0.034	0.033	0.036	0.032	0.023
$E(\hat{\phi}_{1,\,mean})$	6.743	5.696	5.098	5.302	5.178	5.143	5.025	5.030	4.885	5.059	4.972
$sd(\hat{\phi}_{1,\,mean})$	4.754	3.688	1.517	1.533	1.404	0.953	0.844	0.822	0.619	0.509	0.474
$E(\hat{\phi}_{1,\,median})$	5.344	5.178	4.897	5.122	5.070	5.051	4.957	4.969	4.836	5.011	4.950
$sd(\hat{\phi}_{1,\,median})$	3.319	2.620	1.422	1.316	1.314	0.919	0.830	0.816	0.606	0.500	0.472
$E(\hat{w}_1)$	0.404	0.408	0.400	0.403	0.400	0.402	0.404	0.405	0.405	0.401	0.402
$sd(\hat{w}_1)$	0.053	0.045	0.057	0.032	0.029	0.027	0.023	0.025	0.021	0.020	0.016
$E(\hat{\beta}_{0,2})$	1.184	1.068	1.035	1.037	1.003	1.000	1.013	0.997	0.998	0.998	1.001
$sd(\hat{\beta}_{0,2})$	0.472	0.300	0.202	0.214	0.054	0.045	0.046	0.039	0.041	0.036	0.026
$E(\hat{\beta}_{1,2})$	0.370	0.465	0.478	0.482	0.486	0.506	0.489	0.499	0.501	0.497	0.501
$sd(\hat{\beta}_{1,2})$	0.320	0.218	0.133	0.148	0.053	0.041	0.035	0.037	0.038	0.032	0.021
$E(\hat{\beta}_{2,2})$	-0.404	-0.450	-0.483	-0.468	-0.507	-0.498	-0.495	-0.496	-0.500	-0.500	-0.500
$sd(\hat{\beta}_{2,2})$	0.339	0.208	0.134	0.143	0.042	0.040	0.034	0.039	0.030	0.031	0.022
$E(\hat{\phi}_{2,\,mean})$	11.328	15.389	12.474	13.920	12.000	12.553	11.783	11.299	11.850	12.035	10.875
$sd(\hat{\phi}_{2,\,mean})$	6.345	8.833	6.978	7.287	5.616	6.000	4.700	5.353	5.634	5.100	2.612
$E(\hat{\phi}_{2,\,median})$	7.431	10.597	9.580	10.895	9.920	10.426	10.111	9.936	10.499	10.707	10.374
$sd(\hat{\phi}_{2,\,median})$	3.763	5.819	4.498	4.764	3.888	4.121	3.446	4.023	4.240	3.805	2.306
$E(\hat{w}_2)$	0.596	0.592	0.600	0.597	0.600	0.598	0.596	0.595	0.595	0.599	0.598
$sd(\hat{w}_2)$	0.053	0.045	0.057	0.032	0.029	0.027	0.023	0.025	0.021	0.020	0.016

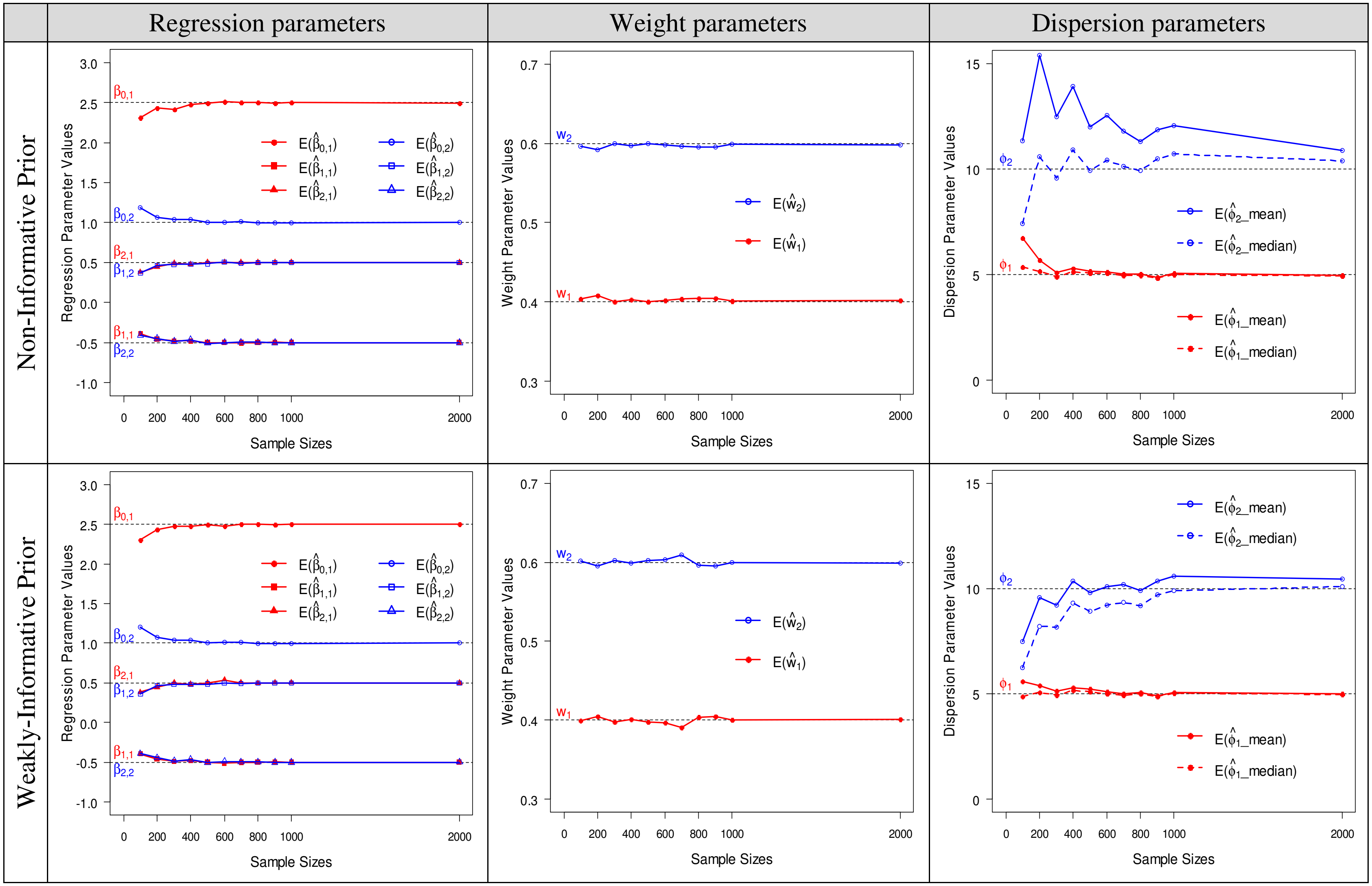


Figure 7.2 Bias trends for model parameters by sample size (High-mean scenario)

Table 7.6 MSEs for dispersion parameters (High-mean, $\phi \sim \Gamma(0.01, 0.01)$)

Sample size	100	200	300	400	500	600	700	800	900	1000	2000
$MSE(\hat{\phi}_{1,\, mean})$	25.64	14.09	2.31	2.44	2.00	0.93	0.71	0.68	0.40	0.26	0.23
$MSE(\hat{\phi}_{1,\, median})$	11.13	6.90	2.03	1.75	1.73	0.85	0.69	0.67	0.39	0.25	0.23
$MSE(\hat{\phi}_{2,\, mean})$	42.02	107.06	54.81	68.47	35.54	42.52	25.27	30.34	35.16	30.15	7.59
$MSE(\hat{\phi}_{2,\, median})$	20.76	34.22	20.40	23.50	15.12	17.16	11.89	16.19	18.23	14.98	5.46

Table 7.7 MSEs for dispersion parameters (High-mean, $\phi \sim \Gamma(0.5, 0.1)$)

Sample size	100	200	300	400	500	600	700	800	900	1000	2000
$MSE(\hat{\phi}_{1,\, mean})$	5.12	4.89	2.11	1.97	1.76	1.14	0.72	0.68	0.39	0.26	0.23
$MSE(\hat{\phi}_{1,\, median})$	3.65	3.72	1.91	1.59	1.61	1.07	0.69	0.66	0.39	0.25	0.23
$MSE(\hat{\phi}_{2,\, mean})$	14.00	11.82	9.96	10.97	10.24	9.05	7.60	8.40	10.52	9.42	4.81
$MSE(\hat{\phi}_{2,\, median})$	19.80	11.98	10.19	8.70	9.13	7.55	6.19	7.42	8.35	7.12	4.00

7.3.2 Moderate-mean value scenario

Many empirical crash data fall in the moderate sample-mean value scenario. Table 7.8 shows the ranges of generated sample means and variances for each sample size. Although the simulation objective was to generate samples ranging $1 < \bar{\mathbf{y}} < 5$, the generated sample means were much narrower with a minimum of 1.93 and a maximum of 3.86 under the parameter setting in Table 7.2.

Table 7.8 Ranges of generated sample means and variances (Moderate-mean)

Sample size		100	200	300	400	500	600	700	800	900	1000	2000
$\bar{\mathbf{y}}$	Min.	1.93	2.07	2.28	2.36	2.54	2.48	2.40	2.38	2.66	2.60	2.66
	Max.	3.80	3.75	3.80	3.79	3.61	3.52	3.86	3.57	3.65	3.52	3.28
$\text{var}(\mathbf{y})$	Min.	9.28	11.02	12.63	14.87	15.59	15.89	18.31	18.75	18.29	19.61	22.73
	Max.	68.98	85.37	143.20	52.85	69.33	67.21	129.90	55.47	67.32	57.55	58.32

Tables 7.9 and 7.10 provide the computed averages and standard deviations for all parameters. It should be noted that when $N = 100$ some of the generated datasets failed to converge or showed a frequent label switching even with the identifiability constraint

(i.e. $w_1 < w_2$) when they were fitted with the FMNB-2 model. This phenomenon was more frequently observed when the non-informative prior was used: there were 26 datasets for $\phi_k \sim \Gamma(0.01, 0.01)$ and 9 datasets for $\phi_k \sim \Gamma(0.5, 0.1)$. Therefore, those datasets were removed from the calculation for $N = 100$. This may also reflect that the use of a non-informative should be avoided when the sample size is very small.

Table 7.9 Averages and standard deviations (Moderate-mean, $\phi_k \sim \Gamma(0.01, 0.01)$)

Sample size	100	200	300	400	500	600	700	800	900	1000	2000
$E(\hat{\beta}_{0,1})$	1.761	1.888	1.951	1.956	1.964	1.988	1.978	1.965	1.982	1.988	1.991
$sd(\hat{\beta}_{0,1})$	0.211	0.155	0.111	0.106	0.094	0.075	0.083	0.066	0.074	0.061	0.045
$E(\hat{\beta}_{1,1})$	-0.572	-0.532	-0.523	-0.504	-0.506	-0.516	-0.496	-0.511	-0.507	-0.508	-0.495
$sd(\hat{\beta}_{1,1})$	0.260	0.132	0.110	0.089	0.072	0.068	0.060	0.053	0.054	0.047	0.035
$E(\hat{\beta}_{2,1})$	0.585	0.522	0.508	0.507	0.504	0.507	0.515	0.515	0.504	0.502	0.496
$sd(\hat{\beta}_{2,1})$	0.229	0.137	0.099	0.078	0.062	0.073	0.065	0.061	0.054	0.053	0.031
$E(\hat{\phi}_{1,mean})$	3.928	5.480	5.300	5.362	4.992	5.533	5.358	5.022	5.337	5.446	5.121
$sd(\hat{\phi}_{1,mean})$	3.073	3.893	2.853	3.906	2.310	1.955	2.376	1.642	1.568	1.848	0.786
$E(\hat{\phi}_{1,median})$	2.348	4.162	4.501	4.707	4.631	5.177	5.080	4.828	5.150	5.266	5.044
$sd(\hat{\phi}_{1,median})$	1.692	2.466	2.137	2.854	1.968	1.750	1.973	1.532	1.470	1.716	0.756
$E(\hat{w}_1)$	0.277	0.228	0.217	0.212	0.208	0.206	0.207	0.209	0.203	0.204	0.202
$sd(\hat{w}_1)$	0.060	0.048	0.034	0.032	0.026	0.023	0.023	0.020	0.019	0.020	0.013
$E(\hat{\beta}_{0,2})$	-0.009	-0.008	-0.002	0.001	0.001	0.006	-0.002	0.006	0.008	-0.001	-0.003
$sd(\hat{\beta}_{0,2})$	0.173	0.108	0.085	0.080	0.073	0.062	0.053	0.051	0.054	0.056	0.032
$E(\hat{\beta}_{1,2})$	0.553	0.535	0.506	0.499	0.504	0.498	0.507	0.500	0.497	0.503	0.499
$sd(\hat{\beta}_{1,2})$	0.148	0.100	0.073	0.071	0.051	0.053	0.040	0.044	0.040	0.038	0.025
$E(\hat{\beta}_{2,2})$	-0.558	-0.511	-0.498	-0.497	-0.502	-0.497	-0.498	-0.501	-0.501	-0.503	-0.503
$sd(\hat{\beta}_{2,2})$	0.138	0.094	0.071	0.056	0.050	0.045	0.049	0.037	0.041	0.039	0.025
$E(\hat{\phi}_{2,mean})$	10.679	13.275	12.199	13.513	14.303	14.013	13.943	14.569	13.037	13.340	13.311
$sd(\hat{\phi}_{2,mean})$	4.295	6.492	5.852	7.144	7.461	7.447	7.448	7.045	7.644	6.498	6.650
$E(\hat{\phi}_{2,median})$	5.841	8.093	7.957	9.209	10.093	10.004	9.970	10.537	10.047	10.144	11.194
$sd(\hat{\phi}_{2,median})$	2.511	4.123	3.526	4.618	4.802	4.900	4.799	4.481	5.090	4.105	4.792
$E(\hat{w}_2)$	0.723	0.772	0.783	0.788	0.792	0.794	0.793	0.791	0.797	0.796	0.798
$sd(\hat{w}_2)$	0.060	0.048	0.034	0.032	0.026	0.023	0.023	0.020	0.019	0.020	0.013

The bias trends for this scenario were similar to the high mean value scenario as depicted in Figure 7.3. The biases associated with the regression parameters and weight parameters are negligible unless the sample size is too small, and they are not affected by the choice of priors on ϕ_k.

Table 7.10 Averages and standard deviations (Moderate-mean, $\phi_k \sim \Gamma(0.5, 0.1)$)

Sample size	100	200	300	400	500	600	700	800	900	1000	2000
$E(\hat{\beta}_{0,1})$	-0.595	-0.534	-0.525	-0.505	-0.508	-0.515	-0.495	-0.511	-0.505	-0.508	1.994
$sd(\hat{\beta}_{0,1})$	0.273	0.130	0.108	0.087	0.071	0.069	0.060	0.053	0.054	0.046	0.045
$E(\hat{\beta}_{1,1})$	0.575	0.526	0.508	0.507	0.505	0.505	0.514	0.517	0.503	0.501	-0.495
$sd(\hat{\beta}_{1,1})$	0.236	0.135	0.097	0.076	0.064	0.075	0.064	0.061	0.054	0.053	0.035
$E(\hat{\beta}_{2,1})$	3.533	4.758	4.968	4.996	4.906	5.528	5.326	5.022	5.370	5.469	0.496
$sd(\hat{\beta}_{2,1})$	1.767	2.287	1.982	2.340	1.825	1.784	1.882	1.452	1.498	1.780	0.032
$E(\hat{\phi}_{1,mean})$	2.538	4.069	4.431	4.577	4.614	5.215	5.091	4.836	5.188	5.294	5.147
$sd(\hat{\phi}_{1,mean})$	1.368	1.926	1.741	2.060	1.668	1.639	1.728	1.373	1.420	1.665	0.766
$E(\hat{\phi}_{1,median})$	0.264	0.224	0.216	0.211	0.207	0.204	0.206	0.208	0.201	0.203	5.079
$sd(\hat{\phi}_{1,median})$	0.063	0.045	0.034	0.031	0.026	0.023	0.023	0.020	0.019	0.019	0.743
$E(\hat{w}_1)$	-0.012	-0.004	0.000	0.004	0.004	0.009	0.000	0.009	0.010	0.001	0.201
$sd(\hat{w}_1)$	0.171	0.108	0.084	0.080	0.073	0.062	0.052	0.051	0.053	0.057	0.012
$E(\hat{\beta}_{0,2})$	0.540	0.534	0.505	0.498	0.504	0.496	0.506	0.499	0.496	0.502	-0.001
$sd(\hat{\beta}_{0,2})$	0.147	0.101	0.073	0.071	0.052	0.052	0.040	0.044	0.040	0.037	0.031
$E(\hat{\beta}_{1,2})$	-0.545	-0.510	-0.497	-0.495	-0.501	-0.496	-0.497	-0.500	-0.500	-0.503	0.499
$sd(\hat{\beta}_{1,2})$	0.139	0.093	0.071	0.056	0.049	0.045	0.049	0.036	0.042	0.040	0.024
$E(\hat{\beta}_{2,2})$	6.184	7.985	8.009	8.907	9.565	9.527	9.497	10.001	9.678	9.836	-0.501
$sd(\hat{\beta}_{2,2})$	1.976	2.605	2.529	2.847	3.073	3.100	3.089	2.971	3.363	2.890	0.025
$E(\hat{\phi}_{2,mean})$	4.707	6.461	6.686	7.553	8.186	8.241	8.256	8.734	8.599	8.736	10.777
$sd(\hat{\phi}_{2,mean})$	1.693	2.332	2.138	2.530	2.632	2.692	2.706	2.595	2.941	2.504	3.078
$E(\hat{\phi}_{2,median})$	0.736	0.776	0.784	0.789	0.793	0.796	0.794	0.792	0.799	0.797	9.891
$sd(\hat{\phi}_{2,median})$	0.063	0.045	0.034	0.031	0.026	0.023	0.023	0.020	0.019	0.019	2.742
$E(\hat{w}_2)$	-0.595	-0.534	-0.525	-0.505	-0.508	-0.515	-0.495	-0.511	-0.505	-0.508	0.799
$sd(\hat{w}_2)$	0.273	0.130	0.108	0.087	0.071	0.069	0.060	0.053	0.054	0.046	0.012

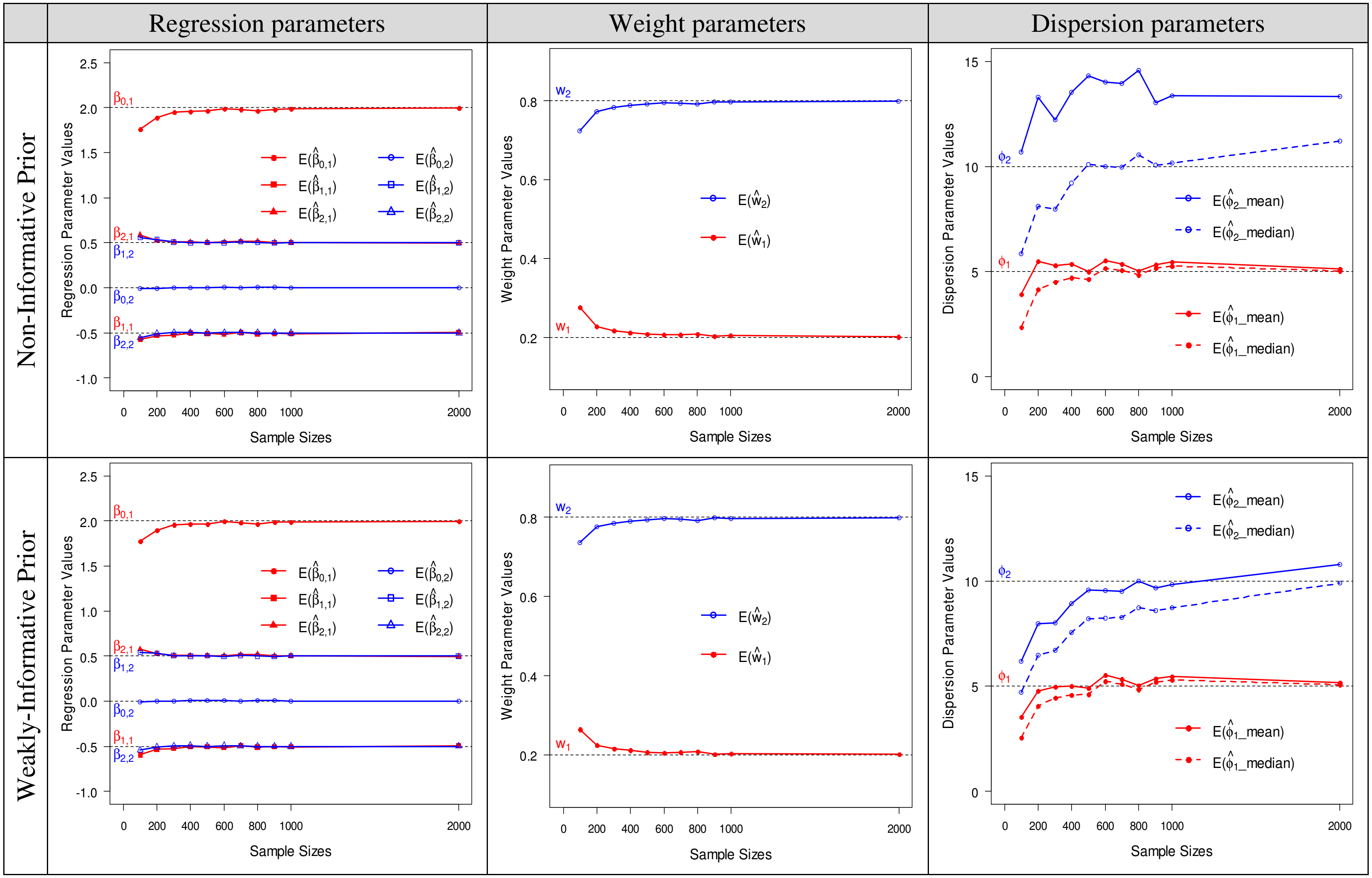


Figure 7.3 Bias trends for model parameters by sample size (Moderate-mean scenario)

For the biases associated with the dispersion parameters, the upward or downward trends are more pronounced, especially for the smaller-mean component. Furthermore, it requires higher sample sizes than the high mean value scenario to obtain the similar amount of bias for components 1 and 2 (about $N = 500$). Obviously the posterior mean with the non-informative prior is not an option. It is interesting to notice that up to $N = 1{,}000$, the bias for the posterior median with the non-informative prior is at its minimum, but as the sample size increases significantly larger (i.e. $N = 2{,}000$), it starts to exhibit the upward bias trend (for component 2). A similar tendency is also observed in the posterior mean with a weakly-informative prior. On the other hand, the posterior median with the weakly-informative prior is consistently lower than the true value up to sample size $N = 2{,}000$. The MSE information for both priors is provided in Tables 7.11 and 7.12, respectively. Even though there is a downward bias in the posterior mean and median when a weakly-informative prior is used, because of the reduced variability in the estimates they are performing better than the posterior median with the non-informative prior in all sample sizes.

Table 7.11 MSEs for dispersion parameters (Moderate-mean, $\phi_k \sim \Gamma(0.01, 0.01)$))

Sample size	100	200	300	400	500	600	700	800	900	1000	2000
$MSE(\hat{\phi}_{1,\,mean})$	10.59	15.37	8.23	15.39	5.34	4.11	5.78	2.70	2.57	3.61	0.63
$MSE(\hat{\phi}_{1,\,median})$	9.89	6.78	4.82	8.23	4.01	3.09	3.90	2.38	2.18	3.01	0.57
$MSE(\hat{\phi}_{2,\,mean})$	18.91	52.88	39.08	63.38	74.18	71.56	71.02	70.51	67.65	53.37	55.19
$MSE(\hat{\phi}_{2,\,median})$	23.60	20.64	16.61	21.95	23.07	24.01	23.03	20.37	25.91	16.87	24.39

Table 7.12 MSEs for dispersion parameters (Moderate-mean, $\phi_k \sim \Gamma(0.5, 0.1)$))

Sample size	100	200	300	400	500	600	700	800	900	1000	2000
$MSE(\hat{\phi}_{1,\,mean})$	5.27	5.29	3.93	5.48	3.34	3.46	3.65	2.11	2.38	3.39	0.61
$MSE(\hat{\phi}_{1,\,median})$	7.93	4.58	3.36	4.42	2.93	2.73	2.99	1.91	2.05	2.86	0.56
$MSE(\hat{\phi}_{2,\,mean})$	18.47	10.85	10.36	9.30	9.64	9.83	9.80	8.83	11.41	8.38	10.08
$MSE(\hat{\phi}_{2,\,median})$	30.88	17.96	15.55	12.38	10.22	10.34	10.36	8.34	10.61	7.87	7.53

7.3.3 Small-mean value scenario

For the small-mean value scenario, much larger sample size was necessary to obtain stable parameter estimates and hence examine the bias properties. When the sample size was below 500, the parameter estimates were very unstable and a lot of generated datasets showed the label switching problem even with the order constraints. Table 7.13 shows the ranges of generated sample means and variances for sample sizes which start from $N = 500$ with a five-hundred step.

Table 7.13 Ranges of generated sample means and variances (Small-mean)

Sample size		500	1000	1500	2000	2500	3000	3500	5000
$\bar{\mathbf{y}}$	Min.	0.67	0.70	0.74	0.75	0.76	0.73	0.76	0.76
	Max.	0.98	0.96	0.890	0.89	0.90	0.87	0.88	0.86
var(**y**)	Min.	1.14	1.40	1.682	1.78	1.82	1.80	1.79	1.94
	Max.	3.46	4.50	3.203	3.52	3.78	2.84	2.96	3.07

Tables 7.14 and 7.15 provide the computed averages and standard deviations for all parameters. It should be noted that when $N = 500$ some of the generated datasets failed to converge or showed a frequent label switching even with the identifiability constraint (i.e. $w_1 < w_2$) when they were fitted with the FMNB-2 model. This phenomenon was more frequently observed when the non-informative prior was used: there were 13 datasets for $\phi_k \sim \Gamma(0.01, 0.01)$ and 5 datasets for $\phi_k \sim \Gamma(0.5, 0.1)$. Therefore, those datasets were removed from the calculation for $N = 500$.

오류! 참조 원본을 찾을 수 없습니다. shows the bias trends for this scenario. The posterior mean estimates associated with the regression parameters and weight parameters appear to be still consistent considering a very small bias, and they are not affected by the choice of priors on ϕ_k like other sample-mean value scenarios.

Table 7.14 Averages and standard deviations (Small-mean, $\phi_k \sim \Gamma(0.01, 0.01)$)

Sample size	500	1000	1500	2000	2500	3000	3500	5000
$E(\hat{\beta}_{0,1})$	0.350	0.417	0.454	0.470	0.465	0.486	0.479	0.482
$sd(\hat{\beta}_{0,1})$	0.197	0.149	0.110	0.093	0.088	0.087	0.074	0.058
$E(\hat{\beta}_{1,1})$	-0.515	-0.498	-0.494	-0.508	-0.499	-0.501	-0.499	-0.506
$sd(\hat{\beta}_{1,1})$	0.128	0.094	0.061	0.061	0.056	0.052	0.041	0.042
$E(\hat{\beta}_{2,1})$	0.507	0.528	0.499	0.507	0.508	0.498	0.499	0.506
$sd(\hat{\beta}_{2,1})$	0.136	0.084	0.075	0.055	0.050	0.049	0.040	0.038
$E(\hat{\phi}_{1,\,mean})$	6.262	5.917	5.903	6.845	5.589	5.559	5.895	5.195
$sd(\hat{\phi}_{1,\,mean})$	4.650	4.347	4.025	3.914	3.091	2.368	2.681	1.747
$E(\hat{\phi}_{1,\,median})$	3.869	4.379	4.694	5.590	4.917	4.969	5.309	4.925
$sd(\hat{\phi}_{1,\,median})$	2.614	2.712	2.575	2.572	2.247	1.788	2.015	1.463
$E(\hat{w}_1)$	0.231	0.221	0.215	0.206	0.209	0.206	0.206	0.205
$sd(\hat{w}_1)$	0.041	0.036	0.031	0.024	0.020	0.020	0.019	0.015
$E(\hat{\beta}_{0,2})$	-1.017	-0.999	-1.005	-1.003	-1.000	-1.005	-1.000	-1.001
$sd(\hat{\beta}_{0,2})$	0.158	0.092	0.084	0.068	0.057	0.059	0.051	0.041
$E(\hat{\beta}_{1,2})$	0.493	0.502	0.506	0.501	0.497	0.500	0.504	0.501
$sd(\hat{\beta}_{1,2})$	0.091	0.066	0.064	0.048	0.042	0.039	0.037	0.032
$E(\hat{\beta}_{2,2})$	-0.526	-0.513	-0.507	-0.499	-0.510	-0.506	-0.496	-0.501
$sd(\hat{\beta}_{2,2})$	0.096	0.068	0.061	0.036	0.039	0.046	0.042	0.030
$E(\hat{\phi}_{2,\,mean})$	11.894	13.503	14.692	14.935	14.427	14.399	14.599	14.340
$sd(\hat{\phi}_{2,\,mean})$	5.192	5.412	7.764	7.706	7.627	7.479	7.501	7.486
$E(\hat{\phi}_{2,\,median})$	6.888	8.335	9.402	9.989	9.882	10.036	10.448	10.619
$sd(\hat{\phi}_{2,\,median})$	3.088	3.230	4.957	4.967	4.908	4.747	4.794	4.712
$E(\hat{w}_2)$	0.769	0.779	0.785	0.794	0.791	0.794	0.794	0.795
$sd(\hat{w}_2)$	0.041	0.036	0.031	0.024	0.020	0.020	0.019	0.015

Table 7.15 Averages and standard deviations (Small-mean, $\phi_k \sim \Gamma(0.5, 0.1)$)

Sample size	500	1000	1500	2000	2500	3000	3500	5000
$E(\hat{\beta}_{0,1})$	0.355	0.428	0.463	0.466	0.468	0.492	0.482	0.483
$sd(\hat{\beta}_{0,1})$	0.169	0.137	0.100	0.089	0.087	0.079	0.071	0.057
$E(\hat{\beta}_{1,1})$	-0.511	-0.496	-0.493	-0.509	-0.499	-0.499	-0.499	-0.507
$sd(\hat{\beta}_{1,1})$	0.133	0.092	0.062	0.063	0.056	0.050	0.041	0.042
$E(\hat{\beta}_{2,1})$	0.504	0.526	0.499	0.508	0.508	0.497	0.499	0.506
$sd(\hat{\beta}_{2,1})$	0.127	0.086	0.073	0.057	0.052	0.049	0.040	0.039
$E(\hat{\phi}_{1,mean})$	4.510	5.057	5.223	5.932	5.312	5.455	5.696	5.163
$sd(\hat{\phi}_{1,mean})$	2.251	2.434	2.196	2.382	2.091	1.816	2.020	1.463
$E(\hat{\phi}_{1,median})$	3.505	4.262	4.568	5.285	4.851	5.025	5.287	4.915
$sd(\hat{\phi}_{1,median})$	1.843	2.006	1.842	2.009	1.785	1.574	1.765	1.319
$E(\hat{w}_1)$	0.230	0.218	0.212	0.207	0.208	0.204	0.205	0.204
$sd(\hat{w}_1)$	0.040	0.033	0.029	0.023	0.019	0.019	0.018	0.015
$E(\hat{\beta}_{0,2})$	-1.018	-0.999	-1.001	-1.002	-0.998	-1.003	-0.997	-0.998
$sd(\hat{\beta}_{0,2})$	0.157	0.095	0.083	0.068	0.058	0.059	0.051	0.043
$E(\hat{\beta}_{1,2})$	0.491	0.499	0.503	0.501	0.496	0.498	0.501	0.499
$sd(\hat{\beta}_{1,2})$	0.089	0.065	0.063	0.048	0.041	0.037	0.036	0.032
$E(\hat{\beta}_{2,2})$	-0.524	-0.512	-0.503	-0.499	-0.508	-0.504	-0.494	-0.499
$sd(\hat{\beta}_{2,2})$	0.090	0.069	0.061	0.038	0.038	0.046	0.041	0.030
$E(\hat{\phi}_{2,mean})$	7.098	8.277	8.827	9.310	9.379	9.522	9.865	10.121
$sd(\hat{\phi}_{2,mean})$	2.048	2.158	3.051	2.987	3.103	3.015	3.055	3.119
$E(\hat{\phi}_{2,median})$	5.553	6.749	7.342	7.857	8.027	8.179	8.569	8.940
$sd(\hat{\phi}_{2,median})$	1.763	1.899	2.642	2.576	2.673	2.611	2.632	2.682
$E(\hat{w}_2)$	0.770	0.782	0.788	0.793	0.792	0.796	0.795	0.796
$sd(\hat{w}_2)$	0.040	0.033	0.029	0.023	0.019	0.019	0.018	0.015

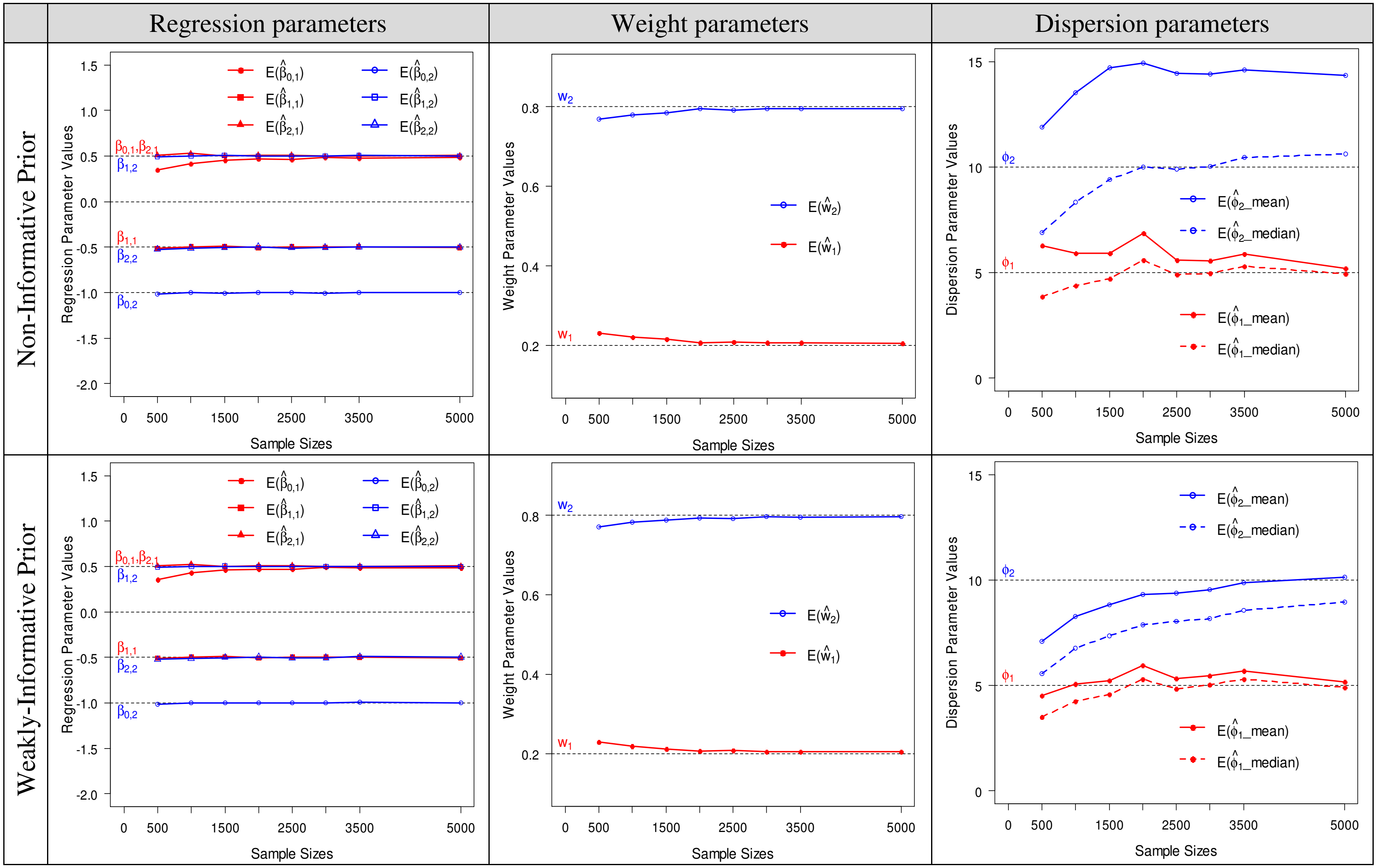


Figure 7.4 Bias trends for model parameters by sample size (Small-mean scenario)

For the biases associated with the dispersion parameters, the posterior mean exhibits a high upward bias for component 1, not to mention for component 2 when the non-informative prior was used. The posterior median seems to be working fine for both components after $N = 1{,}500$, but as the sample size grows the upward bias trend becomes noticeable especially for component 2. For the weakly-informative prior case, the dispersion parameter for component 2 is consistently underestimated for both posterior mean and median. This tendency was already observed in the moderate sample mean value scenario. We can infer from this result that the weakly-informative prior is exercising much influence when the component mean value is low and its sample size is small. It is pulling down the posterior mean toward the prior mean which is 5, which deteriorates the bias properties of posterior mean or median. However, in terms of MSE (Tables 7-16 and 7-17), in spite of the significant underestimation, the summary statistics for the weakly-informative prior perform better than those of the non-informative prior in most sample sizes.

Table 7.16 MSEs for dispersion parameters (Small-mean, $\phi_k \sim \Gamma(0.01, 0.01)$)

Sample size	500	1000	1500	2000	2500	3000	3500	5000
$MSE(\hat{\phi}_{1,\,mean})$	23.22	19.73	17.02	18.73	9.90	5.92	7.99	3.09
$MSE(\hat{\phi}_{1,\,median})$	8.12	7.74	6.72	6.96	5.05	3.20	4.16	2.15
$MSE(\hat{\phi}_{2,\,mean})$	30.54	41.56	82.30	83.74	77.77	75.29	77.41	74.87
$MSE(\hat{\phi}_{2,\,median})$	19.22	13.21	24.93	24.67	24.11	22.54	23.19	22.58

Table 7.17 MSEs for dispersion parameters (Small-mean, $\phi_k \sim \Gamma(0.5, 0.1)$)

Sample size	500	1000	1500	2000	2500	3000	3500	5000
$MSE(\hat{\phi}_{1,\,mean})$	5.31	5.93	4.87	6.54	4.47	3.50	4.57	2.17
$MSE(\hat{\phi}_{1,\,median})$	5.63	4.57	3.58	4.12	3.21	2.48	3.20	1.75
$MSE(\hat{\phi}_{2,\,mean})$	12.62	7.62	10.69	9.40	10.02	9.32	9.35	9.74
$MSE(\hat{\phi}_{2,\,median})$	22.88	14.18	14.05	11.23	11.04	10.13	8.98	8.32

7.4 Chapter Summary and Recommendation

The simulation study conducted on the FMNB-2 model showed that the posterior mean using the non-informative prior exhibited a high bias for the dispersion parameter especially, in the smaller-mean value component. The posterior median, instead, had much better bias properties than the posterior mean, particularly at small sample sizes and small sample-mean values. However, as the sample size increases significantly for both small to moderate mean value scenarios, the posterior median using the non-informative prior also began to exhibit the upward bias trend. This is because as the sample size increases the posterior median is getting closer to the posterior mean which exhibits the upward bias. The use of the weakly-informative prior had the advantage of reducing the variability in the estimates for the posterior mean and median, but it tended to underestimate the true value by pulling the estimates toward its prior mean. As the sample-mean value decreases this tendency was more pronounced.

Based on the results of this chapter, we suggest guidelines about the selection of priors and the corresponding summary statistics to use in terms of their bias properties. The guidelines are tabulated in Table 7.18 for different sample sizes and sample mean values. As indicated in the introduction, since the necessary sample size in FMNB-2 model greatly depends on the current dataset, the sample size ranges suggested in Table 7.18 are for a relatively well-separated data. The minimum sample sizes in each sample mean range (i.e. $N = 300$ for high mean, $N = 500$ for moderate mean, and $N = 1{,}500$ for small mean) was basically determined by the bias associated with the dispersion parameters, but they also minimize the biases in the regression coefficients and the mixing proportions. The bias related to the regression coefficients and the mixing proportions were almost negligible regardless of sample sizes and sample-mean values, as compared to the bias associated with the dispersion parameters.

Table 7.18 Recommended priors and summary statistics in terms of bias properties

Sample Mean Range	Sample Size Range	Recommended Priors	Recommended Summary Statistics
High ($\bar{\mathbf{y}} > 5$)	300 – 1,000	Non-informative prior	Posterior median
		Weakly-informative prior	Posterior mean
	> 1,000	Weakly-informative prior	Posterior median
Moderate ($1 < \bar{\mathbf{y}} < 5$)	500 – 1,000	Non-informative prior	Posterior median
		Weakly-informative prior	Posterior mean
	1,000 – 2,000	Weakly-informative prior	Posterior mean
	> 2,000	Weakly-informative prior	Posterior median
Small ($\bar{\mathbf{y}} < 1$)	1,500 – 3,000	Non-informative prior	Posterior median
	> 3,500	Weakly-informative prior	Posterior mean

CHAPTER VIII

SUMMARY AND CONCLUSIONS

Over-dispersion caused by unobserved heterogeneity is very common in motor vehicle crash data and failure to accommodate such heterogeneity in the model can undermine the validity of the model results. The negative binomial regression model, which is a continuous mixture of Poisson/Gamma distributions, has been a usual choice for accommodating over-dispersion in highway safety study. However, it is often likely that empirical frequencies of crash data do not follow the particular distribution assumed for the Poisson mean rate. In this respect, the primary objective of this research was to examine the applicability of an alternative model formulation that could be used for capturing heterogeneity through the use of finite mixtures of regression models. In finite mixture models, it is assumed that the observations of a sample arise from two or more unobserved components with unknown proportions. The advantage of using a finite mixture model is that it allows the data to determine the true relationships by choosing a finite number of unobserved latent components without making a particular distributional assumption on the mixing variable. The downside of the model is the difficulty in determining the optimal number of components. The final outputs of finite mixture models are the number of components, the proportion of each component and the component-specific regression parameter estimates.

In order to investigate the performance of the finite mixture models in vehicle crash data, finite mixtures of Poisson regression models (FMP-K) and finite mixtures of negative binomial regression models (FMNB-K) were formulated and their respective performances were compared to that of the single negative binomial regression model (NB) using both simulated and empirical crash datasets. For model parameter estimation, a Bayesian approach was adopted since it provides much richer inference than the maximum likelihood approach. For a comparative purpose, the maximum likelihood estimates were also computed where appropriate.

This chapter highlights the main findings from this research and suggests a few recommendations for using the mixture models in highway safety research. This dissertation ends with a discussion on possible directions in which the research can be extended.

8.1 Main Findings

Using simulated datasets, first, it was shown that the CFMP-K model (i.e., a FMP-K model with regression parameters constrained to be the same) could effectively approximate the continuous mixture of Poisson/Gamma model with only a few numbers of mass points and their respective proportions. The necessary number of components depended on the sample size, sample mean and the degree of dispersion. Second, the examples for FMP-2 and FMNB-2 models demonstrated that the single NB regression model was not a viable option in terms of model prediction and parameter interpretation if the source of over-dispersion is due to population heterogeneity. If there is an extra-variation within each component, which may often be the case in crash data, the FMP-2 model was not preferred because such heterogeneity resulted in the under-estimation of standard errors of the model parameters. In both cases, the FMNB-2 model was a good candidate model.

The applications with two empirical crash datasets showed that a two-component finite mixture of NB regression models was quite enough to characterize the randomness of crash occurrence and it provided useful information on features of the population under study. For the intersection crash dataset, the FMNB-2 model did not improve the goodness-of-fit per se. However, it showed a possibility of the different effects of approaching traffic flows on each component, which could not be detected if we used the single aggregate NB model. This information is valuable because we are often interested in assessing the effect of a covariate on crash occurrence from the estimated coefficient. The coefficients estimated from the FMNB-2 model are considered more realistic in that it takes account of the different effects on the different subpopulations. In contrast, the coefficients from the NB model only consider the average effect of a covariate across all intersections in the sample. On the other hand, for the segment crash dataset, the data separation was more distinct and the FMNB-2 (or CFMNB-2) model actually improved the goodness-of-fit. For both cases, the FMP-K model had a tendency to produce too many components, making it difficult to interpret the effects of model parameters. From an application point of view, the FMNB-2 model was considered more useful and parsimonious. However, it may be premature to conclude that crash data can always be approximated with only two components.

In Chapter VI, we applied the developed model (i.e., CFMNB-2 model) for the segment crash dataset to the hotspot identification and the accident medication factor development. Although the difference in the hotspot rankings between the NB and CFMNB-2 models was minor, it is quite possible that the difference can be more pronounced depending on the data under study. In such a case, the ranking results from the CFNB-2 model should be preferred because of the better model specification. This was supported by the simulation study that aimed to demonstrate a high number of false positives and negatives when the mis-specified model was used for identifying hotspots. The simulation study also identified the relationship between the threshold values and the hotspot identification performance criteria. For a judicious use of transportation funds, highway safety managers are recommended to estimate the cost difference resulting from the false positives and the false negatives, and should decide the optimal threshold value based on the trade-off between the two costs. On the other hand, the accident modification factor (AMF) curve equation for the FMNB-2 model was derived. The resultant AMF function for a certain covariate had two good properties over the one from the NB model. The first one was that the safety effect of a covariate was better reflected by the AMF function from the FMNB-2 model, since the model takes into account the differential responsiveness of crash frequency to the covariate. The second one was that the safety effect of a covariate did not increase continuously without a limit, but leveled off after a certain value of the covariate. This made more logical sense. However, there was also a possibility that a U-shaped relationship could be found. This may raise some debate among highway safety community.

Finally, given the superior performance of the FMNB-2 model in crash data, we characterized the bias properties of posterior summary statistics (posterior mean and median) for the dispersion parameters in FMNB-2 model through the simulation study. The result from this study is important if we are to use the calibrated model for future prediction. The biased parameters will degrade the reliability of the predicted values and their confidence intervals as well. While the bias associated with the regression parameters was minimal regardless of sample sizes and sample-mean values considered, the bias for the dispersion parameter was significant. The results showed that the posterior mean using a non-informative prior exhibited a high bias for the dispersion parameter and should be avoided when the dataset contains less than 2,000 observations (even for high sample-mean values). The posterior median showed much better bias properties, particularly at small sample sizes and small sample mean values. However, as the sample size increases, the posterior median using a non-informative prior also began to exhibit an upward bias trend. In such cases, the posterior mean or median with the weakly-informative prior provided a smaller bias. Based on the simulation results, general guidelines were also provided about the choice of priors and the summary statistics to use for different sample sizes and sample mean values. The minimum sample sizes for each sample-mean category were $N = 300$ for high mean, $N = 500$ for moderate mean, and $N = 1{,}500$ for small mean.

8.2 Recommendations

Based on the findings from this research, we are suggesting the following recommendations for using the mixture models in highway safety research:

1. Highway safety analysts should consider the use of finite mixture models before the single aggregate NB model if crash data are suspected to be generated from different sup-populations.

2. If the data are fitted with a single NB regression model and the resulting Pearson χ^2/(degree of freedom) is much larger than 1.0, this indicates that the variance structure in the residuals is not distributed in the negative binomial manner. In this case, the NB model specification should be questioned and the finite mixture models may be good candidate models for this dataset.

3. Even when the goodness-of-fit of the NB model is satisfactory, the consideration of using finite mixture models is still valid since the over-dispersion might have been caused by heterogeneity in the covariates. The NB model ignores such heterogeneity in the data by taking the average effect of each covariate across all observations.

4. When there is a significantly large number of zeros in a dataset, researchers often resorted to the zero-inflated type of models to increase the model fit statistics. Because of the logic problems regarding the crash data generation process inherent in those models, the use of finite mixture models can be considered. In fact, the latter embraces the ZIP or ZINB model as a special case. However, many zeros may result in a very small sample-mean value unless the data are highly dispersed with many large numbers as well. In such a case, the sample size should be sufficiently large enough to obtain the unbiased or less biased parameter estimates.

5. A certain degree of caution has to be exercised when developing the accident modification factor function for an interest covariate. The AMF curve shape produced by a finite mixture model has a better property in that the safety effect of a covariate eventually levels off as the covariate increases significantly from the base condition. However, this is not always the case. A U-shaped curve can

be sometimes produced. The U-shaped relationship was partly supported by some researchers, but is not yet widely accepted by highway safety community.

8.3 Future Research Areas

Although the objectives of this research have been achieved, there are some limitations and valuable extensions that merit further study in the future.

- In this research, we allowed each mixture component to have its own regression coefficients as well as dispersion parameters in the FMNB-K model. This is a general setup, but many variants are possible in which one can appreciate the mixed-effects of model parameters. For example, model parameters can be allowed to vary for all components, vary between groups of components, or to be fixed over all components. In order to encompass all these variants, we may need a more general sampling algorithm than the one provided in Subsection *3.4.4* in Chapter III. During this research we tried only a limited number of variants, such as the CFMP-K model in which only intercepts were allowed to vary across components, and the CFMNB-2 in which some regression parameters in one component were constrained to be zero.
- Along the same line as above, the weight distribution ($\mathbf{w}$) used in both FMP-K and FMNB-K can be generalized by including some covariates in the discrete mixing distribution. As mentioned in Subsection *2.4.2*, the use of varying weight factors was beyond the scope of this research because of estimation complexity. Although there is no guarantee that this generalized model would improve the fit, it is worthwhile to compare it with the constant weight model. The advantage of the varying weight model is that we can identify the covariates that contribute to the separation of data.
- Another interesting research area is to compare model performances between the two-component finite mixture models (FMP-2 and FMNB-2) and the zero-inflated

models (ZIP and ZINB) when data contain many zeros. This can be done with either simulated data or empirical crash data if available. The findings can answer some important issues about adopting zero-inflated models for modeling highway safety data. Recall that if the finite mixture model is restricted to have two components and the mean for one of the components is constrained to be zero, then we have the zero-inflated model.

- Among the ranking criteria for the identification of hotspots, we only used the conditional mean of crash frequency obtained from Equations (6.1) and (6.2). Alternatively, several estimators from the posterior distribution can also be considered including the posterior mean of crash frequency, the potential of accident reduction, and the posterior expectation of ranks (Miranda-Moreno, 2006). If the empirical Bayesian (EB) method is preferred, the derivation of EB estimates for the finite mixture models may be necessary.

- In the simulation study in Chapter VII, we provided the guidelines on the minimum sample sizes for different sample-mean categories. However, these guidelines are based on the results from the limited combinations of simulation design values. To fully understand the bias properties, larger scale simulation studies may be required in the future.

- Finally, despite many advantages of using finite mixture models, there are still several unresolved issues especially for the label switching problem and the determination of optimal number of components. Judging from the literature review on this area, it seems that there is no consensus method for these issues yet. In this research, some of unidentified models (such as FMP-4 or 5 and FMNB-3) were excluded for further analyses and confined the analyses to the models whose label switching problems can be easily corrected by imposing identifiability constraints on the component’s parameters. This may be an obvious limitation of this research and further work should be carried out with more advanced technologies.

REFERENCES

Aitkin, M., 1996. A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and Computing 6 (3), 251-262.

Aitkin, M., 2001. Likelihood and Bayesian analysis of mixtures. Statistical Modelling 1 (4), 287-304.

Anastasopoulos, P., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153-159.

ASHTO, 2001. A Policy on Geometric Design of Highway and Streets. American Association of State Highway and Transportation Officials, Washington, D.C.

Bonneson, J.A., Lord, D., 2005. Role and Application of Accident Modification Factors in the Highway Design Process. Report No. FHWA/TX-05/0-4703-2, Texas Transportation Institute, College Station, Texas.

Bonneson, J.A., Lord, D., Zimmerman, K., Fitzpatrick, K., Pratt, M., 2007. Development of Tools for Evaluating the Safety Implications of Highway Design Decisions. Report No. FHWA/TX-07/0-4703-4, Texas Transportation Institute, College Station, Texas.

Brännäs, K., Rosenqvist, G., 1994. Semiparametric estimation of heterogeneous count data models. European Journal of Operational Research 76 (2), 247-258.

Burnham, K., Anderson, D. 2004. Multimodal inference: understanding AIC and BIC in model selection. Sociological Methods Research 33 (2), 261-304.

Cameron, A., Trivedi, P., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes Methods for Data Analysis (Second Edition). Chapman & Hall/CRC, Boca Raton, Florida

Celeux, G., 1998. Bayesian inference for mixtures: the label-switching problem. In COMPSTAT 98 (pp. 227-232). Payne, R., Green, P. J. (Eds.). Physica, Heidelberg.

Celeux, G, Hurn, M., Robert, C.P., 2000. Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association 95 (451), 957-970.

Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accident Analysis and Prevention 37 (5), 870-881.

Chib, S., Greenberg, E., Winkelmann, R., 1998. Posterior simulation and Bayes factors in panel count data models. Journal of Econometrics 86 (1), 33-54.

Christiansen, C., Morris, C., 1997. Hierarchical Poisson regression models. Journal of the American Statistical Association 92 (438), 618–632.

Chung, H. Loken, E., Schafer, J.L., 2004. Difficulties in drawing inferences with finite-mixture models: a simple example with a simple solution. The American Statistician 58 (2), 152-158.

Clark, S.J., Perry, J.N., 1989. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. Biometrics 45 (1), 309-316.

Conway, R.W., Maxwell, W.L., 1962. A queuing model with state dependent service rates. Journal of Industrial Engineering 12 (2), 132-136.

Dean, C.B., 1994. Modified pseudo-likelihood estimator of the overdispersion parameter in Poisson mixture models. Journal of Applied Statistics 21 (6), 523-532.

Deb, P, Trivedi, P.K., 1997. Demand for medical care by the elderly: a finite mixture approach. Journal of Applied Econometrics 12 (3), 313-336.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39 (1), 1-37.

Diebolt, J. Robert, C.P., 1994. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society, Series B 56 (2), 363-375.

El-Basyouny, K., Sayed T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models. Transportation Research Record 1950, 9-16.

Elvik, R., 2007. State-of-the-Art Approaches to Road Accident Black Spot Management and Safety Analysis of Road Networks. Report 883, Institute of Transport Economics, Oslo, Norway.

Elvik, R., 2008. Comparative analysis of techniques for identifying locations of hazardous roads. Transportation Research Record 2083, 72-75.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90 (430), 577-588.

Famoye, F., 1993. Restricted generalized Poisson regression model. Communications in Statistics – Theory and Methods 22 (5), 1335-1354.

Famoye, F., Wulu, Jr. J.T., Singh, K.P., 2004. On the generalized Poisson regression model with an application to accident data. Journal of Data Science 2 (3), 287-295.

Fitzpatrick, K., Lord, D., Park, B.-J., 2008. Accident modification factors for medians on freeways and multilane highways. Transportation Research Record 2083, 62-71.

Früwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96 (453), 194-209.

Früwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer Series in Statistics, Springer, New York.

Früwirth-Schnatter, S., Früwirth, R., Held, L., Rue, H., 2009. Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. Statistics and Computing 19 (4), 479-482.

Frühwirth-Schnatter, S., Kaufmann, S., 2006. Model-based Clustering of Multiple Time Series. Research Report IFAS. Retrieved January 2009 from http://www.ifas.jku.at.

Galfend, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85 (401), 398-409.

Geedipally, S.R., 2008. Examining the Application of Conway-Maxwell Poisson Model for Analyzing Traffic Crash Data. Ph.D. Dissertation, Department of Civil Engineering, Texas A&M University, College Station, Texas.

Geedipally, S.R., Lord, D. (2008) Effects of the varying dispersion parameter of Poisson-gamma models on the estimation of confidence intervals of crash prediction models. Transportation Research Record 2061, 46-54

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis (Second Edition). Chapman & Hall, London.

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

Geweke, J., 2007. Interpretation and inference in mixture models: simple MCMC works. Computational Statistics & Data Analysis 51 (7), 3529–3550.

Gill, J., 2002. Bayesian Methods: A Social and Behavioral Sciences Approach. Chapman & Hall/CRC, Boca Raton, Florida.

Greene, W., 2007. Limdep, Version 9.0. Econometric Software Inc., Plainview, New York.

Greene, W., 2008. Functional forms for the negative binomial model for count data. Economics Letters 99 (3), 585-590.

Gross, F., Jovanis, P.P., Eccles, K.A., 2009. Safety effectiveness of lane and shoulder width combinations on rural, two-lane, undivided roads. Transportation Research Record 2103, 42-49.

Grün, B., Leisch F., 2007. Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis 51 (11), 5247-5252.

Guikema, S.D., Coffelt, J.P., 2008. A flexible count data regression model for risk analysis. Risk Analysis 28 (1), 213 - 223.

Guo, J.Q., Trivedi, P.K., 2002. Flexible parametric models for long-tailed patent count distributions. Oxford Bulletin of Economics and Statistics 64 (1), 63-82.

Hakkert, A.S, Mahalel, D., 1978. Estimating the number of accidents at intersections from knowledge of the traffic flow on the approaches. Accident Analysis and Prevention 10 (1), 69-79.

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

Hauer, E., 1996. Identification of "Sites with Promise". Transportation Research Record 975, 54-60.

Hauer, E., 1997. Observational Before-after Studies in Road Safety. Pergamon Press, Elsevier Science Ltd., Oxford, UK.

Hauer, E., 2000. Shoulder width, shoulder paving and safety. Unpublished manuscript prepared for the Federal Highway Administration, Toronto, ON. Retrieved April 2009 from http://ca.geocities.com/hauer@rogers.com/Pubs/Shoulderwidth.pdf.

Hauer, E., 2001. Overdispersion in modeling accidents on road sections and in empirical Bayes estimation. Accident Analysis and Prevention 33 (6), 799-808.

Hauer, E., Kononov, J., Allery, B., Griffith, M.S., 2002. Screening the road network for sites with promise. Transportation Research Record 1784, 27-32.

Heckman, J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52 (2), 271-320.

Heydecker, B.G., Wu, J., 2001. Identification of sites for road accident remedial work by Bayesien statistical methods: an example of uncertain inference. Advances in Engineering Software 32 (10), 859-869.

Hilbe, J.M., 2007. Negative Binomial Regression. Cambridge University Press, Cambridge, UK.

Hinde, J., 1982. Compound Poisson regression models. In GLIM 82 (pp. 109-121). Gilchrist, R. (Ed.), New York, Springer.

Ismail, N., Jemain, A.A., 2007. Handling overdispersion with negative binomial and generalized Poisson regression models. Casualty Actuarial Society Forum, Winter, 103-158.

Jasra, A., Holmes, C.C., Stephens, D.A., 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science 20 (1), 50-67.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transportation Planning and Technology 15 (1), 41-58.

Karlis, D., Rahmouni, M., 2007. Analysis of defaulters' behaviour using the Poisson-mixture approach. IMA Journal of Management Mathematics 18 (3), 297-311.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.

Kim, H., Sun, D., and Tsutakawa, R.K., 2002. Lognormal vs. Gamma: extra variations. Biometrical Journal 44 (3), 305–323.

Koop, G., 2003. Bayesian Econometrics. Wiley, Chichester, UK.

Laird, N., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association 73 (364), 805-811.

Land K.C., McCall, P.L., Nagi, D.S., 1996. A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models. Sociological Methods and Research 24 (4), 387-442.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34 (2), 149-161.

Lewis, S.M., Raftery, A.E., 1997. Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. Journal of the American Statistical Association 92 (438), 648-655.

Li, X., Lord, D., Zhang, Y., 2009. Development of accident modification factors for rural frontage road segments in Texas using results from generalized additive models. Unpublished yet.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis and Prevention 40 (4), 1611-1618.

Lloyd-Smith, J.O., 2007. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS ONE v. 2 (2): e180 .

Long, J.S., 1997. Regression Models for Categorical and Limited Dependent Variables. SAGE Publications, Thousand Oaks, California.

Lord, D., 2000. The Prediction of Accidents on Digital Networks: Characteristics and Issues Related to the Application of Accident Prediction Models. Ph.D. Dissertation, Department of Civil Engineering, University of Toronto, Toronto, Ontario.

Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38 (4), 751-766.

Lord, D., Bonneson, J.A., 2005. Calibration of predictive models for estimating the safety of ramp design configurations. Transportation Research Record 1908, 88–95.

Lord, D., Bonneson, J.A., 2007. Development of accident modification factors for rural frontage road segments in Texas. Transportation Research Record 2023, 20-27.

Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I.,Ivan, J.N., Lyon, C., Jonsson, T., 2009. Methodology for Estimating the Safety Performance of Multilane Rural Highways. NCHRP Web-Only Document 126, National Cooperation Highway Research Program, Washington, D.C. Retrieved April 2009 from http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w126.pdf.

Lord, D., Guikema, S.D., Geedipally, S., 2008. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis and Prevention 40 (3), 1123-1134.

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Safety Science 46 (5), 751-770.

Lord, D., Park, P.Y-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accident Analysis and Prevention 40 (4), 1441-1457.

Lord, D., Washington, S.P., Ivan J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37 (1), 35-46.

Lord, D., Washington, S.P., Ivan J.N., 2007. Further notes on the application of zero inflated models in highway safety. Accident Analysis and Prevention 39 (1), 53-57.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10 (4), 325-337.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964–975.

Maher, M.J., Mountain, L.J., 1988. The identification of accident blackspots: a comparison of current methods. Accident Analysis and Prevention 20 (2), 143-151.

Maher, M.J., Summersgill, I., 1996. A comprehensive methodology for the fitting of predictive accident models. Accident Analysis and Prevention 28 (3), 281-296.

Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accident Analysis and Prevention 41 (2), 217-226.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman and Hall, London.

McGuigan, D.R.D., 1981. The use of relationships between road accidents and traffic flow in "blackspot" identification. Traffic Engineering and Control 22 (8/9), 448-451, 453.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley Series in Probability and Statistics, Wiley, New York.

McLean, J., 1996. Review of Accidents and Rural Cross Section Elements Including Roadsides. ARR 297, ARR, ARRB, Victoria, Australia.

Mengersen, K.L., Robert, C.P., 1996. Testing for mixtures: A Bayesian entropic approach (with discussion). In Bayesian Statistics 5 - Proceedings of the Fifth Valencia International Meeting (pp. 255-276). Berger, J.O., Bernardo, J.M., Dawid, A.P., Lindley, D.V., Smith, A.F.M. (Eds.). Oxford University Press, Oxford.

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

Miaou, S-P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis and Prevention 37 (4), 699-720.

Miaou, S-P., Hu, P.S., Wright, T., Rathi, A.K., Davis, S.C., 1992. Relationship between truck accidents and highway geometric design: a Poisson regression approach. Transportation Research Record 1376, 10-18.

Miaou, S-P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention 25 (6), 689-709.

Miaou, S-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transportation Research Record 1840, 31-40.

Miaou, S-P., Song, J.J., and Mallick, B.K., 2003. Roadway traffic crash mapping: a space-time modeling approach. Journal of Transportation and Statistics 6 (1), 33-57.

Miranda-Moreno, L.F., 2006. Statistical models and methods for the identification of hazardous locations for safety improvements. Ph.D. Dissertation, Department of Civil Engineering, University of Waterloo, Canada.

Miranda-Moreno, L.F., Fu, L., Saccomanno, F.F., Labbe, A., 2005. Alternative risk models for ranking locations for safety improvement. Transportation Research Record 1908, 1-8.

Miranda-Moreno, L.F., Lord, D., Fu, L., 2008. Bayesian road safety analysis: incorporation of past experiences and effect of hyper-prior choice. In: Proceedings of TRB 87th Annual Meeting Compendium of Papers DVD, Transportation Research Board, Washington, D.C.

Mitra, S., Washington S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39 (3), 459-468.

NHTSA, 2009. Traffic Safety Facts 2008. National Highway Traffic Safety Administration, National Center for Statistics and Analysis, U.S. Department of Transportation, Washington, D.C.

Oh, J., Lyon, C., Washington, S.P., Persaud, B.N., Bared, J., 2003. Validation of the FHWA crash models for rural intersections: lessons learned. Transportation Research Record 1840, 41–49.

Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway-highway interfaces. Accident Analysis and Prevention 38 (2), 346–356.

Park, B.-J., Lord, D., 2008. Adjustment for the maximum likelihood estimate of the negative binomial dispersion parameter. Transportation Research Record 2061, 9-19.

Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41 (4), 683-691.

Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1-6.

Persaud, B.N., Lord, D., Palminaso, J., 2002. Issues of calibration and transferability in developing accident prediction models for urban intersections. Transportation Research Record 1784, 57-64.

Persaud, B.N., Retting, R.A., Gårder, P.E., Lord, D., 2001. Safety effect of roundabout conversions in the United States: empirical Bayes observational before–after study. Transportation Research Record 1751, 1-8.

Piegorsch, W.W., 1990. Maximum likelihood estimation for the negative binomial dispersion parameter. Biometrics 46 (3), 863-867.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. Retrieved January 2008 from http://www.R-project.org.

Ramaswamy, V., Anderson, E.W., DeSarbo, W.S., 1994. A disaggregate negative binomial regression procedure for count data analysis. Management Science 40 (3), 405-417.

Redner, R.A., Walker, H.F., 1984. Mixture densities maximum likelihood and the EM algorithm. SIAM Review 26 (2), 195-239.

Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixture models with an unknown number of components (with discussion). Journal of the Royal Statistical Society, Series B 59 (4), 731-792.

Roberts, G.O., 1996. Markov chain concepts related to sampling algorithms. In Markov Chain Monte Carlo in Practice (pp. 45-57). Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.). Chapman & Hall, London.

Rossi, P.E., Allenby, G.M., McCulloch, R., 2005. Bayesian Statistics and Marketing. Wiley, Chichester, UK.

SAS Institute Inc., 2002. Version 9 of the SAS System for Windows. Cary, NC.

Scaccia, L., Green, P.J., 2003. Bayesian growth curves using normal mixtures with nonparametric weights. Journal of Computational and Graphical Statistics 12 (2), 308-331.

Sellers, K.F., Shmueli, G., 2008. A Flexible Regression Model for Count Data. Robert H. Smith School Research Paper No. RHS 06-060, Georgetown University, Washington, D.C. Retrieved January 2009 from http://ssrn.com/abstract=1127359.

Shankar, V.N., Albin, R.B., Milton, J.C, Mannering, F.L., 1998. Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. Transportation Research Record 1635, 44-48.

Shankar, V.N., Milton, J., Mannering, F., 1997. Modeling accident frequency as zero-altered probability process: an empirical inquiry. Accident Analysis and Prevention 29 (6), 829-837.

Shanker, V.N., Ulfasson, G.F., Pendyala, R.M., Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41(7), 627-640.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P., 2005. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. Journal of the Royal Statistical Society, Series C 54 (1), 127-142.

Simar, L., 1976. Maximum likelihood estimation of a compound Poisson process. Annals of Statistics 4 (6), 1200-1209.

Son, H., Kweon, Y-J., Park, B., 2009. Development of crash prediction models with individual vehicular data. In: Proceedings of TRB 88th Annual Meeting Compendium of Papers DVD, Transportation Research Board, Washington, D.C.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 64 (4), 583-639.

Stephens, M., 2000a. Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. The Annals of Statistics 28 (1), 40-74.

Stephens, M., 2000b. Dealing with label switching in mixture models. Journal of the Royal Statistical Society, Series B 62 (4), 795-809.

Tanner, T., Wong, W., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82 (398), 528-549.

Titterington, D.M., Smith, A.F.M. Makov, U.E., 1985. Statistical Analysis of Finite Mixture Distributions. Wiley Series in Probability and Statistics, Wiley, New York.

Toft, N., Innocent, G.T., Mellor, D.J., Reid, S.W.J., 2006. The gamma-Poisson model as a statistical method to determine if micro-organisms are randomly distributed in a food matrix. Food Microbiology 23 (1), 90-94.

Tong, J., Lord, D., 2007. Investigating the application of beta-binomial models in highway safety. Canadian Multidisciplinary Road Safety Conference XVII, June 3-8, 2007, Montreal.

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31 (2&3), 221-229.

Van Dongen, S., 2006. Prior specification in Bayesian statistics: three cautionary tales. Journal of Theoretical Biology 242 (7), 90-100.

Viallefont, V., Richardson, S., Green, P.J., 2002. Bayesian analysis of Poisson mixtures. Journal of Nonparametric Statistics 14 (1&2), 181-202.

Wang, P., Cockburn, I.M., Puterman, M.L., 1998. Analysis of patent data: a mixed Poisson regression model approach. Journal of Business and Economic Statistics 16 (1), 27-41.

Wang, W., Famoye, F., 1997. Modeling household fertility decisions with generalized Poisson regression. Journal of Population Economics 10 (3), 273-283.

Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16 (3), 275-289.

Washington, S., Oh, J., 2006. Bayesian methodology incorporating expert judgment for ranking countermeasure effectiveness under uncertainty: Example applied to at grade railroad crossings in Korea. Accident Analysis and Prevention 38 (2), 234-247.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC, Boca Raton, Florida.

Washington, S.P., Persaud, B.N., Lyon, C., Oh, J., 2005. Validation of Accident Models for Intersections. Report No. FHWA-RD-03-037. Federal Highway Administration, Washington, D.C.

Wedderburn, R., 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61 (3), 439-447.

Wedel, M., DeSarbo, W.S., Bult, J.R., Ramaswamy, V., 1993. A latent class Poisson regression model for heterogeneous count data. Journal of Applied Econometrics 8 (4), 397-411.

Winkelmann, R., 2008. Econometric Analysis of Count Data (Fifth Edition). Springer-Verlag, Berlin.

Wood, G.R., 2002. Generalized linear accident models and goodness of fit testing. Accident Analysis and Prevention 34 (1), 417-427.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural networks: an empirical analysis. Accident Analysis and Prevention 39 (5), 922-933.

Zhang, Y., Ye, Z., Lord, D., 2007. Estimating the dispersion parameter of the negative binomial distribution for analyzing crash data using a bootstrapped maximum likelihood method. Transportation Research Record 2019, 15-21.

APPENDIX A

DERIVATION OF NEGATIVE BINOMIAL DISTRIBUTION FROM A POISSON-GAMMA MIXTURE

The negative binomial distribution can be derived from two approaches.

Derivation 1:

$$y_i \sim Pois\ (\lambda_i)$$

$$\sim Pois\ (\mu_i \cdot e^{\varepsilon_i})$$

$$e^{\varepsilon_i} \sim Gamma\ (\phi, \phi)$$

$$p(y_i) = \int_0^\infty Pois(y_i \mid \mu_i \cdot e^{\varepsilon_i}) \cdot g(e^{\varepsilon_i}) d\varepsilon_i$$

$$= \int_0^\infty \frac{(\mu_i \cdot e^{\varepsilon_i})^{y_i} e^{-\mu_i \cdot e^{\varepsilon_i}}}{y_i!} \cdot \frac{(\phi)^\phi}{\Gamma(\phi)} (e^{\varepsilon_i})^{\phi-1} e^{-\phi e^{\varepsilon_i}} d\varepsilon_i$$

$$= \frac{(\phi)^\phi (\mu_i)^{y_i}}{y_i! \Gamma(\phi)} \int_0^\infty e^{\varepsilon_i (y_i + \phi - 1)} e^{-(\mu_i + \phi) e^{\varepsilon_i}} d\varepsilon_i$$

Here, the inside of the integral, $e^{\varepsilon_i (y_i + \phi - 1)} e^{-(\mu_i + \phi) e^{\varepsilon_i}}$ is a kernel of $Gamma\ (y_i + \phi, \mu_i + \phi)$ for e^{ε_i}. Thus, the following is true.

$$\frac{(\mu_i + \phi)^{y_i + \phi}}{\Gamma(y_i + \phi)} \int_0^\infty e^{\varepsilon_i (y_i + \phi - 1)} e^{-(\mu_i + \phi) e^{\varepsilon_i}} d\varepsilon_i = 1$$

$$\int_0^\infty e^{\varepsilon_i (y_i + \phi - 1)} e^{-(\mu_i + \phi) e^{\varepsilon_i}} d\varepsilon_i = \frac{\Gamma(y_i + \phi)}{(\mu_i + \phi)^{y_i + \phi}}$$

Therefore, $p(y_i) = \frac{(\phi)^\phi (\mu_i)^{y_i}}{y_i! \Gamma(\phi)} \frac{\Gamma(y_i + \phi)}{(\mu_i + \phi)^{y_i + \phi}}$

$$= \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1) \Gamma(\phi)} \left(\frac{\phi}{\phi + \mu_i} \right)^\phi \left(\frac{\mu_i}{\phi + \mu_i} \right)^{y_i}$$

Derivation 2:

$$y_i \sim Pois\ (\lambda_i)$$

$$\lambda_i \sim Gamma\ (\phi, \phi / \mu_i)$$

$$\begin{aligned} p(y_i) &= \int_0^\infty Poiss(y_i \mid \lambda_i) \cdot g(\lambda_i) d\lambda_i \\ &= \int_0^\infty \frac{\lambda_i^{y_i} e^{-\lambda_i}}{\Gamma(y_i + 1)} \cdot \frac{(\phi / \mu_i)^\phi}{\Gamma(\phi)} \lambda_i^{\phi - 1} e^{-(\phi / \mu_i)\lambda_i} d\lambda_i \\ &= \frac{(\phi / \mu_i)^\phi}{\Gamma(y_i + 1)\Gamma(\phi)} \int_0^\infty \lambda_i^{y_i + \phi - 1} e^{-(1 + \phi / \mu_i)\lambda_i} d\lambda_i \end{aligned}$$

Here, the inside of the integral, $\lambda_i^{y_i + \phi - 1} e^{-(1 + \phi / \mu_i)\lambda_i}$ is a kernel of $Gamma(y_i + \phi, 1 + \phi / \mu_i)$ for λ_i. Thus, the following is true:

$$\frac{(1 + \phi / \mu_i)^{y_i + \phi}}{\Gamma(y_i + \phi)} \int_0^\infty \lambda_i^{y_i + \phi - 1} e^{-(1 + \phi / \mu_i)\lambda_i} d\lambda_i = 1$$

$$\int_0^\infty \lambda_i^{y_i + \phi - 1} e^{-(1 + \phi / \mu_i)\lambda_i} d\lambda_i = \frac{\Gamma(y_i + \phi)}{(1 + \phi / \mu_i)^{y_i + \phi}}$$

Therefore, $p(y_i) = \dfrac{(\phi / \mu_i)^\phi}{\Gamma(y_i + 1)\Gamma(\phi)} \dfrac{\Gamma(y_i + \phi)}{(1 + \phi / \mu_i)^{y_i + \phi}}$

$$= \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \mu_i} \right)^\phi \left(\frac{\mu_i}{\phi + \mu_i} \right)^{y_i}$$

The marginal mean and variance of y_i are obtained from the following relationships.

$$E(y_i) = E\{E(y_i \mid \lambda_i)\} = E(\lambda_i) = \mu_i$$

$$Var(y_i) = E\{Var(y_i \mid \lambda_i)\} + Var\{E(y_i \mid \lambda_i)\} = E\{\lambda_i\} + Var\{\lambda_i\} = \mu_i + \mu_i^2 / \phi$$

Alternatively, the negative binomial distribution can also be obtained by assuming λ_i has a gamma distribution with shape $\phi\mu_i$ and scale ϕ (McCullagh and Nelder, 1989). Analogous to the derivation above, it can be shown that $p(y_i)$ is expressed as:

$$p(y_i) = \frac{\Gamma(y_i + \phi\mu_i)}{\Gamma(y_i + 1)\Gamma(\phi\mu_i)} \left(\frac{\phi}{\phi + 1} \right)^{\phi\mu_i} \left(\frac{1}{\phi + 1} \right)^{y_i}$$

In this case, using iterated expectation and variance, it can be shown that the marginal mean and variance of y_i are:

$$E(y_i) = \mu_i$$

$$Var(y_i) = \mu_i \left[1 + \frac{1}{\phi} \right]$$

This is the NB1 model termed by Cameron and Trivedi (1998) and corresponds to Equation (2.13) in Chapter II.

APPENDIX B

METROPOLIS-HASTINGS ALGORITHM

A Metropolis-Hastings (MH) is a generic MCMC algorithm that generates samples from a probability distribution, using a full joint density function, $\pi(\boldsymbol{\theta})$ of unknown parameters $\boldsymbol{\theta}$. The algorithm relies on a so-called *proposal distribution* (q) and consists of the following steps:

1. Set a starting value, $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \cdots, \theta_p^{(0)})$. Set $t = 1$.
2. Generate a candidate value from a proposal distribution:
   $\boldsymbol{\theta}^{(c)} = (\theta_1^{(c)}, \cdots, \theta_p^{(c)}) \sim q(\cdot \mid \boldsymbol{\theta}^{(t-1)})$
3. Compute the ratio $R = \dfrac{\pi(\boldsymbol{\theta}^{(c)})q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^{(c)})}{\pi(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^{(c)} \mid \boldsymbol{\theta}^{(t-1)})}$.
4. Set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(c)}$ with probability $\alpha(\boldsymbol{\theta}^{(c)} \mid \boldsymbol{\theta}^{(t-1)}) = \min(1, R)$

   Otherwise, $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$
5. Set $t = t + 1$ and return to step 2 until enough samples are obtained.

In Step 4, if R is greater than 1, the candidate will be accepted with probability of 1. If $0 < R < 1$, the comparison is done between R and some random probability draw (e.g. $u \sim U(0,1)$). If $R > u$, the candidate is accepted, otherwise it is rejected. Especially in a Random-Walk Metropolis algorithm, a symmetric proposal distribution (usually, a normal distribution) is considered i.e., one satisfying $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t-1)}) = q(\mid \boldsymbol{\theta} - \boldsymbol{\theta}^{(t-1)} \mid)$. With such a proposal, $q(\boldsymbol{\theta}^{(c)} \mid \boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^{(c)})$ and thus the acceptance probability in Step 4 becomes $\alpha(\boldsymbol{\theta}^{(c)} \mid \boldsymbol{\theta}^{(t-1)}) = \min\left(1, \dfrac{\pi(\boldsymbol{\theta}^{(c)})}{\pi(\boldsymbol{\theta}^{(t-1)})}\right)$. It turns out that the series of $\boldsymbol{\theta}$ generated in this manner forms a Markov chain whose stationary distribution is the desired posterior distribution $\pi(\boldsymbol{\theta})$.

APPENDIX C

SIMULATION RESULTS FOR EXAMPLE 1 (TABLE 4.5)

For small-mean value and phi=0.5

N=50 (mean=0.76, variance=2.3)

	True Value	NB	FMP-2	
			Comp 1	Comp 2
β_0	-0.5	-1.158	-11.718*	-0.217*
β_1	0.5	1.736	1.219	
β_2	-0.5	-0.133*	-0.149*	
ϕ	0.5	0.585	-	
w	1	1		
-2LL	-	97.1		
AIC	-	105.1	106.7	
BIC	-	112.8	116.3	

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=0.78, variance=3.1)

	True Value	NB	FMP-2	
			Comp 1	Comp 2
β_0	-0.5	-0.533	-11.623*	0.142*
β_1	0.5	0.645	0.561	
β_2	-0.5	-0.969	-0.911	
ϕ	0.5	0.649	-	
w	1	1	0.468	0.532
-2LL	-	113.6	112.5	
AIC	-	121.6	122.5	
BIC	-	129.3	132.0	

N=500 (mean=0.76, variance=2.6)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	-0.529	0.795	-1.475	1.071	-0.456	-3.018				
β_1	0.5	0.471	0.513		0.459						
β_2	-0.5	-0.452	-0.429		-0.484						
ϕ	0.5	0.497	-		-						
w	1	1	0.184	0.816	0.101	0.430	0.468				
-2LL	-	1096.3	1108.4		1088.5						
AIC	-	1104.3	1118.4		1102.5						
BIC	-	1121.1	1139.5		1132.0						

N=1000 (mean=0.78, variance=3.9)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	-0.512	0.910	-1.474	2.191	0.779	-1.514				
β_1	0.5	0.433	0.464		0.391						
β_2	-0.5	-0.568	-0.504		-0.482						
ϕ	0.5	0.473	-		-						
w	1	1	0.170	0.830	0.065	0.255	0.680				
-2LL	-	2185.8	2242.3		2211.7						
AIC	-	2193.8	2252.3		2225.7						
BIC	-	2213.5	2276.9		2260.1						

For small-mean value and phi=2

N=50 (mean=0.76, variance=1.1)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	-0.524	-0.7095	0.974*							
β_1	0.5	0.729	0.874								
β_2	-0.5	-0.313*	-0.287*								
ϕ	2	8.9*	-								
w	1	1	0.874	0.126							
-2LL	-	108.2	107.9								
AIC	-	116.2	117.9								
BIC	-	123.9	127.5								

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=0.48, variance=0.6)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	0.177	0.262*	-1.184							
β_1	0.5	0.172*	0.081*								
β_2	-0.5	0.188	-0.545								
ϕ	2	2.80	-								
w	1	1	0.123	0.877							
-2LL	-	177.4	176.7								
AIC	-	185.4	186.7								
BIC	-	195.8	199.8								

N=500 (mean=0.74, variance=1.9)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	-0.604	0.201	-1.080							
β_1	0.5	0.464	0.450								
β_2	-0.5	-0.488	-0.480								
ϕ	2	2.121	-								
w	1	1	0.256	0.744							
-2LL	-	1069.6	1070.1								
AIC	-	1077.6	1080.1								
BIC	-	1094.4	1101.2								

N=1000 (mean=0.78, variance= 1.6)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	0.529	0.313	-1.013							
β_1	0.5	0.527	0.529								
β_2	-0.5	-0.513	-0.532								
ϕ	2	2.058	-								
w	1	1	0.231	0.769							
-2LL	-	2189.6	2186.2								
AIC	-	2197.6	2196.2								
BIC	-	2217.2	2220.8								

For small-mean value and phi=5

N=50 (mean=0.80, variance=1.2)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	-0.367*	-0.503	0.965*							
β_1	0.5	0.230*	0.230*								
β_2	-0.5	-0.619	-0.676								
ϕ	5	6.9*	-								
w	1	1	0.838	0.162							
-2LL	-	112.0	111.8								
AIC	-	120.0	121.8								
BIC	-	127.6	131.3								

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=0.82, variance=1.5)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	-0.560	-2.546	-0.298*							
β_1	0.5	0.715	0.749								
β_2	-0.5	-0.409	-0.525								
ϕ	5	2.72	-								
w	1	1	0.287	0.713							
-2LL	-	222.7	220.5								
AIC	-	230.7	230.5								
BIC	-	241.1	243.5								

N=500 (mean=0.79, variance=1.5)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	-0.586	-0.268*	-1.347							
β_1	0.5	0.484	0.494								
β_2	-0.5	-0.577	-0.566								
ϕ	5	4.86	-								
w	1	1	0.527	0.473							
-2LL	-	1068.0	1066.9								
AIC	-	1076.0	1076.9								
BIC	-	1092.9	1097.9								

N=1000 (mean=0.77, variance=1.3)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	-0.5	-0.512	-0.193*	-1.203							
β_1	0.5	0.482	0.486								
β_2	-0.5	-0.450	-0.503								
ϕ	5	5.10	-								
w	1	1	0.512	0.488							
-2LL	-	2152.8	2151.6								
AIC	-	2160.8	2161.6								
BIC	-	2180.5	2186.2								

For moderate mean value and phi=0.5

N=50 (mean=2.86, variance=31.3)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.656	1.964	-0.121*	2.024	0.487	-3.721*				
β_1	0.5	0.694	0.747		0.819						
β_2	-0.5	-0.672	-0.864		-0.622						
ϕ	0.5	0.475	-		-						
w	1	1	0.167	0.833	0.147	0.471	0.382				
-2LL	-	187.6	203.5		180.5						
AIC	-	195.7	213.5		194.5						
BIC	-	203.3	223.0		207.8						

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=3.64, variance=102.0)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.741	1.843	-0.397	1.915	0.305*	-1.870				
β_1	0.5	0.630	0.863		0.839						
β_2	-0.5	-0.689	-0.823		-0.791						
ϕ	0.5	0.540	-		-						
w	1	1	0.228	0.772	0.199	0.408	0.393				
-2LL	-	406.5	432.6		398.1						
AIC	-	414.5	442.6		412.1						
BIC	-	424.9	455.6		430.3						

N=500 (mean=3.32, variance=33.9)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	1.005	2.195	-0.132*	2.343	0.938	-1.478	2.699	1.920	0.592	-2.089
β_1	0.5	0.385	0.307		0.344			0.312			
β_2	-0.5	-0.456	-0.387		-0.459			-0.373			
ϕ	0.5	0.465	-		-			-			
w	1	1	0.235	0.765	0.165	0.357	0.478	0.061	0.169	0.385	0.384
-2LL	-	2138.0	2345.5		2184.8			2129.8			
AIC	-	2146.0	2355.5		2198.8			2147.9			
BIC	-	2162.8	2376.5		2228.3			2185.9			

N=1000 (mean=3.56, variance=43.2)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.974	2.226	0.014*	2.455	1.215	-0.872	2.558	1.658	0.600	-1.781
β_1	0.5	0.535	0.344		0.536			0.537			
β_2	-0.5	-0.553	-0.351		-0.471			-0.451			
ϕ	0.5	0.506	-		-			-			
w	1	-	0.223	0.777	0.116	0.331	0.552	0.087	0.161	0.359	0.393
-2LL	-	4250.2	4813.3		4362.4			4268.8			
AIC	-	4258.2	4823.3		4376.5			4286.8			
BIC	-	4277.8	4847.8		4410.8			4330.9			

For moderate mean value and phi=2

N=50 (mean=2.54, variance=7.5)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.604	0.670	-0.926*							
β_1	0.5	0.664	0.767								
β_2	-0.5	-0.679	-0.743								
ϕ	2	7.32	-								
w	1	1	0.876	0.124							
-2LL	-	171.7	168.4								
AIC	-	179.7	178.4								
BIC	-	187.3	188.0								

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=2.40, variance=4.4)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.718	1.489	0.267							
β_1	0.5	0.252	0.267								
β_2	-0.5	-0.467	-0.470								
ϕ	2	6.06	-								
w	1	1	0.133	0.867							
-2LL	-	369.4	368.4								
AIC	-	377.4	378.4								
BIC	-	387.8	391.4								

N=500 (mean=3.46, variance=21.6)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.903	1.696	0.433	1.803	0.778	-0.434				
β_1	0.5	0.554	0.538		0.521						
β_2	-0.5	-0.515	-0.479		-0.462						
ϕ	2	1.992	-		-						
w	1	1	0.247	0.4332	0.175	0.582	0.244				
-2LL	-	2109.5	2153.9		2121.4						
AIC	-	2117.5	2163.9		2135.4						
BIC	-	2134.3	2185.0		2164.9						

N=1000 (mean=3.39, variance=18.0)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.981	1.610	0.356	1.942	1.133	-0.048*	1.976	1.256	0.396	-1.618
β_1	0.5	0.4842	0.468		0.502			0.485			
β_2	-0.5	-0.475	-0.459		-0.461			-0.468			
ϕ	2	1.904	-		-			-			
w	1	1	0.353	0.647	0.121	0.476	0.403	0.100	0.367	0.429	0.104
-2LL	-	4291.7	4383.9		4308.9			4292.3			
AIC	-	4299.7	4393.9		4322.9			4310.3			
BIC	-	4319.4	4418.5		4357.3			4354.5			

For moderate mean value and phi=5

N=50 (mean=3.5, variance=12.8)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.918	1.266	0.675							
β_1	0.5	0.760	0.753								
β_2	-0.5	-0.554	-0.550								
ϕ	5	14.2	-								
w	1	1	0.365	0.635							
-2LL	-	189.5	188.9								
AIC	-	197.5	198.9								
BIC	-	205.2	208.5								

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=3.49, variance=10.8)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.951	1.414	0.708							
β_1	0.5	0.561	0.555								
β_2	-0.5	-0.480	-0.464								
ϕ	5	7.75	-								
w	1	1	0.288	0.712							
-2LL	-	401.9	401.7								
AIC	-	409.9	411.7								
BIC	-	420.3	424.8								

N=500 (mean=3.66, variance=17.8)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	0.994	1.393	0.574							
β_1	0.5	0.512	0.516								
β_2	-0.5	-0.467	-0.435								
ϕ	5	5.260	-								
w	1	1	0.438	0.562							
-2LL	-	2065.8	2073.7								
AIC	-	2073.8	2083.7								
BIC	-	2090.6	2104.7								

N=1000 (mean=3.53, variance=17.3)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1	1.001	1.438	0.589							
β_1	0.5	0.497	0.501								
β_2	-0.5	-0.487	-0.495								
ϕ	5	4.742	-								
w	1	1	0.386	0.614							
-2LL	-	4119.8	4130.1								
AIC	-	4127.8	4140.1								
BIC	-	4147.5	4164.6								

For high mean value and phi=0.5

N=50 (mean=5.5, variance=113.3)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.336	2.892	0.563	2.699	1.205	-2.553				
β_1	0.5	0.717	0.078*		0.873						
β_2	-0.5	-0.563	-0.829		-0.517						
ϕ	0.5	0.438	-		-						
w	1	1	0.184	0.816	0.143	0.468	0.389				
-2LL	-	239.5	292.9		230.9						
AIC	-	247.5	302.9		244.9						
BIC	-	255.1	312.5		258.3						

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=6.1, variance=127.3)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.463	2.391	-0.045*	2.962	1.703	-0.769	3.238	2.107	0.862	-1.530
β_1	0.5	0.676	0.693		0.788			0.479			
β_2	-0.5	-0.464	-0.173		-0.086*			-0.244			
ϕ	0.5	0.440	-		-			-			
w	1	1	0.347	0.653	0.126	0.332	0.542	0.061	0.303	0.216	0.420
-2LL	-	512.8	656.8		527.7			515.9			
AIC	-	520.8	666.8		541.7			533.9			
BIC	-	531.2	679.9		559.9			557.3			

N=500 (mean=7.5, variance=170.5)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.669	2.706	0.519	3.058	1.917	-0.285	3.118	2.236	1.310	-0.799
β_1	0.5	0.535	0.344		0.456			0.483			
β_2	-0.5	-0.544	-0.405		-0.427			-0.398			
ϕ	0.5	0.487	-		-			-			
w	1	1	0.294	0.706	0.134	0.321	0.545	0.113	0.191	0.249	0.446
-2LL	-	2723.2	3480.2		2871.5			2779.7			
AIC	-	2731.2	3490.2		2885.5			2797.7			
BIC	-	2748.1	3511.2		2915.0			2835.6			

NOTE: FMP-5: AIC=2769.9, BIC=2816.3

N=1000 (mean=7.5, variance=238.7)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.703	2.730	0.420	3.076	1.867	-0.296	3.273	2.469	1.323	-1.007
β_1	0.5	0.487	0.420		0.472			0.404			
β_2	-0.5	-0.586	-0.587		-0.591			-0.604			
ϕ	0.5	0.489	-		-			-			
w	1	1	0.274	0.726	0.129	0.345	0.526	0.078	0.172	0.345	0.405
-2LL	-	5444.5	7151.4		6021.7			5610.4			
AIC	-	5452.5	7161.4		6035.7			5628.4			
BIC	-	5472.1	7185.9		6070.1			5672.5			

For High mean value and phi=2

N=50 (mean= 5.7, variance=36.9)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.416	1.814	1.086							
β_1	0.5	0.565	0.507								
β_2	-0.5	-0.765	-0.727								
ϕ	2	8.11	-								
w	1	1	0.395	0.605							
-2LL	-	221.5	221.2								
AIC	-	229.5	231.2								
BIC	-	237.1	240.8								

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=6.3, variance=73.8)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.399	2.269	1.016	2.302	1.330	0.102				
β_1	0.5	0.654	0.545		0.589						
β_2	-0.5	-0.674	-0.542		-0.521						
ϕ	2	2.069	-		-						
w	1	1	0.216	0.784	0.175	0.576	0.249				
-2LL	-	513.1	543.7		522.9						
AIC	-	521.1	553.7		536.9						
BIC	-	531.6	566.8		555.2						

N=500 (mean=7.4, variance=97.5)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.654	2.250	1.129	2.570	1.732	0.605	2.624	2.002	1.492	0.448
β_1	0.5	0.501	0.496		0.514			0.499			
β_2	-0.5	-0.542	-0.627		-0.492			-0.510			
ϕ	2	2.149	-		-			-			
w	1	1	0.324	0.676	0.132	0.355	0.513	0.100	0.248	0.374	0.278
-2LL	-	2730.4	2915.7		2758.5			2744.4			
AIC	-	2738.4	2925.7		2772.5			2762.4			
BIC	-	2755.3	2946.7		2802.0			2800.3			

NOTE: FMP-5: AIC=2769.9, BIC=2816.3

N=1000 (mean=7.2, variance=78.5)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.689	2.325	1.079	2.452	1.598	0.573	2.771	2.199	1.379	0.386
β_1	0.5	0.505	0.476		0.455			0.490			
β_2	-0.5	-0.519	-0.517		-0.491			-0.529			
ϕ	2	1.852	-		-			-			
w	1	1	0.341	0.659	0.244	0.406	0.351	0.061	0.272	0.405	0.262
-2LL	-	5505.6	5745.9		5581.9			5507.0			
AIC	-	5513.6	5755.9		5595.9			5525.0			
BIC	-	5533.2	5780.4		5630.3			5569.2			

NOTE: FMP-5: AIC=5515.0, BIC=5569.0

For High mean value and phi=5

N=50 (mean=6.9, variance= 33.9)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.688	1.896	1.317							
β_1	0.5	0.594	0.560								
β_2	-0.5	-0.561	-0.547								
ϕ	5	15.64	-								
w	1	1	0.575	0.425							
-2LL	-	234.6	232.9								
AIC	-	242.6	242.9								
BIC	-	250.3	252.5								

NOTE: * indicates the non-significance at 0.05 level.

N=100 (mean=6.5, variance=26.4)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.643	2.162	1.465							
β_1	0.5	0.355	0.388								
β_2	-0.5	-0.543	-0.545								
ϕ	5	10.32	-								
w	1	1	0.192	0.808							
-2LL	-	489.4	486.7								
AIC	-	497.4	496.7								
BIC	-	507.8	509.8								

N=500 (mean=7.4, variance=66.0)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.697	2.093	1.269	2.499	1.937	1.161				
β_1	0.5	0.519	0.492		0.530						
β_2	-0.5	-0.487	-0.490		-0.496						
ϕ	5	5.220	-		-						
w	1	1	0.425	0.575	0.050	0.482	0.468				
-2LL	-	2596.1	2610.6		2591.4						
AIC	-	2604.1	2620.6		2605.4						
BIC	-	2621.0	2641.7		2634.9						

NOTE: FMP-5: AIC=2769.9, BIC=2816.3

N=1000 (mean=7.3, variance=63.2)

	True Value	NB	FMP-2		FMP-3			FMP-4			
			Comp 1	Comp 2	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 4
β_0	1.7	1.702	2.087	1.271	2.274	1.702	0.903	2.484	2.066	1.573	0.789
β_1	0.5	0.522	0.503		0.529			0.528			
β_2	-0.5	-0.522	0.494		-0.517			-0.532			
ϕ	5	4.740	-		-			-			
w	1	1	0.439	0.561	0.194	0.537	0.268	0.055	0.279	0.467	0.199
-2LL	-	5189.6	5276.7		5196.2			5187.5			
AIC	-	5197.6	5286.7		5210.2			5205.5			
BIC	-	5217.2	5311.2		5244.6			5249.6			

APPENDIX D

EXAMPLES OF R CODES FOR FMP-2 AND FMNB-2 MODELS

This appendix provides example codes for generating the FMP-2 and FMNB-2 random variates, and for implementing the MCMC sampling method for each model. The MCMC algorithm for a single negative binomial regression is available from an R function (*rnegbinRw*) within the R package called a "Bayesm" (Rossi, 2008). The *rnegbinRw* function implements a Random-Walk Metropolis algorithm for the NB regression model. Therefore, the codes were modified to implement the MCMC sampling procedure for the finite mixture regression model used in Chapter III (Data augmentation and Gibbs sampling).

1. Generating FMP-2 random variates and MCMC sampling for FMP-2 model

```
library(bayesm)

## Generating FMP-2 random variates ##
set.seed(1)
N=500
b=matrix(c(2,-0.5,0.5,0,0.5,-0.5), nrow=3, ncol=2)
X=cbind(rep(1,N),rnorm(N,mean=0,sd=1),rnorm(N,mean=0,sd=1))
offset=c(rep(0,N))
xbeta=X%*%b
m=exp(xbeta)

set.seed(1)
w=rbinom(N, 1, .2)
y=w*rpois(N,m[,1])+(1-w)*rpois(N,m[,2])
table(y); mean(y); var(y)
plot(table(y),col='red',ylab='Frequency',cex.lab=1.2)

## MCMC sampling for FMP-2 model ##

nobs=N
nvar=ncol(X)
ncomp=2

# Data info #
Data=list(y=y,X=X,offset=offset,nobs=nobs,nvar=nvar,ncomp=ncomp)

# Prior info #
betabar=rep(0,nvar)
```

```
A=0.01*diag(nvar)
alpha=c(rep(1,ncomp))
weight=c(rep(1/ncomp,ncomp))
z=rmultinom(nobs,size=1,weight)
Prior=list(betabar=betabar,A=A,alpha=alpha,weight=weight,z=z)

# MCMC implementation #
R=5000;keep=1;burnin=(R/keep)/2+1;B=burnin;E=R/keep
s_beta1=2.5; s_beta2=3.3  # need to be adjusted to induce a good mix
Mcmc=list(R=R,keep=keep,s_beta1=s_beta1,s_beta2=s_beta2)
source("C:/FMP_K2_unconst.txt")
out.fmp=mcmcpoismix(Data,Prior,Mcmc)

# Acceptance rate check #
out.fmp$acceptrbeta1; out fmp$acceptrbeta2
```

The “FMP_K2_unconst.txt” file above contains the following function (mcmcpoismix) for implementing an MCMC sampling for the FMP-2 model. The following is an example in which no identifiability constraint is placed on the parameters.

```
mcmcpoismix=function(Data,Prior,Mcmc){
      llpois=function(par,X,y,nvar){
            beta = par[1:nvar]
            mean = exp(X %*% beta+offset)
            out = .Internal(dpois(y,mean,TRUE))
            return(sum(out))
      }
      llpoismix=function(beta1,beta2,X,y,weight){
            mean1=exp(X%*%beta1+offset)
            out1=dpois(y,mean1)
            mean2=exp(X%*%beta2+offset)
            out2=dpois(y,mean2)
            sum1=weight[1]*out1+weight[2]*out2
            sum2=sum(log(sum1))
            return(sum2)
      }
      lpostbeta1=function(beta,X,y,z,betabar,A){
            out=log(dpois(y,exp(X%*%beta+offset)))
            residual=as.vector(beta-betabar)
            sumlpostbeta1=z[1,]%*%out-0.5*(t(residual)%*%A%*%residual)
            return(sumlpostbeta1)
      }
      lpostbeta2=function(beta,X,y,z,betabar,A){
            out=log(dpois(y,exp(X%*%beta+offset)))
            residual=as.vector(beta-betabar)
            sumlpostbeta2=z[2,]%*%out-0.5*(t(residual)%*%A%*%residual)
            return(sumlpostbeta2)
      }
      postz=function(beta,ncomp,X,y,weight){
            avg=exp(X%*%beta+offset)
            z=matrix(0,nrow=ncomp,ncol=nobs)
            for(i in 1:nobs){
```

```
            num=weight*dpois(y[i],avg[i,])
            num=ifelse(num<1e-100,1e-100,num)
            z[,i]=rmultinom(1,size=1,prob=num)
        }
        return(z)
    }
    postweight=function(z,ncomp,alpha){
        nalpha=NULL
        for (i in 1:ncomp){
            nalpha[i]=sum(z[i,])+alpha[i]
        }
        weight=rdirichlet(nalpha)
        return(weight)
    }

    X=Data$X;y=Data$y;nobs=Data$nobs;nvar=Data$nvar;ncomp=Data$ncomp
    offset=Data$offset;betabar=Prior$betabar;A=Prior$A
    alpha=Prior$alpha;weight=Prior$weight;z=Prior$z
    R=Mcmc$R;keep=Mcmc$keep;s_beta1=Mcmc$s_beta1;s_beta2=Mcmc$s_beta2

    cat(" ", fill = TRUE)
   cat("Starting Random Walk Metropolis Sampler for Poisson Regression",
fill = TRUE)
    fsh()

    par=rep(0,nvar)
    mle = optim(par,llpois,X=X,y=y,nvar=nvar,method="L-BFGS-B",
       upper=c(rep(Inf,nvar)),hessian=TRUE,control=list(fnscale=-1))
    fsh()

    beta_mle=mle$par[1:nvar]
    varcovinv = -mle$hessian

    betacvar1 = s_beta1 * solve(varcovinv[1:nvar, 1:nvar])
    betaroot1 = t(chol(betacvar1))
    betacvar2 = s_beta2 * solve(varcovinv[1:nvar, 1:nvar])
    betaroot2 = t(chol(betacvar2))
    cat("beta_mle = ", beta_mle, fill = TRUE)
    fsh()

    beta=matrix(c(beta_mle,beta_mle),nrow=nvar,ncol=ncomp)
    beta1=beta[,1]; beta2=beta[,2]

    nacceptbeta1=0; acceptrbeta1=0
    nacceptbeta2=0; acceptrbeta2=0

    betadraw=matrix(double(floor(R/keep)*(nvar*ncomp)),ncol=nvar*ncomp)
    weightdraw = matrix(double(floor(R/keep)*(ncomp)),ncol=ncomp)
    llike=rep(0,floor(R/keep))
    itime = proc.time()[3]
    cat(" ", fill = TRUE)
    cat("MCMC Iteration (est time to end - min) ", fill = TRUE)
    fsh()

    for (r in 1:R) {
        tempweight=postweight(z,ncomp,alpha)
```

```
            newbeta1=beta1+betaroot1%*%rnorm(nvar)
            oldlpostbeta1=lpostbeta1(beta1,X,y,z,betabar,A)
            newlpostbeta1=lpostbeta1(newbeta1,X,y,z,betabar,A)
            ldiff=newlpostbeta1-oldlpostbeta1
            acc=min(1,exp(ldiff))

            if (acc<1){unif=runif(1)} else {unif=0}
            if (unif<=acc) {
                  beta1=newbeta1
                  nacceptbeta1=nacceptbeta1+1
            }

            newbeta2=beta2+betaroot2%*%rnorm(nvar)
            oldlpostbeta2=lpostbeta2(beta2,X,y,z,betabar,A)
            newlpostbeta2=lpostbeta2(newbeta2,X,y,z,betabar,A)
            ldiff=newlpostbeta2-oldlpostbeta2
            acc=min(1,exp(ldiff))

            if (acc<1){unif=runif(1)} else {unif=0}
            if (unif<=acc) {
                  beta2=newbeta2
                  nacceptbeta2=nacceptbeta2+1
            }

            beta=matrix(c(beta1,beta2),nrow=nvar,ncol=ncomp) # Update beta
            z=postz(beta,ncomp,X,y,weight)  # Update z

            if (r%%100 == 0) {
                  ctime = proc.time()[3]
                  timetoend = ((ctime - itime)/r) * (R - r)
                  cat(" ",r," (",round(timetoend/60, 1),")",fill=TRUE)
                  fsh()
            }
            if(r%%keep==0){
            mkeep=r/keep
            betadraw[mkeep,]=beta
            weightdraw[mkeep,]=weight
            llike[mkeep]=llpoismix(beta1,beta2,X,y,weight)
            }
      }
      ctime = proc.time()[3]
      cat("  Total Time Elapsed: ", round((ctime - itime)/60, 2), "\n")

      return(list(llike=llike,betadraw=betadraw,weightdraw=weightdraw,
      acceptrbeta1=nacceptbeta1/R*100,acceptrbeta2=nacceptbeta2/R*100))
}
```

2. Generating FMNB-2 random variates and MCMC sampling for FMNB-2 model

```
library(bayesm)

## Generating FMNB-2 random variates ##
set.seed(1)
N=500
b=matrix(c(2,-0.5,0.5,0,0.5,-0.5), nrow=3, ncol=2)
phi_true=c(5,10)
X=cbind(rep(1,N),rnorm(N,mean=0,sd=1),rnorm(N,mean=0,sd=1))
offset=c(rep(0,N))
xbeta=X%*%b
m=exp(xbeta)

set.seed(11)
w=rbinom(N, 1, 0.2)
y=w*rnbinom(N,mu=m[,1],size=phi_true[1])+(1-
w)*rnbinom(N,mu=m[,2],size=phi_true[2])
table(y); mean(y); var(y)
plot(table(y),col='red',ylab='Frequency',cex.lab=1.2)

## MCMC sampling for FMNB-2 model ##
nobs=N; nvar=ncol(X); ncomp=2

# Data info #
Data=list(y=y,X=X,offset=offset,nobs=nobs,nvar=nvar,ncomp=ncomp)

# Prior info #
betabar=rep(0,nvar)
A=0.01*diag(nvar)
a=0.01;b=0.01
alpha=c(rep(1,ncomp))
weight=c(rep(1/ncomp,ncomp))
z=rmultinom(nobs,size=1,weight)
Prior=list(betabar=betabar,A=A,a=a,b=b,alpha=alpha,weight=weight,z=z)

# MCMC implementation #
R=5000;keep=1;burnin=(R/keep)/2+1;B=burnin;E=R/keep
s_beta1=1.2;s_beta2=1.8;s_phi1=300;s_phi2=60  # need to be adjusted
Mcmc=list(R=R,keep=keep,s_beta1=s_beta1,s_beta2=s_beta2,s_phi1=s_phi1,s_phi
2=s_phi2)
source("C:/FMNB_K2_const_w.txt")
out.fmnb=mcmcnbmix(Data,Prior,Mcmc)

# Acceptance rate check #
out.fmnb$acceptrbeta1;out.fmnb$acceptrbeta2
out.fmnb$acceptrphi1;out.fmnb$acceptrphi2
```

The “FMNB_K2_const_w.txt” file above contains the following function (mcmcnbmix) for implementing an MCMC sampling for the FMNB-2 model. The following is an example in which the identifiability constraint is placed on the weight parameter (w).

```
mcmcnbmix=function(Data,Prior,Mcmc){
      llnegbin=function(par,X,y,nvar){
            beta = par[1:nvar]
            phi = exp(par[nvar + 1]) + 1e-50
            mean = exp(X %*% beta+offset)
            prob = phi/(phi + mean)
            prob = ifelse(prob < 1e-100, 1e-100, prob)
            out = .Internal(dnbinom(y, phi, prob, TRUE))
            return(sum(out))
      }
      llnbmix=function(beta1,beta2,phi,X,y,weight){
            lambda1=exp(X%*%beta1+offset)
            p1=phi[1]/(phi[1]+lambda1)
            p1=ifelse(p1<1e-100,1e-100,p1)
            out1=dnbinom(y,phi[1],p1)
            lambda2=exp(X%*%beta2+offset)
            p2=phi[2]/(phi[2]+lambda2)
            p2=ifelse(p2<1e-100,1e-100,p2)
            out2=dnbinom(y,phi[2],p2)
            sum1=weight[1]*out1+weight[2]*out2
            sum2=sum(log(sum1))
            return(sum2)
      }
      lpostbeta1=function(beta,phi,X,y,z,betabar,A){
            mean=exp(X %*% beta+offset)
            prob=phi/(phi + mean)
            prob=ifelse(prob < 1e-100, 1e-100, prob)
            out=.Internal(dnbinom(y, phi, prob, TRUE))
            residual=as.vector(beta-betabar)
            sumlpostbeta1=z[1,]%*%out-0.5*(t(residual)%*%A%*%residual)
            return(sumlpostbeta1)
      }
      lpostbeta2=function(beta,phi,X,y,z,betabar,A){
            mean = exp(X %*% beta+offset)
            prob = phi/(phi + mean)
            prob = ifelse(prob < 1e-100, 1e-100, prob)
            out = .Internal(dnbinom(y, phi, prob, TRUE))
            residual=as.vector(beta-betabar)
            sumlpostbeta2=z[2,]%*%out-0.5*(t(residual)%*%A%*%residual)
            return(sumlpostbeta2)
      }
      lpostphi1=function(beta,phi,X,y,z,a,b){
            mean = exp(X %*% beta+offset)
            prob = phi/(phi + mean)
            prob = ifelse(prob < 1e-100, 1e-100, prob)
            out = .Internal(dnbinom(y, phi, prob, TRUE))
            sumlpostphi1=z[1,]%*%out+(a-1)*log(phi)-b*phi
            return(sumlpostphi1)
      }
      lpostphi2=function(beta,phi,X,y,z,a,b){
            mean = exp(X %*% beta+offset)
            prob = phi/(phi + mean)
            prob = ifelse(prob < 1e-100, 1e-100, prob)
            out = .Internal(dnbinom(y, phi, prob, TRUE))
            sumlpostphi2=z[2,]%*%out+(a-1)*log(phi)-b*phi
            return(sumlpostphi2)
```

```
    }
    postz=function(beta,phi,ncomp,X,y,weight){
        avg=exp(X%*%beta+offset)
        z=matrix(0,nrow=ncomp,ncol=nobs)
        for(i in 1:nobs){
            num=weight*dnbinom(y[i],mu=avg[i,],size=phi)
            num=ifelse(num<1e-100,1e-100,num)
            z[,i]=rmultinom(1,size=1,prob=num)
        }
        return(z)
    }
    postweight=function(z,ncomp,alpha){
        nalpha=NULL
        for (i in 1:ncomp){
            nalpha[i]=sum(z[i,])+alpha[i]
        }
        weight=rdirichlet(nalpha)
        return(weight)
    }

    X=Data$X;y=Data$y;nobs=Data$nobs;nvar=Data$nvar;ncomp=Data$ncomp
    offset=Data$offset;betabar=Prior$betabar;A=Prior$A;a=Prior$a
    b=Prior$b;alpha=Prior$alpha;weight=Prior$weight;z=Prior$z;R=Mcmc$R
    keep=Mcmc$keeps_beta1=Mcmc$s_beta1;s_beta2=Mcmc$s_beta2
    s_phi1=Mcmc$s_phi1;s_phi2=Mcmc$s_phi2

    cat(" ",fill = TRUE)
    cat("Starting Random Walk Metropolis Sampler for NB Regression",
fill = TRUE)
    fsh()

    par=rep(0,(nvar+1))
    mle=optim(par,llnegbin,X=X,y=y,nvar=nvar,method="L-BFGS-B",upper=
c(rep(Inf,nvar),log(1e+08)),hessian=TRUE,control=list(fnscale=-1))
    fsh()

    beta_mle=mle$par[1:nvar]
    phi_mle = exp(mle$par[nvar + 1])
    varcovinv = -mle$hessian

    betacvar1 = s_beta1 * solve(varcovinv[1:nvar, 1:nvar])
    betaroot1 = t(chol(betacvar1))
    phicvar1 = s_phi1/varcovinv[nvar + 1, nvar + 1]
    phicroot1 = sqrt(phicvar1)

    betacvar2 = s_beta2 * solve(varcovinv[1:nvar, 1:nvar])
    betaroot2 = t(chol(betacvar2))
    phicvar2 = s_phi2/varcovinv[nvar + 1, nvar + 1]
    phicroot2 = sqrt(phicvar2)

    beta=matrix(c(beta_mle,beta_mle),nrow=nvar,ncol=ncomp)
    beta1=beta[,1]; beta2=beta[,2]
    phi =rep(phi_mle,ncomp); phi1=phi[1]; phi2=phi[2]

    nacceptbeta1=0; acceptrbeta1=0
    nacceptbeta2=0; acceptrbeta2=0
```

```
nacceptphi1=0; acceptrphi1=0
nacceptphi2=0; acceptrphi2=0

betadraw=matrix(double(floor(R/keep)*(nvar*ncomp)),ncol=nvar*ncomp)
phidraw=matrix(double(floor(R/keep)*(ncomp)),ncol=ncomp)
weightdraw=matrix(double(floor(R/keep)*(ncomp)),ncol=ncomp)
llike=rep(0,floor(R/keep))

itime = proc.time()[3]
cat(" ", fill = TRUE)
cat("MCMC Iteration (est time to end - min) ", fill = TRUE)
fsh()

for (r in 1:R) {
      tempweight=postweight(z,ncomp,alpha)

      newbeta1=beta1+betaroot1%*%rnorm(nvar)
      oldlpostbeta1=lpostbeta1(beta1,phi1,X,y,z,betabar,A)
      newlpostbeta1=lpostbeta1(newbeta1,phi1,X,y,z,betabar,A)
      ldiff=newlpostbeta1-oldlpostbeta1
      acc=min(1,exp(ldiff))

      if (acc<1){unif=runif(1)} else {unif=0}
      if (unif<=acc) {
            beta1=newbeta1
            nacceptbeta1=nacceptbeta1+1
      }

      logphi1=rnorm(1,mean=log(phi1),sd=phicroot1)
      oldlpostphi1=lpostphi1(beta1,phi1,X,y,z,a,b)
      newlpostphi1=lpostphi1(beta1,exp(logphi1),X,y,z,a,b)
      ldiff=newlpostphi1-oldlpostphi1
      acc=min(1,exp(ldiff))

      if (acc<1){unif=runif(1)} else {unif=0}
      if (unif<=acc) {
            phi1=exp(logphi1)
            nacceptphi1=nacceptphi1+1
      }

      newbeta2=beta2+betaroot2%*%rnorm(nvar)
      oldlpostbeta2=lpostbeta2(beta2,phi2,X,y,z,betabar,A)
      newlpostbeta2=lpostbeta2(newbeta2,phi2,X,y,z,betabar,A)
      ldiff=newlpostbeta2-oldlpostbeta2
      acc=min(1,exp(ldiff))

      if (acc<1){unif=runif(1)} else {unif=0}
      if (unif<=acc) {
            beta2=newbeta2
            nacceptbeta2=nacceptbeta2+1
      }

      logphi2=rnorm(1,mean=log(phi2),sd=phicroot2)
      oldlpostphi2=lpostphi2(beta2,phi2,X,y,z,a,b)
      newlpostphi2=lpostphi2(beta2,exp(logphi2),X,y,z,a,b)
      ldiff=newlpostphi2-oldlpostphi2
```

```
            acc=min(1,exp(ldiff))

            if (acc<1){unif=runif(1)} else {unif=0}
            if (unif<=acc) {
                  phi2=exp(logphi2)
                  nacceptphi2=nacceptphi2+1
            }

# Permutation based on the constraint #
            tempbeta=matrix(c(beta1,beta2),nrow=nvar,ncol=ncomp)
            tempphi=c(phi1,phi2)
            tempz=postz(tempbeta,tempphi,ncomp,X,y,tempweight) #updating z
            if (tempweight[1]>tempweight[2]){
                  beta=tempbeta; phi=tempphi; weight=tempweight; z=tempz

            }
            else{
                  beta=matrix(c(beta2,beta1),nrow=nvar,ncol=ncomp)
                  phi=c(phi2,phi1)
                  weight=c(tempweight[2],tempweight[1])
                  z=matrix(c(tempz[2,],tempz[1,]),nrow=ncomp,ncol=nobs,
byrow=T)
            }
            beta1=beta[,1]; beta2=beta[,2] # updating beta
            phi1=phi[1]; phi2=phi[2] # updating phi

            if (r%%100==0){
                  ctime=proc.time()[3]
                  timetoend=((ctime-itime)/r)*(R-r)
                  cat(" ",r," (", round(timetoend/60, 1), ")", fill=TRUE)
            fsh()
            }

            if(r%%keep==0){
                  mkeep=r/keep
                  betadraw[mkeep,]=beta
                  phidraw[mkeep,]=phi
                  weightdraw[mkeep,]=weight
                  llike[mkeep]=llnbmix(beta1,beta2,phi,X,y,weight)
            }
      }
      ctime = proc.time()[3]
      cat("  Total Time Elapsed: ",round((ctime-itime)/60,2),"\n")

      return(list(llike=llike,betadraw=betadraw,phidraw=phidraw,weightdraw
=weightdraw,acceptrbeta1=nacceptbeta1/R*100,acceptrbeta2=nacceptbeta2/R*100
,acceptrphi1=nacceptphi1/R*100,acceptrphi2=nacceptphi2/R*100))
}
```

VITA

Byung Jung Park received a B.E. (1997) degree in urban engineering and an M.S. (1999) degree in transportation engineering both from Seoul National University. In August 2006, he came to Texas A&M University to pursue a Ph.D. in transportation engineering at the Zachry Department of Civil Engineering. Prior to coming to Texas A&M University, he was a research scientist with the Highway Research Division at the Korea Transport Institute for five years.

During his Ph.D. studies, Mr. Park maintained a perfect 4.0 GPA academic record while taking advanced courses in civil engineering, statistics and industrial engineering. He received the Jacobs Engineering Scholarship during 2008 – 2009 and was nominated as the best doctoral student in the Zachry Department of Civil Engineering in 2009. He is also a member of Phi Kappa Phi honor society by election of the chapter at Texas A&M University. His research ability was recognized at the Texas Transportation Institute in 2008, as he was awarded the Keese-Wootan Transportation Fellowship.

Mr. Park has a particular interest in applying advanced statistical theories and techniques to transportation applications. While the majority of his applications have focused on highway safety analysis, he has a keen interest in transportation modeling in planning and operation as well. His dissertation topic was selected for presentation at the Doctoral Student Research in Transportation Operations and Traffic Control at the 88th Annual Meeting of Transportation Research Board in January, 2009. He has published or submitted several papers to academic journals based on his dissertation.

Mr. Park may be reached at Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843-3136. His email address is soldie71@gmail.com.