Robust estimation and inference in finite mixtures of generalized linear models

By

Jing Shen

(Under the direction of Daniel B. Hall)

Abstract

This dissertation studies robust estimation methods for finite mixtures of generalized linear models (FMGLMs). FMGLMs have been proven useful for modelling data arising from heterogeneous populations. Maximum likelihood (ML) estimation is usually an attractive approach for fitting such models. However, it is well known that the ML estimator (MLE) can be very unstable when the data have outliers, when there is poor separation between mixture components, or under certain other types of model violations. In this dissertation, we study two types of robust methods for such models. Chapter 3 studies minimum Hellinger distance (MHD) estimation methods in FMGLM context, and we propose approaches on both marginal and conditional density-based definitions of the Hellinger distance. We discuss the practical feasibility of employing these methods and examine empirically their robustness properties. Asymptotic properties of these estimators are also discussed. Simulations and examples show that they are competitive to the MLE when the data are correctly specified, but more robust when the data are not. In addition, the zero inflated regression model, a special case of the class of FMGLMs, is also considered in detail. In chapter 4, we propose another type of robust method for zero inflated models, which we term as the robust expectation solution (RES) method. It is designed to downweight outliers and yields an estimator with bounded influence function. Consistency and asymptotic normality of this estimator are established. Robustness properties of the RES approach are presented, as well as simulation-based comparisons between this approach and both ML estimation and marginal MHD.

Index words: Conditional density; Double kernel methods; EM algorithm; Expectation solution algorithm; Mallow's class; Minimum Hellinger distance; M-estimator; Mixture models; Robustness; Zero-inflated binomial; Zero-inflated Poisson.

Robust estimation and inference in finite mixtures of generalized linear models

by

Jing Shen

B.S., Beijing Polytechnic University, China, 2001

M.S., The University of Georgia, U.S., 2002

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

Athens, Georgia

2006

Robust estimation and inference in finite mixtures of generalized linear models

by

Jing Shen

Approved:

Major Professor: Daniel B. Hall

Committee: Lynne Seymour
Jaxk Reeves
T. N. Sriram
Xiangrong Yin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2006

Acknowledgements

I am deeply indebted to Dr. Daniel B. Hall, my major professor, who inspires and leads me towards the path of the research career, and has given me tremendous help all along the way. His expert guidance and kind encouragement are significantly important to my dissertation and my career. I am sincerely grateful for everything he has done for me.

I want to express my appreciation to Dr. Jaxk Reeves, who gives me so many great opportunities to explore different fields in statistics and encourages me to pursue my achievements in this career.

I am truly grateful to Dr. Lynne Seymour, Dr. T. N. Sriram and Dr. Xiangrong Yin, my committee members. Their assistance and suggestions have led to improvement in the quality of my dissertation.

Also I would like to thank all the faculty members, staff and students in the Department of Statistics. Their kindness and help during my five wonderful years at UGA is appreciated.

I also want to thank my parents and brother for their love and support. They have given me endless wishes from the day I was born.

Finally, I express my deepest gratitude to my dear husband, Ping Eric Yeung. Thank him for his love, support and encouragement from the first day we met. Without him, there won't be any success in my life.

Table of Contents

Chapter 1

Introduction

Finite mixtures of generalized linear models (FMGLMs) provide a useful tool to analyze data arising from a heterogeneous population. They can handle situations where a simple parametric distribution is unable to provide a satisfactory model for local variation in the observed data. In addition, an important special case of FMGLMs occurs when one component is a degenerate distribution with point mass of one at zero. Such models are known as zero inflated (ZI) models and are useful for analyzing data with extra zeros. In the last decade, there has been increasing interested in FMGLMs. Jansen (1993) showed that by adopting a simple expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977), the mixture problem can be split into two simpler non-mixture problems. Maximum likelihood (ML) estimation is an attractive approach to fitting such models because of the consistency and efficiency of ML estimators and the computational convenience of implementation via the EM algorithm. However, when outliers and other types of data contamination exist or when the components of the mixture are not well separated, the ML estimator is known to be extremely unstable.

Recently, robust alternatives to ML estimation for certain mixture models have been studied in the literature. The aim of robust statistics is to describe the structure which best fits the bulk of the data without distortion due to the effects a few anomalous data points or features. There exists a great variety of approaches to robust estimation. Huber (1964) introduced the M-estimator which became a very popular

robust method in literature. Hampel (1968) proposed the infinitesimal approach, which is based on the influence function. The influence function itself became one of the most useful heuristic tools to quantify the robustness of a statistic.

Minimum distance estimation forms another subclass of robust methods. The most popular distance metrics include the Kolmogorov-Smirnov distance, the Cramér-von Mises distance, and the Hellinger distance. Among the minimum distance methods, the minimum Hellinger distance (MHD) approach, proposed by Beran (1977) has been shown to be robust to certain types of violations of model assumptions and is asymptotically equivalent to the ML estimator (Donoho and Liu, 1988). The MHD approach has also been discussed in the context of finite mixtures of discrete data distributions by Lindsay (1994), Cutler and Cordero-Brana (1996), Karlis and Xekalaki (2001), and Lu et al. (2003). Lindsay discussed how the MHD estimator balances efficiency when the model has been appropriately chosen and robustness when it has not. Cutler and Cordero-Brana considered MHD estimation for finite mixture models when the exact forms of the component densities are unknown in detail but thought to be close to members of some parametric family. Karlis and Xekalaki considered MHD estimation in the case of finite mixtures of Poisson distributions, while Lu et al. added regression structure to this context.

The goal of my dissertation is to develop robust estimation methods for mixture regression models. In this dissertation we extend MHD estimation to a general context of FMGLMs and we also propose a more efficient MHD estimator based on the conditional density. The idea of this minimum conditional Hellinger distance (MCHD) method is to minimize the Hellinger distance between a nonparametric conditional density and the model density. Simulations are carried out to study the estimation properties of (marginal) MHD estimation, MCHD estimation and ML for data with and without outliers. The results show that the MCHD method performs comparably to the MLE when the data have no outliers, but offers much greater

robustness otherwise. For a special subclass of FMGLMs, ZI regression models, we propose another robust estimation method by taking advantage of the mixture structure, and adapting the standard ML EM algorithm, which we refer to as the RES method. This approach is more closely related to M-estimation and can be extended to general FMGLMs easily. The asymptotic properties are investigated under some regularity conditions. Simulation results indicate that the RES method is useful in downweighting outliers in the data compared to ML, and also outperforms marginal Hellinger distance based methods in several respects.

The organization of this dissertation is as follows. Chapter 2 gives the literature background of FMGLMs, ZI models, MHD estimation, kernel density estimation and M estimation. Chapter 3 studies the MHD estimation methods including unconditional (marginal) and conditional MHD estimation for FMGLMs. Chapter 4 proposes the RES method and contrasts it with MHD approach, focusing on the ZI regression context. Examples involving veterinary heart arhythmia data, and data related to a study of aggressive behavior among sixth grade school children are used to illustrate the methods in chapters 3 and 4. Chapter 5 presents future extensions of this dissertation work.

Chapter 2

Literature Review

This chapter provides a literature review on finite mixture models, as well as some robust statistical methods and techniques that will be needed in subsequent chapters.

2.1 Finite Mixtures of Generalized Linear Models

FMGLMs have proven useful for modelling data arising from a heterogeneous population. The basic definition is as follows. Let $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ denote observations where y_i represents the observed value of Y_i, and $\boldsymbol{x}_i = (\boldsymbol{x}_i^{(m)^T}, \boldsymbol{x}_i^{(r)^T})^T$ denotes a vector of explanatory variables or covariates. Usually, the first element of $\boldsymbol{x}_i^{(m)^T}$ and $\boldsymbol{x}_i^{(r)^T}$ is 1, corresponding to an intercept. Here the superscript (m) denotes covariates associated with the mixing probability and (r) denotes covariates related to the component means, which sometimes may contain the same variables. The unobserved mixing process assumes that Y_i can come from any one of c states, where c is finite but possibly unknown. Let $Z_{ij} = 1$ if observation i comes from component j, 0 otherwise, where $j = 1, \ldots, c$, and let $P(Z_{ij} = 1) = p_{ij}$, where $\sum_{j=1}^{c} p_{ij} = 1$. We assume that these mixing probabilities are related to covariates via a generalized logit type model of the following form, although other link functions could easily be accommodated:

$$p_{ij} = p_{ij}(\boldsymbol{x}_i^{(m)}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}_j' \boldsymbol{x}_i^{(m)})}{1 + \sum_{k=1}^{c-1} \exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i^{(m)})} \quad j = 1, \ldots, c-1, \tag{2.1.1}$$

where

$$p_{ic} = p_{ic}(\boldsymbol{x}_i^{(m)}, \boldsymbol{\beta}) = 1 - \sum_{j=1}^{c-1} p_{ij}.$$

In addition, we assume exponential dispersion family densities for each component of the mixture; that is,

$$f_j(y_i|\boldsymbol{x}_i^{(r)}, a_i, \boldsymbol{\alpha}_j) \equiv G(y_i|a_i, \eta_{ij}) = \exp\{\frac{y_i\eta_{ij} - b(\eta_{ij})}{a_i(\phi)} + c(y_i)\}. \tag{2.1.2}$$

Then the probability function of Y_i is

$$f(y_i|\boldsymbol{x}_i^{(r)}, \boldsymbol{x}_i^{(m)}, a_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{c} p_{ij} G(y_i|a_i, \eta_{ij}). \tag{2.1.3}$$

We refer to the class of models defined in this way as FMGLMs. Such models have received considerable attention recently.

2.1.1 Introduction to FMGLMs

Jansen (1993) was among the first to consider the FMGLM class in full generality and presented a DNA example where a finite mixture of ten Weibull distribution was adopted. Wang (1996) discussed mixtures of Poisson regression models with covariate dependent rates. Wang (1998) also studied mixtures of logistic regression models that accounted for extra-binomial variation through a mixture structure as well as by including covariates in the binomial parameters, the mixing probabilities or both. He argued that such models provided a more interpretable alternative to other approaches for dealing with extra-binomial variation, such as beta-binomial and quasi-likelihood models. He applied the mixture logistic regression model to analyze data from a study in evolutionary biology reported in McCullagh and Nelder (1989, p.143). This study investigated whether three adult *tribolium* (a type of beetle) species have developed an evolutionary advantage to recognize and avoid eggs of their own species while foraging. The issue being tested is whether the adult species exhibit preference for eggs of the other species. Fitting the mixture models

here provided a more general mean-variance relationship than the beta-binomial or quasi-likelihood models, and also provided some interesting insights into the data. Xiao et al. (1998) and several other more recent papers have used FMGLMs to model length of stay in a hospital. In this context a two-component mixture is motivated by the assumption of short-stay and long-stay patients who have different health care resource consumption patterns. Other recent papers on FMGLMs include Rosen, Jiang, and Tanner (2000) who considered marginal FMGLMs for clustered heterogeneous data and Hall and Wang (2004) who considered random effect versions of such models.

A closely related line of research is that on nonparametric generalized linear models (NPGLMs). NPGLMs can be thought of as GLMs with random effects in which the continuous random effects distribution is replaced by a discrete one with a finite number of estimated mass points. The result is a finite mixture of GLMs. NPGLMs have been developed by many authors (Follmann and Lambert, 1989; Hinde and Wood, 1987; Dietz, 1992; Aitkin, 1999).

In fitting mixture models, Jansen (1993) adopted the EM algorithm (Dempster, Laird and Rubin 1977) . This makes it possible to fit FMGLMs by iteratively applying standard GLM estimation methods. Moreover, standard statistical packages can be used to implement the computational work. A general procedure, which requires specification of the GLM for the mixing proportions and specification of the GLM for the mixing distribution can be easily written. The mixture problem can be considered as one of many examples in which the data can be viewed as incomplete. Mixture data are incomplete in the sense that each observation's component of origin z_{ij} is unobserved, or missing, where $z_{ij} = 1$ if the observation i comes from the jth component, $z_{ij} = 0$ otherwise. Suppose that the probability density of the

ith observation is (2.1.3). Then the complete loglikelihood data $(\boldsymbol{y}, \boldsymbol{z})$ is

$$\sum_{i=1}^{n}\sum_{j=1}^{c} z_{ij}\log\{p_{ij}(\boldsymbol{\beta})\} + \sum_{i=1}^{n}\sum_{j=1}^{c} z_{ij}\log G(y_i|\boldsymbol{\alpha}_j), \tag{2.1.4}$$

Maximum likelihood estimates may be obtained by applying the EM algorithm to (2.1.4). Each iteration consists of two steps. First, in the E-step, z_{ij} is replaced in (2.1.4) by its expected value given the current parameter estimates, which is $z^*_{ij} = \frac{p_{ij}(\boldsymbol{\beta})G(y_i|\boldsymbol{\alpha}_j)}{\sum_{j=1}^{c} p_{ij}(\boldsymbol{\beta})G(y_i|\boldsymbol{\alpha}_j)}$. Next, in the M-step, new parameter estimates are obtained by maximizing the resulting quantity with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Note that the quantity to be maximized splits into two terms where the first term is a function of the mixing proportions only, and the second term is a function of the parameters of the component distributions only. So the M- step can be split into two M-steps for standard non-mixture problems. The first M-step is solved by fitting a GLM for multinomial data to the "data" z^*_{ij}. The second one is solved by fitting c GLMs to the response variable, where the jth of these GLMs have weights $z^*_{1j}, \ldots, z^*_{nj}$.

2.1.2 Zero Inflated Models

The zero inflated (ZI) regression model is a special subclass of FMGLMs where one of the components is taken to be a degenerate distribution, having mass of 1 at 0. The other component is a non-degenerate distribution such as the Poisson, binomial, negative binomial or other form depending on the situation. For example, when manufacturing equipment is properly aligned, defects may be nearly impossible. But when it is misaligned, defects may occur according to a Poisson($\boldsymbol{\lambda}$) distribution. For such data that also have covariates, Lambert (1992) proposed the zero inflated Poisson (ZIP) regression model. In ZIP regression model, the response vector is $\boldsymbol{y} = (y_1, \ldots, y_n)'$, where y_i is the observed value of the random variable Y_i. We assume that Y_i's are independent where

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i; \\ \text{Poisson}(\lambda_i) & \text{with probability } 1 - p_i. \end{cases}$$

Moreover, the parameters $\boldsymbol{p} = (p_1, ..., p_n)^T$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)^T$ are modelled through canonical link GLMs as $\text{logit}(\boldsymbol{p}) = \boldsymbol{G\gamma}$ and $\log(\boldsymbol{\lambda}) = \boldsymbol{B\beta}$, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are regression parameters, and $\boldsymbol{G}$ and $\boldsymbol{B}$ are corresponding design matrices that pertain to the probability of the zero state and the Poisson mean, respectively. The log-likelihood function is written as

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{y}) &= \sum_{y_i=0} \log\{e^{\boldsymbol{G}_i\boldsymbol{\gamma}} + \exp(-e^{\boldsymbol{B}_i\boldsymbol{\beta}})\} + \sum_{y_i>0} (y_i \boldsymbol{B}_i\boldsymbol{\beta} - e^{\boldsymbol{B}_i\boldsymbol{\beta}}) \\ &\quad - \sum_{y_i>0} \log(y_i!) - \sum_{i=1}^{n} \log(1 + e^{\boldsymbol{G}_i\boldsymbol{\gamma}}), \end{aligned} \tag{2.1.5}$$

where $\boldsymbol{B}_i$ and $\boldsymbol{G}_i$ are the ith rows of design matrices $\boldsymbol{B}$ and $\boldsymbol{G}$.

As with FMGLMs, the EM algorithm is a convenient way to obtain the MLE. Suppose we knew which zeros came from the degenerate distribution and which came from the Poisson distribution; that is, suppose we could observe $z_i = 1$ when y_i is from degenerate state, and $z_i = 0$ when y_i is from the Poisson state. Then the log-likelihood for the complete data $(\boldsymbol{y}, \boldsymbol{z})$ would be

$$\begin{aligned}\ell^c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z}) &= \sum_{i=1}^{n} [z_i \boldsymbol{G}_i\boldsymbol{\gamma} - \log(1 + e^{\boldsymbol{G}_i\boldsymbol{\gamma}})] + \sum_{i=1}^{n} (1 - z_i)[y_i \boldsymbol{B}_i\boldsymbol{\beta} - e^{\boldsymbol{B}_i\boldsymbol{\beta}} - \log(y_i!)] \\ &= \ell^c_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z}) + \ell^c_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{z}). \end{aligned} \tag{2.1.6}$$

This log-likelihood is easy to maximize, because $\ell^c_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z})$ and $\ell^c_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{z})$ can be maximized separately. With the EM algorithm, the log-likelihood (2.1.5) is maximized iteratively by alternating between estimating z_i by its expectation under the current estimates of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ (E step) and then, with the z_i fixed at their expected values from the E step, maximizing $\ell^c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z})$ (M step), until the estimated $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ converges and iteration stops.

Hall (2000) extended Lambert's methodology to an upper bounded count situation, thereby obtaining a zero inflated binomial (ZIB) model. In the ZIB model, the Poisson($\boldsymbol{\lambda}$) component is replaced by a binomial(m, π) component and instead of modelling $\boldsymbol{\lambda}$, we model $\text{logit}(\pi) = \boldsymbol{B\beta}$.

2.2 Minimum Hellinger Distance Estimation

Minimum distance estimation methods have received considerable attention in the literature. Beran (1977) first considered the Hellinger distance as the basis of estimation. This method is asymptotically efficient under a specific regular parametric family of densities and is robust in a small Hellinger metric neighborhood of the given family. In this section, we will briefly review this method.

2.2.1 Definition of MHD Estimation

Beran (1977) proposed MHD estimation for iid observations. The Hellinger distance is defined as follows. Given a random sample $y_1, \ldots, y_n$, let $f_{\boldsymbol{\theta}}$ denote the probability density assumed for the data according to the model under consideration, which depends on a possibly unknown and vector-valued parameter $\boldsymbol{\theta}$, and let f_n denote a nonparametric density estimate. Let $\|.\|_p$ denote the $\mathbf{L}^p$ norm defined by

$$\|h\|_p = \{\int |h|^p\}^{1/p}.$$

The Hellinger distance is

$$H^2(f_{\boldsymbol{\theta}}, f_n) = \|f_{\boldsymbol{\theta}}^{1/2} - f_n^{1/2}\|_2^2 = 2 - 2\int f_{\boldsymbol{\theta}}^{1/2} f_n^{1/2} d\mu.$$

The MHD estimator of $\boldsymbol{\theta}$ is defined as

$$T(f_n) = \arg\min_{\boldsymbol{\theta}\in\Theta} H^2(f_{\boldsymbol{\theta}}, f_n). \tag{2.2.7}$$

Many papers have been published that investigate the properties and applications of MHD estimation. Beran (1977) used the α-influence function (IF_α) to examine the robustness of the MHD estimator. For $\boldsymbol{\theta} \in \Theta$, let $f_{\alpha,\boldsymbol{\theta},\delta} = (1-\alpha) f_{\boldsymbol{\theta}}(x) + \alpha\delta(x)$ and

$$IF_\alpha = \frac{T(f_{\alpha,\boldsymbol{\theta},\delta}) - \boldsymbol{\theta}}{\alpha}, \tag{2.2.8}$$

where $\delta(x)$ is a contamination component. The (ordinary) influence function can be obtained as $\alpha \to 0$. Since IF_α may not converge uniformly to the influence function in δ, it is necessary to examine IF_α rather than the influence function to assess the robustness of an estimator. Beran showed that, for any $\alpha \in (0, 1)$, the IF_α is a bounded function of δ. Hence $T(.)$ is robust at $f_{\boldsymbol{\theta}}$ with $100\alpha\%$ contamination. Donoho and Liu (1987) established that MHD estimation is "automatically" robust in the sense that it optimizes certain quantitative measures of "robustness". Among many notions of robustness, they identified two quantitative measures. First, stability of variance, which means the asymptotic variance of the estimator should stay small, uniformly over a neighborhood of the true model. Huber (1964) showed that M estimators can be designed to satisfy this criterion in an optimal fashion. Second, stability of quantity estimated, which means the quantity being estimated should change as little as possible, uniformly over a neighborhood of the true model. They showed that minimum distance estimators are robust with respect to this latter notion of robustness and when the distance is a Hellinger distance, the estimator has the smallest possible sensitivity to contamination of the model among Fisher-consistent functionals.

Lindsay (1994) investigated how the MHD estimator and its relatives balance efficiency and robustness. This paper defined the residual adjustment function (based on the Pearson's residual), and showed that this function carries the relevant information about the trade-off between efficiency and robustness. MHD estimation is a procedure that essentially downweights large Pearson's residuals, whereas M estimation downweights influential (on the MLE) observations with some sacrifice of first-order efficiency (Hampel, 1974).

2.2.2 MHD Estimation for Count Data

Simpson (1987) extended MHD estimation to models with countable support, for example, models for non-negative integer-valued counts. He pointed out that MHD estimation gives little weight to counts that are improbable relative to the model, and is asymptotically equivalent to the ML estimator when the model is correct. He proposed the MHD estimator for discrete data as follows. Let N_y be the frequency of y among $y_1, \ldots, y_n$, and let f_n be the empirical density function. That is,

$$f_n(y) = N_y/n, \quad y \in U,$$

where U is the set of all possible value of y. Then the MHD estimation is equivalent to maximizing $\rho_n(\boldsymbol{\theta}) = \sum_{y\in U} f_n^{1/2}(y) f_{\boldsymbol{\theta}}^{1/2}(y)$. This yields the standardized estimating equation

$$\rho_n(\boldsymbol{\theta})^{-1} \sum_{y\in U} f_n^{1/2}(y) f_{\boldsymbol{\theta}}^{1/2}(y) \frac{\partial \log f_{\boldsymbol{\theta}}(y)}{\partial \boldsymbol{\theta}} = 0. \tag{2.2.9}$$

Here, $\frac{f_n^{1/2}(y) f_{\boldsymbol{\theta}}^{1/2}(y)}{\rho_n(\boldsymbol{\theta})}$ is called the MHD weight for y. The equation is standardized so that $\sum_{y\in U} \frac{f_n^{1/2}(y) f_{\boldsymbol{\theta}}^{1/2}(y)}{\rho_n(\boldsymbol{\theta})} = 1$.

The consistency and asymptotic normality of the MHD estimator were also discussed by Simpson (1987). Write $H^2(\boldsymbol{\theta}; G) = \|g^{1/2} - f_{\boldsymbol{\theta}}^{1/2}\|_2^2$, where G is the true underlying distribution of the data and g is the density corresponding to G. The MHD functional T solves

$$H^2(T(G); G) = \min_{t\in\Theta} H^2(t; G)$$

if a solution exists. Suppose $f_{\boldsymbol{\theta}}(x)$ is continuous in $\boldsymbol{\theta}$ for each x, then for each $G \in \mathbf{G}$, Simpson proved that $T(G)$ exists. In addition if $T(G)$ is unique, then $\|g_n^{1/2} - g^{1/2}\|_2 \to 0$ implies that $T(G_n) \to T(G)$ as $n \to \infty$. Under some assumed smoothness conditions on the model, we can get $\dot{H}(t; G) = -2 \int \dot{s}_t g^{1/2} d\mu$, and $\ddot{H}(t; G) = -2 \int \ddot{s}_t g^{1/2} d\mu$, where $s_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}^{1/2}$ and $\dot{s}_{\boldsymbol{\theta}} = \partial s/\partial t$. The MHD estimator T_n is

then a solution of $\dot{H}(t; G_n) = 0$. In the discrete case, G_n is the empirical distribution function. Also if $\dot{H}(t; G)$ has a zero solution $\boldsymbol{\theta}$ interior to Θ, $\dot{H}(\boldsymbol{\theta}, G)$ is nonsingular, and $\dot{s}_{\boldsymbol{\theta}} \in \mathbf{L}^1$. Then $T_n \to \boldsymbol{\theta}$ in probability implies that $n^{1/2}(T_n - \boldsymbol{\theta}) \to N_d(0, V_{\boldsymbol{\theta}})$ in law as $n \to \infty$, where $V_{\boldsymbol{\theta}} = \frac{1}{4}\ddot{H}(\boldsymbol{\theta}, G)^{-1} i(\boldsymbol{\theta}) \ddot{H}(\boldsymbol{\theta}; G)^{-1}$ and $i(\boldsymbol{\theta}) = 4 \int \dot{s}_{\boldsymbol{\theta}} \dot{s}'_{\boldsymbol{\theta}} d\mu$. If $G \equiv F_{\boldsymbol{\theta}}$, then $V_{\boldsymbol{\theta}} = i(\boldsymbol{\theta})^{-1}$.

Simpson (1989) also provide the theoretical basis to make inference based on the MHD estimator. For fixed $\boldsymbol{\theta}_0 \in \Theta$ and τ, let $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \tau n^{-1/2}$ then under some smoothness conditions and $G \equiv f_{\boldsymbol{\theta}_n}$, as $n \to \infty$,

$$8n\{\rho(f_{\hat{\boldsymbol{\theta}}}, f_n) - \rho(f_{\hat{\boldsymbol{\theta}}_0}, f_n)\} - 2\log\{L_n(\tilde{\boldsymbol{\theta}})/L_n(\tilde{\boldsymbol{\theta}}_0)\} \to 0. \tag{2.2.10}$$

That is, Simpson (1989) established the asymptotic equivalence between the likelihood ratio statistic and a similar quantity based upon the Hellinger distance objective function, which thus can be used to form tests and confidence regions.

2.2.3 MHD Estimation for Finite Mixture Models

More recently, several researchers have extended MHD estimation to the finite mixture model context when the data are subject to contamination or poor separation. Woodward, Whitney, and Eslinger (1994) considered MHD estimation for finite mixtures, concentrating on the problem of estimating the mixing proportions in the mixture of two normals. Their results indicated that the MHD estimator obtains full efficiency at the fitting model while performing comparably with the minimum distance estimator based on Cramer-von Mises distance. Cutler et al. (1996) also discussed MHD estimation for finite mixture models, but in contrast to Woodward et al. (1994), they considered all parameters (mixing proportions and component means and variance) to be of interest. They proposed the HMIX algorithm to get the MHD estimator. The algorithm comprises a sequence of weighted one-component likelihood problems and is somewhat similar to the EM algorithm.

Karlis and Xekalaki (2000) investigated MHD estimation for finite Poisson mixture models. Their work extended MHD methodology to the semi-parametric case, where the number of support points is not known a priori. They developed a test procedure referred to as the Hellinger deviance test (HDT) for testing the Poisson assumption against a mixture Poisson assumption. The HDT is the Hellinger distance analogue of the likelihood ratio test for parametric inference which was first proposed in a non-mixture contest as (2.2.10) by Simpson. The HDT statistic is defined as follows,

$$HDT = 4n[H_0 - H_1],$$

where $H_i, i = 0, 1$, are the minimized Hellinger distances (objective functions) for the distribution under the null hypothesis and the alternative hypothesis respectively. Under some regularity conditions, the asymptotic distribution of the HDT is a χ^2 distribution with degrees of freedom equal to the difference in the numbers of parameters under the two hypotheses. However, the regularity conditions are not satisfied in some cases; in particular, they do not hold when testing the number of components in the mixture. For such situations they proposed the use of a bootstrap test; i.e., construction of the null distribution via parametric bootstrap. The test statistic measures the reduction in the Hellinger distance if one new component is added. The scheme allows the testing of H_0 : the data come from a k-component Poisson mixture, against H_1 : the data come from a (k+1)-component Poisson mixture. Since the influence of an outlier on this distance is much less than on the likelihood, the test based on the Hellinger distance is expected to be more robust. In addition, they reported the critical points for the null distribution of the test statistics derived via simulation, which make the HDT a convenient procedure to use for finite mixtures of Poissons.

Lu et al. (2003) extended the MHD method to finite mixtures of Poisson regression models. For model with covariates, an intuitive way to estimate $\boldsymbol{\theta}$ based on the Hellinger distance is to minimize the distance between the conditional densities, $f_n(y|\boldsymbol{x})$, a nonparametric conditional density and $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, the parametric density according to the model. To avoid introduction of a nonparametric conditional density estimate into the Hellinger distance criterion, however, they proposed the MHD estimator based on the distance between the empirical marginal densities $f_n(y)$ and $f_{\boldsymbol{\theta}}(y)$, with respect to $\boldsymbol{\theta}$,

$$H^2(f_{\boldsymbol{\theta}}(.), f_n(.)) = \int (f_{\boldsymbol{\theta}}^{1/2}(y) - f_n^{1/2}(y))^2 dy. \tag{2.2.11}$$

Monte Carlo simulation showed that the resulting marginal MHD estimator is robust and the finite sample bias is relatively small and decreases with the sample size. It performs better in comparison to the ML estimator when the mixture components are not well separated or when some mixing proportions are near zero and also when there is contamination in the responses.

2.3 Kernel Density Estimation

Kernel density estimation is one of the most widely used nonparametric density estimation methods. Given an iid random sample $y_1, \ldots, y_n$ from an unknown density f, the kernel density estimator is defined as

$$\hat{f}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b} K(\frac{y - y_i}{b}),$$

where K is a symmetric density function, and the scale factor b is called the bandwidth. The normal referencing rule, i.e. assuming f is a Gaussian density with standard deviation σ and replacing σ by sample standard deviation s_y (Silverman, 1986, p.45), yields

$$\hat{b} = \Big[\frac{8\pi \int K^2(y)dy}{3\{\int y^2 K(y)dy\}^2}\Big]^{1/5} s_y n^{-1/5}.$$

Parzen (1962) shows that $E[\hat{f}_n(y) - f(y)]^2 \to 0$ provided $h_n \to 0$, and $nh_n \to \infty$ as $n \to \infty$. Common examples of kernel functions include the Gaussian kernel and the Epanechnikov kernel $K(z) = \frac{1}{c}p(\frac{z}{c})$ where

$$p(z) = \begin{cases} \frac{3}{4}(1-z^2) & |z| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Recently, the estimation of conditional densities via kernel density has been developed in the literature. Rosenblatt (1969) first proposed that the estimation of $f(y|x)$ can be viewed as a nonparametric regression of $K_b(y - Y_i)$ on $\{X_i\}$. As $b_n \to 0$, from a standard Taylor argument,

$$E\{K_{b_n}(y - Y)|X = x\} \simeq f(y|x).$$

The estimator of $f(y|x)$ is called Nadaraya-Watson estimator and defined as

$$\hat{f}_{NW}(y|x) = \sum_{i=1}^{n} K_{b_n}(y - Y_i) w_i^{NW}(x),$$

where $w_i^{NW}(x) = \frac{W_{hn}(x-X_i)}{\sum_{i=1}^{n} W_{hn}(x-X_i)}$, K_{b_n} and W_{h_n} are two kernel functions, and b_n and h_n are the bandwidths (or smoothing parameters) corresponding to W and K respectively.

Fan et al. (1996) suggested using a locally polynomial regression function to estimate $f(y|x)$ and its partial derivative with respect to x. By Taylor expansion at x, we have

$$\begin{aligned} E\{K_b(Y - y)|X = z\} &\simeq f(y|z) \\ &\simeq f(y|x) + \dot{f}(y|x)^T(z - x) + \frac{1}{2}(z - x)^T \ddot{f}(y|x)(z - x) \\ &\equiv \beta_0 + \beta_1^T(z - x) + \beta_2^T \text{vec}\{(z - x)(z - x)^T\}, \end{aligned}$$

Then

$$\hat{f}(y|x) = \hat{\beta}_0 = \sum_{i=1}^{N} K_{b_n}(y - Y_i) w_i(x),$$

where $w_i(x) = \frac{W_{hn}(X_i-x)\{T_{n,2}-(X_i-x)T_{n,1}\}}{(T_{n,0}T_{n,2}-T_{n,1}^2)}$, with $T_{n,j} = \sum_{i=1}^{N} W_{h_n}(X_i - x)(X_i - x)^j$, $(j = 0, 1, 2)$. They also proposed a method for choosing the smoothing parameters, b_n and h_n, which uses the residual squares criteria proposed by Fan and Gijbels (1995). Hall et al. (1999) proposed a modified Nadaraya-Watson estimator for the conditional distribution function which is always non-negative and shares the same first order asymptotic properties as the local polynomial estimator. Hyndman et al. (2002) proposed another method for estimating conditional densities by adapting the locally polynomial function. It produces estimators that are always non-negative. If a local polynomial regression function i.e. $\sum_{j=0}^{r} \theta_j(X_i - x)^j$, is used, Hyndman's estimator reduces to that of Fan et al (1996). They also propose a practical two-step bandwidth selection algorithm. First, the optimal bandwidth b is chosen to minimize the weighted integrated mean squared error (IMSE) defined by $\int\int E\{\hat{f}(y|x) - f(y|x)\}^2 f^2(x)dxdy$. Second, given the bandwidth b, the bandwidth h can be found through the cross-validation technique (Fan and Gijbels, 1996, p.45).

Gooijer et al. (2003) proposed a re-weighted Nadaraya-Watson (RNW) estimator. They have demonstrated that asymptotically the RNW estimator preserves the superior large sample bias property of the local linear smoother of the conditional density proposed in the literature such as Rosenblatt (1969) and Fan et al. (1996). Fan et al. (2004) extended the technique of cross-validation to develop a consistent bandwidth selection rule and applied to their locally polynomial estimator (Fan et al. 1996).

All the conditional density estimators described above are defined based upon a single covariate x but they have straightforward extensions to the multivariate case in theory. However, with increasing dimension, the kernel density methods become hard to implement in practice. In addition, the curse of dimensionality becomes a problem and the performance of these methods suffers. Very recently, however, Hall et al. (2005) proposed a method for approximating the conditional distribution function of a random variable Y given a dependent random d-dimensional vector

$\boldsymbol{x}$ using a dimension reduction technique. Instead of estimating the distribution of $Y|\boldsymbol{x}$, they proposed to estimate $Y|\boldsymbol{\theta}^T\boldsymbol{x}$, where the vector $\boldsymbol{\theta}$ is selected so that the estimation is optimal under a least-squares criterion. More specifically, first they obtain an estimate $\hat{\boldsymbol{\theta}}$ by using a "leave two out" technique. Let

$$T^{[k]}_{-i,-j}(\boldsymbol{\theta}) = \frac{1}{(n-2)h}\sum_{i_1:i_1\neq i,j} K\{\frac{\boldsymbol{\theta}^T(\boldsymbol{x}_i-\boldsymbol{x}_{i_1})}{h}\}\{\frac{\boldsymbol{\theta}^T(\boldsymbol{x}_i-\boldsymbol{x}_{i_1})}{h}\}^k,$$

$$w_{i_1,-i,-j}(\boldsymbol{\theta}) = K\{\frac{\boldsymbol{\theta}^T(\boldsymbol{x}_i-\boldsymbol{x}_{i_1})}{h}\}\times\{T^{[2]}_{-i,-j}(\boldsymbol{\theta}) - \frac{\boldsymbol{\theta}^T(\boldsymbol{x}_i-\boldsymbol{x}_{i_1})}{h}T^{[1]}_{-i,-j}(\boldsymbol{\theta})\}.$$

$$\hat{F}_{-i,-j}(y|\boldsymbol{\theta}^T\boldsymbol{x}_i) = \{\sum_{i_1:i_1\neq i,j} w_{i_1,-i,-j}(\boldsymbol{\theta})I(Y_{i1}\leq y)\}\times\{\sum_{i_{1:}i_1\leq i,j} w_{i_1,-i,-j}(\boldsymbol{\theta})\}^{-1}.$$

Here $\hat{F}_{-i,-j}(y|\boldsymbol{\theta}^T\boldsymbol{x}_i)$ is a local linear estimator of $F(y|\boldsymbol{\theta}^T\boldsymbol{x}_i)$ based on data pairs other than the ith and jth; and $\frac{1}{n-1}\sum_{i:i\neq j,\boldsymbol{x}_i\in A}\hat{F}_{-i,-j}(y|\boldsymbol{\theta}^T\boldsymbol{x}_i)$ is an estimator of $\pi_{\boldsymbol{\theta}}(A,B)$ when $B=(-\infty,y]$. Let $\hat{F}_{-j}(A,y)$ be the proportion of the $n-1$ values of $(\boldsymbol{x}_i,y_i)$, for $i\neq j$, which satisfy $(\boldsymbol{x}_i,y_i)\in A\times(-\infty,y]$, and let $S(\boldsymbol{\theta},A) = \sum_{j=1}^n\{\hat{F}_{-j}(A,Y_j) - \frac{1}{n-1}\sum_{i:i\neq j,\boldsymbol{x}_i\in A}\hat{F}_{-i,-j}(Y_j|\boldsymbol{\theta}^T\boldsymbol{x}_i)\}^2$. They choose $\hat{\boldsymbol{\theta}}$ to minimize $S(\boldsymbol{\theta}) = \int S(\theta,A)d\mu(A)$ over $\boldsymbol{\theta}\in\Theta$. In practice, the integration is typically replaced by a sum over a class of selected balls. The integration is approximated by a series $S(\boldsymbol{\theta}) = \frac{1}{L}\sum_{i=1}^L S(\boldsymbol{\theta},A_i)$, where the $A_i's$ are spheres of radius r contained within R. They recommended choosing the bandwidth h by bootstrapping based on an approximating parametric model (Hall et al. 1999) and a arbitrary large value L. Their results show that the choice of L has little effect on final results. The estimator of the conditional distribution function of Y given $\hat{\boldsymbol{\theta}}^T\boldsymbol{x}$ is shown to be first order equivalent to its counterpart when the true value of $\boldsymbol{\theta}$ is known. Therefore we can use the one dimensional methods described above to estimate $f(y|\boldsymbol{\theta}^T\boldsymbol{x})$.

2.4 M Estimation

Huber (1964) introduced a flexible class of estimators, "M estimators", which have the properties of consistency and asymptotic normality under certain reg-

ularity conditions. These estimators are a generalization of the MLE. Suppose $X \sim$ a distribution F with density $f(x)$. The M estimator T_n is defined as $T_n = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho(x_i, \boldsymbol{\theta})$. If ρ has derivative Ψ, T_n is obtained at the solution of $\sum_{i=1}^{n} \Psi(x_i; \boldsymbol{\theta}) = 0$. Typically, the functions ρ and Ψ are chosen to downweight the contributions of extreme observations.

M-estimators for the linear regression context have been studied by Moronna and Yohai (1981). The class of estimators studied by these authors are written $(T_n, \hat{\sigma})$ is defined as the solution of the following two equations,

$$\sum_{i=1}^{n} \eta\{\boldsymbol{x}_i, (y - \boldsymbol{\theta}^T \boldsymbol{x}_i)/\sigma\}\boldsymbol{x}_i = \mathbf{0},$$

$$\sum_{i=1}^{n} \chi(|y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i|/\sigma) = 0.$$

There have been several proposals for choosing η and χ. Typically, however, η may be written in the form

$$\eta(x, r) = w(x)\psi(rv(x)),$$

for appropriate functions $\psi : R \to R$, $w : R^p \to R^+$, and $v : R^p \to R^+$. Examples of proposed choices of the $w(x)$, $v(x)$ that have been proposed in the literature are listed in Table 2.1. As for the scale σ, a popular choice is $\chi(r) = \psi(r)^2 - \beta$, where β is a constant chosen to make $\int \chi(x)\Phi(x)dx = 0$.

Table 2.1: Overview of some useful functions applied in η

Type	$w(x)$	$v(x)$
Huber (1973)	$w(x) = 1$	$v(x) = 1$
Mallows's	-	$v(x) = 1$
Andrews's	$w(x) = 1$	-
Hill et al(1977)	$w(x) = v(x)$	
Schweppe (1971)	$v(x) = 1/w(x)$	

The functional $T(F)$ corresponding to the M-estimator is the solution of

$$\int \eta(x, y - x^T T(F))x dF(x, y) = 0.$$

The influence function of T at a distribution F is given by

$$IF(x, y; T, F) = \eta(x, y - x^T T(F))M^{-1}(\eta, F)x,$$

where $M(\eta, F) = \int \eta'(x, y - x^T T(F))xx^T dF(x, y)$. Moronna and Yohai (1981) showed that, under certain regularity conditions, these estimators are consistent and asymptotically normally distribued with asymptotic covariance matrix

$$\begin{aligned} V(T, F) &= \int IF(x, y; T, F)IF^T(x, y; T, F)dF(x, y) \\ &= M^{-1}(\eta, F)Q(\eta, F)M^{-1}(\eta, F), \end{aligned}$$

where $Q(\eta, F) = \int \eta^2(x, y - x^T T(F))xx^T dF(x, y)$.

Recently, several authors have extended M estimation to the generalized linear model context and beyond. Preisser and Qaqish (1999) proposed a class of robust estimators in the more general setting of the generalized estimating equations (GEEs). Cantoni and Ronchetti (2001) studied robust inference for GLMs. They derived the asymptotic distribution of tests based on the quasi deviance function corresponding to a robust estimating function. Adimari and Ventura (2001) also studied robust inference for GLMs and derived a robust quasi-profile log likelihood function to make inference.

Other than the M-estimator, various robust estimators have been proposed by researchers. For example, Hodges and Lehmann (1963) proposed the so called R-estimator by using the idea of rank statistics. Bickel (1973) introduced the L-estimator, which minimizes an L_q norm of the residuals. Rousseeuw (1984) introduced the least median of square (LMS) estimator, defined by $\min_\theta \text{median}_i(r_i)^2$, where r_i is the Pearson's residuals. Rousseeuw (1984) adjusted the LMS method and

introduced the least trimmed squares (LTS) estimator, given by $\min_\theta \sum_{i=1}^{h} r_{(i)}^2$ where $r_{(i)}$, $i = 1, \ldots, n$ are the ordered Pearson's residuals and h determines the amount of trimming in the estimator. Both the LMS and LTS estimators are defined by minimizing a robust measure of the Person's residuals. In this dissertation, however, we will focus on minimum Hellinger distance estimation and M-estimation, and develop appropriate robust estimators for FMGLMs that fall into these traditions.

2.5 References

[1] Adimari, G. and Ventura, L. (2001). Robust inference for generalized linear models with application to logistic regression. *Statistics and Probability Letters*, **55**, 413-419.

[2] Aitkin, M. A. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics,* **55**, 117-128.

[3] Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics,* **46**, 683-705.

[4] Beran, R. (1977). Minimum Hellinger distance for parametric models. *The Annals of Statistics,* **5**, 445-463.

[5] Böhning, D., Schlattmann, P. and Lindsay, B. G. (1992). Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithm. *Biometrics,* **48**, 283-303.

[6] Böhning, D. and Seidel, W. (2002). Editorial: recent developments in mixture models. *Computational Statistics and Data Analysis,* **41**, 349-357.

[7] Bickel, P.J. (1973). On some analogues to linear combination of order statistics in the linear model. *The Annals of Statistics,* **1**, 597-616.

[8] Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association,* **96**, 1022-1030.

[9] Cutler, A., Cordero-Brana, O. (1996). Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association,* **91**, 1716-1724.

[10] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B,* **39**, 1-38.

[11] Dietz, E. (1992). Estimation of heterogeneity - A GLM approah. In *Advances in GLIM and Statistical Modelling, Lecture Notes in Statistics*, L. Fahrmeir, F. Francis, R. Gilchrist, and G. Tutz (eds), 66-72. Berlin: Springe' Verlag.

[12] Donoho, D. L. and Liu, R. C. (1988). The "automatic" robustness of minimum distance functionals. *The Annals of Statistics,* **16**, 552-586.

[13] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of Royal Statistics, Series B,* **57**, 371-394.

[14] Fan, J and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman and Hall, New York.

[15] Fan, J., Yao. Q. and Tong H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika,* **83**, 189-206.

[16] Fan J. and Y. T. (2004). A crossvalidation method for estimating conditional densities. *Biometrika,* **91**, 819-834.

[17] Follmann, D. A. and Lambert, D. (1989). Generalizing logistic regression non-parametrically. *Journal of the American Statistical Association,* **84**, 295-300.

[18] Follmann, D. A. and Lambert, D. (1991). Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference,* **27**, 375-381.

[19] Gooijer, J and Zerom, D. (2003). On conditional density estimation. *Statistical Neerlandica,* **57**, 159-176.

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

[21] Hall, D. B. and Wang, L. Mixtures of generalized linear mixed-effects models for cluster-correlated data. *Statistical Modelling: An International Journal.* Accepted, subject to revision.

[22] Hall, P., Wolff, R. C. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of American Statistical Association,* **94**, 154-163.

[23] Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *The Annal of Statistics,* **33**, 1404-1421.

[24] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association,* **69**, 383-393.

[25] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986) *Robust Statistics.* New York: John Wiley and Sons.

[26] Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association,* **89**, 897-904.

[27] Hili, O. (1999). On the estimation of β-ARCH models. *Statistics and Probability Letters,* **45**, 285-293.

[28] Hill, R.W. (1977). Robust regression when there are outliers in the carriers. Ph.D. thesis. Harvard University, Cambridge.

[29] Hinde, J.P. and Wood, A.T.A (1987). Binomial variance component models with a nonparametric assumption concerning random effects. In *Longitudinal Data Analysis*, R. Crouchley (ed.) Averbury, Aldershot, Hants.

[30] Hodges, J.L. Jr. and Lehmann, E.L. (1963). Estimates of location based on rank tests. *The Annal of Mathematics Statistics,* **34**, 598-611.

[31] Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics,* **35**, 73-101

[32] Huber, P. J. (1972). Robust statistics: a review. *The Annals of Mathematical Statistics,* **43**, 1041-1067.

[33] Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Nonparametric Statistics,* **14**, 259-278.

[34] Jansen, R. C. (1993). Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics,* **49**, 227-231.

[35] Karlis, D. and Xekalaki, E. (2001). Robust inference for finite Poisson mixtures. *Journal of Statistical Planning and Inference,* **93**, 93-115.

[36] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics,* **34**, 1-14.

[37] Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics,* **22**, 1081-1114.

[38] Lindsay, B. G. (1995). Mixture Models: theory, geometry and applications regional conference series in probability and statistics, Vol. 5. Institute of Mathematical Statistics and Am. Statist. Assoc., Hayward, California.

[39] Lu. Z., Hui. Y. and Lee, A. (2003) Minimum Hellinger distance estimation for finite mixtures of Poisson regression models and its applications. *Biometrics,* **59**, 1016-1026.

[40] Maronna, R.A. and Yohai, V.J. (1981). Asymptotic Behavior of General M-estimates for Regression and Scale with Random Carriers. *Z. Wahrsch. verw. Geb.,* **58**, 7-20.

[41] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, London: Chapman and Hall.

[42] McLachlan G. and Peel D. (2001). *Finite Mixture Models*. New York: Wiley.

[43] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**, 1065-1076.

[44] Pregibon, D. (1982). The Consultant's Forum: Resistant Fits for Some Commonly Used Logistic Models with Medical Applications. *Biometrics,* **3**, 485-498.

[45] Preisser J. S. and Qaqish, B. F. (1999). *Biometrics*, **55**,574-579.

[46] Rosen, O, Jiang, W.X. and Tanner, M.A (2000). Mixtures of marginal models. *Biometrika,* **87**, 391-404.

[47] Rosenblatt, M. (1969). Conditional pobability density and regression estimators. *Multivariate Analysis II* New York: academic Press, 25-31.

[48] Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association,* **79**, 871-880.

[49] Rousseeuw, P.J. and Yohai, V. (1984). Robust regression by means of S-estimators. *Robust an Nonlinear Time Series Analysis*. Springer, New York.

[50] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[51] Simpson, D. G. (1987). Minimum Hellinger distance estimation for analysis of count data. *Journal of the American Statistical Association,* **82**, 802-807.

[52] Simpson, D. G. (1989). Hellinger deviance tests: efficiency, breakdown points, and examples. *Journal of the American Statistical Association,* **84**, 107-113.

[53] Wang, P., Puterman, M. L., Cockburn, I. and Le, N. (1996). Mixed poisson regression models with covariate dependent rates. *Biometrics,* **52**, 391-400.

[54] Wang, P. and Puterman M. L. (1998). Mixed logistic regression models. *American Statistical Association and the International Biometric Society Journal of Agricultural, Biological, and Environmental Statistics,* **3**, 175-200.

[55] Woodward, W. A., Whitney P. and Eslinger, P. (1994). Minimum Hellinger distance estimation of mixture proportions. *Journal of Statistical Planning and Inference,* **48**, 303-319.

[56] Xiao, J.G., Douglas, D., Lee, A.H. and Vemuri, S.R. (1998). A discordancy test approach to identify outliers of length of hospital stay. *Statistics in Medicine,* **17**, 2199-2206.

Chapter 3

Minimum Hellinger Distance Estimation for Finite Mixtures of Generalized Linear Models

3.1 Introduction

Finite mixtures of generalized linear models (FMGLMs) provide a useful class of models for analyzing continuous and discrete (e.g., count-valued) data that exhibit heterogeneity relative to some exponential dispersion family distributions such as the binomial, Poisson, or normal. Maximum likelihood (ML) estimation is an attractive approach to fitting such models because of the consistency and efficiency of ML estimators and the computational convenience of implementation via the EM algorithm. However, when outliers and other types of data contamination exist or when the components of the mixture are not well separated, the ML estimator is known to be extremely unstable. Recently, robust alternatives to ML estimation for certain mixture models have been studied in the literature. The goal of this paper is to develop robust, Hellinger distance-based estimation methods for FMGLMs.

Beran (1977) proposed minimum Hellinger distance (MHD) estimation for independent and identically distributed (iid) parametric models. The MHD estimator has been shown to be robust to certain types of violations of model assumptions and is asymptotically equivalent to the ML estimator (Donoho and Liu, 1988). Simpson (1987) focused on MHD estimation for discrete data, where the model is allowed to have countably infinite support e.g., in models for unbounded counts. His paper showed that MHD estimation provides an effective means of estimation for count

data that are prone to outliers. The MHD approach has also been discussed in the context of finite mixtures of discrete data distributions by Lindsay (1994), Cutler and Cordero-Brana (1996), Karlis and Xekalaki (2001), and Lu et al. (2003). Lindsay discussed how the MHD estimator balances efficiency when the model has been appropriately chosen and robustness when it has not. Cutler and Cordero-Brana considered MHD estimation for finite mixture models when the exact forms of the component densities are unknown in detail but thought to be close to members of some parametric family. Karlis and Xekalaki considered MHD estimation in the case of finite mixtures of Poisson distributions, while Lu et al. added regression structure to this context.

In this paper, we extend Lu et al.'s work by studying MHD estimation in the more general context of FMGLMs, and propose a more efficient MHD estimator, which we term the minimum conditional Hellinger distance (MCHD) estimator (MCHDE). In addition, an important special case of FMGLMs occurs when one component is a degenerate distribution with point mass of one at zero. Such models are known as zero inflated (ZI) regression models (Lambert, 1992; Hall, 2000) and are useful for analyzing data with extra zeros relative to some exponential family distributions such as the Poisson. We will discuss MHD estimation in this important subclass of models in details as well.

The organization of this paper is as follows. Section 3.2 gives the literature background of FMGLMs and MHD estimation. Section 3.3 and 3.4 describes MHD estimation methods for FMGLMs, based on marginal and conditional density respectively. Simulation results are presented in section 3.5 to compare these two methods to each other and to ML estimation. Section 3.6 examines the robustness of these two methods through the α influence function. Two examples are given in section 3.7 for illustration of the methodology. Finally, some discussion is provided in section 3.8.

3.2 Background on FMGLMs

Generalized linear models (GLMs) have proved to be very useful in a wide array of application areas (McCullagh and Nelder, 1989). GLMs are based upon the assumption that the data are generated from an exponential family distribution such as the binomial, Poisson, or normal. However, in practice, data often exhibit patterns of variability inconsistent with such standard distributional assumptions. For example, excess variance, or overdispersion, relative to an exponential family distribution often occurs. In some cases, overdispersion has a specific form in which the data may be seen as arising from a heterogeneous population composed of two or more subclasses of subjects whose data follow distributions from the same parametric family with possible different values of parameters. Such models are sometimes called latent class models, and are examples of FMGLMs.

Let $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ denote observations where y_i represents the observed value of Y_i, and $\boldsymbol{x}_i = (\boldsymbol{x}_i^{(m)^T}, \boldsymbol{x}_i^{(r)^T})^T$ denotes a vector of explanatory variables or covariates. Usually, the first element of $\boldsymbol{x}_i^{(m)^T}$ and $\boldsymbol{x}_i^{(r)^T}$ is 1, corresponding to an intercept. Here the superscript (m) denotes covariates associated with the mixing probability and (r) denotes covariates related to the component means, which sometimes may share common variables. The unobserved mixing process assumes that Y_i can come from any one of c states, where c is finite but possibly unknown. Let $Z_{ij} = 1$ if observation i comes from component j, 0 otherwise, where $j = 1, \ldots, c$, and let $P(Z_{ij} = 1) = p_{ij}$ where $\sum_{j=1}^{c} p_{ij} = 1$. We assume that these mixing probabilities are related to covariates via a generalized logit type model of the following form, although other link functions could easily be accommodated

$$p_{ij} = p_{ij}(\boldsymbol{x}_i^{(m)}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}_j' \boldsymbol{x}_i^{(m)})}{1 + \sum_{k=1}^{c-1} \exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i^{(m)})} \quad j = 1, \ldots, c-1, \tag{3.2.1}$$

where

$$p_{ic} = p_{ic}(\boldsymbol{x}_i^{(m)}, \boldsymbol{\beta}) = 1 - \sum_{j=1}^{c-1} p_{ij}.$$

In addition, we assume exponential dispersion family densities for each component of the mixture; that is,

$$f_j(y_i|\boldsymbol{x}_i^{(r)}, a_i, \boldsymbol{\alpha}_j) \equiv G(y_i|a_i, \eta_{ij}) = \exp\{\frac{y_i\eta_{ij} - b(\eta_{ij})}{a_i(\phi)} + c(y_i)\}. \tag{3.2.2}$$

Then the density function of Y_i is

$$f(y_i|\boldsymbol{x}_i^{(r)}, \boldsymbol{x}_i^{(m)}, a_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{c} p_{ij} G(y_i|a_i, \eta_{ij}). \tag{3.2.3}$$

We refer to the class of models defined in this way as FMGLMs. Such models have received considerable attention recently. In addition, an important special case of FMGLMs occurs when one component is a degenerate distribution with point mass of one at zero. Such models are known as zero inflated (ZI) regression models. Lambert (1992) proposed the zero inflated Poisson (ZIP) regression model. In ZIP regression, the response vector is $\boldsymbol{y} = (y_1, \ldots, y_n)'$, where y_i is the observed value of the random variable Y_i. We assume the Y_i's are independent where

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i; \\ \text{Poisson}(\lambda_i) & \text{with probability } 1 - p_i. \end{cases}$$

Moreover, the parameters $\boldsymbol{p} = (p_1, ..., p_N)^T$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_N)^T$ are modelled through canonical link GLMs as $\text{logit}(\boldsymbol{p}) = \boldsymbol{\beta}'\boldsymbol{x}^{(m)}$ and $\log(\boldsymbol{\lambda}) = \boldsymbol{\alpha}'\boldsymbol{x}^{(r)}$, where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are regression parameters, and $\boldsymbol{x}^{(m)}$ and $\boldsymbol{x}^{(r)}$ are corresponding design matrices which pertain to the probability of the zero state and the Poisson mean, respectively. Other ZI regression models can be defined similarly. For example, in the zero inflated binomial model (Hall, 2000), the Poisson($\boldsymbol{\lambda}$) component is replaced by a binomial(m,π) component and instead of modelling $\boldsymbol{\lambda}$, one model, $\text{logit}(\pi) = \boldsymbol{\alpha}'\boldsymbol{x}^{(r)}$.

Maximum likelihood (ML) estimation is an attractive approach to fitting FMGLMs because of the consistency and efficiency of ML estimators and the

computational convenience of implementation via the EM algorithm. However, when outliers and other types of data contamination exist or when the components of the mixture are not well separated, the ML estimator is known to be extremely unstable.

Recently, several researchers have extended MHD estimation to the finite mixture model context when the data are subject to contamination or poor separation. Woodward, Whitney, and Eslinger (1994) considered MHD estimation for finite mixtures, concentrating on the problem of estimating the mixing proportions in the mixture of two normals. Their results indicated that the MHD estimator obtains full efficiency at the fitting model while performing comparably with the minimum distance estimator based on Cramer-von Mises distance away from the true model. Karlis and Xekalaki (2000) investigated MHD estimation for finite Poisson mixture models. Lu et al. (2003) extended the MHD method to finite mixtures of Poisson regression models. For model with covariates, an intuitive way to estimate $\boldsymbol{\theta}$ based on Hellinger distance is to minimize the distance between an empirical and model-based conditional densities, $f_n(y|\boldsymbol{x})$ and $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, respectively. That is, we define the conditional Hellinger distance as

$$H_c^2(f_n(.|.), f_{\boldsymbol{\theta}}(.|.)) = \int\int (f_n^{1/2}(y|\boldsymbol{x}) - f_{\boldsymbol{\theta}}^{1/2}(y|\boldsymbol{x}))^2 f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} dy. \tag{3.2.4}$$

To avoid introduction of a nonparametric conditional density estimate into the Hellinger distance criterion, however, Lu et al. (2003) proposed the MHD estimator based on the distance between the marginal densities $f_n(y)$ and $f_{\boldsymbol{\theta}}(y)$,

$$H^2(f_{\boldsymbol{\theta}}, f_n) = \int (f_{\boldsymbol{\theta}}^{1/2}(y) - f_n^{1/2}(y))^2 dy.$$

Here, the $f_n(y)$ indicates an empirical, non-parametric density estimate based only on the data, whereas the subscript $f_{\boldsymbol{\theta}}(y)$ indicates the parametric model-based marginal density. Monte Carlo simulation showed that the resulting marginal MHD estimator

is robust and its finite sample bias is relatively small and decreases with the sample size. It performs better in comparison to the ML estimator when the mixture components are not well separated or when some mixing proportions are near zero and also when there is contamination of the response, i.e. outliers in y.

3.3 Marginal MHD Estimation for Discrete FMGLMs

First we extend Lu et al.'s approach based on $H^2(.,.)$ to a more general context of FMGLMs for discrete data. We limit attention to finite mixtures of discrete data GLMs because in this context, it is reasonable to use the empirical frequency function of the data as the nonparametric density on which to base the MHD criterion of estimation. Outside of this class, other nonparametric density estimators could be used to define the Hellinger distance; however, we defer discussion of this possibility to section 4 where we introduce the conditional MHD approach.

Let $f_n(y)$ be the empirical frequency function defined by

$$f_n(y) = N_y/n, \quad y \in U, \tag{3.3.5}$$

where U is the set of all possible values of y. Let $f_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)$ denote the probability defined as in (3.2.3), where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$. Take $\boldsymbol{x}$ as the combined vector of all observed covariates, which we condition on and consider fixed.

3.3.1 Estimation Based on Marginal Densities

For a moment consider the iid case and suppose we know the form of $f_{\boldsymbol{\theta}}(y)$. Then the MHD estimator $\hat{\boldsymbol{\theta}}$ would be the root of

$$\sum_{y \in U} \frac{f_n^{1/2}(y)}{f_{\boldsymbol{\theta}}^{1/2}(y)} \dot{f}_{\boldsymbol{\theta}}(y) = n \sum_{i=1}^{n} \frac{f_{\boldsymbol{\theta}}^{1/2}(y_i)}{f_n^{1/2}(y_i)} \frac{\partial \log f_{\boldsymbol{\theta}}(y_i)}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{3.3.6}$$

Lindsay et al. (1992) introduced Lindsay's residual function $r(y) = \frac{f_n(y)}{f_{\boldsymbol{\theta}}(y)} - 1$, as a more appropriate quantity than Pearson's residual to assess goodness of fit in a

mixture context. Notice that (3.3.6) can be written as

$$\sum_{i=1}^{n} \frac{1}{\{1+r(y_i)\}^{1/2}} \frac{\partial \log f_{\boldsymbol{\theta}}(y_i)}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

From this representation, it can be seen that MHD downweights observations that have large Lindsay's residuals in the estimating equation.

In the non-iid situation with which we are concerned, however, $f_{\boldsymbol{\theta}}(y)$ must be computed from a conditional density $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ through

$$f_{\boldsymbol{\theta}}(y) = \int f_{\boldsymbol{\theta}}(y|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}. \tag{3.3.7}$$

When $f_{\boldsymbol{X}}(\boldsymbol{x})$ is unknown, or the integration is impractical due to the high dimension of the covariate vector $\boldsymbol{x}$, the objective function (3.3.6) is unavailable. Lu et al. (2003) used a consistent estimator $f_{\boldsymbol{\theta},n}(y)$ to replace $f_{\boldsymbol{\theta}}(y)$, which is defined by

$$f_{\boldsymbol{\theta},n}(y) = \frac{1}{n}\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i, a_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{c}\sum_{i=1}^{n} \frac{p_{ij}}{n} G(a_i, y, \boldsymbol{\alpha}_j' \boldsymbol{x}_i^{(r)}). \tag{3.3.8}$$

Then Lu et al.'s MHD estimator is defined as follows:

$$T(f_n) = \arg\min_{\boldsymbol{\theta}\in\Theta} H^2(f_{\boldsymbol{\theta},n}, f_n);$$

i.e., it is the maximizer of

$$\rho(\boldsymbol{\theta}) = \sum_{y\in U} f_n^{1/2} f_{\boldsymbol{\theta},n}^{1/2}(y).$$

While $f_{\boldsymbol{\theta},n}$ defined above is a consistent estimator of $f_{\boldsymbol{\theta}}$, it may not be the best estimator of (3.3.7). Other classes of estimators can be considered which are not only consistent but also more efficient than $f_{\boldsymbol{\theta},n}$. For example, we can use Monte Carlo integration to estimate (3.3.7). Generate $\boldsymbol{x}_j, j = 1, \ldots, B$ from $\hat{f}(\boldsymbol{x})$, where B is a large number, say 1000, and $\hat{f}(\boldsymbol{x})$ is a kernel density estimate of $\boldsymbol{x}$. This approach has been termed as the smoothed bootstrap (e.g., Chernick, 1999; Efron, 1982). Let

$$\hat{f}_{\boldsymbol{\theta}}(y) = \frac{1}{B}\sum_{j=1}^{B} f_{\boldsymbol{\theta}}(y|\boldsymbol{x}_j), \tag{3.3.9}$$

which is used as an estimator of $f_{\boldsymbol{\theta}}(y)$. Alternatively, we can use bootstrap resampling to estimate (3.3.7). Resample from the empirical distribution of the $\boldsymbol{x}_i's$, say $\hat{F}$. Suppose $(\boldsymbol{x}^{\star}_{j1}, \ldots, \boldsymbol{x}^{\star}_{jn})$ is a bootstrap sample, where $j = 1, \ldots, B$. Then let

$$\tilde{f}_{\boldsymbol{\theta}}(y) = \frac{1}{B}\sum_{j=1}^{B}\frac{1}{n}\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(y|\boldsymbol{x}^{\star}_{ji}). \tag{3.3.10}$$

Accordingly, we can adjust the MHD estimator by replacing $f_{\boldsymbol{\theta},n}$ by either (3.3.9) or (3.3.10). Such improvements on the basic marginal MHD estimation approach will be pursued elsewhere. In the current paper, we focus our attention here and present the algorithm below in terms of the estimator defined by (3.3.8). The algorithm can be altered to utilize the other estimators proposed above in an obvious way.

3.3.2 MHD Estimating Equations

Now we establish the estimating equations. The MHD estimator of $\boldsymbol{\theta}$ is $\arg\max_{\boldsymbol{\theta}\in\Theta}\rho_n(\boldsymbol{\theta})$, where $\rho_n(\boldsymbol{\theta})$ is

$$\begin{aligned}\rho_n(\boldsymbol{\theta}) &= \sum_{y\in U} f_n^{1/2}(y)\Big[\frac{1}{n}\sum_{i=1}^{n}\{\sum_{j=1}^{c-1}\frac{\exp(\boldsymbol{\beta}_j'\boldsymbol{x}_i^{(m)})}{1+\sum_{k=1}^{c-1}\exp(\boldsymbol{\beta}_k'\boldsymbol{x}_i^{(m)})}G(a_i, y, \boldsymbol{\alpha}_j'\boldsymbol{x}_i^{(r)}) \\ &+ \frac{G(a_i, y, \boldsymbol{\alpha}_c'\boldsymbol{x}_i^{(r)})}{1+\sum_{k=1}^{c-1}\exp(\boldsymbol{\beta}_k'\boldsymbol{x}_i^{(m)})}\}\Big]^{1/2}.\end{aligned}$$

Taking the partial derivatives with respective to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and setting them equal to $\mathbf{0}$, we get the following set of equations:

$$\begin{aligned}2\frac{\rho_n(\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}_j} &= \sum_{y\in U} f_n^{1/2}(y) f_{\boldsymbol{\theta},n}^{1/2}(y)\frac{1}{f_{\boldsymbol{\theta},n}(y)}\times\{\sum_{i=1}^{n}\frac{p_{ij}}{n}G(a_i, y, \boldsymbol{\alpha}_j'\boldsymbol{x}_i^{(r)})\frac{y}{a_i}\boldsymbol{x}_i^{(r)} \\ &\quad -\sum_{i=1}^{n}\frac{p_{ij}}{n}G(a_i, y, \boldsymbol{\alpha}_j'\boldsymbol{x}_i^{(r)})\frac{E_{j|i}}{a_i}\boldsymbol{x}_i^{(r)}\} = \mathbf{0},\end{aligned}$$

where $E_{j|i}$ is the mean of jth component given the covariates $\boldsymbol{x}^{(r)}$, and

$$\begin{aligned}2\frac{\rho_n(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_k} &= \sum_{y\in U} f_n^{1/2} f_{\boldsymbol{\theta},n}^{1/2}\frac{1}{f_{\boldsymbol{\theta},n}(y)}\Big[\frac{1}{n}\sum_{i=1}^{n}\{p_{ik}(1-p_{ik})\boldsymbol{x}_i^{(m)}G(a_i, y, \boldsymbol{\alpha}_k'\boldsymbol{x}_i^{(r)}) \\ &\quad -p_{ik}\sum_{j=1,\neq k} p_{ij}\boldsymbol{x}_i^{(m)}G(a_i, y, \boldsymbol{\alpha}_j'\boldsymbol{x}_i^{(r)})\}\Big] = \mathbf{0},\end{aligned} \tag{3.3.11}$$

where $k = 1, \ldots, c-1$ and p_{ij} is defined by (3.2.1). To simplify the above two equations, define the MHD weight for y as

$$v_{\boldsymbol{\theta},n}(y) = f_n^{1/2}(y) f_{\boldsymbol{\theta},n}^{1/2}(y)/\rho_n(\boldsymbol{\theta}).$$

Note that, as $n \to \infty$, the MHD weight converges in probability to the probability of observing y. Also let

$$w_{ij}(a_i, y, \boldsymbol{\theta}) = \frac{p_{ij}}{n} G(a_i, y, \boldsymbol{\alpha}_j' \boldsymbol{x}_i^{(r)})/f_{\boldsymbol{\theta},n}(y).$$

Note also that this quantity converges in probability to the probability that the ith observation comes from the jth component given the value of y under the fitting model. Then the first partial derivative equation (3.3.11) can be written as follows:

$$\sum_{y \in U} v_{\boldsymbol{\theta},n}(y) \{ \sum_{i=1}^{n} w_{ij} (\frac{y}{a_i} - \frac{E_{j|i}}{a_i}) \boldsymbol{x}_i^{(r)} \} = 0. \tag{3.3.12}$$

Recall that $G(a_i, y, \boldsymbol{\alpha}_j' \boldsymbol{x}_i^{(r)})$ has the form of (2.1.2), so $E_{j|i} = \dot{b}(\boldsymbol{\alpha}_j' \boldsymbol{x}_i^{(r)})$. We can solve this equation by using iteratively re-weighted estimating equations as follows. That is, when $\|\boldsymbol{\alpha}^{(l+1)} - \boldsymbol{\alpha}^{(l)}\|$ is small, we have,

$$\begin{aligned} \sum_{i=1}^{n} \sum_{y \in U} v_{\boldsymbol{\theta}^{(l)},n}(y) w_{ij}^{(l)} \frac{y}{a_i} \boldsymbol{x}_i^{(r)} \quad &\simeq \quad \sum_{i=1}^{n} \sum_{y \in U} v_{\boldsymbol{\theta}^{(l)},n} w_{ij}^{(l)} \frac{E_{j|i}^{(l)}}{a_i} \boldsymbol{x}_i^{(r)} \\ &\quad + \sum_{i=1}^{n} \sum_{y \in U} v_{\boldsymbol{\theta}^{(l)},n}(y) w_{ij}^{(l)} \frac{Var_{j|i}^{(l)}}{a_i^2} \boldsymbol{x}_i^{(r)} (\boldsymbol{\alpha}_j^{(l+1)} - \boldsymbol{\alpha}_j^{(l)}), \end{aligned}$$

where $\frac{Var_{j|i}}{a_i} = \ddot{b}(\boldsymbol{\alpha}_j' \boldsymbol{x}_i^{(r)})$. This leads to the updating formula

$$\boldsymbol{\alpha}_j^{(l+1)} = \boldsymbol{\alpha}_j^{(l)} + (\boldsymbol{X}^{(r)\prime} \boldsymbol{W}_j^{(l)} \boldsymbol{X}^{(r)})^{-1} \boldsymbol{X}^{(r)\prime} \boldsymbol{W}_j^{(l)} \boldsymbol{D}_j^{(l)},$$

where $\boldsymbol{W}_j^{(l)}$ is a diagonal matrix with elements $\{\boldsymbol{W}_j^{(l)}\}_{(i,i)} = \sum_{y \in U} v_{\boldsymbol{\theta}^{(l)},n} w_{ij}^{(l)} \frac{Var_{j|i}^{(l)}}{a_i^2}$, and

$$\{\boldsymbol{D}_j^{(l)}\}_{(i)} = \frac{\sum_{y \in U} \frac{v_{\boldsymbol{\theta}^{(l)},n} w_{ij}^{(l)}(y)}{\sum_{y \in U} v_{\boldsymbol{\theta}^{(l)},n} w_{ij}^{(l)}(y)} \frac{y}{a_i} - \frac{E_{j|i}^{(l)}}{a_i}}{\frac{Var_{j|i}^{(l)}}{a_i^2}} \quad j = 1, \ldots, c.$$

Here a_i is the scale parameter in the GLMs, for example $a_i = 1$ for Poisson distribution and $a_i = \frac{1}{m}$ for the binomial(m, p). Note that the coefficient in $\boldsymbol{W}_j^{(i,i)}$, $\sum_{y\in U} v_{\boldsymbol{\theta},n}(y)w_{ij}(y)$, converge in probability to the probability of the value of the ith random variable drawn from the jth component as n goes to infinity. Similarly, the second partial derivative equation (3.3.11) would be written as

$$\sum_{y\in U} v_{\boldsymbol{\theta},n}(y)\{\sum_{i=1}^{n}(w_{ik}(1-p_{ik})\boldsymbol{x}_i^{(m)} - p_{ik}\boldsymbol{x}_i^{(m)}\sum_{j=1,\neq k}^{c} w_{ij})\} = 0 \quad k = 1,\ldots,c-1.$$

When $\|\boldsymbol{\beta}_k^{(l+1)} - \boldsymbol{\beta}_k^{(l)}\|$ is small , we can obtain:

$$\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)\sum_{i=1}^{n} w_{ik}^{(l)}\boldsymbol{x}_i^{(m)} \simeq \sum_{i=1}^{n}\sum_{j=1}^{c}\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)w_{ij}^{(l)}\{\text{expit}(\boldsymbol{\beta}_k^{(l)\prime}\boldsymbol{x}_i^{(m)})\} +$$

$$\sum_{i=1}^{n}\sum_{j=1}^{c}\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)w_{ij}^{(l)}\text{expit}(\boldsymbol{\beta}_k^{(l)\prime}\boldsymbol{x}_i^{(m)})\{1-\text{expit}(\boldsymbol{\beta}_k^{(l)\prime}\boldsymbol{x}_i^{(m)})\}\boldsymbol{x}_i^{(m)}\boldsymbol{x}_i^{(m)\prime}(\boldsymbol{\beta}_k^{(l+1)} - \boldsymbol{\beta}_k^{(l)}),$$

where $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$. Then we can update $\boldsymbol{\beta}$ through

$$\boldsymbol{\beta}_k^{(l+1)} = \boldsymbol{\beta}_k^{(l)} + (\boldsymbol{X}^{(m)\prime}\boldsymbol{W}_k^{(l)}\boldsymbol{X}^{(m)})^{-1}\boldsymbol{X}^{(m)\prime}\boldsymbol{W}_k^{(l)}\boldsymbol{S}_k^{(l)},$$

where

$$\{\boldsymbol{W}_k^{(l)}\}_{(i,i)} = \sum_{j=1}^{c}\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)w_{ij}^{(l)}Var^{(l)}(Z_{ik}),$$

and

$$\{\boldsymbol{S}_k^{(l)}\}_{(i)} = \frac{\frac{\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)w_{ik}^{(l)}}{\sum_{j=1}^{c}\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)w_{ij}^{(l)}} - E^{(l)}(Z_{ik})}{Var^{(l)}(Z_{ik})}, \quad k = 1,\ldots,c-1.$$

Recall that Z_{ik} is the unobserved random variable indicating whether the i^{th} observation comes from the k^{th} component. The coefficient in $\boldsymbol{W}_k^{(l)}$, $\sum_{j=1}^{c}\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)w_{ij}^{(l)}$, converges in probability to the probability of observing the ith observation, i.e. to 1, as n $\to \infty$.

3.3.3 MHD Fitting Algorithm

The MHD fitting algorithm can be summarized in terms of the following four steps:

Step (1) : Given $\boldsymbol{\theta}^{(l)} = \{\boldsymbol{\alpha}^{(l)}, \boldsymbol{\beta}^{(l)}\}$, calculate

$$v^{(l)}_{\boldsymbol{\theta},n}(y) = f^{1/2}_n(y) f^{1/2}_{\boldsymbol{\theta}^{(l)},n}(y)/\rho_n(\boldsymbol{\theta}^{(l)}),$$

and

$$w^{(l)}_{ij} = \frac{p^{(l)}_{ij}}{n} G(a_i, y, \boldsymbol{\alpha}^{(l)\prime}_j \boldsymbol{x}^{(r)}_i)/f^{(l)}_{\boldsymbol{\theta},n}(y), \quad y \in U,$$

where

$$p^{(l)}_{ij} = \frac{\exp(\boldsymbol{\beta}^{(l)\prime}_j \boldsymbol{x}^{(m)}_i)}{1 + \sum_{k=1}^{c-1} \exp(\boldsymbol{\beta}^{(l)\prime}_k \boldsymbol{x}^{(m)}_i)}, \quad \rho_n(\boldsymbol{\theta}^{(l)}) = \sum_{y \in U} f^{1/2}_n(y) f^{1/2}_{\boldsymbol{\theta}^{(l)},n}(y).$$

Step (2): Update $\boldsymbol{\alpha}$ by

$$\boldsymbol{\alpha}^{(l+1)}_j = \boldsymbol{\alpha}^{(l)}_j + (\boldsymbol{X}^{(r)\prime} \boldsymbol{W}^{(l)}_j \boldsymbol{X}^{(r)})^{(-1)} \boldsymbol{X}^{(r)} \boldsymbol{W}^{(l)}_j \boldsymbol{D}^{(l)}_j,$$

where $\boldsymbol{W}^{(l)}_j$ is a diagonal matrix with components:

$$\{\boldsymbol{W}^{(l)}_j\}_{(i,i)} = \sum_{y \in U} v^{(l)}_{\boldsymbol{\theta},n}(y) w^{(l)}_{ij}(y) Var^{(l)}_{j|i}/a^2_i,$$

and

$$\{\boldsymbol{D}^{(l)}_j\}_{(i)} = \frac{\sum_{y \in U} \left(\frac{v^{(l)}_{\boldsymbol{\theta},n}(y) w^{(l)}_{ij}(y)}{\sum_{y \in U} v^{(l)}_{\boldsymbol{\theta},n}(y) w^{(l)}_{ij}(y)} \frac{y}{a_i} \right) - \frac{E^{(l)}_{j|i}}{a_i}}{Var^{(l)}_{j|i}/a^2_i}, \quad j = 1, \ldots, c.$$

Step (3): Update $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{(l+1)}_k = \boldsymbol{\beta}^{(l)}_k + (\boldsymbol{X}^{(m)\prime} \boldsymbol{W}^{(l)}_k \boldsymbol{X}^{(m)})^{(-1)} \boldsymbol{X}^{(m)\prime} \boldsymbol{W}^{(l)}_k \boldsymbol{S}^{(l)}_k,$$

where

$$\{\boldsymbol{W}^{(l)}_k\}_{(i,i)} = \sum_{j=1}^{c} \sum_{y \in U} v^{(l)}_{\boldsymbol{\theta},n}(y) w^{(l)}_{ij}(y) Var^{(l)}(Z_{ik}),$$

and

$$\{\boldsymbol{S}^{(l)}\}_{(i)} = \frac{\frac{\sum_{y \in U} v^{(l)}_{\boldsymbol{\theta},n}(y) w^{(l)}_{ik}(y)}{\sum_{j=1}^{c} \sum_{y \in U} v^{(l)}_{\boldsymbol{\theta},n}(y) w^{(l)}_{ij}(y)} - E^{(l)}(Z_{ik})}{Var^{(l)}(Z_{ik})}, \quad k = 1, \ldots, c-1.$$

Step (4): If a convergence criterion has been obtained then stop, otherwise return to step (1). Any of several convergence criteria could be used here. For example, if

$\|\boldsymbol{\alpha}_j^{(l+1)} - \boldsymbol{\alpha}_j^{(l)}\| \leq \epsilon$ and $\|\boldsymbol{\beta}_k^{(l+1)} - \boldsymbol{\beta}_k^{(l)}\| \leq \epsilon$ then stop, where ϵ is some suitably small constant (e.g., 1e-6).

The above algorithm is based on the canonical link of the GLMs. If $\eta = \eta(\mu)$ is not the canonical link function, we can modify step (2) of the above algorithm as follows:

$$\boldsymbol{\alpha}_j^{(l+1)} = \boldsymbol{\alpha}_j^{(l)} + (\boldsymbol{X}^{(r)'}\boldsymbol{W}_j^{(l)}\boldsymbol{X}^{(r)})^{(-1)}\boldsymbol{X}^{(r)}\boldsymbol{W}_j^{(l)}\boldsymbol{D}_j^{(l)},$$

where $\boldsymbol{W}_j^{(l)}$ is a diagonal matrix with components:

$$\{\boldsymbol{W}_j^{(l)}\}_{(i,i)} = \sum_{y\in U} v_{\boldsymbol{\theta},n}^{(l)}(y) w_{ij}^{(l)}(y) \times \frac{1}{Var_{j|i}^{(l)}(y)}(\frac{\partial \mu_i}{\partial \eta_i})^2,$$

and

$$\{\boldsymbol{D}_j^{(l)}\}_{(i)} = \Big[\sum_{y\in U}\{\frac{v_{\boldsymbol{\theta},n}^{(l)}(y) w_{ij}^{(l)}(y)}{\sum_{y\in U} v_{\boldsymbol{\theta},n}^{(l)}(y) w_{ij}^{(l)}(y)}\frac{y}{a_i}\} - \frac{E_{j|i}^{(l)}}{a_i}\Big] \times \frac{\partial \eta_i}{\partial \mu_i}, \quad j = 1, \ldots, c.$$

It is instructive to compare the algorithm for minimizing the MHD criterion to ML estimation via the EM algorithm. For FMGLMs, the latter approach leads to the following E and M steps.

E step: Estimate p_{ij} by its conditional mean $p_{ij}^{(l)} = E(Z_{ij}|y_i, \boldsymbol{\beta}^{(l)}, \boldsymbol{\alpha}^{(l)})$ under current estimates of the regression parameters.

M step for $\boldsymbol{\alpha}$: Update $\hat{\boldsymbol{\alpha}}$ via

$$\boldsymbol{\alpha}_j^{(l+1)} = \boldsymbol{\alpha}_j^{(l)} + (\boldsymbol{X}^{(r)'}\boldsymbol{W}_j^{(l)}\boldsymbol{X}^{(r)})^{(-1)}\boldsymbol{X}^{(r)'}\boldsymbol{W}_j^{(l)}\boldsymbol{D}_j^{(l)},$$

where $\{\boldsymbol{W}_j^{(l)}\}_{(i,i)} = p_{ij}^{(l)}\frac{Var_{j|i}^{(l)}(Y)}{a_i^2}$, and $\{\boldsymbol{D}_j^{(l)}\}_{(i)} = \frac{y_i/a_i - E_{j|i}^{(l)}(Y)/a_i}{Var_{j|i}^{(l)}(Y)/a_i^2}$.

M step for $\boldsymbol{\beta}$: Update $\hat{\boldsymbol{\beta}}$ via

$$\boldsymbol{\beta}_k^{(l+1)} = \boldsymbol{\beta}_k^{(l)} + (\boldsymbol{X}^{(m)'}\boldsymbol{W}_k^{(l)}\boldsymbol{X}^{(m)})^{(-1)}\boldsymbol{X}^{(m)'}\boldsymbol{W}_k^{(l)}\boldsymbol{S}_k^{(l)},$$

where $\{\boldsymbol{W}_k^{(l)}\}_{(i,i)} = Var_k^{(l)}(Z_{ik})$, and $\{\boldsymbol{S}_k^{(l)}\}_{(i)} = \frac{p_{ik}^{(l)} - p_{ik}}{Var_k^{(l)}(Z_{ik})}$. Comparing the EM updating formulas with the corresponding ones for MHD estimation we see some striking similarities. Both approaches involve iteratively updating $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ using

weights that are recomputed at each iteration. These weights involve the mixing probabilities in the case of ML, and the weight $\sum_{y\in U} v_{\boldsymbol{\theta},n}(y)w_{ij}$ in the case of MHD estimation. In each case, though, the weight quantifies the probability that the value of the random variable Y_i is drawn from the jth component. However, MHD estimation assigns less weight to those points that are outlying in the sense of large Lindsay's residuls. It also sets the weights $\sum_{j=1}^{c}\sum_{y\in U} v_{\boldsymbol{\theta},n}(y)w_{ij}(y)$ instead of 1 to the mixture component indicator (missing data) Z_{ij}.

3.3.4 MHDE asymptotic properties

Theorem Under the assumption of Theorem 1 in Beran (1977), Θ is a compact subset, $\theta_1 \neq \theta_2 \Rightarrow f_{\theta_1} \neq f_{\theta_2}$ on a set of positive Lebesgue measure,$f_\theta(y)$ is continuous in θ, and $\sum_{y=0}^{\infty} f_\theta^{1/2}(y) < \infty$, we have

$$\tilde{T}_n \overrightarrow{p} \quad \theta,$$

where $\tilde{T}_n$ satisfies $\tilde{H}(\tilde{T}_n, f_n) = \mathrm{Min}_t \tilde{H}(t, f_n)$,

$$\tilde{H}(t, f_n) = \sum_{y=0}^{\infty}(f_n^{1/2}(y) - f_{n,t}^{1/2}(y))^2 = \|f_n^{1/2} - f_{n,t}^{1/2}\|,$$

f_n is the empirical density of $(Y_1, \ldots, Y_n)$ and $f_{n,t} = \frac{1}{n}\sum_{i=1}^{n} f_t(y|X_i)$.

Proof: (1) Show that $\mathrm{E}_X\|f_{n,\tilde{T}_n}^{1/2} - f_\theta^{1/2}\| \to 0$.

$$\begin{aligned} E_X\|f_{n,\tilde{T}_n}^{1/2} - f_\theta^{1/2}\| &\leq E_X(2\|f_{n,\tilde{T}_n}^{1/2} - f_n^{1/2}\| + 2\|f_n^{1/2} - f_\theta^{1/2}\|) \\ &\leq E_X(2\|f_{n,\theta}^{1/2} - f_n^{1/2}\| + 2\|f_n^{1/2} - f_\theta^{1/2}\|) \\ &\leq E_X(4\|f_{n,\theta}^{1/2} - f_\theta^{1/2}\| + 4\|f_\theta^{1/2} - f_n^{1/2}\| + 2\|f_n^{1/2} - f_\theta^{1/2}\|) \end{aligned}$$

$$\begin{aligned} E_X\|f_\theta^{1/2} - f_n^{1/2}\| &= \sum_{y=0}^{\infty} E_X(f_\theta^{1/2} - f_n^{1/2})^2 \\ &\leq \sum_{y=0}^{\infty} E_X|f_\theta - f_n| \end{aligned}$$

$$\begin{aligned} &\leq \sum_{y=0}^{\infty}\{E_X(f_\theta - f_n)^2\}^{1/2} \\ &\leq \sum_{y=0}^{\infty}\{\frac{1}{n}f_\theta(1-f_\theta)\}^{1/2} \\ &\leq \frac{1}{\sqrt{n}}\sum_{y=0}^{\infty} f_\theta^{1/2} \to 0. \end{aligned}$$

$$\begin{aligned} E_X\|f_{n,\theta}^{1/2} - f_\theta^{1/2}\| &\leq \sum_{y=0}^{\infty} E_X[\{\frac{1}{n}\sum f_\theta(y|X_i)\}^{1/2} - f_\theta^{1/2}]^2 \\ &\leq \sum_{y=0}^{\infty}[E_X\{\frac{1}{n}\sum f_\theta(y|X_i) - f_\theta\}^2]^{1/2} \\ &= \sum_{y=0}^{\infty}\frac{1}{\sqrt{n}}[E_X\{f_\theta(y|X_1) - f_\theta\}^2]^{1/2} \\ &\leq \frac{1}{\sqrt{n}}\sum_{y=0}^{\infty}(E_X f_\theta^2(y|X_1))^{1/2} \\ &\leq \frac{1}{\sqrt{n}}\sum_{y=0}^{\infty}(E_X f_\theta(y|X_1))^{1/2} \\ &= \frac{1}{\sqrt{n}}\sum_{y=0}^{\infty} f_\theta^{1/2} \to 0. \end{aligned}$$

(2) $\tilde{T}_n \to \theta$.

By Markov Inequality,

$$\|f_{n,\tilde{T}_n}^{1/2} - f_\theta^{1/2}\| \to 0, \text{i.p.}$$

So that $\Longrightarrow \forall N_1 \subset N, \exists N_2 \subset N$, s.t.

$$\|f_{n,\tilde{T}_n}^{1/2} - f_\theta^{1/2}\| \to 0, \text{a.s. } n \subset N_2.$$

By the Theorem 1 of Beran (1977), $\tilde{T}_n \to \theta$ a.s , $n \in N_2$. By Lemma2 of Chow and Teicher (1997), $\tilde{T}_n \to \theta$ in probability.¶

Lu et al (2003) proposed to extend Simpson's result (1987) to the finite mixture of Poisson cases, i.e., under certain regularity conditions, $\hat{\boldsymbol{\theta}}^{MHD}$ has an asymptotic normal distribution. Moreover,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{MHD} - \boldsymbol{\theta}) \to N(\mathbf{0}, V_{\boldsymbol{\theta}})$$

in law as $n \to \infty$, where $V_{\boldsymbol{\theta}} = \frac{1}{4}\ddot{H}(\boldsymbol{\theta}; F)^{-1} i(\boldsymbol{\theta}) \ddot{H}(\boldsymbol{\theta}; F)^{-1}$, and F is the true underlying distribution. If $F \equiv F_{\boldsymbol{\theta}}$, then $V_{\boldsymbol{\theta}} = i(\boldsymbol{\theta})^{-1}$. The asymptotic variance of the MHDE can be estimated by $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}} = \hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})^{-1}$, where $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) = \sum_{y \in U} \left\{ \hat{\dot{l}}_{\boldsymbol{\theta}}(y) \hat{\dot{l}}^{T}_{\boldsymbol{\theta}}(y) f_n(y) - \frac{\partial^2 f_{\boldsymbol{\theta},n}(y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}$, and $\hat{\dot{l}}_{\boldsymbol{\theta}}(y) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta},n}(y) |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

This marginal approach could be extended beyond discrete FMGLMs, but this would require replacing $f_n(y)$ by a kernel density estimate which would defeat the simplicity of the marginal approach (the simplicity of being able to look at the empirical frequency function as the appropriate nonparametric density estimate). So, if it is necessary to introduce a kernel density estimate, then we might as well work with a nonparametric estimate for the conditional density function, then proceed to the MCHDE section and present it as applicable to the full class (discrete and continuous) of FMGLMs. We think that the more natural approach in a regression context such as ours is to work with the conditional Hellinger distance criterion.

Moreover, Lu et al.'s approach complicates the issue of model identifiability (Lu et al. 2003). Although these authors cite previous literature on the identifiability of finite mixtures of Poisson models to support their approach, these results apply only to the situation in which the mixing probability is constant. In many regression contexts in which FMGLMs are natural models, this assumption is restrictive. FMGLMs are most natural as a tool to account for underlying population heterogeneity or latent class structure. In many contexts, the available data may include covariates on which the class membership probability depends. For example, in section 6 we analyze heart arhythmia data from a sample of clinically normal dogs. The motivation for a Poisson FMGLM in that problem is the hypothesis that this sample of "normal" animals is contaminated with a subset of misdiagnosed dogs that have a genetic defect for cardiomyopathy. The mixing probability here represents the probability that a particular animal has the genetic defect and this probability may depend upon other clinical factors available to the analyst, such as gender. In this

and many other natural settings for FMGLMs, it is appealing to specify a regression structure for the mixing probability in the model. Unfortunately, with the marginal MHD approach identifiability is not guaranteed in such models, and in our simulation studies and other experiences, we have frequently encountered identifiability problems when the mixing probability is not constant. Fortunately, a conditional MHD approach avoids these identifiability problems, is more appealing in a regression context, is more widely applicable to non-discrete FMGLMS, and, as we will see, often offers greater efficiency than the marginal MHD approach. Therefore, we propose to minimize the conditional Hellinger distance as the basis of estimation.

3.4 Minimum Conditional Hellinger Distance Estimation for FMGLMs

Several researchers have proposed non-parametric estimators of conditional densities (Fan et al. 1996; Hall et al. 1999; Hyndman et al. 2002). These papers have dealt with the general conditional density estimation problem, but Gooijer and Zerom have recently applied these approaches to a mixture context. For certain cases of FMGLMs, if suitable nonparametric conditional density estimates can be found, it is natural to use MCHD estimation. We expect that the MCHDE should be more efficient than the MHDE, since we don't need to estimate the marginal densities. Instead, we estimate the nonparametric conditional density directly. Moreover, we avoid the identifiability problems that may arise in the marginal MHD approach.

If we know the conditional density $f_n(y|\boldsymbol{x})$, we can minimize (3.2.4) directly. Clearly, it's equivalent to maximize

$$\rho_n^c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{n} \sum_{y \in U} f_n^{1/2}(y|\boldsymbol{x}_i) f_{\boldsymbol{\theta}}^{1/2}(y|\boldsymbol{x}_i).$$

Then the MCHD estimator is the solution of

$$\dot{\rho}_n^c(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{y \in U} f_n^{1/2}(y|\boldsymbol{x}_i) f_{\boldsymbol{\theta}}^{-1/2} \dot{f}_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i) = 0.$$

To calculate the MHD estimator of $\boldsymbol{\theta}$, we can adjust the algorithm of section 3.3, by redefining $v_{\boldsymbol{\theta},n}(y)w_{ij}(a_i, y, \boldsymbol{\theta})$ in **step (1)** as follows:

$$v_{\boldsymbol{\theta},n}(y|\boldsymbol{x}_i) = f_n^{1/2}(y|\boldsymbol{x}_i)f_{\boldsymbol{\theta}}^{1/2}(y|\boldsymbol{x}_i)/\rho_n^c(\boldsymbol{\theta}),$$

and

$$w_{ij}(a_i, y, \boldsymbol{\theta}|\boldsymbol{x}_i) = \frac{\frac{p_{ij}}{n}G(a_i, y, \boldsymbol{\alpha}_j'\boldsymbol{x}_i^{(r)})}{f_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)}.$$

Simulation results in Section 4 show that for FMGLMs, the MCHD estimator is more efficient than the MHD estimator of the regression parameters, especially for the parameters related to the component means. In the following, we will demonstrate how to estimate $f_n(y|\boldsymbol{x})$ in several cases.

Case 1: Categorical Covariates

For a model with categorical explanatory variables, we can estimate the nonparametric density $f_n(y|\boldsymbol{x}_i)$ in each covariate class separately by $N_{y|\boldsymbol{x}_i}/N(\boldsymbol{x}_i)$, where $N(\boldsymbol{x}_i)$ is the frequency of $\boldsymbol{x}_i$ among $\boldsymbol{x}$, and $N_{y|\boldsymbol{x}_i}$ is the frequency of y among $y_1, \ldots, y_n$ with covariates $\boldsymbol{x}_i$.

Case 2: A Single Continuous Covariate

For models with a single continuous covariate, we use the nonparametric conditional density estimation proposed by Hyndman et al. (2003) as our nonparametric conditional density estimate in (3.2.4). In their papers, they suggested that $f_n(y|x_i)$ can be estimated through a local regression function of $K_b(Y_i - y)$ on x_i, where $K_b(u) = b^{-1}K(u/b)$ with $K(.)$ being a symmetric density function on R. They proposed non-negative estimators as follows. Let

$$R(\boldsymbol{\theta}; x, y) = \sum_{i=1}^{n}\{K_b(Y_i - y) - A(X_i - x, \boldsymbol{\theta})\}^2 W_h(X_i - x), \tag{3.4.13}$$

where $A(x, \boldsymbol{\theta}) = \exp(\sum_{j=0}^{r} \boldsymbol{\theta}_j x^j)$, and $W_h(u) = h^{-1}W(u/h)$, $W(.)$ is a kernel function. Then

$$\hat{f}(y|x) = \exp(\hat{\boldsymbol{\theta}}_0),$$

where $\hat{\boldsymbol{\theta}}_{xy} = (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_r)^T$ minimizes $R(\boldsymbol{\theta}; x, y)$. Clearly, the performance of MCHDE depends on the choice of bandwidths. We apply the bandwidth selection algorithm proposed by Hyndman et al. (2003) to choose the optimal bandwidths of b and h. This algorithm combines two steps.

(1) Select the smoothing parameter b using the normal reference rule. That is, assume both the conditional distribution and the marginal distribution are normally distributed with $f(y|x) = \frac{1}{\sigma}\phi(\frac{y-d_0-d_1(x-\mu)}{\sigma})$ and $g(x) = \frac{1}{\nu}\phi(\frac{x-\mu}{\nu})$ where $\phi(x) = \frac{1}{2}\exp(-\frac{x^2}{2})$, we can get $\hat{h} \approx 0.916(\nu\sigma^5/n|d_1|^5)^{1/6}$ and $\hat{b} = 1.05|d_1|\hat{h}$.

(2) Given this value of $\hat{b}$, find value of h is a standard nonparametric problem of regression $K_b(Y_i - y)$ on X_i. Therefor, the cross-validation technique (Fan and Gijbels, 1996 p.45) can be adapted to update the bandwidth $\hat{h}$.

In practice, we can use Hyndman's R library *Hdrcde* which is already built into this algorithm, to form the estimator of $f_n(y|x_i)$, as well as the optimal bandwidth $\hat{b}$ and $\hat{h}$. The MCHD estimator is then computed following the same steps as given above for the marginal case.

In theory, for the general case, we obtain get the conditional density as above no matter what the dimension of the covariate vector is. Replacing scalars X_i, x by the vectors $\boldsymbol{X}_i, \boldsymbol{x}$ in (3.4.13), the estimator of $f(y|\boldsymbol{x})$ would be $\exp(\hat{\boldsymbol{\theta}}_0)$, where $\hat{\boldsymbol{\theta}}_{\boldsymbol{x}y} = (\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_r)$ minimizes $R(\boldsymbol{\theta}; \boldsymbol{x}, y)$, where $A(\boldsymbol{x}, \boldsymbol{\theta}) = \exp(\sum_{j=0}^{r} \boldsymbol{\theta}_j^T \boldsymbol{x}^j)$. However due to the curse of dimensionality, this estimator rapidly loses efficiency as the dimension increases. In practice, this approach to estimation is appropriate only for small dimensions.

Case 3: High dimension continuous covariates

Recently, Hall et al. (2005) proposed a method for approximating the conditional distribution function of a random variable Y given a dependent random d-dimensional vector $\boldsymbol{x}$ using a dimension reduction technique. Instead of estimating the distribution of $Y|\boldsymbol{x}$, they proposed to estimate $Y|\boldsymbol{\theta}^T\boldsymbol{x}$, where the vector $\boldsymbol{\theta}$ is selected so that the estimation is optimal under a least-squares criterion. More specifically, first they estimate $\hat{\boldsymbol{\theta}}$ by using the "leave two out" technique. Let

$$T^{[k]}_{-i,-j}(\boldsymbol{\theta}) = \frac{1}{(n-2)h}\sum_{i_1:i_1\neq i,j} K\{\frac{\boldsymbol{\theta}^T(\boldsymbol{x}_i-\boldsymbol{x}_{i_1})}{h}\}\{\frac{\boldsymbol{\theta}^T(\boldsymbol{x}_i-\boldsymbol{x}_{i_1})}{h}\}^k,$$

$$w_{i_1,-i,-j}(\boldsymbol{\theta}) = K\{\frac{\boldsymbol{\theta}^T(\boldsymbol{x}_i-\boldsymbol{x}_{i_1})}{h}\}\times\{T^{[2]}_{-i,-j}(\boldsymbol{\theta}) - \frac{\boldsymbol{\theta}^T(\boldsymbol{x}_i-\boldsymbol{x}_{i_1})}{h}T^{[1]}_{-i,-j}(\boldsymbol{\theta})\}.$$

$$\hat{F}_{-i,-j}(y|\boldsymbol{\theta}^T\boldsymbol{x}_i) = \{\sum_{i_1:i_1\neq i,j} w_{i_1,-i,-j}(\boldsymbol{\theta})I(Y_{i1}\leq y)\}\times\{\sum_{i_{1:}i_1\leq i,j} w_{i_1,-i,-j}(\boldsymbol{\theta})\}^{-1}.$$

Here $\hat{F}_{-i,-j}(y|\boldsymbol{\theta}^T\boldsymbol{x}_i)$ is a local linear estimator of $F(y|\boldsymbol{\theta}^T\boldsymbol{x}_i)$ based on data pairs other than the ith and jth; and $\frac{1}{n-1}\sum_{i:i\neq j,\boldsymbol{x}_i\in A}\hat{F}_{-i,-j}(y|\boldsymbol{\theta}^T\boldsymbol{x}_i)$ is an estimator of $\pi_{\boldsymbol{\theta}}(A,B)$ when $B=(-\infty,y]$. Let $\hat{F}_{-j}(A,y)$ be the proportion of the $n-1$ values of $(\boldsymbol{x}_i,y_i)$, for $i\neq j$, which satisfy $(\boldsymbol{x}_i,y_i)\in A\times(-\infty,y]$, and let $S(\boldsymbol{\theta},A)=\sum_{j=1}^n\{\hat{F}_{-j}(A,Y_j) - \frac{1}{n-1}\sum_{i:i\neq j,\boldsymbol{x}_i\in A}\hat{F}_{-i,-j}(Y_j|\boldsymbol{\theta}^T\boldsymbol{x}_i)\}^2$. They choose $\hat{\boldsymbol{\theta}}$ to minimize $S(\boldsymbol{\theta})=\int S(\boldsymbol{\theta},A)d\mu(A)$ over $\boldsymbol{\theta}\in\Theta$. They recommended choosing the bandwidth h by bootstrapping based on an approximating parametric model (Hall et al. 1999). The estimator of the conditional distribution function of Y given $\hat{\boldsymbol{\theta}}^T\boldsymbol{x}$ is shown to be first order equivalent to its counterpart when the true value of $\boldsymbol{\theta}$ is known. Therefore we can use the one-dimensional method described above to estimate $f(y|\boldsymbol{\theta}^T\boldsymbol{x})$.

Note that in a model with continuous covariates and additional categorical covariates, we can combine the approaches of case 1 and case 3 to handle a much broader class of models.

In the iid case, Beran (1977) defined the MHDE $T(\hat{g}_n)$ to be the minimizer of $H^2=\int\{\hat{g}_n^{1/2}(y)-f_{\boldsymbol{\theta}}^{1/2}(y)\}^2dy$, where $\hat{g}_n(y)$ is a kernel density estimator. Under

certain regularity condition, if $H^2 \to 0$, the limiting distribution of $\sqrt{n}\{T(\hat{g}_n) - \boldsymbol{\theta}\}$ is $N(0, \{4\int \dot{s}_{\boldsymbol{\theta}}(y)\dot{s}^T_{\boldsymbol{\theta}}(y)dy\}^{-1})$, where $s_{\boldsymbol{\theta}}(y) = f^{1/2}_{\boldsymbol{\theta}}(y)$.

Conjecture This result can be extended to the regression context, i.e., MCHD case. Recall that

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{MCHD} &= \arg\min_{\boldsymbol{\theta}} H_c^2(f_n(.|.), f_{\boldsymbol{\theta}}(.|.)) \\ &\equiv \arg\min_{\boldsymbol{\theta}} \int\int (f_n^{1/2}(y|\boldsymbol{x}) - f_{\boldsymbol{\theta}}^{1/2}(y|\boldsymbol{x}))^2 f_X(\boldsymbol{x})d\boldsymbol{x}dy,\end{aligned}$$

where $f_n(y|\boldsymbol{x})$ is defined depending on the different cases of $\boldsymbol{x}$ as described above. Under certain regularity conditions, the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}^{MCHD} - \boldsymbol{\theta})$ should go to $N(0, \Lambda^{-1}_{\boldsymbol{\theta}})$ as $n \to \infty$, where $\Lambda_{\boldsymbol{\theta}} = 4\int\int \dot{S}_{\boldsymbol{\theta}}(y|\boldsymbol{x})\dot{S}^T_{\boldsymbol{\theta}}(y|\boldsymbol{x})dy f_X(\boldsymbol{x})d\boldsymbol{x}$, and $S_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = f^{1/2}_{\boldsymbol{\theta}}(y|\boldsymbol{x})$. Then, the asymptotic variance of $\hat{\boldsymbol{\theta}}^{MCHD}$ can be estimated by $\hat{\Lambda}_{\hat{\boldsymbol{\theta}}} = \frac{4}{n}\sum_{i=1}^n \int \dot{S}_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_i)\dot{S}^T_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_i)dy$.

A formal proof of this extension is hard to establish, because (1) the property of $\int\int (f_n^{1/2}(y|\boldsymbol{x}) - f_{\boldsymbol{\theta}}^{1/2}(y|\boldsymbol{x}))^2 dy f_X(\boldsymbol{x})d\boldsymbol{x} \to 0$ in probability as $n \to \infty$ is hard to obtain without further assumptions, on the distribution of $\boldsymbol{X}$, and (2) the regularity conditions affected by bandwidths are not clear. However, simulation results in the next section confirm that our conjecture is reasonable, so that the estimated variance of the MCHDE is very close to the sample variance when n is large.

3.5 Simulation Study

The aim of the following set of simulations is to assess the performance of MHD method and MCHD method compared with ML methods for mixture models. Separate studies are conducted to verify and compare the properties of these three methods, under three scenarios: correct model specification, data generated from the model but with outliers in the response variable, and data generated from the model but with outliers in the covariates. Simulation study 1 followed a 2 x 2 design in which we consider two types of finite mixture models (2-component binomial and

ZIP), and two sample sizes (100 and 200). In addition, studies 2 and 3 considered two additional factors: degree of mean separation between the components of the finite mixture (low and high), and level of mixing, corresponding to even (50:50) or skewed (10:90) mixing of the components. For the binomial mixture models, independent data are generated from a model of the form

$$Y_i \sim \begin{cases} \text{Binomial}(m, \mu_{1i}), & \text{with probability } p_i, \\ \text{Binomial}(m, \mu_{2i}), & \text{with probability } 1 - p_i, \end{cases}$$

where

$$\begin{aligned} \mu_{ji} &= \text{expit}(\alpha_{j0} + \alpha_{j1} x_{1i}), \quad j = 1, 2 \\ p_i &= p, \end{aligned}$$

x_{1i} is a random variable generated from a uniform(0,1) distribution. The other type of FMGLMs considered is a ZIP model, in which independent data are generated according to

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

where

$$\begin{aligned} \mu_i &= \exp(\alpha_{20} + \alpha_{21} x_{1i}), \\ p_i &= \text{expit}(\beta_{10} + \beta_{11} x_{1i}), \end{aligned}$$

where again the x_{1i}'s are iid $U(0, 1)$.

Various settings of the true parameters are used which correspond to different levels of component separation. The specific values of these parameters are listed along with the simulation results in Table 3.1-3.18. For each setting, 500 data sets were generated from the model, with or without the addition of outliers depending on the aim of each study. Bias, mean square error (MSE), and size of the Wald test for the hypothesis that the parameter is equal to the corresponding true value are reported. In the case of the binomial FMGLM, the model assumes constant

mixing probability, so all three estimation methods, MHD, MCHD, and ML, are compared with each other. In the ZIP model, we allow the mixing probability to have a regression structure and the resulting model is not identifiable under marginal MHD estimation. Therefore, in this setting we consider only the conditional MHD and ML estimation methods.

3.5.1 Study 1: Finite Sample Bias and Precision for Model Without Contaminations

The aim of study 1 is to explore how the MHD estimator (MHDE) and MCHD estimator (MCHDE) work for data without any contamination compared with ML estimator (MLE) in terms of bias, MSE, size of Wald test as well as the estimate of the asymptotic variance.

Two component binomial model

Firstly, we use the MHD method and MCHD method described above to do some simulation studies involving two component mixtures of binomial distributions with regression structure as an illustration. Here, we let the mixing probability to be a constant 0.5. We generated the mixture binomial data with sample size $n = 100$ and 200, binomial denominator $m = 30$ by choosing $(\alpha_{10}, \alpha_{11}) = (1.0, 0.5)$ and $(\alpha_{20}, \alpha_{21}) = (-1.0, 0.5)$. Bias, MSE, and size of the Wald test for equality with the true value were calculate for each model parameter. In addition, the average estimate of asymptotic variance and finite sample variance (Avar) of the parameter estimates are also listed in Table 3.1.

From the results summarized in Table 3.1, several conclusions can be drawn: (1) MHDE, MCHDE and MLE all have low bias when the model is correctly specified, and this bias decreases as we increase the sample size from 100 to 200. (2) Among these three estimators, the MLE is the best in terms of bias and MSE. However,

the MCHDE is very competitive with the MLE and both of them are significantly better than the MHDE. (3) The estimate of the asymptotic variance under the MCHDE method is very close to the sample variance for all parameters, whereas the estimate under the MHD method is quite different, especially for the parameters corresponding to component means. (4) Both the MCHDE and the MLE have the size of the Wald based test close to the nominal size 0.05 while the MHDE always deflates the Wald test size for the component means.

ZIP model

Here, we compare MCHDE and MLE in the ZIP regression context. The true parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are listed in Table 3.2, which correspond to the true ZIP models with relatively large Poisson component means (i.e., well separated cases). Two sample sizes, $n = 100$ and $n = 200$ are considered. Bias, MSE, and Wald test size are calculated. Estimates of asymptotic variance and finite sample variance are also listed in Table 3.2. Table 3.2 exhibits a similar pattern as Table 3.1. Specifically, when no outliers are present, both estimators exhibit low bias, with greater efficiency seen for the MLE.

3.5.2 Study 2: models with outliers in y

In study 2, we want to explore the performance of MHDE and MCHDE compared with MLE for the case where outliers occur in the response vector.

Two component binomial model

In this section, we will compare MHD, MCHD and ML estimation methods for data generated from a two component binomial model with constant mixing probability p and contamination in y. For each data set generated, 5% of the responses were selected at random and replaced by 30, the denominator of the binomial distribution,

Table 3.1: Two component binomial data with constant $p = 0.5$ without any contamination, n=100 and 200

Parameters	Bias	MSE	Size	Var	Estimate of Avar
$n = 100$					
MHDE					
$p = 0.5$	0.0008	0.0030	0.08	0.0030	0.0025
$\alpha_{10} = 1.0$	0.1069	0.0426	0.02	0.0315	0.0848
$\alpha_{11} = 0.5$	-0.2247	0.1410	0.02	0.0914	0.3242
$\alpha_{20} = -1$	0.1079	0.0321	0.01	0.0207	0.0576
$\alpha_{21} = 0.5$	-0.2317	0.1206	0.00	0.0676	0.2189
MCHDE					
$p = 0.5$	-0.0007	0.0030	0.06	0.0030	0.0025
$\alpha_{10} = 1.0$	0.0205	0.0162	0.06	0.0159	0.0159
$\alpha_{11} = 0.5$	-0.0452	0.0500	0.07	0.0486	0.0513
$\alpha_{20} = -1$	-0.0018	0.0122	0.02	0.0123	0.0140
$\alpha_{21} = 0.5$	-0.0043	0.0351	0.01	0.0354	0.0405
MLE					
$p = 0.5$	-0.0027	0.0026	0.06	0.0026	0.0025
$\alpha_{10} = 1.0$	0.0020	0.0162	0.07	0.0167	0.0163
$\alpha_{11} = 0.5$	-0.0074	0.0574	0.08	0.0579	0.0536
$\alpha_{20} = -1$	-0.0273	0.0144	0.03	0.0138	0.0143
$\alpha_{21} = 0.5$	0.0412	0.0437	0.04	0.0425	0.0416
$n = 200$					
MHDE					
$p = 0.5$	-0.0004	0.0014	0.05	0.0014	0.0013
$\alpha_{10} = 1.0$	0.0728	0.0227	0.02	0.0175	0.0718
$\alpha_{11} = 0.5$	-0.1496	0.0888	0.03	0.0667	0.2894
$\alpha_{20} = -1$	0.0844	0.0226	0.02	0.0156	0.0387
$\alpha_{21} = 0.5$	-0.1477	0.0764	0.02	0.0549	0.1518
MCHDE					
$p = 0.5$	-0.0003	0.0013	0.05	0.0013	0.0013
$\alpha_{10} = 1.0$	0.0102	0.0075	0.03	0.0075	0.0077
$\alpha_{11} = 0.5$	-0.0203	0.0252	0.05	0.0249	0.0251
$\alpha_{20} = -1$	0.0159	0.0067	0.05	0.0065	0.0068
$\alpha_{21} = 0.5$	-0.0193	0.0161	0.05	0.0158	0.0198
MLE					
$p = 0.5$	-0.0014	0.0012	0.04	0.0012	0.0013
$\alpha_{10} = 1.0$	0.0012	0.0076	0.04	0.0076	0.0078
$\alpha_{11} = 0.5$	-0.0010	0.0270	0.05	0.0270	0.0254
$\alpha_{20} = -1$	0.0034	0.0072	0.05	0.0072	0.0069
$\alpha_{21} = 0.5$	0.0015	0.0186	0.05	0.0186	0.0199

Table 3.2: ZIP regression model without any contamination, n=100 and 200

Parameters	Bias	MSE	Size	Var	Estimate of Avar
$n = 100$					
MCHDE					
$\beta_{10} = -1$	-0.2418	0.2460	0.05	0.1894	0.2185
$\beta_{11} = 0.5$	0.2685	0.6271	0.06	0.5606	0.6155
$\alpha_{20} = 2$	0.0699	0.0100	0.18	0.0052	0.0056
$\alpha_{21} = 1$	-0.1593	0.0360	0.21	0.0108	0.0150
MLE					
$\beta_{10} = -1$	-0.106	0.2042	0.04	0.1950	0.2078
$\beta_{11} = 0.5$	0.1649	0.5964	0.06	0.5750	0.5935
$\alpha_{20} = 2$	-0.0072	0.0057	0.02	0.0057	0.0060
$\alpha_{21} = 1$	0.0105	0.0157	0.07	0.0158	0.0150
$n = 200$					
MCHDE					
$\beta_{10} = -1$	-0.2127	0.1492	0.10	0.1050	0.1052
$\beta_{11} = 0.5$	0.2090	0.3761	0.10	0.3358	0.3000
$\alpha_{20} = 2$	0.0744	0.0080	0.30	0.0025	0.0026
$\alpha_{21} = 1$	-0.1576	0.0290	0.45	0.0043	0.0070
MLE					
$\beta_{10} = -1$	-0.0696	0.1149	0.08	0.1111	0.0993
$\beta_{11} = 0.5$	0.1532	0.3749	0.12	0.3549	0.2857
$\alpha_{20} = 2$	0.0070	0.0027	0.03	0.0026	0.0029
$\alpha_{21} = 1$	-0.0131	0.0065	0.02	0.0063	0.0075

as outliers. The parameters p and $\boldsymbol{\alpha}$ were specified as listed in Tables 3.3-3.6. Two values, 0.5 and 0.9, were considered for the mixing probability p, corresponding to even and uneven mixing of the components. The regression parameters $\boldsymbol{\alpha}$ were chosen to make the two components' mean either "well-separated" (Tables 3.3 and 3.4) or "poorly separated" (Tables 3.5 and 3.6), respectively. We also set $n = 100$ and $n = 200$ to investigate the effect of sample size. In each table of results (Tables 3.3-3.6), we report the bias, MSE, and size of the Wald test of equality to the true value, for each parameter to assess the performance of the three methods. As expected, the MCHDE exhibits less bias and smaller MSE than the MLE for most of the parameters when contamination is present in the response. Surprisingly, we do not observe the clear evidence of robustness of the MHDE when the two components are well-separated.

From the results above, we conclude that in the presence of outliers in y: (1) the MCHDE improved dramatically upon ML estimation in the cases we examined. (2) the MHDE is more robust than the MLE for the cases where the two components are poorly-separated. In such case, the MCHDE is superior to the MHDE for almost all the parameters except α_{10} which is the intercept corresponding to the component with larger mean. (3) Increasing the sample size from $n = 100$ to 200 has the expected effect of decreasing bias and MSE for all parameters across all three methods.

ZIP regression model

Next, we compare MCHD and ML estimation methods for data generated from a ZIP regression model with non-constant mixing probability and contamination in y. For each data set generated, 5% of the responses were selected at random and replaced by $y + 25$, as outliers. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ were specified as listed in Tables 3.7-3.10. The regression parameters $\boldsymbol{\alpha}$ were chosen to make the two components' mean either "well-separated" (Tables 3.7 and 3.8) or "poorly separated" (Tables 3.9

Table 3.3: Well separated two component binomial data with constant $p = 0.5$ and outliers in y, n= 100 and 200

	MCHDE			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$n = 100$									
$p = 0.5$	0.004	0.0032	0.08	0.005	0.0032	0.08	0.020	0.0028	0.06
$\alpha_{10} = 1.0$	0.044	0.0221	0.11	0.037	0.0421	0.06	0.192	0.0550	0.31
$\alpha_{11} = 0.5$	-0.025	0.0620	0.08	-0.014	0.1653	0.13	-0.106	0.0732	0.09
$\alpha_{20} = -1$	0.000	0.0125	0.03	0.118	0.0343	0.01	-0.022	0.0154	0.05
$\alpha_{21} = 0.5$	-0.001	0.0364	0.02	-0.247	0.1213	0.00	0.046	0.0479	0.03
$n = 200$									
$p = 0.5$	0.002	0.0011	0.02	0.002	0.0012	0.03	0.018	0.0013	0.05
$\alpha_{10} = 1.0$	0.021	0.0095	0.09	-0.033	0.0279	0.13	0.161	0.0338	0.41
$\alpha_{11} = 0.5$	0.020	0.0311	0.07	0.136	0.1494	0.26	-0.044	0.0315	0.05
$\alpha_{20} = -1$	0.021	0.0069	0.04	0.090	0.0255	0.01	0.014	0.0074	0.03
$\alpha_{21} = 0.5$	-0.023	0.0174	0.04	-0.160	0.0857	0.01	-0.005	0.0198	0.04

Table 3.4: Well separated two component binomial data with constant $p = 0.9$ and outliers in y, n= 100 and 200

	MCHDE			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$n = 100$									
$p = 0.9$	0.019	0.0015	0.21	0.034	0.0022	0.48	0.002	0.0010	0.07
$\alpha_{10} = 1.0$	0.027	0.0098	0.08	0.045	0.0276	0.07	0.105	0.0195	0.17
$\alpha_{11} = 0.5$	-0.009	0.0286	0.07	-0.033	0.1213	0.14	-0.058	0.0330	0.05
$\alpha_{20} = -1$	0.097	0.0761	0.04	0.174	0.0827	0.00	0.187	2.3798	0.08
$\alpha_{21} = 0.5$	-0.145	0.2246	0.03	-0.339	0.2319	0.00	-0.201	3.8661	0.10
$n = 200$									
$p = 0.9$	0.015	0.0007	0.18	0.022	0.0011	0.37	0.007	0.0005	0.08
$\alpha_{10} = 1.0$	0.022	0.0045	0.03	-0.009	0.0114	0.11	0.095	0.0125	0.35
$\alpha_{11} = 0.5$	0.002	0.0114	0.04	0.073	0.0558	0.22	-0.034	0.0120	0.02
$\alpha_{20} = -1$	0.041	0.0373	0.02	0.150	0.0496	0.00	0.017	0.0578	0.04
$\alpha_{21} = 0.5$	-0.069	0.1077	0.02	-0.298	0.1655	0.00	-0.020	0.1661	0.08

Table 3.5: Poorly separated two component binomial data with constant $p = 0.5$ and outliers in y, n= 100 and 200

	MCHDE			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$n = 100$									
$p = 0.5$	-0.142	0.057	0.33	-0.133	0.055	0.30	-0.435	0.192	0.11
$\alpha_{10} = 1.0$	0.875	3.508	0.20	-0.358	0.387	0.02	17.642	411.606	0.74
$\alpha_{11} = 0.5$	-0.311	1.521	0.04	4.567	93.107	0.08	-0.544	151.014	0.91
$\alpha_{20} = 0.5$	0.133	0.041	0.28	0.199	0.073	0.08	0.244	0.071	0.25
$\alpha_{21} = 0.5$	-0.100	0.062	0.03	-0.163	0.123	0.03	-0.042	0.032	0.90
$n = 200$									
$p = 0.5$	-0.144	0.054	0.31	-0.140	0.057	0.32	-0.442	0.197	0.96
$\alpha_{10} = 1.0$	-0.108	3.285	0.26	-0.370	0.363	0.03	15.436	308.093	0.23
$\alpha_{11} = 0.5$	-0.228	0.820	0.04	4.240	82.174	0.08	2.219	96.690	0.02
$\alpha_{20} = 0.5$	0.070	0.019	0.26	0.095	0.026	0.02	0.229	0.057	0.94
$\alpha_{21} = 0.5$	-0.024	0.017	0.02	-0.112	0.092	0.14	-0.010	0.012	0.04

Table 3.6: Poorly separated two component binomial data with constant $p = 0.9$ and outliers in y, n= 100 and 200

	MCHDE			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$n = 100$									
$p = 0.9$	-0.146	0.092	0.14	-0.033	0.028	0.05	-0.734	0.646	0.86
$\alpha_{10} = 1.0$	0.237	0.592	0.10	-0.039	0.049	0.07	18.663	1393.650	0.26
$\alpha_{11} = 0.5$	0.002	0.191	0.02	0.293	1.617	0.18	-3.032	3411.062	0.09
$\alpha_{20} = 0.5$	0.189	0.110	0.14	0.362	0.188	0.01	0.955	33.741	0.84
$\alpha_{21} = 0.5$	-0.087	0.155	0.01	-0.321	0.172	0.00	-0.444	43.680	0.07
$n = 200$									
$p = 0.9$	-0.087	0.048	0.08	-0.034	0.030	0.05	-0.838	0.711	0.96
$\alpha_{10} = 1.0$	0.163	0.385	0.05	-0.071	0.027	0.02	15.128	306.632	0.22
$\alpha_{11} = 0.5$	-0.010	0.112	0.02	0.374	1.412	0.19	1.228	119.144	0.00
$\alpha_{20} = 0.5$	0.121	0.072	0.10	0.274	0.133	0.07	0.447	0.206	0.96
$\alpha_{21} = 0.5$	-0.074	0.120	0.02	-0.248	0.143	0.00	-0.011	0.014	0.07

and 3.10), respectively. Two values of $\boldsymbol{\beta}$ were considered corresponding to low and moderate levels of zero inflation. We also set $n = 100$ and $n = 200$ to investigate the effect of sample size.

With respect to $\boldsymbol{\beta}$, MCHD estimation is similar to ML when models are well-separated, but does better under poor separation. With respect to $\boldsymbol{\alpha}$, the MCHDE exhibits dramatically better performance than the MLE which has unacceptable levels of bias and MSE in all cases. Sample size has the expected effect of decreasing bias and improving efficiency for both methods, but the MCHD approach maintains a clear advantage even when $n = 200$.

Table 3.7: Well separated and moderate level zero inflated Poisson distribution with outliers in y, n= 100 and 200

	MCHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$n = 100$						
$\beta_{10} = -1.0$	-0.231	0.2492	0.05	-0.195	0.2418	0.06
$\beta_{11} = 0.5$	0.224	0.6021	0.05	0.184	0.6089	0.05
$\alpha_{20} = 2$	0.074	0.0107	0.17	0.217	0.0524	0.84
$\alpha_{21} = 1$	-0.142	0.0303	0.19	-0.187	0.0516	0.37
$n = 200$						
$\beta_{10} = -1.0$	-0.146	0.1050	0.04	-0.077	0.0903	0.03
$\beta_{11} = 0.5$	0.001	0.2561	0.03	-0.015	0.2622	0.04
$\alpha_{20} = 2$	0.061	0.0062	0.19	0.196	0.0412	0.99
$\alpha_{21} = 1$	-0.127	0.0210	0.29	-0.157	0.0320	0.49

3.5.3 Study 3: Models with Outliers in $\boldsymbol{x}$

Another type of model violation occurs when there are outliers in the covariate vector $\boldsymbol{x}$. Such points can be hard to detect and will typically have high leverage with the potential to severely influence estimates and inference. In this study, we

Table 3.8: Well separated zero and low level inflated Poisson distribution with outliers in y, n= 100 and 200

	MCHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$n = 100$						
$\beta_{10} = -2.0$	-0.249	0.3988	0.01	-0.182	0.3900	0.01
$\beta_{11} = 0.5$	0.155	0.8960	0.01	0.099	0.9715	0.00
$\alpha_{20} = 2$	0.078	0.0096	0.21	0.191	0.0399	0.85
$\alpha_{21} = 1$	-0.145	0.0267	0.20	-0.158	0.0350	0.31
$n = 200$						
$\beta_{10} = -2.0$	-0.197	0.2588	0.03	-0.113	0.2388	0.03
$\beta_{11} = 0.5$	0.070	0.6829	0.06	0.053	0.7056	0.07
$\alpha_{20} = 2$	0.070	0.0068	0.25	0.174	0.0325	0.96
$\alpha_{21} = 1$	-0.138	0.0232	0.47	-0.136	0.0240	0.49

Table 3.9: Poorly separated and moderate level zero inflated Poisson distribution with outliers in y, n= 100 and 200

	MCHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$n = 100$						
$\beta_{10} = -1.0$	-0.180	0.2813	0.02	0.293	0.2738	0.09
$\beta_{11} = 0.5$	0.012	0.6839	0.02	-0.428	0.7901	0.10
$\alpha_{20} = 0.6$	0.047	0.0257	0.05	0.883	0.7904	1.00
$\alpha_{21} = 1$	-0.159	0.0753	0.07	-0.719	0.5430	0.98
$n = 200$						
$\beta_{10} = -1.0$	-0.108	0.1227	0.00	0.353	0.2140	0.23
$\beta_{11} = 0.5$	-0.122	0.3023	0.01	-5.019	0.5209	0.18
$\alpha_{20} = 0.6$	0.040	0.0108	0.02	0.815	0.6681	1.00
$\alpha_{21} = 1$	-0.118	0.0330	0.04	-0.590	0.3600	1.00

Table 3.10: Poorly separated and low level zero inflated Poisson distribution with outliers in y, n= 100 and 200

	MCHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$n = 100$						
$\beta_{10} = -2.0$	-0.046	0.5768	0.03	0.664	0.7293	0.24
$\beta_{11} = 0.5$	-0.234	1.3114	0.03	-0.930	1.7790	0.12
$\alpha_{20} = 0.6$	0.070	0.0212	0.04	0.790	0.6319	1.00
$\alpha_{21} = 1$	-0.157	0.0592	0.03	-0.674	0.4755	0.99
$n = 200$						
$\beta_{10} = -2.0$	0.024	0.3492	0.04	0.687	0.6834	0.48
$\beta_{11} = 0.5$	-0.270	0.8666	0.03	-0.882	1.3590	0.28
$\alpha_{20} = 0.6$	0.062	0.0110	0.04	0.728	0.5335	1.00
$\alpha_{21} = 1$	-0.116	0.0307	0.06	-0.550	0.3116	1.00

want to explore if the MCHD and MHD estimators are more robust than the MLE in the sense of protection against outliers in $\boldsymbol{x}$.

Two component binomial model

To investigate the robustness to outliers in the covariates $\boldsymbol{x}$, we generate data from the models described in section 5.2 and summarized in Tables 3-6 again. Instead of adding contamination in y, we create outliers in $\boldsymbol{x}$, by randomly choosing 1% of the observations, and replacing the covariate x_1 by $x_1 + 3$, leaving the response y unchanged. All the results are listed in Tables 3.11-3.14.

In all cases, MCHD estimation has less bias, smaller MSE and closer to nominal size than ML estimation. For the well-separated cases (Tables 3.11-3.12), MHDE does not exhibit clear robustness, while for the poorly-separated cases, the MHDE

is more robust than the MLE. Generally, we can observe the usual positive effect of sample size on efficiency.

Table 3.11: Well separated two component binomial data with constant $p = 0.5$ and outliers in $\boldsymbol{x}$, n= 100 and 200

	MCHDE			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$n = 100$									
$p = 0.5$	-0.002	0.003	0.06	-0.002	0.003	0.08	-0.006	0.004	0.05
$\alpha_{10} = 1.0$	0.029	0.017	0.09	0.159	0.043	0.00	0.016	0.018	0.10
$\alpha_{11} = 0.5$	-0.062	0.056	0.12	-0.332	0.154	0.00	-0.039	0.064	0.11
$\alpha_{20} = -1.0$	0.007	0.015	0.10	0.090	0.028	0.02	0.055	0.022	0.27
$\alpha_{21} = 0.5$	-0.022	0.047	0.18	-0.198	0.105	0.02	-0.133	0.083	0.54
$n = 200$									
$p = 0.5$	-0.004	0.001	0.05	-0.007	0.001	0.05	-0.008	0.001	0.04
$\alpha_{10} = 1.0$	0.013	0.009	0.11	0.133	0.031	0.00	0.017	0.010	0.11
$\alpha_{11} = 0.5$	-0.027	0.030	0.17	-0.275	0.123	0.00	-0.036	0.037	0.18
$\alpha_{20} = -1.0$	0.025	0.008	0.16	0.073	0.020	0.04	0.089	0.020	0.46
$\alpha_{21} = 0.5$	-0.034	0.022	0.19	-0.125	0.068	0.03	-0.195	0.068	0.64

ZIP regression model

For ZIP models, we added outliers in $\boldsymbol{x}$, by randomly choosing 1% of the observations, and replacing the covariate x_1 by $x_1 + 3$, leaving the response y unchanged. The corresponding results are summarized in Tables 3.15-3.18.

In these tables, MCHD and ML perform similarly with respect to $\boldsymbol{\beta}$. This result is sensible, since there is only a small amount of contamination in x which is of a form that does not obscure the mixture structure much. With respect to $\boldsymbol{\alpha}$ however, the MCHDE has less bias, smaller MSE and closer to nominal size than ML estimation. Generally speaking, these results are consistent with those of Tables 3.11-3.14.

Table 3.12: Well separated two component binomial data with constant $p = 0.9$ and outliers in $\boldsymbol{x}$, n= 100 and 200

	MCHDE			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$n = 100$									
$p = 0.9$	0.016	0.001	0.19	0.031	0.002	0.42	-0.005	0.001	0.09
$\alpha_{10} = 1.0$	0.035	0.011	0.13	0.144	0.030	0.02	0.037	0.012	0.13
$\alpha_{11} = 0.5$	-0.075	0.042	0.22	-0.298	0.119	0.02	-0.084	0.051	0.26
$\alpha_{20} = -1.0$	0.053	0.063	0.05	0.156	0.077	0.00	0.010	0.711	0.08
$\alpha_{21} = 0.5$	-0.065	0.136	0.06	-0.292	0.194	0.00	0.005	0.196	0.10
$n = 200$									
$p = 0.9$	0.012	0.001	0.14	0.020	0.001	0.30	-0.002	0.001	0.05
$\alpha_{10} = 1.0$	0.029	0.005	0.12	0.122	0.022	0.00	0.036	0.007	0.14
$\alpha_{11} = 0.5$	-0.056	0.019	0.15	-0.252	0.088	0.00	-0.075	0.026	0.20
$\alpha_{20} = -1.0$	0.020	0.026	0.04	0.136	0.047	0.00	0.021	0.028	0.07
$\alpha_{21} = 0.5$	-0.028	0.077	0.08	-0.262	0.144	0.00	-0.051	0.081	0.19

Table 3.13: Poorly separated two component binomial data with constant $p = 0.5$ and outliers in $\boldsymbol{x}$, n= 100 and 200

	MCHDE			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$n = 100$									
$p = 0.5$	-0.016	0.012	0.01	-0.016	0.009	0.00	0.041	0.061	0.23
$\alpha_{10} = 1.0$	-0.012	0.046	0.05	0.113	0.039	0.02	-0.159	0.177	0.27
$\alpha_{11} = 0.5$	-0.004	0.161	0.20	-0.295	0.127	0.00	0.240	0.489	0.24
$\alpha_{20} = 0.5$	0.117	0.040	0.09	0.189	0.054	0.25	0.198	0.175	0.28
$\alpha_{21} = 0.5$	-0.157	0.104	0.23	-0.275	0.104	0.00	-0.350	0.222	0.66
$n = 200$									
$p = 0.5$	-0.009	0.005	0.01	-0.006	0.004	0.02	0.081	0.040	0.22
$\alpha_{10} = 1.0$	0.011	0.024	0.04	0.118	0.032	0.01	-0.104	0.085	0.30
$\alpha_{11} = 0.5$	-0.015	0.105	0.18	-0.258	0.124	0.00	0.501	0.190	0.27
$\alpha_{20} = 0.5$	0.064	0.019	0.10	0.099	0.021	0.02	0.098	0.062	0.22
$\alpha_{21} = 0.5$	-0.126	0.061	0.30	-0.191	0.061	0.00	-0.307	0.244	0.79

Table 3.14: Poorly separated two component binomial data with constant $p = 0.9$ and outliers in $\boldsymbol{x}$, n= 100 and 200

	MCHDE			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$n = 100$									
$p = 0.9$	-0.071	0.022	0.02	-0.015	0.007	0.01	-0.162	0.15	0.20
$\alpha_{10} = 1.0$	0.033	0.018	0.07	0.147	0.035	0.00	-0.082	0.683	0.09
$\alpha_{11} = 0.5$	-0.032	0.072	0.13	-0.306	0.117	0.02	0.012	0.489	0.22
$\alpha_{20} = 0.5$	0.200	1.121	0.05	0.323	0.156	0.08	0.275	0.301	0.31
$\alpha_{21} = 0.5$	-0.096	0.216	0.06	-0.301	0.132	0.00	-0.301	1.151	0.40
$n = 200$									
$p = 0.9$	-0.036	0.010	0.01	0.002	0.001	0.00	-0.164	0.088	0.21
$\alpha_{10} = 1.0$	0.035	0.010	0.04	0.102	0.017	0.06	-0.041	0.048	0.23
$\alpha_{11} = 0.5$	-0.045	0.032	0.11	-0.230	0.074	0.00	0.115	0.174	0.20
$\alpha_{20} = 0.5$	0.171	0.087	0.07	0.213	0.096	0.04	0.319	0.248	0.43
$\alpha_{21} = 0.5$	-0.147	0.109	0.14	-0.204	0.111	0.00	-0.308	0.325	0.61

Table 3.15: Well separated and moderate level zero inflated Poisson distribution with outliers in $\boldsymbol{x}$, n= 100 and 200

	MCHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$n = 100$						
$\beta_{10} = -1.0$	-0.380	0.3307	0.07	0.069	0.1449	0.04
$\beta_{11} = 0.5$	0.373	0.6571	0.06	-0.191	0.3799	0.01
$\alpha_{20} = 2$	0.066	0.0094	0.18	0.255	0.0945	0.72
$\alpha_{21} = 1$	-0.148	0.0328	0.18	-0.507	0.3680	0.73
$n = 200$						
$\beta_{10} = -1.0$	-0.140	0.1031	0.03	0.142	0.0807	0.06
$\beta_{11} = 0.5$	0.159	0.2764	0.03	-0.313	0.2736	0.09
$\alpha_{20} = 2$	0.048	0.0048	0.15	0.295	0.0998	0.90
$\alpha_{21} = 1$	-0.122	0.0200	0.26	-0.591	0.3933	0.90

Table 3.16: Well separated and low level zero inflated Poisson distribution with outliers in $\boldsymbol{x}$, n= 100 and 200

	MCHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$n = 100$						
$\beta_{10} = -2.0$	-0.374	0.4710	0.00	0.129	0.2281	0.04
$\beta_{11} = 0.5$	0.250	0.8733	0.01	-0.372	0.5670	0.00
$\alpha_{20} = 2$	0.072	0.0086	0.17	0.304	0.1074	0.89
$\alpha_{21} = 1$	-0.146	0.0275	0.20	-0.590	0.3960	0.89
$n = 200$						
$\beta_{10} = -2.0$	-0.221	0.2811	0.04	0.138	0.1287	0.06
$\beta_{11} = 0.5$	0.251	0.7516	0.07	-0.322	0.3679	0.05
$\alpha_{20} = 2$	0.055	0.0050	0.19	0.309	0.1026	0.95
$\alpha_{21} = 1$	-0.124	0.0196	0.33	-0.610	0.3948	0.95

Table 3.17: Poorly separated and moderate level zero inflated Poisson distribution with outliers in $\boldsymbol{x}$, n= 100 and 200

	MCHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$n = 100$						
$\beta_{10} = -1.0$	-0.419	0.1011	0.02	0.104	0.4061	0.08
$\beta_{11} = 0.5$	0.267	0.6066	0.03	-0.236	0.9990	0.06
$\alpha_{20} = 0.6$	0.047	0.0315	0.14	0.234	0.1120	0.63
$\alpha_{21} = 1$	-0.175	0.1061	0.17	-0.452	0.3660	0.60
$n = 200$						
$\beta_{10} = -1.0$	-0.187	0.1406	0.01	0.285	0.1812	0.20
$\beta_{11} = 0.5$	0.172	0.2429	0.01	-0.476	0.4395	0.20
$\alpha_{20} = 0.6$	0.016	0.0124	0.04	0.322	0.1265	0.89
$\alpha_{21} = 1$	-0.079	0.0327	0.05	-0.607	0.4309	0.90

Table 3.18: Poorly separated and low level zero inflated Poisson distribution with outliers in $\boldsymbol{x}$, n= 100 and 200

	MCHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$n = 100$						
$\beta_{10} = -2.0$	-0.304	0.7229	0.04	0.275	0.9322	0.14
$\beta_{11} = 0.5$	0.071	1.4009	0.03	-0.621	2.4453	0.08
$\alpha_{20} = 0.6$	0.079	0.0281	0.17	0.289	0.1184	0.80
$\alpha_{21} = 1$	-0.176	0.0893	0.17	-0.549	0.3998	0.78
$n = 200$						
$\beta_{10} = -2.0$	-0.180	0.4544	0.06	0.463	0.5247	0.33
$\beta_{11} = 0.5$	0.170	0.8705	0.03	-0.739	1.2196	0.20
$\alpha_{20} = 0.6$	0.034	0.0099	0.03	0.350	0.1319	0.97
$\alpha_{21} = 1$	-0.018	0.0255	0.08	-0.647	0.4403	0.97

3.6 Robustness Studies

The robustness of the MHD estimator for finite mixture models is discussed in the literature for various modelling contexts, including Beran (1977), Cutler et al. (1996), Karlis and Xekalaki (2000) and Lu et al. (2003). In this section, we investigate the robustness of MHD estimation methods for FMGLMs through an examination of the α-influence function (IF_α) defined as follows. In general, assume a parametric model with density $f_{\boldsymbol{\theta}}(x)$, $\boldsymbol{\theta} \in \Theta$, and estimator $T(f_{\boldsymbol{\theta}})$. Let $f_{\alpha,\boldsymbol{\theta},\delta}(x) = (1-\alpha)f_{\boldsymbol{\theta}}(x)+\alpha\delta(x)$ where $\delta(x)$ is a contamination function. Then the α-influence function quantifies the effect of an α-level degree of contamination on the relative estimation error as follows:

$$IF_\alpha = \frac{T(f_{\alpha,\boldsymbol{\theta},\delta}) - \boldsymbol{\theta}}{\alpha}.$$

If IF_α is a bounded function of δ such that $\lim_{\delta\to\infty} IF_\alpha = 0$, then the functional T is robust at $f_{\boldsymbol{\theta}}$ against $100\alpha\%$ contamination by gross errors at arbitrary large value δ.

Although it is hard to get a closed expression of the IF_α in most cases, we can calculate it numerically for specific examples. Once again, we use the a two-component binomial regression model and a ZIP regression model as examples to quantify the robustness of our estimation methods via the IF_α. For the two-component binomial model, we add a third component $\delta \sim$ binomial$(m, 1)$ (a degenerate distribution with point mass 1 at m) with a contamination rate α. We assume the same regression structure as in the simulation studies. Recall from section 3.5 that we assume constant mixing probability and simple linear regressions in each component mean with covariate vector $\boldsymbol{x}_i = (1, x_{1i})$ where $x_{1i} \sim U(0, 1)$. Then we set the sample size to $n = 200$ and $m = 30$, and choose true parameters $p = 0.5$, $\boldsymbol{\alpha}_1 = (1, 0.5)'$, and $\boldsymbol{\alpha}_2 = (0.5, 0.5)'$. Then we can calculate IF_α through

$$IF_\alpha(\delta; T, f) = \frac{T((1-\alpha)f + \alpha\delta) - T(f)}{\alpha}.$$

Figure 3.1 displays a plot of $IF_\alpha(\delta; T, f)$ for α ranging from 0.01 to 0.15. The curves of MHDE and MCHDE are very close. When the data have a small portion of contamination, the IF_α of MHDE and MCHDE are much smaller than that of MLE for most of the parameters. These results indicate the robustness of MHDE and MCHDE.

Next we add a third component $\delta_p \sim$ binomial(m, p) where we allow p to vary from 0 to 1 and we fix the contamination rate and sample size at 5% and $n = 200$, respectively. Figure 3.2 plots $IF_{0.05}\{\delta_p; T, f\}$ versus p. When the third component parameter p is less than 0.3 or greater than 0.8, the α-influence function for the MHDE and MCHDE are smaller than that of the MLE. This implies MHDE and MCHDE are robust to points severely deviating from the true model.

Similarly, we also examine the IF_α plot for a ZIP model. A data set with $n =$ 200 was are generated in the same way as described in section 5. We examined the α-influence function for this model, under the addition of a third component

$\delta \sim$ Poisson(20) with a contamination rate α, letting α vary from 0.01 to 0.15. The IF_α is calculated at each value of α (see Figure 3.3). Finally, to plot Figure 3.4, we randomly select 5% of the data and replace them by an arbitrary number P, where P varies from 0 to 20. In both Figures 3.3 and 3.4, the MCHDE and MLE have similar curves with respect to $\boldsymbol{\beta}$, the parameter corresponding to the mixing probability. This reflects the fact that the outliers we added did not obscure the mixture structure much. That is, these outliers rarely changed the pattern of occurrence of "excess zeros" in the data set. However, with respect to $\boldsymbol{\alpha}$, which contains parameters corresponding to component means, the MCHDE is much more robust than the MLE.

For the sake of brevity, we present only the IF_α curves for the mixture models with poor-separation. In fact, we also calculate the α-influence function for other parameter settings. All those curves mirror the results in the simulation study; that is, when the two components are well separated, only the MCHDE shows clear robustness to the MLE, when they are poorly separated, both the MHDE and MCHDE are more robust than the MLE.

3.7 Examples

The results of the previous section indicate that the MCHDE is more robust than MLE in a variety of settings. This section illustrates this property with two examples.

3.7.1 Example 1

In this section, we analyze data from a cohort study carried out at the College of Veterinary medicine, University of Georgia which characterize ambulatory electrocardiographic results of overtly healthy Doberman Pinschers. The selected subjects include

114 (58 male, 56 female) overtly healthy Doberman Pinschers without echocardiographic evidence of cardiac disease. Among all measures, heart rates, ventricular premature contractions (VPCs), age and sex are recorded. As shown in Calvert et al (2000), VPC rate measured on the initial Holter recording was associated with subsequent development of dilated cardiomyopathy. One objective of the original study was to explore the associations between VPC rate and age and sex. Figure 3.5 displays a histogram of the VPC counts for the 106 animals who had fewer than 35 VPCs. In addition, 8 animals (omitted in the histogram) had much larger counts equal to 58, 65, 98, 144, 326, 492, 568 and 1894/24 hours. Figure 3.5 suggests some heterogeneity in this sample of dogs where there appears to be a subgroup of animals for whom VPCs are very rare events (counts of 0, 1 or perhaps 2/24 hours). In addition, there is clearly a sizable subgroup of animals who experienced VPCs more commonly as well as a few animals with extremely large VPC counts. Given this pattern in the data and the fact that cardiomyopathy is a genetic abnormality affecting a subset of the Doberman Pinscher population, we initially considered a two-component mixture structure for these data. The components of this mixture were hypothesized to correspond to genetically normal and abnormal subpopulations. In addition, we considered the 8 extremely large counts to be outliers, representing contamination of our sample by a few atypical animals falling outside of the population of primary interest (clinically normal dogs).

However, after initial analysis of the data via two-component Poisson GLMs, it became clear that there is additional heterogeneity in these VPC counts that must be accounted for. Essentially, these models grouped all of the counts that were ≤ 10 into one component, with a second, much larger component centered in the low 20's. This resulted in very large Lindsay's residuals at 0 and at values between 5 and 10. These results strongly suggested the need for a three-component model to capture additional heterogeneity among the single digit counts while still allowing a larger

mean component to explain the sizable number of observed values between 15 and 30 in the data. It is worth noting that some researchers may use additional components to capture the 8 outliers. However, in this example, since the 8 extremely large counts are spread out widely, we could not fit all these data well by adding only one or two components additional. So we are stick to the three-component model and treat those extremely large values as potential outliers. Age and gender were considered as covariates in both the mixing probability and mean specifications of these models. In addition, the robust estimation methods discussed in this paper were used to downweight the influence of the few extremely large counts present here.

Specifically, we assume that y_i, the number of VPCs for the i^{th} subject, follows a three component Poisson mixture with mean μ_{ik} and mixing probability p_{ik} for the k^{th} component, where $k = 1, 2, 3$. The mixing probabilities of p_k are modelled via generalized logit links

$$\log(\frac{p_{i1}}{1 - p_{i1} - p_{i2}}) = \beta_{10} + \beta_{11}\mathrm{I}(female)_i + \beta_{12}\mathrm{Age}_i$$
$$\log(\frac{p_{i2}}{1 - p_{i1} - p_{i2}}) = \beta_{20} + \beta_{21}\mathrm{I}(female)_i + \beta_{22}\mathrm{Age}_i$$

while the μ_{ik}'s are modelled as

$$\log(\mu_{i1}) = \alpha_{10} + \alpha_{11}\mathrm{I}(female)_i + \alpha_{12}\mathrm{Age}_i,$$

$$\log(\mu_{i2}) = \alpha_{20} + \alpha_{21}\mathrm{I}(female)_i + \alpha_{22}\mathrm{Age}_i,$$

$$\log(\mu_{i3}) = \alpha_{30} + \alpha_{31}\mathrm{I}(female)_i + \alpha_{32}\mathrm{Age}_i.$$

Parameter estimates and standard errors for this model fit with both MCHD and ML estimation appear in Table 3.19. Note that the marginal MHD approach is omitted here because of the dependence of the mixing probabilities on covariates. Clearly, parameter estimates under the MCHD approach are quite different to those obtained via ML.

Table 3.19: Parameters Estimates and Standard Errors (SE) for the VPC data

	MCHDE	MLE
Parameters	Estimate (SE)	Estimate (SE)
β_{10}	1.4686 (0.6219)	2.7678 (1.0668)
β_{11}	1.6620 (0.5648)	1.4386 (1.1408)
β_{12}	0.0212 (0.1105)	-0.0775 (0.1762)
β_{20}	0.6896 (0.6564)	1.4377 (1.1401)
β_{21}	1.6651 (0.6002)	1.4167 (1.1971)
β_{22}	-0.0896 (0.1170)	-0.0846 (0.1923)
α_{10}	-1.1750 (0.5842)	-0.8927 (0.3845)
α_{11}	0.0064 (0.4879)	-1.2970 (0.2605)
α_{12}	0.0145 (0.0929)	0.2345 (0.0277)
α_{20}	1.6115 (0.3024)	3.4935 (0.1317)
α_{21}	-0.4003 (0.2775)	-1.2827 (0.0886)
α_{22}	0.0220 (0.0572)	0.0173 (0.0102)
α_{30}	3.1216 (0.2768)	6.9152 (0.1278)
α_{31}	-0.0535 (0.3253)	1.1526 (0.0411)
α_{32}	-0.0004 (0.0487)	-0.1738 (0.0231)

To compare the MCHD and ML fits of the model, we use Lindsay's residual function

$$r(y;\hat{\boldsymbol{\theta}}) = \frac{f_n(y)}{f_{\hat{\boldsymbol{\theta}},n}(y)} - 1,$$

where $f_n(y)$ and $f_{\boldsymbol{\theta},n}$ are defined in (3.3.5) and (3.3.8) respectively. Figure 3.6 is the residual plot for the majority of the data (less than 50) under MCHD and ML methods. It shows that the MCHDE fits the data better than MLE. Based on the results of MCHD approach, this model can be reduced. The mixing probabilities of p_k are modelled via generalized logit links

$$\log(\frac{p_{i1}}{1-p_{i1}-p_{i2}}) = \beta_{10} + \beta_{11}\mathrm{I}(female)_i$$

$$\log(\frac{p_{i2}}{1-p_{i1}-p_{i2}}) = \beta_{20} + \beta_{21}\mathrm{I}(female)_i$$

while the μ_{ik}'s are modelled as

$$\log(\mu_{i1}) = \alpha_i \quad i = 1, 2, 3.$$

Parameter estimates and standard errors for this reduced model fit with MCHD appear in Table 3.20. This results lead to the conclusion that female dog are less likely having large VPC rate.

Table 3.20: Parameters Estimates and Standard Errors (SE) of the reduced model for the VPC data

	MCHDE	
Parameters	Estimate	SE
β_{10}	2.0172	(0.5771)
β_{11}	1.7272	(0.4875)
β_{20}	1.2207	(0.6301)
β_{21}	1.4957	(0.5306)
α_1	-1.0577	(0.2273)
α_2	1.5316	(0.1322)
α_3	3.1071	(0.0896)

3.7.2 Example 2

As a second example of the usefulness of robust methods for finite mixture models, we consider zero inflated regression models for data from the Multisite Violence Prevention Project (MVPP). This study, conducted by investigators from four US universities in cooperation with the Centers for Disease Control and Prevention, was designed to investigate approaches for reducing violent and aggressive behaviors among middle school aged children. The study utilized a randomized complete block design involving 37 schools randomized to a 4 treatment structure within each of four blocks corresponding to the sites of the universities participating in the project. Included among the outcomes measured via teacher surveys was a 30-day recall of the

number of insults that teachers received from their students. Although the study was conducted longitudinally, with each teacher generating data at several measurement occasions, here we analyze just the baseline data to determine whether there were pre-existing differences between the insult rates across the four treatment groups. For simplicity, we also restrict attention to just one of the four sites involved in the study, from which 86 teachers' data were available. Table 3.21 summarizes these insult counts by treatment.

Table 3.21: Frequency of the Insults Number Received from Students among 86 Teachers in 4 Treatment Groups

	No. of insults received from students								
Treatment	0	1	2	3	4	5	7	10	30
Treatment 1	5	5	2	1	2	3	0	1	0
Treatment 2	8	2	3	2	1	0	1	1	1
Treatment 3	11	3	10	1	0	2	0	0	0
Treatment 4	15	3	1	1	0	1	0	0	0

From this table it is apparent that a very large proportion of the teachers reported 0 insults. However, there are also large frequencies of insult counts that are larger than 0 indicating possible zero inflation in these data. In addition, there are a few teachers who reported very large numbers of insults (7,10 and 30) which are clearly outlying relative to the main portion of the data and which may strongly affect inferences on the treatment group. Given this data structure and experimental design, a natural model to consider here is a zero inflated Poisson analysis of variance type model.

Specifically, we assume that y_{ij}, the number of insults for the ith teacher in the jth treatment, follows a ZIP distribution with Poisson mean $\log(\mu_{ij}) = \alpha_j$, and mixing probability p_{ij} with model $\log(\frac{p_{ij}}{1-p_{ij}}) = \beta_j$, $j = 1, 2, 3, 4$.

Table 3.22: Parameters Estimates and Standard Errors (SE) for MVPP

	MCHDE	MLE
Parameters	Estimate (SE)	Estimate (SE)
β_1	-1.2351 (0.7313)	-1.2250 (0.6132)
β_2	-0.2933 (0.5753)	-0.3251 (0.4665)
β_3	-0.7224 (0.6366)	-0.8206 (0.5665)
β_4	0.6012 (0.6766)	0.6570 (0.5556)
α_1	0.7842 (0.2083)	1.1198 (0.1617)
α_2	0.6900 (0.2603)	1.7737 (0.1250)
α_3	0.4405 (0.2467)	0.6524 (0.2036)
α_4	0.2194 (0.4665)	0.5951 (0.3449)

Parameter estimates and standard errors for these models appear in Table 3.22. From these results, we can find that MCHDE and MLE are similar with respect to $\boldsymbol{\beta}$, but quite different for $\boldsymbol{\alpha}$. Under ML estimation, the μ_j are severely affected by the few large values present, especially for treatment 2.

As with example 1, we use the Lindsay's residual to compare the MCHD and ML fits of the model. The MCHDE is treating the large values as outliers. It downweights them and fits the model based mostly upon the smaller values. Therefor, we expected the Lindsay's residuals to be large at 30, 10, 7 and perhaps even at 5. Figure 3.7 is the Lindsay's residual plot for the model under MCHD and ML methods (Only observations less than 7 are plotted). It shows that the MCHDE fits the data better than MLE for majority of the data except those outliers.

3.8 Discussion

In this thesis, we've considered robust estimation methods for FMGLMs based on the Hellinger distance. Firstly, we extended Lu et al.'s marginal MHD approach, and then proposed a new conditional MHD method. We've seen that the latter is more general in the sense that it applies to FMGLMs with continuous components, and it applies to situations where the mixing probability is non-constant. However, it also has limitations/drawbacks in that it requires conditional density estimation, which requires bandwidth selection, and which can break down or become infeasible for high dimensional covariates. Based on the simulation results presented here MCHD estimation and MHD are both more robust than ML estimation, but MCHD has a substantial advantage over MHD with respect to MSE and size of asymptotic Wald-based inference.

In all the methods proposed above, the number of components in the mixture is fixed. In our case we assume that we have prior knowledge about how many components should be used. In practical application, the question of how to estimate the number of components is a very important question for fitting FMGLMs, especially when no prior knowledge is available. A large amount of literature exists on this topic including Lindsay (1992, 1995), Böhning (1992, 2000) and Peel (2000). Schlattmann, (2000) developed an inferential Bootstrap approach for estimating the number of components in a mixture. Karlis and Xekalaki (2000) proposed a robust alternative based on MHD estimates. A robust estimation for mixture complexity was discussed by Woo and Sriram (2006). Other approaches are based upon the Akaike information criterion, Bayesian information criterion and their relatives (see McLachlan and Peel, 2000).

3.9 References

[1] Aitkin, M. A. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics,* **55**, 117-128.

[2] Beran, R. (1977). Minimum Hellinger distance for parametric models. *The Annals of Statistics,* **5**, 445-463.

[3] Böhning, D., Schlattmann, P. and Lindsay, B. G. (1992). Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithm. *Biometrics,* **48**, 283-303.

[4] Böhning, D. and Seidel, W. (2002). Editorial: recent developments in mixture models. *Computational Statistics and Data Analysis,* **41**, 349-357.

[5] Chow, Y.S. and Teicher, H. (1997). *Probability Theory.* Springer.

[6] Calvert, A.C., Jacobs, G.J., Smith, D.D., Rathbun, S.L. and Pickus, C.W. (2000). Association between results of ambulatory electrocardiography and development of cardiomyopathy during long-term follow-up of Doberman Pinschers. *Journal of the American Vet Medicine Association,* **216**, 34-39.

[7] Cutler, A., Cordero-Brana, O. (1996). Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association,* **91**, 1716-1724.

[8] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B,* **39**, 1-38.

[9] Dietz, E. (1992). Estimation of heterogeneity - A GLM approah. In *Advances in GLIM and Statistical Modelling, Lecture Notes in Statistics*, L. Fahrmeir, F. Francis, R. Gilchrist, and G. Tutz (eds), 66-72. Berlin: Springe Verlag.

[10] Donoho, D. L. and Liu, R. C. (1988). The "automatic" robustness of minimum distance functionals. *The Annals of Statistics,* **16**, 552-586.

[11] Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.*

[12] Fan, J and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman and Hall, New York.

[13] Follmann, D. A. and Lambert, D. (1989). Generalizing logistic regression non-parametrically. *Journal of the American Statistical Association,* **84**, 295-300.

[14] Gooijer, J and Zerom, D. (2003). On conditional density estimation. *Statistical Neerlandica,* **57**, 159-176.

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

[16] Hall, D. B. and Wang, L. Mixtures of generalized linear mixed-effects models for cluster-correlated data. *Statistical Modelling: An International Journal.* Accepted, subject to revision.

[17] Hall, P., Wolff, R. C. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of American Statistical Association,* **94**, 154-163.

[18] Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *The Annal of Statistics,* **33**, 1404-1421.

[19] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association,* **69**, 383-393.

[20] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986) *Robust Statistics.* New York: John Wiley and Sons.

[21] Hinde, J.P. and Wood, A.T.A (1987). Binomial variance component models with a nonparametric assumption concerning random effects. In *Longitudinal Data Analysis*, R. Crouchley (ed.) Averbury, Aldershot, Hants.

[22] Huber, P. J. (1972). Robust statistics: a review. *Annals of Mathematical Statistics,* **43**, 1041-1067.

[23] Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Nonparametric Statistics,* **14**, 259-278.

[24] Jansen, R. C. (1993). Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics,* **49**, 227-231.

[25] Karlis, D. and Xekalaki, E. (2001). Robust inference for finite Poisson mixtures. *Journal of Statistical Planning and Inference,* **93**, 93-115.

[26] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics,* **34**, 1-14.

[27] Lindsay, B. G. and Roeder, K. (1992). Residual diagnostics for mixture models. *Journal of the American Statistical Association,* **87**, 785-794.

[28] Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics,* **22**, 1081-1114.

[29] Lindsay, B. G. (1995). Mixture Models: theory, geometry and applications regional conference series in probability and statistics, Vol. 5. *Institute of Mathematical Statistics and Am. Statist. Assoc.* Hayward, California

[30] Lu. Z., Hui. Y. and Lee, A. (2003) Minimum Hellinger distance estimation for finite mixtures of Poisson regression models and its applications. *Biometrics,* **59**, 1016-1026.

[31] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, London: Chapman and Hall.

[32] McLachlan G. and Peel D. (2001). *Finite Mixture Models*. New York: Wiley.

[33] Pregibon, D. (1982). The Consultant's Forum Resistant Fits for Some Commonly Used Logistic Models with Medical Applications. *Biometrics,* **3**, 485-498.

[34] Rosen, O, Jiang, W.X. and Tanner, M.A (2000). Mixtures of marginal models. *Biometrika,* **87**, 391-404.

[35] Simpson, D. G. (1987). Minimum Hellinger distance estimation for analysis of count data. *Journal of the American Statistical Association,* **82**, 802-807.

[36] Simpson, D. G. (1989). Hellinger deviance tests: efficiency, breakdown points, and examples. *Journal of the American Statistical Association,* **84**, 107-113.

[37] Wang, P. and Puterman M. L. (1998). Mixed logistic regression models. *American Statistical Association and the International Biometric Society Journal of Agricultural, Biological, and Environmental Statistics,* **3**, 175-200.

[38] Woo, Mi-Ja. and Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of American Statistical Association*, To Appear.

[39] Woodward, W. A., Whitney P. and Eslinger, P. (1994). Minimum Hellinger distance estimation of mixture proportions. *Journal of Statistical planning and inference,* **48**, 303-319.

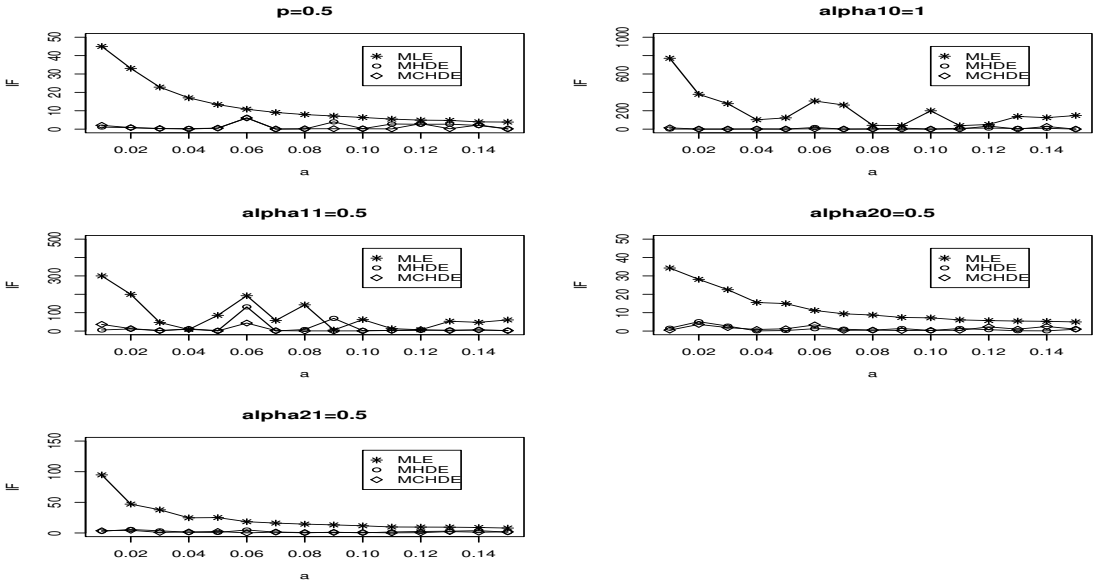

Figure 3.1: α-influence function for a two component binomial distribution with $100\alpha\%$ contaminations of a degenerate distribution δ =binomial$(30, 1)$.

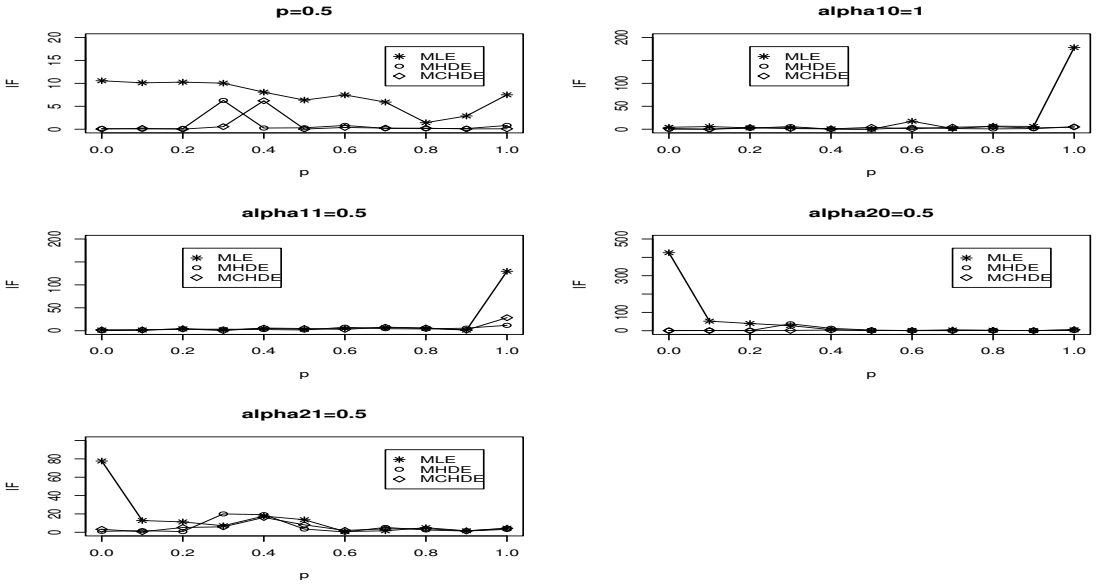

Figure 3.2: α-influence function for a two component binomial distribution with 5% contaminations of a binomial$(30, p)$ distribution.

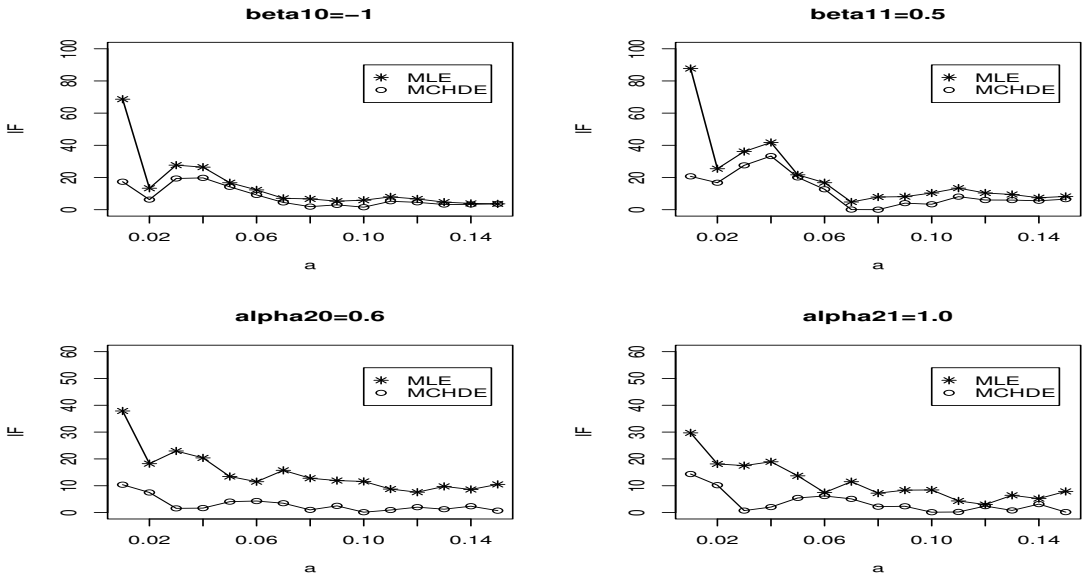

Figure 3.3: α-influence function for a ZIP distribution with $100\alpha\%$ contaminations of Poisson(20).

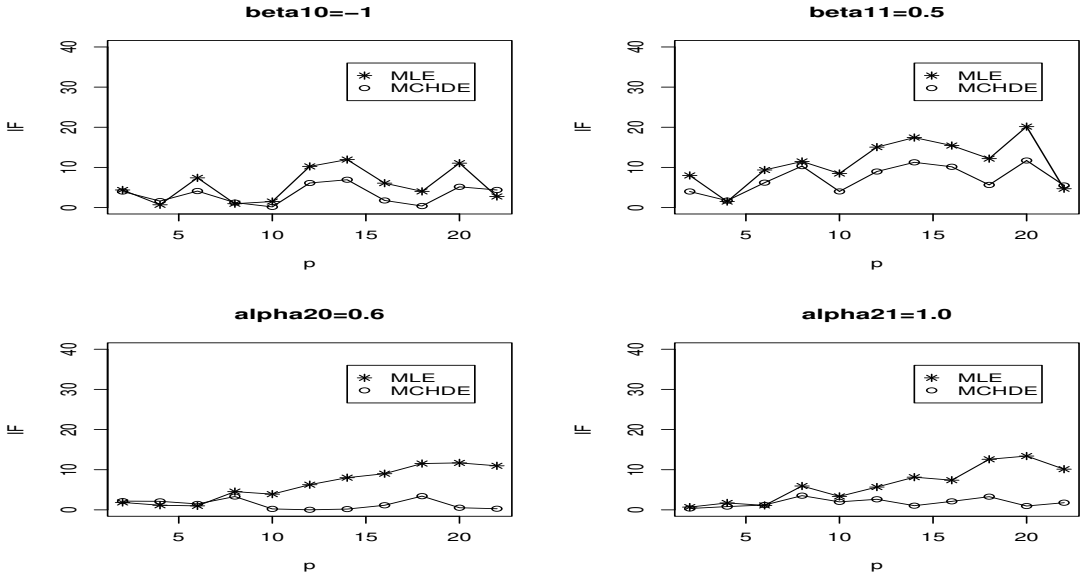

Figure 3.4: α-influence function for a ZIP distribution with 5% contaminations of value P.

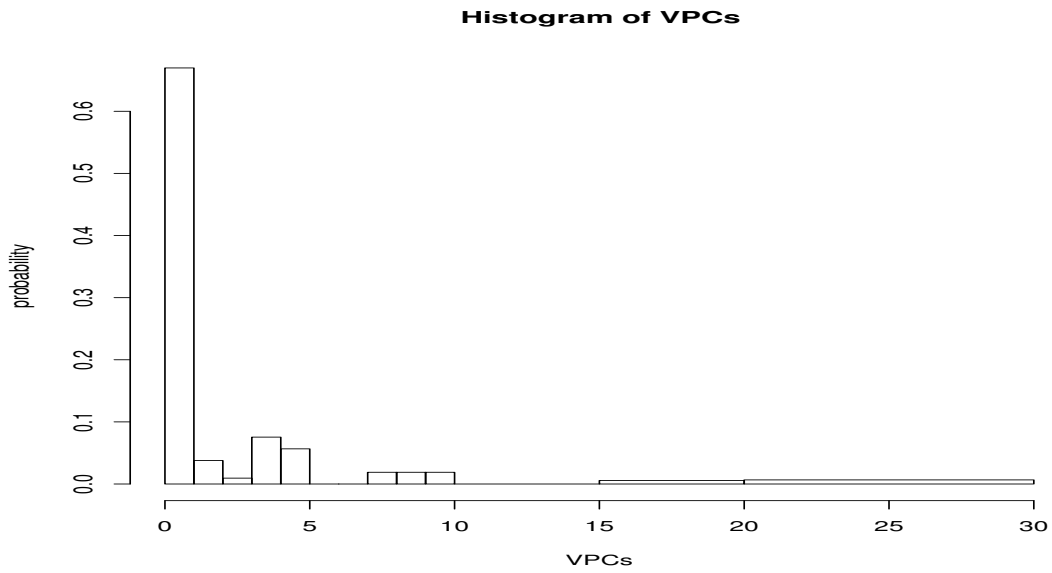

Figure 3.5: Histogram of VPCs, excluding 8 extremely large counts

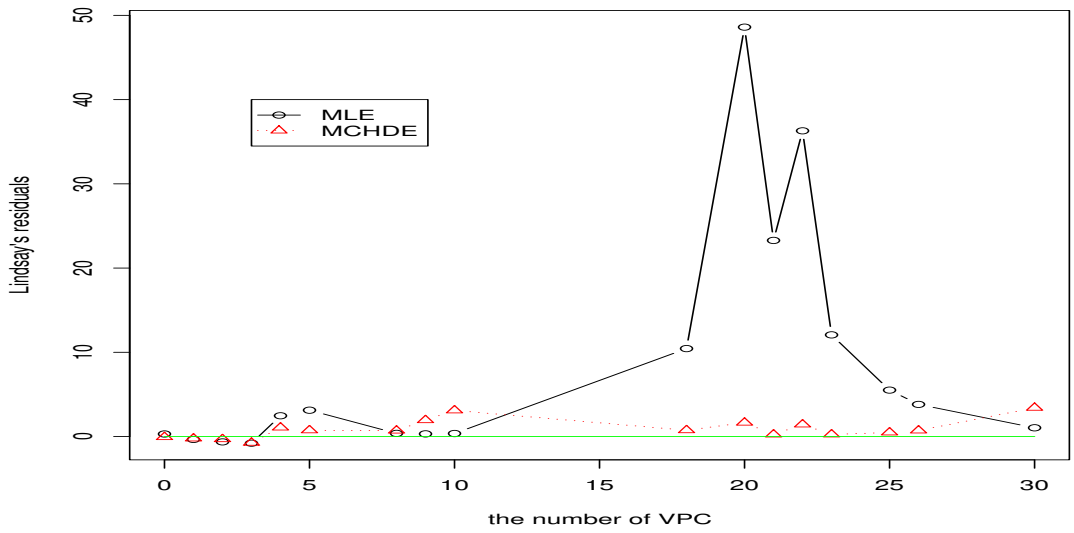

Figure 3.6: Lindsay's residuals under MCHDE and MLE for the VPC example.

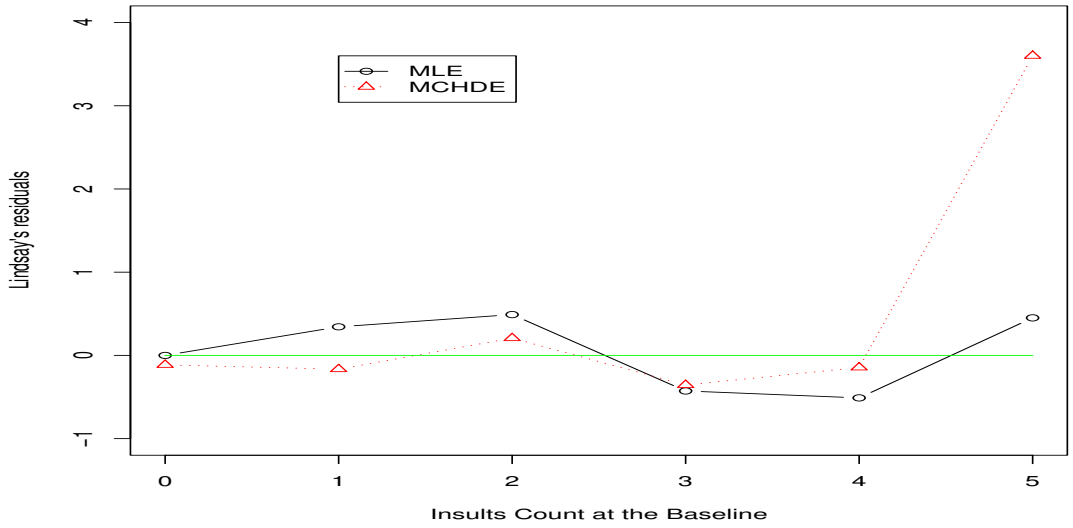

Figure 3.7: Lindsay's residuals under MCHDE and MLE for the MVPP example.

Chapter 4

Robust Estimation for Zero-Inflated Regression Models

4.1 Introduction

The goal of this paper is to develop robust estimation for ZI models, an important subclass of mixture models. In general, the issue of robustness in a finite mixture model can be confounded with the choice of the number of components in the mixture. That is, it is difficult to distinguish, both conceptually and practically, between a g_1 component mixture model with outliers and a $g_2 > g_1$ component mixture model in which additional components have been introduced corresponding to the outliers. In the ZI regression context this complication typically does not arise since the two component size of the mixture can usually be assumed, and is not part of the estimation problem. We study two types of robust estimation for ZI models: first, a minimum Hellinger distance (MHD) method based on the approach of Lu et al (2003); and second, an M-estimation type approach which is implemented via a robustified EM algorithm and which we call robust expectation-solution (RES).

Count data with many zeros and relatively large non-zero values are common in a wide variety of disciplines. This phenomenon can be handled by a two-component mixture where one of the components is taken to be a degenerate distribution, having mass one at zero. The other component is a non-degenerate distribution such as the Poisson, binomial, negative binomial or other form depending on the situation. For example, when manufacturing equipment is operating properly, defects may be nearly impossible. But when it is configured incorrectly, defects may occur

according to a Poisson($\boldsymbol{\mu}$) distribution. For such data, Lambert (1992) proposed the zero inflated Poisson (ZIP) regression model. In ZIP regression, the response vector is $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, where y_i is the observed value of the random variable Y_i. The Y_i's are assumed independent where

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

Moreover, the parameters $\boldsymbol{p} = (p_1, ..., p_N)^T$ and $\boldsymbol{\mu} = (\mu_1, ..., \mu_N)^T$ are modelled through canonical link generalized linear models (GLMs) as $\text{logit}(\boldsymbol{p}) = \boldsymbol{G\gamma}$ and $\log(\boldsymbol{\mu}) = \boldsymbol{B\beta}$, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are regression parameters, and $\boldsymbol{G}$ and $\boldsymbol{B}$ are corresponding design matrices which pertain to the probability of the zero state and the Poisson mean, respectively. The log-likelihood function for this model can be written as:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{y}) &= \sum_{y_i=0} \log\{e^{\boldsymbol{G}_i\boldsymbol{\gamma}} + \exp(-e^{\boldsymbol{B}_i\boldsymbol{\beta}})\} + \sum_{y_i>0} (y_i \boldsymbol{B}_i\boldsymbol{\beta} - e^{\boldsymbol{B}_i\boldsymbol{\beta}}) \\ &\quad - \sum_{y_i>0} \log(y_i!) - \sum_{i=1}^{N} \log(1 + e^{\boldsymbol{G}_i\boldsymbol{\gamma}}), \end{aligned} \tag{4.1.1}$$

where $\boldsymbol{B}_i$ and $\boldsymbol{G}_i$ are the ith rows of design matrices $\boldsymbol{B}$ and $\boldsymbol{G}$. Although this loglikelihood can be maximized directly, a particularly convenient method to obtain the MLE is to capitalize on the mixture structure of the problem and use the EM algorithm.

Hall (2000) extended Lambert's model and methodology to an upper bounded count situation, thereby obtaining a zero inflated binomial (ZIB) regression model. In the ZIB model, the Poisson(μ_i) component is replaced by a binomial(m_i, π_i) component and instead of modeling $\boldsymbol{\mu}$, we model $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$ through a logit link: $\text{logit}(\boldsymbol{\pi}) = \boldsymbol{B\beta}$. Here, y_i is assumed to have an interpretation as the number of successes out of m_i independent identical trials.

We can define ZI models in a general way as follows. For $i = 1, \ldots, N$, let

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i; \\ ED\{\eta_i(\boldsymbol{\beta}), a_i(\phi)\} & \text{with probability } 1 - p_i. \end{cases} \tag{4.1.2}$$

where ED is the cumulative distribution function of the non-degenerate component of the mixture. In addition, we assume ED is in the exponential dispersion family with probability density function of the form

$$g\{y_i; \eta_i(\boldsymbol{\beta}), a_i(\phi)\} = \exp\{\frac{y_i\eta_i - b(\eta_i)}{a_i(\phi)} + c(y_i)\}, \tag{4.1.3}$$

where $\eta_i = \boldsymbol{B}_i^T\boldsymbol{\beta}$. The mean and variance functions for this component are $\mu_i = \dot{b}(\eta_i)$ and $\nu(\mu_i) = \ddot{b}(\eta_i)a_i(\phi)$, respectively. In addition, $a_i(\phi)$ is the GLM weight function which, in general, may involve a scale parameter ϕ as well as known weights; in particular $a_i = 1$ for Poisson distribution and $a_i = \frac{1}{m_i}$ for binomial(m_i, π_i). We assume that the mixing probability is related to covariates in $\boldsymbol{G}_i$ via a logit model of the form $p_i = p(\boldsymbol{G}_i^T\boldsymbol{\gamma}) = \frac{\exp(\boldsymbol{G}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\gamma})}$. This mixture model implies a marginal probability density for the observed response given by $p_i I(y_i = 0)+(1-p_i)g(y_i; \boldsymbol{\beta}, \phi)$.

Both in mixture models and more generally, the robustness of the MLE has been studied extensively. It is well known that the MLE can be unstable when the data have contamination points. Recently, MHD estimation has received considerable attention as a robust alternative to ML in mixture models. Most of this literature has focused on the independent and identically distributed (iid) case (Cutler et al. 1996; Karlis and Xekalaki, 2001). Outside the iid framework, MHD has received very limited attention. However, a recent paper by Lu et al. (2003) proposes an MHD approach for finite mixtures of Poisson regression models, a class of models in which the ZIP model falls. These authors present simulation results that suggest that their approach performs very well relative to ML in the presence of outliers and/or poor separation between the mixture components.

As we will see, however, this approach has a limited domain of application because of identifiability problems that can arise when the mixing probability

depends upon covariates, which is typical in applications of ZI regression. In addition, MHD approaches are not particularly effective with respect to decreasing the effect of abnormal covariates (i.e., high leverage points). Furthermore, the asymptotics of MHD estimation in the regression context are difficult to establish.

Because of these limitations, we propose another approach, which we term robust expectation solution (RES) estimation, and which is more closely related to M-estimation. Huber (1964) proposed M-estimation as a generalization of ML in which the score function in the likelihood equation is replaced by an estimating function typically chosen to downweight the contributions of extreme observations. Recently, several authors have extended M estimation to the generalized linear model context and beyond (Preisser and Qaqish, 1999; Cantoni and Ventura, 2001; Adimari and Ventura, 2001).

The organization of this paper is as follows. Section 4.2 describes the MHD estimation method for ZI models and addresses the identifiability problem that arises with this approach, and section 4.3 presents the RES methodology. Simulation results are presented in section 4.4 to compare these methods with ML estimation, and section 4.5 gives an example to illustrate the methodology which involves data from a study of aggressive behavior among middle school-aged children. Finally, some concluding remarks are provided in section 4.6.

4.2 MHDE in ZI regression models

Let $f_n(y)$ be the empirical frequency function defined by

$$f_n(y) = N_y/n, \quad y \in U, \tag{4.2.4}$$

where N_y is the number of observations having value y, U is the sample space for Y. Let $f_{\boldsymbol{\theta}}(y|\boldsymbol{X}_i)$ denote the probability density corresponding to (4.1.2), where

$\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$. Take $\boldsymbol{X}$ as the combined matrix of all observed covariates $(\boldsymbol{G}, \boldsymbol{B})$, which we condition on and consider fixed. Lu et al. (2003) based estimation of finite mixtures of Poisson regression models on the unconditional Hellinger distance:

$$H^2(f_n, f_{\boldsymbol{\theta}}) = \int (f_n(y)^{1/2} - f_{\boldsymbol{\theta}}(y)^{1/2})^2 dy, \tag{4.2.5}$$

where $f_{\boldsymbol{\theta}}(y)$ denotes the marginal probability of observing y. In what follows, we extend Lu et al.'s approach based on $H^2(.,.)$ to the ZI context.

4.2.1 Estimation Based on Marginal Densities

For a moment consider the iid case and suppose we know the form of $f_{\boldsymbol{\theta}}(y)$. Then the MHD estimator $\hat{\boldsymbol{\theta}}$ would be the root of

$$\sum_{y \in U} \frac{f_n^{1/2}(y)}{f_{\boldsymbol{\theta}}^{1/2}(y)} \dot{f}_{\boldsymbol{\theta}}(y) = n \sum_{i=1}^{n} \frac{f_{\boldsymbol{\theta}}^{1/2}(y_i)}{f_n^{1/2}(y_i)} \frac{\partial \log f_{\boldsymbol{\theta}}(y_i)}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{4.2.6}$$

Lindsay et al. (1992) introduced Lindsay's residual function $r(y) = \frac{f_n(y)}{f_{\boldsymbol{\theta}}(y)} - 1$, as a more appropriate quantity than Pearson's residual to assess goodness of fit in a mixture context. Notice that (4.2.6) can be written as

$$\sum_{i=1}^{n} \frac{1}{\{1 + r(y_i)\}^{1/2}} \frac{\partial \log f_{\boldsymbol{\theta}}(y_i)}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

From this representation, it can be seen that MHD downweights observations that have large Lindsay's residuals in the estimating equation.

In the non-iid situation with which we are concerned, however, $f_{\boldsymbol{\theta}}(y)$ must be computed from a conditional density $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ through $f_{\boldsymbol{\theta}}(y) = \int f_{\boldsymbol{\theta}}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) d\boldsymbol{x}$. When $f_X(\boldsymbol{x})$ is unknown, or the integration is impractical due to the high dimension of the covariate vector $\boldsymbol{x}$, the objective function (4.2.6) is unavailable. Lu et al. (2003) used a consistent estimator $f_{\boldsymbol{\theta},n}(y)$ to replace $f_{\boldsymbol{\theta}}(y)$, which is defined by

$$f_{\boldsymbol{\theta},n}(y) = \frac{1}{n} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i), \quad y \in U. \tag{4.2.7}$$

While improvements on this estimator are possible, in this paper we adopt Lu et al.'s method, and extend it to the ZI models defined as (4.1.2). This leads to an MHD estimator defined as follows

$$\hat{\theta}_{\text{MHD}} = \arg\min_{\boldsymbol{\theta}\in\Theta} H^2(f_{\boldsymbol{\theta},n}, f_n),$$

i.e., it is the maximizer of

$$\rho_n(\boldsymbol{\theta}) = \sum_{y\in U} f_n^{1/2} f_{\boldsymbol{\theta},n}^{1/2}(y).$$

4.2.2 MHD Estimating Equations

Now we establish the estimating equations.

The MHD estimator of $\boldsymbol{\theta}$ is $\arg\max_{\boldsymbol{\theta}\in\Theta}\rho_n(\boldsymbol{\theta})$, where $\rho_n(\boldsymbol{\theta})$ is

$$\sum_{y\in U} f_n^{1/2}(y)\Big[\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{\exp(\boldsymbol{G}_i\boldsymbol{\gamma})}{1+\exp(\boldsymbol{G}_i\boldsymbol{\gamma})}I(y_j=0)+\frac{g\{y_i;\eta_i(\boldsymbol{\beta}),a_i(\phi)\}}{1+\exp(\boldsymbol{G}_i\boldsymbol{\gamma})}\Big\}\Big]^{1/2}. \quad (4.2.8)$$

Taking partial derivatives with respective to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, and setting them equal 0, we obtain the following set of equations:

$$\begin{aligned} 2\frac{\rho_n(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}} &= \sum_{y\in U} f_n^{1/2}(y) f_{\boldsymbol{\theta},n}^{1/2}(y)\frac{1}{f_{\boldsymbol{\theta},n}(y)}\Big[\sum_{i=1}^{n}\frac{p_i}{n}g\{y_i;\eta_i(\boldsymbol{\beta}),a_i(\phi)\}\frac{y}{a_i(\phi)}\boldsymbol{B}_i \\ &\quad -\sum_{i=1}^{n}\frac{p_i}{n}g\{y_i;\eta_i(\boldsymbol{\beta}),a_i(\phi)\}\frac{E_i(\boldsymbol{\beta})}{a_i(\phi)}\boldsymbol{B}_i\Big]=0, \end{aligned}$$

where $E_i(\boldsymbol{\beta}) = E\{y_i;\eta_i(\boldsymbol{\beta}),a_i(\phi)\}$; and

$$\begin{aligned} 2\frac{\rho_n(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}} &= \sum_{y\in U} f_n^{1/2} f_{\boldsymbol{\theta},n}^{1/2}\frac{1}{f_{\boldsymbol{\theta},n}(y)}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big\{p_i(1-p_i)I(y_i=0)\boldsymbol{G}_i \\ &\quad -p_i(1-p_i)g\{y_i;\eta_i(\boldsymbol{\beta}),a_i(\phi)\}\boldsymbol{G}_i\Big\}\Big]=0. \end{aligned}$$

Adopting the iteratively reweighted computational technique proposed by Basu and Lindsay (2004), and letting

$$v_{\boldsymbol{\theta},n}(y) = f_n^{1/2}(y) f_{\boldsymbol{\theta},n}^{1/2}(y)/\rho_n(\boldsymbol{\theta}),$$

$$w_{i1}(a_i, y, \boldsymbol{\theta}) = \frac{p_i}{n} I(y_i = 0)/f_{\boldsymbol{\theta},n}(y),$$

and

$$w_{i2}(a_i, y, \boldsymbol{\theta}) = \frac{1 - p_i}{n} g\{y_i; \eta_i(\boldsymbol{\beta}), a_i(\phi)\}/f_{\boldsymbol{\theta},n}(y),$$

we can solve for $\boldsymbol{\theta}$ by the following two update equations:

$$\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} + (\boldsymbol{B}'\boldsymbol{W}^{(l)}_{\boldsymbol{\beta}}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{W}^{(l)}_{\boldsymbol{\beta}}\boldsymbol{D}^{(l)}_{\boldsymbol{\beta}}, \tag{4.2.9}$$

$$\boldsymbol{\gamma}^{(l+1)} = \boldsymbol{\gamma}^{(l)} + (\boldsymbol{G}'\boldsymbol{W}^{(l)}_{\boldsymbol{\gamma}}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}^{(l)}_{\boldsymbol{\gamma}}\boldsymbol{S}^{(l)}_{\boldsymbol{\gamma}}. \tag{4.2.10}$$

Here $\boldsymbol{W}^{(l)}_{\boldsymbol{\beta}}$ is diagonal matrix with ith diagonal elements equal to $\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y) w_{i2} Var_i(\boldsymbol{\beta}^{(l)})(y)/a_i^2$, $Var_i(\boldsymbol{\beta}^{(l)}) = var\{y_i; \eta_i(\boldsymbol{\beta}^{(l)}), a_i(\phi)\}$, and

$$\boldsymbol{D}^{(i)}_{\boldsymbol{\beta}^{(l)}} = \Big[\sum_{y\in U}\Big\{\frac{v_{\boldsymbol{\theta}^{(l)},n}(y) w_{i2}(y)}{\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y) w_{i2}(y)} \frac{y}{a_i(\phi)}\Big\} - \frac{E_i(\boldsymbol{\beta}^{(l)})}{a_i(\phi)}\Big]/\frac{Var_i(\boldsymbol{\beta}^{(l)})}{a_i^2(\phi)}.$$

Similarly, $\boldsymbol{W}_{\boldsymbol{\gamma}^{(l)}}$ is diagonal with ith diagonal elements $\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)(w_{i1} + w_{i2}) var^{(l)}(Z_i)$, and

$$\boldsymbol{S}^{(i)}_{\boldsymbol{\gamma}^{(l)}} = \Big\{\frac{\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y) w_{i1}}{\sum_{y\in U} v_{\boldsymbol{\theta}^{(l)},n}(y)(w_{i1} + w_{i2})} - E_{\boldsymbol{\gamma}^{(l)}}(Z_i)\Big\}/var_{\boldsymbol{\gamma}^{(l)}}(Z_i),$$

where Z_i is the indicator of whether the i^{th} observations comes from the zero stage. Estimation proceeds by iterating between (4.2.9) and (4.2.10) until a convergence criterion has been obtained. Note that we have presented the equations here assuming the canonical GLM link functions have been used in the ZI regression model. This is usual, but not necessary as equations (4.2.9) and (4.2.10) can be easily modified to accommodate other valid links.

4.2.3 Asymptotics of MHD Estimation in ZI regression

Under certain regularity conditions, Simpson proved that the MHDE is consistent and asymptotically normal in the iid Poisson case. This result has been extended to finite mixtures of Poisson distributions by Karlis and Xekalaki (2001). When

regression structure is added to this framework, the asymptotics of MHD estimation become much more problematic. Lu et al (2003) argue that the results of the earlier authors should extend to the finite mixture of Poisson regression context, of which the ZIP model is a special case. These authors propose an asymptotic variance estimator of the form $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}} = \hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})^{-1}$, where $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) = \sum_{y\in U}\Big\{\hat{\dot{l}}_{\boldsymbol{\theta}}(y)\hat{\dot{l}}^{T}_{\boldsymbol{\theta}}(y)f_n(y) - \frac{\partial^2 f_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\Big\}$, and $\hat{\dot{l}}_{\boldsymbol{\theta}}(y) = \frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta},n}(y)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. Specifically, for ZIP models,

$$\begin{aligned}
\hat{\dot{l}}_{\hat{\boldsymbol{\theta}}}(y) &= \frac{1}{\hat{f}_{\hat{\boldsymbol{\theta}},n}(y)}\frac{\partial \hat{f_{\boldsymbol{\theta},n}}(y)}{\partial\boldsymbol{\theta}} \\
&= \frac{1}{\hat{f}_{\hat{\boldsymbol{\theta}},n}(y)}\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)\Big\{I(y_i=0) - f_{\text{Pois}}(y_i|\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\Big\}\boldsymbol{G}_i, \\
&\quad \frac{1}{n}\sum_{i=1}^{n}(1-\hat{p}_i)f_{\text{Pois}}(y_i|\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\Big\{y_i - \exp(\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\Big\}\boldsymbol{B}_i\Big).
\end{aligned}$$

$$\frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \begin{pmatrix} \frac{\partial^2 \hat{f_{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} & \frac{\partial^2 \hat{f_{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^T} \\ \frac{\partial^2 \hat{f_{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^T} & \frac{\partial^2 \hat{f_{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \end{pmatrix}$$

with

$$\frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} = \frac{1}{n}\sum_{i=1}^{n}\Big\{I(y_i=0) - f_{\text{Pois}}(y_i|\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\Big\}\hat{p}_i(1-\hat{p}_i)(1-2\hat{p}_i)\boldsymbol{G}_i\boldsymbol{G}_i^T,$$

$$\frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \frac{1}{n}(1-\hat{p}_i)f_{\text{Pois}}(y_i|\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\Big\{(y_i - \exp(\boldsymbol{B}_i\hat{\boldsymbol{\beta}}))^2 - \exp(\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\Big\}\boldsymbol{B}_i\boldsymbol{B}_i^T,$$

and

$$\frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^T} = (\frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}})^T = -\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)f_{\text{Pois}}(y_i|\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\boldsymbol{B}_i\boldsymbol{G}_i^T.$$

For ZIB models:

$$\begin{aligned}
\hat{\dot{l}}_{\hat{\boldsymbol{\theta}}}(y) &= \frac{1}{\hat{f}_{\hat{\boldsymbol{\theta}},n}(y)}\frac{\partial \hat{f_{\boldsymbol{\theta},n}}(y)}{\partial\boldsymbol{\theta}} \\
&= \frac{1}{\hat{f}_{\hat{\boldsymbol{\theta}},n}(y)}\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)\Big\{I(y_i=0) - f_{\text{bino}}(y_i|m_i,\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\Big\}\boldsymbol{G}_i, \\
&\quad \frac{1}{n}\sum_{i=1}^{n}(1-\hat{p}_i)f_{\text{bino}}(y_i|m_i,\boldsymbol{B}_i\hat{\boldsymbol{\beta}})\Big\{y_i - m_i\frac{\exp(\boldsymbol{B}_i\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{B}_i\hat{\boldsymbol{\beta}})}\Big\}\boldsymbol{B}_i\Big).
\end{aligned}$$

with

$$\frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} = \frac{1}{n} \sum_{i=1}^{n} \Big\{ I(y_i = 0) - f_{\text{bino}}(y_i | m_i, \boldsymbol{B}_i \hat{\boldsymbol{\beta}}) \Big\} \hat{p}_i (1 - \hat{p}_i)(1 - 2\hat{p}_i) \boldsymbol{G}_i \boldsymbol{G}_i^T,$$

$$\begin{aligned} \frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \frac{1}{n} (1 - \hat{p}_i) f_{\text{bino}}(y_i | m_i, \boldsymbol{B}_i \hat{\boldsymbol{\beta}}) \\ &\times \Big[\Big\{ y_i - m_i \frac{\exp(\boldsymbol{B}_i \hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{B}_i \hat{\boldsymbol{\beta}})} \Big\}^2 - m_i \frac{\exp(\boldsymbol{B}_i \hat{\boldsymbol{\beta}})}{\{1 + \exp(\boldsymbol{B}_i \hat{\boldsymbol{\beta}})\}^2} \Big] \boldsymbol{B}_i \boldsymbol{B}_i^T, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} &= (\frac{\partial^2 \hat{f}_{\hat{\boldsymbol{\theta}}}(y)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}})^T = -\frac{1}{n} \sum_{i=1}^{n} \hat{p}_i (1 - \hat{p}_i) f_{\text{bino}}(y_i | m_i, \boldsymbol{B}_i \hat{\boldsymbol{\beta}}) \\ &\times \Big\{ y_i - m_i \frac{\exp(\boldsymbol{B}_i \hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{B}_i \hat{\boldsymbol{\beta}})} \Big\} \boldsymbol{B}_i \boldsymbol{G}_i^T. \end{aligned}$$

However, formal proofs for asymptotic properties of MHD estimators in the non-iid set-up have not yet been established.

4.2.4 Identifiability

The issue of identifiability of finite mixture models has attracted considerable attention in the literature (Teicher, 1960; Jiang and Tanner, 1999 etc.), but most of this discussion has centered on the likelihood function and has assumed constant mixing probability in the model. When using MHD estimation via marginal densities rather than ML, however, the class of identifiable models is more restricted. In addition, ZI regression models allow a regression structure $\text{logit}(p_i) = \boldsymbol{G}_i^T \boldsymbol{\gamma}$, which invalidates many of the existing results and adds complexity to the identifiability question. For ZI models with non-constant mixing probability it is not hard to find simple non-identifiable models based on the unconditional (marginal) density. For example, let

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i; \\ \text{Poisson}(\mu) & \text{with probability } 1 - p_i. \end{cases}$$

and suppose that p_i depends on X_i where X is binary variable,

$$X_i = \begin{cases} 0, & i = 1, 2, .., \frac{n}{2}; \\ 1, & i = \frac{n}{2} + 1, ..., n. \end{cases}$$

It follows that $p_i = \text{expit}(\gamma_1) \equiv \text{P}_1$ if $i = 1, 2, ..., n/2$, otherwise $p_i = \text{expit}(\gamma_2) \equiv \text{P}_2$. The marginal density is

$$f_{\theta,n}(y) = \frac{1}{2}(P_1 + P_2)I(y = 0) + \{1 - \frac{1}{2}(P_1 + P_2)\}\frac{e^{-\mu}\mu^y}{y!},$$

which is clearly not identifiable.

Necessary and sufficient identifiability conditions in the class of models defined by (4.2.8) are difficult to establish. However, in our simulation studies we encountered near singular Hessian matrices for the MHD criterion for many of the models we considered that have non-constant mixing probability. In the next section we consider the RES approach which does not lead to the same identifiability problems encountered with MHD estimation.

4.3 The Robust Expectation Solution Approach for ZI Regression

4.3.1 The RES Algorithm

In ZI models, as in other mixture models, the EM algorithm is a particularly convenient approach for computing MLEs (e.g., Lambert, 1992; Hall, 2000). This algorithm is set up by introducing "missing data" into the problem. In particular, suppose we knew which zeros came from the degenerate distribution (the zero state); and which came from the non-degenerate distribution (the non-zero state); that is, suppose we could observe $z_i = 1$ when y_i is from zero state, and $z_i = 0$ when y_i is from the non-zero state. Then the log-likelihood for the complete data $(\boldsymbol{y}, \boldsymbol{z})$ would be:

$$\ell^c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z}) \quad = \quad \sum_{i=1}^{N} \left\{ z_i \boldsymbol{G}_i^T \boldsymbol{\gamma} - \log(1 + e^{\boldsymbol{G}_i^T \boldsymbol{\gamma}}) \right\}$$

$$
\begin{aligned}
& + \sum_{i=1}^{N} (1 - z_i) \Big\{ \frac{y_i \boldsymbol{B}_i^T \boldsymbol{\beta} - b(\boldsymbol{B}_i^T \boldsymbol{\beta})}{a_i(\phi)} + c(y_i) \Big\} \\
= \quad & \ell^c_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z}) + \ell^c_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{z}),
\end{aligned}
$$

where $\boldsymbol{z} = (z_1, \ldots, z_N)^T$. This log-likelihood is easy to maximize, because $\ell^c_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z})$ and $\ell^c_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{z})$ can be maximized separately with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively. With the EM algorithm, the log-likelihood of model (4.1.2) is maximized iteratively by alternating between estimating z_i by its conditional expectation under the current estimates of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ (E step) and then, with the z_i fixed at their expected values from the E step, maximizing $\ell^c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z})$ (M step), until the estimated $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ converges and iteration stops.

In more detail, the EM algorithm begins with starting values $(\boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)})$ and proceeds iteratively via the following three steps.

E step. Estimate z_i by its conditional mean $z_i^{(r)}$ under the current estimates $\boldsymbol{\gamma}^{(r)}$ and $\boldsymbol{\beta}^{(r)}$

$$
\begin{aligned}
z_i^{(r)} &= P(\text{zero state} | y_i, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\beta}^{(r)}) \\
&= \frac{P(y_i | \text{zero state}) P(\text{zero state})}{P(y_i | \text{zero state}) P(\text{zero state}) + P(y_i | \text{Poisson state}) P(\text{Poisson state})}.
\end{aligned}
$$

M Step for $\boldsymbol{\gamma}$. Find $\boldsymbol{\gamma}^{(r+1)}$ by maximizing $\ell^c_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{z}^{(r)})$. This can be accomplished by fitting a binomial logistic regression of $\boldsymbol{z}^{(r)}$ on design matrix $\boldsymbol{G}$ with binomial denominator equal one.

M Step for $\boldsymbol{\beta}$. Find $\boldsymbol{\beta}^{(r+1)}$ by maximizing $\ell^c_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{z}^{(r)})$. It is equivalent to solve the estimating equation

$$
\sum_{i=1}^{n} (1 - z_i^{(r)}) \frac{y_i - \mu(\boldsymbol{B}_i^T \boldsymbol{\beta})}{\nu(\boldsymbol{B}_i^T \boldsymbol{\beta})} \frac{\partial \mu(\eta_i)}{\partial \eta_i} \Big|_{\eta_i = \boldsymbol{B}_i^T \boldsymbol{\beta}} \boldsymbol{B}_i = 0. \tag{4.3.11}
$$

In the RES approach, We propose to replace the estimating equation (4.3.11) from the M step of the EM algorithm with a robustified estimating equation. Essentially, we follow Cantoni and Ronchetti in the specific form of that estimating func-

tion. Specifically, we suggest that $\boldsymbol{\gamma}^{(r+1)}$ be found by solving

$$\frac{1}{n}\sum_{i=1}^{n}(1-Z_i^{(r)})\omega(B_i)\Big\{\psi_c\Big(\frac{y_i-\mu(\boldsymbol{B}_i\boldsymbol{\beta}^{(r)})}{\sqrt{\nu(\boldsymbol{B}_i\boldsymbol{\beta}^{(r)})}}\Big)-o_i(\boldsymbol{\beta}^{(r)},c)\Big\}\frac{\frac{\partial\mu(\boldsymbol{B}_i\boldsymbol{\beta}^{(r)})}{\partial\eta_i}\boldsymbol{B}_i}{\sqrt{\nu(\boldsymbol{B}_i\boldsymbol{\beta}^{(r)})}}=0, \quad (4.3.12)$$

where

$$\psi_c(x)=\begin{cases} x, & |x|<c; \\ c, & \text{otherwise .} \end{cases} \quad (4.3.13)$$

and

$$o_i(\boldsymbol{\beta},c)=E\psi_c\Big(\frac{y_i-\mu(\boldsymbol{B}_i\boldsymbol{\beta})}{\sqrt{\nu(\boldsymbol{B}_i\boldsymbol{\beta})}}\Big).$$

Here, $\omega(\boldsymbol{B}_i)$ is a function to downweight large leverage points. A simple choice for $\omega(\boldsymbol{B}_i)$ is $\sqrt{1-h_i}$, where h_i is the ith diagonal element of $\boldsymbol{H}=\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}$. More sophisticated choices for $\omega(.)$ based on, e.g., Mahalanobis distance, are available (Cantoni and Ronchetti, 2001). The choice of c controls the trade-off between robustness and efficiency. Selecting a value for c is not an easy problem, but in a simpler context Cantoni and Ronchetti (2001) have derived connections between the value of c and the asymptotic distribution of quasi-likelihood ratio test statistics in a neighborhood of the model.

The ML EM algorithm involves iteratively fitting a GLM with a weighted version of the standard ML estimating equations for a GLM, where the weights are recomputed at each iteration in the E step. In the RES algorithm we instead iteratively fit a GLM with a weighted version of Cantoni & Ronchetti's M estimating equations for fit the GLM. In fact, our S step can be performed by using their S-PLUS routine for robust estimation in GLMs. Alternatively, a Newton-Raphson approach can be implemented. For ZI Poisson and ZI binomial models, we can obtain a closed form expression for the o_i given c. Let $j_1=\lfloor\mu_i-c\sqrt{\nu^{1/2}(\mu_i)}\rfloor$ and $j_2=\lfloor\mu_i+c\sqrt{\nu^{1/2}(\mu_i)}\rfloor$ where $\lfloor r\rfloor$ represent the integer no larger then r. Then in the Poisson case, $o_i=c\{P(Y_i>j_2+1)-P(Y_i\le j_1)\}+\frac{\mu_i}{\sqrt{\mu_i}}\{P(Y_i=j_1)-P(Y_i=j_2)\}$

and in the binomial case $o_i = c\{P(Y_i > j_2 + 1) - P(Y_i \leq j_1)\} + \frac{\mu_i}{\nu^{1/2}(\mu_i)}\{P(j_1 \leq \tilde{Y}_i \leq j_2 - 1) - P(j_1 + 1 \leq Y_i \leq j_2)\}$, where $\tilde{Y}_i \sim B(m_i - 1, p_i)$.

For the cases of primary interest, the ZIP and ZIB regression models, the non-degenerate component of the mixture involves a known scale parameter $\phi = 1$. The most prominent example when ϕ is unknown occurs for the ZI normal regression model, where ϕ corresponds to the error variance. In that case, this parameter must also be estimated in the S step, which can be done by a modification of the standard M-estimator of the error variance in the linear regression problem. In particular, we suggest the estimator

$$\phi^{(r+1)} = \frac{1}{\sum_{i=1}^{n} a(1 - z_i^{(r)})\omega(\boldsymbol{B}_i) - p} \sum_{i=1}^{n} (1 - z_i^{(r)})\omega(\boldsymbol{B}_i)\psi_c^2\Big(\frac{y_i - \mu(\boldsymbol{B}_i^T \boldsymbol{\beta}^{(r)})}{\sigma^{(r)}}\Big)\phi^{(r)},$$

where $a = E\psi_c^2$ and p is the dimension of $\boldsymbol{\beta}$ (cf. Huber, 2004, equation (7.8)). Note that the RES algorithm for the ZI normal regression model is discussed here mainly for completeness sake; it is more easily handled by recognizing that for a continuous non-degenerate distribution, the mixture component to which each observation belongs is easily identified. Therefore, the likelihood of the model factors into terms corresponding to the zero and non-zero responses, and the entire model can be fit by separately modeling (i) the non-zero observations with robust regression and (ii) a vector of indicators for whether or not each observation is zero with logistic regression.

4.3.2 Starting Values

To facilitate convergence in the RES algorithm , it is necessary to start with a good initial value. Rousseeuw (1984) suggested the least median of squares (LMS) estimator or least trimmed squares (LTS) as a high-breakdown starting value for the iterative computation of M-estimators. Rosseeuw (1984) gives the definitions of

LMS and LTS for linear regression models. The former is defined as

$$\hat{\theta}^{LMS} = \arg\min_{\theta} \text{median}_i (r_i)^2,$$

where r_i is the Pearson residual. Similarly, LTS is defined as

$$\hat{\theta}^{LTS} = \arg\min_{\theta} \sum_{i=1}^{h} r_{(i)}^2$$

where $r_{(i)}$, $i = 1, \ldots, n$ are the ordered Pearson residuals and h determines the amount of trimming in the estimator.

Christmann (1998) extends LMS and LTS to categorical regression models. The idea is to transform the discrete data model into an approximately linear regression with normal errors by the delta method, and use the LMS or LTS methods for the transformed data set.

We propose using LTS for the positive data (e.g., the positive Poisson or binomial data) to get an initial guess for $\boldsymbol{\beta}$. Let P be a random variable with mean $\mu(\boldsymbol{B\beta})$ and variance $\nu(\boldsymbol{B\beta})$. Define

$$\tilde{y} = \frac{\mu^{-1}(P)}{g^*(P)\sqrt{\nu(\mu^{-1}(P))}},$$

and

$$\tilde{B} = \frac{1}{g^*(P)\sqrt{\nu(\mu^{-1}(P))}} B,$$

where $g^*(P) = \frac{\partial(\mu^{-1}(P))}{\partial P}$. Then $\tilde{Y}$ follows approximately a linear regression model with covariates $\tilde{B}$.

For the ZIP model, P is a truncated Poisson(λ) random variable with $\log(\lambda) = \boldsymbol{B\beta}$,

$$\mu(\boldsymbol{B\beta}) = \frac{\lambda}{1 - \exp(-\lambda)},$$

$$\nu(\boldsymbol{B\beta}) = \frac{\lambda\{1 - \exp(-\lambda) - \lambda\exp(-\lambda)\}}{(1 - \exp(-\lambda))^2},$$

and

$$g^* = \frac{(1-\exp(-\lambda))^2}{\lambda\{1-\exp(-\lambda)-\lambda\exp(-\lambda)\}}.$$

For the ZIB model, P is a truncated binomial(m,p) random variable divided by m with $\text{logit}(p) = \boldsymbol{B}\boldsymbol{\beta}$, so we then have

$$\mu(\boldsymbol{B}\boldsymbol{\beta}) = \frac{p}{1-(1-p)^m},$$

$$\nu(\boldsymbol{B}\boldsymbol{\beta}) = \frac{\frac{p(1-p)}{m}\{1-(1-p)^m - mp(1-p)^{m-1}\}}{\{1-(1-p)^m\}^2},$$

and

$$g^* = \frac{\{1-(1-p)^m\}^2}{\frac{p(1-p)}{m}\{1-(1-p)^m - mp(1-p)^{m-1}\}}.$$

The method is applied by setting $P = \mu(\boldsymbol{B}\boldsymbol{\beta})$ and solving for $\mu^{-1}(P)$. In both cases, $\mu^{-1}(P)$ and $g^*(P)$ can only be calculated numerically. We can take the estimator of $\boldsymbol{\beta}$ through LTS for the transformed data as our initial guess. For initial values for the mixing probability model we may follow Lambert's suggestion and use

$$\hat{P}_0 = \frac{\sum_{i=1}^n \{I(y_i=0) - e^{-\exp(\boldsymbol{B}_i\boldsymbol{\beta}^{(0)})}\}}{n},$$

and

$$\hat{P}_0 = \frac{\sum_{i=1}^n \left[I(y_i=0) - \left\{\frac{\exp(\boldsymbol{B}_i\boldsymbol{\beta}^{(0)})}{1+\exp(\boldsymbol{B}_i\boldsymbol{\beta}^{(0)})}\right\}^m\right]}{n},$$

the observed average probability of an excess 0 for ZIP and ZIB respectively.

4.3.3 Influence Function

The influence function (IF) is a useful and popular tool for quantifying the degree of robustness of a statistic by measuring the potential effect of an additional observation. The classical ML estimating equations for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ can be written as $\frac{1}{n}\sum_{i=1}^n \{E_{\boldsymbol{\theta}}(z_i|y_i) - \frac{\exp(\boldsymbol{G}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\gamma})}\}\boldsymbol{G}_i = \boldsymbol{0}$, and $\frac{1}{n}\sum_{i=1}^n \{1 - E_{\boldsymbol{\theta}}(z_i|y_i)\}\frac{y_i - \mu(\boldsymbol{B}_i^T\boldsymbol{\beta})}{a_i(\phi)}\boldsymbol{B}_i = \boldsymbol{0}$, where the expectation is with respect to z_i given y_i. The influence function (IF)

of $\hat{\boldsymbol{\beta}}_{MLE}$, the MLE with respect to $\boldsymbol{\beta}$ for the ZI model, quantifies the influence of one additional observation y_j corresponding to the random variable Y_j drawn from model (4.1.2). This function is given by

$$IF_{MLE}(y_j) = \frac{\{1 - E_{\boldsymbol{\theta}}(z_j|Y_j)\}\{y_j - \mu(\boldsymbol{B}_j^T\boldsymbol{\beta})\}\boldsymbol{B}_j}{-E_{\boldsymbol{\theta}}\Big(\frac{\partial}{\partial\boldsymbol{\beta}}\Big[\{1 - E_{\boldsymbol{\theta}}(z_j|Y_j)\}\{Y_j - \mu(\boldsymbol{B}_j^T\boldsymbol{\beta})\}\Big]\boldsymbol{B}_j\Big)}. \tag{4.3.14}$$

As can be seen in (4.3.14), the influence of an outlier on the ML estimator is proportional to the score function and is, therefore, unbounded in general. The estimating functions underlying the RES method are

$$\frac{1}{n}\sum_{i=1}^{n}\{E_{\boldsymbol{\theta}}(z_i|y_i) - \frac{\exp(\boldsymbol{G}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{G}_i^T\boldsymbol{\gamma})}\}\boldsymbol{G}_i = \mathbf{0}. \tag{4.3.15}$$

$$\frac{1}{n}\sum_{i=1}^{n}(1-E_{\boldsymbol{\theta}}(z_i|y_i))\omega(\boldsymbol{B}_i)\Big\{\psi_c\Big(\frac{y_i - \mu(\boldsymbol{B}_i^T\boldsymbol{\beta})}{\sqrt{\nu(\boldsymbol{B}_i^T\boldsymbol{\beta})}}\Big) - o_i(\boldsymbol{\beta}, c)\Big\}\frac{\partial\mu(\eta_i)/\partial\eta_i}{\sqrt{\nu(\boldsymbol{B}_i^T\boldsymbol{\beta})}}\boldsymbol{B}_i = \mathbf{0}. \tag{4.3.16}$$

Let $\Psi(\boldsymbol{\beta}, y_i) = \omega(\boldsymbol{B}_i)\Big\{\psi_c\Big(\frac{y_i-\mu(\boldsymbol{B}_i^T\boldsymbol{\beta})}{\sqrt{\nu(\boldsymbol{B}_i^T\boldsymbol{\beta})}}\Big) - o_i(\boldsymbol{\beta}, c)\Big\}\frac{\partial\mu(\eta_i)/\partial\eta_i}{\sqrt{\nu(\boldsymbol{B}_i^T\boldsymbol{\beta})}}\boldsymbol{B}_i$; then the IF of $\hat{\boldsymbol{\beta}}_{RES}$ is

$$IF_{RES}(y_j) = \frac{\{1 - E_{\boldsymbol{\theta}}(z_j|Y_j)\}\Psi(\boldsymbol{\beta}, y_j)}{-E_{\boldsymbol{\theta}}\Big(\frac{\partial}{\partial\boldsymbol{\beta}}\Big[\{1 - E_{\boldsymbol{\theta}}(z_j|Y_j)\}\Psi(\boldsymbol{\beta}, Y_j)\Big]\Big)}. \tag{4.3.17}$$

The IF of RES estimator is bounded because the score equation Ψ is bounded, so the supremum of the absolute value is bounded. Therefore, $\hat{\boldsymbol{\beta}}_{RES}$ is so called *B-robust* (Hampel et al., 1981).

4.3.4 Asymptotics

For simplicity, we combine (4.3.15) and (4.3.16) and rewrite them as one equation,

$$U(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{\theta}}\{\boldsymbol{s}_i(y_i, z_i, \boldsymbol{\theta})|y_i\} = \mathbf{0} \tag{4.3.18}$$

with the expectation respect to z_i given y_i. Here, $\boldsymbol{s}_i(y_i, z_i, \boldsymbol{\theta}) = (\boldsymbol{s}_{i1}(y_i, z_i, \boldsymbol{\theta})^T$, $\boldsymbol{s}_{i2}(y_i, z_i, \boldsymbol{\theta})^T)^T$, with $\boldsymbol{s}_{i1}(y_i, z_i, \boldsymbol{\theta}) = \{z_i - \text{expit}(\boldsymbol{G}_{\text{i}}^{\text{T}}\boldsymbol{\gamma})\}\boldsymbol{G}_{\text{i}}$, and

$$\boldsymbol{s}_{i2}(y_i, z_i, \boldsymbol{\theta}) = (1 - z_i)\omega(\boldsymbol{B}_i)\{\psi_c(\frac{Y_i - \mu(\boldsymbol{B}_i^T\boldsymbol{\beta})}{\sqrt{\nu(\boldsymbol{B}_i^T\boldsymbol{\beta})}}) - o_i(\boldsymbol{\beta}, c)\}\frac{\partial\mu(\eta_i)/\partial\eta_i}{\sqrt{\nu(\boldsymbol{B}_i^T\boldsymbol{\beta})}}\boldsymbol{B}_i.$$

Rosen et al. (2001) show that under certain regularity conditions, if there exists a point $\hat{\boldsymbol{\theta}}$ such that $\lim_{r\to\infty} \boldsymbol{\theta}^{(r)} = \hat{\boldsymbol{\theta}}$ where $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ is a sequence generated by the expectation solution algorithm, then $\hat{\boldsymbol{\theta}}$ satisfies: (1) $U(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) = \mathbf{0}$, and (2) $U(\boldsymbol{\theta}; \boldsymbol{y})$ is an unbiased estimating function, satisfying $E_{\boldsymbol{\theta}}\{U(\boldsymbol{\theta}; \boldsymbol{y})\} = \mathbf{0}$ for all $\boldsymbol{\theta}$.

The conditions of the above theory are easily verified for RES algorithm (see Appendix on page 117). Therefore, if the RES algorithm converges, it converges to a solution to $\hat{\boldsymbol{\theta}}$ of an unbiased estimating equation. Moreover under mild regularity conditions (e.g., Carroll, Ruppert and Stefanski, 1995, §A.3), the RES estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}^T, \hat{\boldsymbol{\beta}}^T)^T$ is consistent: $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$, almost surely; and asymptotically normal: $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{V})$. Here, $\boldsymbol{V} = \boldsymbol{M}^{-1}\boldsymbol{Q}\boldsymbol{M}^{-1}$ where $\boldsymbol{Q} = E\{U(\boldsymbol{\theta}; \boldsymbol{y})U(\boldsymbol{\theta}; \boldsymbol{y})^T\}$ and $\boldsymbol{M} = -E\{\dot{U}(\boldsymbol{\theta}; \boldsymbol{y})\}$, where $\dot{\boldsymbol{U}} = \frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{U}$. The asymptotic variance of $\hat{\boldsymbol{\theta}}$ can be estimated by $V_n = M_n^{-1} Q_n M_n^{-1}$ at $\hat{\boldsymbol{\theta}}$, where $M_n = -\frac{1}{n}\sum_{i=1}^n E_{\boldsymbol{\theta}}\{\dot{s}_i(y_i, z_i, \boldsymbol{\theta})|y_i\}$, and $Q_n = \frac{1}{n}\sum_{i=1}^n [E_{\boldsymbol{\theta}}\{s_i(y_i, z_i, \boldsymbol{\theta})|y_i\} \times E_{\boldsymbol{\theta}}\{s_i(y_i, z_i, \boldsymbol{\theta})|y_i\}^T]$.

4.4 Simulations

The aim of the following simulations is to assess performance under model misspecification and/or poor separation of the mixture components. Three separate simulation studies were conducted in which we compare the ML, MHD and RES estimation methods in the context of ZI regression with outliers. Because of non-identifiability problems with the MHD method, we restrict attention to the case of constant mixing probability in study 1, and consider only the RES and ML approaches for non-constant mixing probability in simulation studies 2 and 3. In studies 1 and 2, the specific form of model violation considered is the presence of outlying values in the response, whereas in study 3 we consider outlying values in the explanatory variables (high leverage points). In all three simulation studies, both

ZIP and ZIB models are considered, as well as two different degrees of separation between the mixture components and two sample sizes.

All simulations involving the ZIB model were patterned after the structure of a data set reported in Hall and Berenhaut (2002) regarding alligator egg hatch rates. The data came from a study in which alligator egg nests were monitored over several consecutive years at two sites in Florida. Therefore, data were simulated from ZIB models of the form

$$Y_i \sim \begin{cases} 0, & \text{with probability } p_i; \\ \text{Binomial}(m_i, \mu_i), & \text{with probability } 1 - p_i. \end{cases}$$

where

$$\begin{aligned} \mu_i &= \text{expit}(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}), \\ p_i &= \begin{cases} p, & \text{in simulation study 1,} \\ \text{expit}(\gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{3i}), & \text{in simulation studies 2, 3.} \end{cases} \end{aligned}$$

x_{ji} is an indicator for whether observation i was taken from site j, $j = 1, 2$, x_{3i} is the year of the ith observation, and x_{4i} is a random variable generated from a uniform(0,1) distribution.

The simulations involving the ZIP model were patterned after a data set presented by Ridout et al. (1998) which contains the results of an experiment in which apple tree roots were propagated under eight different treatments corresponding to the 2×4 combinations of chemical medium and light level. The covariates here were defined similarly to those in the ZIB model above. Let x_1, x_2 be indicators for the two propagation media, let x_3 take values 1–4 corresponding to the 4 increasing light levels, and again let $x_4 \sim \text{uniform}(0, 1)$. Then the model under which the data were generated takes the form

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

where

$$\mu_i = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}),$$

$$p_i = \begin{cases} p, & \text{in simulation study 1,} \\ \text{expit}(\gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{3i}), & \text{in simulation studies 2, 3.} \end{cases}$$

The data were generated under various settings of the model parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, chosen to correspond to low versus high levels of zero inflation combined with low versus high levels of separation between the mixture components. For every parameter setting, 500 data sets were generated from the model, with outliers added depending upon the particular model type and data contamination scheme under consideration. Bias, mean square error (MSE), and size of the Wald test for equality with the true value were calculated for each model parameter. In addition, we provided the MSE for $\zeta \equiv \frac{1}{n}\sum_{i=1}^{n}(1-p_i)\mu_i$, the average marginal mean according to the model, to assess the performance of the different estimation methods.

4.4.1 Study 1: ZI Regression with Constant p and Outliers in y

In study 1, we compare ML, MHD and RES estimation methods for data generated from the ZI regression model with constant p and contamination in y. For each data generated set, 10% of the responses were selected at random and replaced by outliers. In the ZIB model, these responses were replaced by m, the binomial denominator, and in the ZIP case, they were replaced by $y + 15$. The parameters p and $\boldsymbol{\beta}$ were specified as list in Tables 4.1-4.8. The regression parameter $\boldsymbol{\beta}$ was chosen to make the non-degenerate component's mean either large or small, which we refer to as the "well-separated" (Tables 4.3, 4.4, 4.7, and 4.8) and "poorly separated" (Tables 4.1, 4.2, 4.5, and 4.6) cases, respectively. Two values, 0.1 and 0.3, were considered for the mixing probability p, corresponding to low and moderate levels of zero inflation. We also set $n = 64$ and $n = 200$ to investigate the effect of sample size. In each table of results (Tables 4.1-4.8), we report the bias, MSE, and size of the Wald test of equality to the true value, for each parameter to assess the performance of the three methods.

Tables 4.1 and 4.2 are the results for poorly separated ZIB models. As expected, both RES and MHDE exhibit less bias and smaller MSE for ζ and for all components of $\boldsymbol{\beta}$ than MLE, but perform similarly with respect to p. RES has better size (closer to 0.05) than MHDE and MLE. In the well-separated ZIB cases (Tables 4.3 and 4.4), again, both RES and MHDE generally yield estimators of $\boldsymbol{\beta}$ and ζ with smaller bias and MSE than the MLEs, but produced comparable estimators for p.

Tables 5 through 8 repeat the simulations above but in the ZIP context and exhibit a similar pattern to the ZIB results. Both RES and MHDE perform better than MLE in bias, MSE of $\boldsymbol{\beta}$ and ζ. In addition, when $p = 0.1$, both of them are consistently better in p, but very close to MLE when $p = 0.3$.

From the results above, we conclude that in the presence of outliers in y: (1) Both RES and MHDE improved dramatically upon ML estimation, which had unacceptable levels of bias and MSE in the cases we examined. (2) With respect to $\boldsymbol{\beta}$, the presence of outliers severely inflated the size of Wald tests under ML estimation and, in most cases, deflated the size of these tests under MHD estimation. With few exceptions the MHD method led to sizes of 0.01 or 0.00 for the $\boldsymbol{\beta}$ parameters. In contrast, the RES approach typically led to mildly inflated test sizes, which, in many cases, were close to nominal. With respect to p, the size comparisons among the three methods are less consistent. Generally speaking, the size of Wald tests for p were strongly inflated under ML estimation, but much less severely affected under RES and MHD estimation, which performed roughly the same. (3) As expected, increasing sample size from $n = 64$ to $n = 200$ generally has the effect of decreasing bias and MSE for all parameters across all three methods. This effect may have been limited somewhat by the fact that the proportion of outliers stayed constant, which also may explain the lack of a clear effect of sample size on Wald test size.

Table 4.1: ZIB poorly separated data with constant zero state probability $p = 0.1$, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$p = 0.1$	-0.013	0.0013	0.01	0.006	0.0017	0.07	-0.013	0.0013	0.18
$\beta_1 = -1.0$	0.050	0.0122	0.07	-0.004	0.0021	0.00	0.298	0.0976	0.89
$\beta_2 = -0.8$	0.056	0.0142	0.07	0.097	0.0182	0.01	0.259	0.0764	0.80
$\beta_3 = -0.1$	0.028	0.0026	0.08	0.047	0.0036	0.00	0.165	0.0287	0.99
$\beta_4 = 0.1$	-0.073	0.0261	0.05	-0.051	0.0038	0.00	-0.378	0.1592	0.83
ζ		1.5118			1.0902			17.8055	
$n = 200$									
$p = 0.1$	-0.010	0.0005	0.04	0.003	0.0005	0.02	-0.010	0.0005	0.12
$\beta_1 = -1.0$	0.071	0.0097	0.09	-0.004	0.0005	0.00	0.419	0.1787	1.00
$\beta_2 = -0.8$	0.071	0.0089	0.17	-0.041	0.0046	0.00	0.374	0.1431	1.00
$\beta_3 = -0.1$	-0.003	0.0004	0.01	0.019	0.0009	0.01	0.019	0.0007	0.12
$\beta_4 = 0.1$	-0.005	0.0087	0.04	-0.022	0.0010	0.00	-0.092	0.0144	0.22
ζ		0.8132			0.2774			14.6069	

Table 4.2: ZIB poorly separated data with constant zero state probability $p = 0.3$, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$p = 0.3$	-0.039	0.0044	0.07	0.018	0.0047	0.10	-0.038	0.0044	0.10
$\beta_1 = -1.0$	0.062	0.0165	0.06	-0.003	0.0025	0.00	0.372	0.1545	0.89
$\beta_2 = -0.8$	0.073	0.0019	0.08	-0.101	0.0185	0.00	0.322	0.1205	0.83
$\beta_3 = -0.1$	0.039	0.0040	0.08	0.0497	0.0037	0.00	0.194	0.0404	1.00
$\beta_4 = 0.1$	-0.098	0.0377	0.06	-0.053	0.0038	0.00	-0.451	0.2399	0.81
ζ		2.3027			1.3197			20.6134	
$n = 200$									
$p = 0.3$	-0.030	0.0017	0.18	0.010	0.0011	0.00	-0.030	0.0017	0.20
$\beta_1 = -1.0$	0.086	0.0143	0.08	-0.004	0.0009	0.00	0.500	0.2560	1.00
$\beta_2 = -0.8$	0.085	0.0122	0.15	-0.056	0.0063	0.00	0.450	0.2086	1.00
$\beta_3 = -0.1$	0.002	0.0006	0.01	0.026	0.0013	0.00	0.028	0.0015	0.21
$\beta_4 = 0.1$	-0.006	0.0133	0.06	-0.029	0.0014	0.00	-0.099	0.0222	0.26
ζ		1.3672			0.3665			16.9054	

Table 4.3: ZIB well separated data with constant zero state probability $p = 0.1$, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$p = 0.1$	-0.012	0.0015	0.00	0.010	0.0023	0.18	-0.012	0.0015	0.19
$\beta_1 = 0.5$	0.044	0.0099	0.05	0.008	0.0016	0.00	0.139	0.0275	0.33
$\beta_2 = 0.7$	0.046	0.0109	0.03	-0.103	0.0166	0.00	0.125	0.0231	0.26
$\beta_3 = -0.1$	0.033	0.0027	0.13	0.056	0.0041	0.00	0.086	0.0088	0.67
$\beta_4 = 0.1$	-0.077	0.0246	0.04	-0.058	0.0042	0.00	-0.192	0.0554	0.37
ζ		3.1062			3.1650			8.0338	
$n = 200$									
$p = 0.1$	-0.0101	0.0005	0.05	0.003	0.0005	0.03	-0.0103	0.0005	0.14
$\beta_1 = 0.5$	0.065	0.0070	0.05	0.002	0.0005	0.00	0.206	0.0446	0.97
$\beta_2 = 0.7$	0.066	0.0082	0.17	-0.046	0.0045	0.00	0.191	0.0403	0.93
$\beta_3 = -0.1$	0.001	0.0005	0.01	0.024	0.0011	0.00	0.194	0.0004	0.04
$\beta_4 = 0.1$	-0.014	0.0066	0.02	-0.027	0.0012	0.00	-0.050	0.0091	0.14
ζ		1.5805			0.8680			5.5705	

Table 4.4: ZIB well separated data with constant zero state probability $p = 0.3$, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$p = 0.3$	-0.025	0.0032	0.04	0.043	0.0063	0.10	-0.025	0.0033	0.06
$\beta_1 = 0.5$	0.065	0.0147	0.06	0.012	0.0026	0.00	0.185	0.0472	0.44
$\beta_2 = 0.7$	0.067	0.0144	0.03	-0.106	0.0177	0.00	0.166	0.0368	0.36
$\beta_3 = -0.1$	0.042	0.0035	0.08	0.141	0.0045	0.00	0.105	0.0128	0.72
$\beta_4 = 0.1$	-0.1006	0.0325	0.05	-0.062	0.0046	0.00	-0.237	0.0832	0.40
ζ		5.4738			6.5393			11.2489	
$n = 200$									
$p = 0.3$	-0.031	0.0019	0.11	0.011	0.0014	0.06	-0.03	0.0019	0.23
$\beta_1 = 0.5$	0.082	0.0110	0.04	0.001	0.0008	0.00	0.258	0.0695	1.00
$\beta_2 = 0.7$	0.086	0.0133	0.15	-0.058	0.0063	0.01	0.245	0.0655	0.97
$\beta_3 = -0.1$	0.001	0.0007	0.02	0.030	0.0015	0.00	0.007	0.0007	0.11
$\beta_4 = 0.1$	-0.018	0.0077	0.02	-0.032	0.0016	0.00	-0.065	0.0117	0.11
ζ		3.6528			1.6205			9.2044	

Table 4.5: ZIP poorly separated data with constant zero state probability $p = 0.1$, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$p = 0.1$	-0.001	0.0018	0.03	0.008	0.0023	0.01	0.028	0.0026	0.03
$\beta_1 = 0.5$	0.146	0.0506	0.10	0.044	0.0805	0.02	0.592	0.3685	0.96
$\beta_2 = 1.0$	0.090	0.0398	0.06	-0.012	0.0222	0.01	0.366	0.1492	0.72
$\beta_3 = 0.2$	0.059	0.0089	0.10	-0.036	0.0067	0.00	0.087	0.0104	0.31
$\beta_4 = 0.1$	-0.017	0.1030	0.08	-0.011	0.0092	0.00	-0.432	0.2240	0.53
ζ		0.5936			0.3396			2.8334	
$n = 200$									
$p = 0.1$	-0.0001	0.0007	0.03	-0.001	0.0008	0.00	0.033	0.0018	0.22
$\beta_1 = 0.5$	0.157	0.0450	0.14	-0.061	0.0619	0.02	0.792	0.0633	1.00
$\beta_2 = 1.0$	0.132	0.0328	0.19	-0.054	0.0223	0.02	0.561	0.3205	1.00
$\beta_3 = 0.2$	-0.008	0.0014	0.02	0.013	0.0038	0.01	-0.087	0.0082	0.87
$\beta_4 = 0.1$	-0.018	0.0127	0.04	0.050	0.0342	0.02	-0.110	0.0240	0.11
ζ		0.2395			0.2199			2.2850	

Table 4.6: ZIP poorly separated data with constant zero state probability $p = 0.3$, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$p = 0.3$	-0.022	0.0043	0.03	0.010	0.0047	0.06	-0.004	0.0035	0.07
$\beta_1 = 0.5$	0.201	0.0826	0.11	-0.006	0.1663	0.24	0.695	0.5094	0.97
$\beta_2 = 1.0$	0.138	0.0660	0.06	-0.061	0.0483	0.12	0.431	0.2103	0.69
$\beta_3 = 0.2$	0.069	0.0130	0.05	-0.026	0.0118	0.26	0.084	0.0114	0.21
$\beta_4 = 0.1$	-0.055	0.0452	0.04	0.031	0.0488	0.00	-0.459	0.2639	0.52
ζ		0.6972			0.4208			2.7595	
$n = 200$									
$p = 0.3$	-0.0100	0.0006	0.04	-0.0057	0.0006	0.09	-0.0094	0.0006	0.20
$\beta_1 = 0.5$	0.1247	0.0228	0.10	-0.1505	0.0314	0.03	0.2974	0.0927	1.00
$\beta_2 = 1.0$	0.1674	0.0157	0.23	-0.0462	0.0054	0.00	0.2144	0.0497	0.96
$\beta_3 = 0.2$	-0.0096	0.0019	0.03	-0.0764	0.0065	0.00	-0.0319	0.0015	0.36
$\beta_4 = 0.1$	-0.0140	0.0227	0.03	0.1151	0.0154	0.00	-0.0526	0.0098	0.11
ζ		0.3042			0.2881			2.4882	

Table 4.7: ZIP well separated data with constant zero state probability $p = 0.1$, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$p = 0.1$	-0.015	0.0016	0.01	0.005	0.0016	0.04	-0.014	0.0017	0.23
$\beta_1 = 1.5$	0.071	0.0215	0.13	0.033	0.0790	0.04	0.200	0.0523	0.43
$\beta_2 = 2.0$	0.066	0.0151	0.05	0.034	0.0484	0.01	0.135	0.0269	0.30
$\beta_3 = 0.2$	0.025	0.0023	0.04	0.015	0.0039	0.00	0.036	0.0028	0.14
$\beta_4 = 0.1$	-0.013	0.0326	0.09	-0.072	0.1135	0.01	0.168	0.0459	0.20
ζ		1.4067			1.5983			3.2599	
$n = 200$									
$p = 0.1$	-0.010	0.0006	0.04	-0.006	0.0006	0.09	-0.009	0.0006	0.20
$\beta_1 = 1.5$	0.092	0.0152	0.12	-0.151	0.0314	0.03	0.297	0.0927	1.00
$\beta_2 = 2.0$	0.079	0.0101	0.16	-0.046	0.0054	0.00	0.214	0.0497	0.96
$\beta_3 = 0.2$	-0.008	0.0006	0.02	0.076	0.0065	0.00	-0.032	0.0015	0.36
$\beta_4 = 0.1$	-0.002	0.0093	0.05	0.115	0.0154	0.00	-0.053	0.0098	0.11
ζ		0.928			2.8812			2.4882	

Table 4.8: ZIP well separated data with constant zero state probability $p = 0.3$, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MHDE			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
$p = 0.3$	-0.026	0.0033	0.01	0.022	0.0043	0.06	-0.026	0.0033	0.05
$\beta_1 = 1.5$	0.095	0.0303	0.08	-0.039	0.0453	0.19	0.231	0.0712	0.47
$\beta_2 = 2.0$	0.068	0.0226	0.07	-0.021	0.0251	0.03	0.139	0.0349	0.31
$\beta_3 = 0.2$	0.026	0.0033	0.03	0.008	0.0036	0.07	0.032	0.0034	0.14
$\beta_4 = 0.1$	-0.025	0.0399	0.07	0.069	0.0178	0.00	-0.158	0.0495	0.16
ζ		1.4769			0.9128			2.9648	
$n = 200$									
$p = 0.3$	-0.023	0.0015	0.07	-0.006	0.0011	0.04	-0.023	0.0014	0.09
$\beta_1 = 1.5$	0.137	0.0267	0.18	0.108	0.0236	0.01	0.353	0.1320	1.00
$\beta_2 = 2.0$	0.108	0.0172	0.18	-0.010	0.0045	0.00	0.250	0.0675	0.96
$\beta_3 = 0.2$	-0.013	0.0009	0.04	0.062	0.0049	0.00	-0.04	0.0023	0.42
$\beta_4 = 0.1$	-0.019	0.0099	0.02	0.084	0.0093	0.00	-0.079	0.0147	0.15
ζ		0.8140			1.6412			2.3701	

4.4.2 Study 2: ZI Regression with Nonconstant p and Outliers in y

For data generated from ZI regression with nonconstant p, we restricted our attention to RES and ML estimation methods. Data were contaminated in the same way as described in study 1. True values for the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are specified in Tables 4.9-4.12. Again, two sample size $n = 64$ and 200 are examined in this study.

In Table 4.9, results are presented corresponding to estimation of data with small proportion of zeros and well separated ZIB models. As expected, RES demonstrates less bias and MSE than ML approach for $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and ζ. RES also performs much better in size. Table 4.10 gives result for data with a large portion of zeros and poorly separated ZIB models. Here we find that RES parameter estimates of $\boldsymbol{\gamma}$ exhibit similar bias, MSE and size compared with ML, but RES does significantly better in $\boldsymbol{\beta}$ and ζ. Tables 4.11 and 4.12 represent analogues of Tables 4.9 and 4.10 for the ZIP context and exhibit similar patterns of results.

In general, we can conclude that RES method is much more accurate than ML in estimating $\boldsymbol{\beta}$ when data have outliers in y. In estimating $\boldsymbol{\gamma}$, RES method improves upon ML estimation when the components are poorly separated. Whereas the size of Wald tests was extremely far from nominal in many cases under ML estimation, the RES approach led to Wald tests with fairly accurate observed sizes. Sample size has the expected effect of decreasing bias and improving efficiency for both methods, but the RES approach maintains a clear advantage even when $n = 200$.

4.4.3 Study 3: ZI Regression with Nonconstant p and Outliers in $\boldsymbol{x}$

Again, to avoid non-identifiability issues with the MHD approach for non-constant p, we restrict our attention to RES and ML methods only. To create these outliers, we randomly chose about 1% of the observations (one point for $n = 64$ and two points for $n = 200$), and replaced the covariate value x_4 by $x_4 + 3$, leaving the response y

Table 4.9: ZIB with complex mixing probability with small portion of zero state and well separated binomial component, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$\gamma_1 = -2$	-0.1930	4.8539	0.04	-0.9359	14.9168	0.04
$\gamma_2 = -1.5$	-0.3360	4.0044	0.03	-0.3468	4.6700	0.02
$\gamma_3 = -0.1$	-0.2340	3.0315	0.07	-0.2541	3.6168	0.02
$\beta_1 = 0.5$	0.0452	0.0107	0.07	0.1424	0.0289	0.34
$\beta_2 = 0.7$	0.0537	0.0124	0.05	0.1415	0.0283	0.32
$\beta_3 = -0.1$	0.0332	0.0027	0.08	0.0859	0.0089	0.67
$\beta_4 = 0.1$	-0.0643	0.0261	0.03	-0.1989	0.0582	0.37
ζ		6.9065			12.1800	
$n = 200$						
$\gamma_1 = -2$	-0.1743	0.2626	0.06	-0.1743	0.2626	0.06
$\gamma_2 = -1.5$	-0.2109	0.2387	0.10	-0.2109	0.2387	0.09
$\gamma_3 = -0.1$	-0.0013	0.0437	0.06	-0.0013	0.0437	0.06
$\beta_1 = 0.5$	0.0658	0.0072	0.04	-0.2082	0.0457	0.97
$\beta_2 = 0.7$	0.0669	0.0086	0.11	0.2037	0.0453	0.97
$\beta_3 = -0.1$	0.001	0.0005	0.01	0.0043	0.0005	0.05
$\beta_4 = 0.1$	-0.0196	0.0068	0.02	-0.0495	0.0096	0.12
ζ		3.1125			7.4134	

Table 4.10: ZIB with complex mixing probability with large portion of zero states and poorly separated binomial component, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$\gamma_1 = -1$	-0.116	0.3106	0.02	-0.116	0.3106	0.01
$\gamma_2 = -0.5$	-0.1276	0.4169	0.08	-0.1276	0.4169	0.05
$\gamma_3 = -0.1$	-0.0788	0.0928	0.06	-0.0788	0.0928	0.05
$\beta_1 = -1.0$	0.0530	0.014	0.07	0.3526	0.1380	0.89
$\beta_2 = -0.8$	0.0803	0.0200	0.08	0.3550	0.1418	0.90
$\beta_3 = -0.1$	0.0369	0.0040	0.09	-0.2576	0.0367	0.99
$\beta_4 = 0.1$	-0.0095	0.0402	0.06	-0.01047	0.2467	0.77
ζ		3.5607			21.9422	
$n = 200$						
$\gamma_1 = -1$	-0.2112	0.1518	0.08	-0.2118	0.1518	0.08
$\gamma_2 = -0.5$	-0.1476	0.0986	0.07	-0.1476	0.0986	0.08
$\gamma_3 = -0.1$	0.0016	0.0256	0.06	0.0016	0.0256	0.06
$\beta_1 = 1.0$	0.0980	0.0137	0.09	0.4935	0.249	1.00
$\beta_2 = -0.8$	0.0867	0.0143	0.11	0.5023	0.2580	1.00
$\beta_3 = -0.1$	-0.0026	0.0006	0.01	0.0134	0.0008	0.15
$\beta_4 = 0.1$	-0.0083	0.0117	0.04	-0.11195	0.0237	0.31
ζ		1.6727			17.0735	

Table 4.11: ZIP with complex mixing probability with small portion of zero states and well separated Poisson component, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$\gamma_1 = -2$	-0.2071	1.1644	0.04	-1.4009	4.6692	0.01
$\gamma_2 = -1.5$	-0.2269	1.0986	0.04	-0.7024	4.5802	0.02
$\gamma_3 = -0.1$	-0.0660	0.2400	0.05	-0.1968	3.8604	0.05
$\beta_1 = 1.5$	0.0068	0.0299	0.04	0.2013	0.0533	0.42
$\beta_2 = 2.0$	0.0137	0.0259	0.06	0.1348	0.0272	0.26
$\beta_3 = 0.2$	0.0271	0.0030	0.06	0.0358	0.0028	0.11
$\beta_4 = 0.1$	-0.0105	0.0172	0.02	-0.1657	0.0453	0.19
ζ		1.810			3.6752	
$n = 200$						
$\gamma_1 = -2$	-0.1422	0.2098	0.01	-0.1929	0.2378	0.03
$\gamma_2 = -1.5$	-0.1635	0.2310	0.10	-0.1681	0.2533	0.10
$\gamma_3 = -0.1$	-0.0148	0.0479	0.07	0.0036	0.0393	0.02
$\beta_1 = 1.5$	0.1021	0.0166	0.17	0.3089	0.1002	0.99
$\beta_2 = 2.0$	0.0833	0.0119	0.14	0.2277	0.0559	0.95
$\beta_3 = 0.2$	-0.0077	0.0007	0.05	-0.0347	0.0017	0.39
$\beta_4 = 0.1$	-0.0058	0.0073	0.02	-0.064	0.0195	0.09
ζ		0.8155			2.5899	

Table 4.12: ZIP with complex mixing probability with large portion of zero states and poorly separated Poisson component, with 10% outliers, $n = 64$ and 200

$n = 64$	RES			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$\gamma_1 = -1$	-0.0487	0.5379	0.06	0.2821	0.5860	0.12
$\gamma_2 = -0.5$	-0.0968	0.3638	0.02	0.0837	0.3740	0.06
$\gamma_3 = -0.1$	-0.1006	0.1153	0.07	-0.1391	0.1161	0.08
$\beta_1 = 0.5$	0.1007	0.1285	0.04	0.6900	0.5033	0.95
$\beta_2 = 1.0$	0.1059	0.0877	0.05	0.4706	0.2450	0.79
$\beta_3 = 0.2$	0.0539	0.0079	0.02	0.0682	0.0113	0.15
$\beta_4 = 0.1$	-0.0322	0.0591	0.02	-0.4475	0.2435	0.49
ζ		0.7201			3.0567	
$n = 200$						
$\gamma_1 = -1$	0.0135	0.1204	0.06	0.2966	0.1780	0.14
$\gamma_2 = -0.5$	-0.0799	0.0817	0.04	0.0268	0.0752	0.01
$\gamma_3 = -0.1$	-0.0225	0.0180	0.06	-0.0677	0.0249	0.05
$\beta_1 = 0.5$	0.2555	0.0887	0.18	0.8841	0.8245	1.00
$\beta_2 = 1.0$	0.2154	0.0634	0.13	0.6790	0.4683	1.00
$\beta_3 = 0.2$	-0.0376	0.0032	0.04	-0.1118	0.0153	0.95
$\beta_4 = 0.1$	-0.026	0.0244	0.05	-0.0582	0.0436	0.15
ζ		0.3285			2.3419	

and other covariates unchanged. Like in study 2, two levels of mixture separation and two sample sizes were considered.

Tables 4.13 and 4.14 gives results for well separated ZIB data with small portion of zeros and for poorly separated ZIB data with a large portion of zeros, respectively. In these tables, RES and ML perform similarly with respect to $\boldsymbol{\gamma}$. This results is sensible, since there is only a small amount of contamination in x which is of a form which does not obscure the mixture structure much. With respect to $\boldsymbol{\beta}$ and ζ however, RES has less bias, smaller MSE and closer to nominal size than ML estimation. Generally speaking, these results are replicated in Tables 4.15 and 4.16, which contain the corresponding results for the ZIP setting. Across all of these tables, the usual positive effect of sample size on efficiency is observed.

4.5 Example

To illustrate the use of robust methods for zero inflated regression models, we consider data from the Multisite Violence Prevention Project (MVPP), a study designed to reduce violent and aggressive behaviors among middle school aged children conducted by investigators from four US universities and the Centers for Disease Control and Prevention. The study utilized a randomized complete block design involving 37 schools randomized to a 4 treatment structure within each of four blocks corresponding to the sites of the universities participating in the project. Included among the outcomes measured via teacher surveys in the study was a 30-day recall of number of insults received from students. Although the study was longitudinal with data collected twice annually over 4 years, we avoid this aspect of the design by examining only data from the first post-treatment measurement occasion, treating the pre-treatment, or baseline, response as a covariate in our analysis. For simplicity,

Table 4.13: ZIB with complex mixing probability with small portion of zero states and well separated binomial component, with 1% moderate outliers having abnormal covariates, $n = 64$ and 200

$n = 64$	RES			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$\gamma_1 = -2$	-0.097	0.6819	0.06	-0.097	0.6819	0.05
$\gamma_2 = -1.5$	-0.129	0.6262	0.03	-0.129	0.6279	0.03
$\gamma_3 = -0.1$	0.029	0.1077	0.04	0.030	0.1077	0.02
$\beta_1 = 0.5$	0.046	0.0126	0.07	0.240	0.0739	0.76
$\beta_2 = 0.7$	0.022	0.0107	0.04	0.274	0.0453	0.52
$\beta_3 = -0.1$	-0.005	0.0017	0.06	-0.003	0.0017	0.06
$\beta_4 = 1.0$	-0.073	0.0350	0.05	-0.527	0.3438	0.12
ζ		7.3990			9.7052	
$n = 200$						
$\gamma_1 = -2$	-0.036	0.2538	0.06	-0.036	0.2538	0.06
$\gamma_2 = -1.5$	-0.121	0.1916	0.03	-0.121	0.1926	0.03
$\gamma_3 = -0.1$	0.006	0.0401	0.03	0.006	0.0402	0.03
$\beta_1 = 0.5$	0.020	0.0044	0.04	0.179	0.0376	0.82
$\beta_2 = 0.7$	0.022	0.0049	0.04	0.166	0.0324	0.79
$\beta_3 = -0.1$	-0.001	0.0005	0.07	-0.006	0.0005	0.05
$\beta_4 = 1.0$	-0.049	0.0090	0.01	-0.377	0.1562	0.93
ζ		2.4616			3.4021	

Table 4.14: ZIB with complex mixing probability with large portion of zero states and poorly separated binomial component, with 1% moderate outliers having abnormal covariates, $n = 64$ and 200

$n = 64$	RES			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$\gamma_1 = -1$	-0.1248	0.4646	0.07	-0.1248	0.4646	0.06
$\gamma_2 = -0.5$	0.0173	0.2215	0.02	0.0173	0.2215	0.01
$\gamma_3 = -0.1$	0.0054	0.0611	0.04	0.0054	0.0611	0.05
$\beta_1 = -1.0$	0.0449	0.0176	0.06	0.2078	0.0772	0.65
$\beta_2 = -0.8$	0.0446	0.0157	0.04	0.1801	0.0565	0.61
$\beta_3 = -0.1$	0.0000	0.0014	0.01	-0.0002	0.0014	0.04
$\beta_4 = 1.0$	-0.1146	0.0591	0.05	-0.4594	0.3436	0.67
ζ		3.6389			5.2225	
$n = 200$						
$\gamma_1 = -1$	0.0168	0.1315	0.08	-0.0789	0.1289	0.07
$\gamma_2 = -0.5$	-0.0071	0.838	0.02	-0.0617	0.1053	0.06
$\gamma_3 = -0.1$	0.0001	0.0231	0.05	0.0152	0.0273	0.09
$\beta_1 = -1.0$	0.0492	0.0075	0.03	1.1195	0.0586	0.88
$\beta_2 = -0.8$	0.0494	0.0069	0.01	0.2181	0.0570	0.88
$\beta_3 = -0.1$	-0.0031	0.0004	0.03	-0.0058	0.0005	0.02
$\beta_4 = 1.0$	-0.1059	0.0238	0.01	-0.4348	0.2255	0.91
ζ		1.4222			3.5329	

Table 4.15: ZIP with complex mixing probability with small portion of zero and well separated Poisson component, with 1% moderate outliers having abnormal covariates, $n = 64$ and 200

$n = 64$	RES			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$\gamma_1 = -2$	-0.3530	0.8194	0.03	-0.3463	0.8133	0.01
$\gamma_2 = -1.5$	-0.1821	0.7093	0.06	00.1796	0.7072	0.04
$\gamma_3 = -0.1$	0.0744	0.1473	0.06	0.0727	0.1470	0.03
$\beta_1 = 1.5$	0.0981	0.02718	0.10	0.4781	0.3041	0.79
$\beta_2 = 2.0$	0.0791	0.0210	0.09	0.3995	0.2126	0.79
$\beta_3 = 0.2$	-0.0095	0.0015	0.07	-0.0511	0.0047	0.42
$\beta_4 = 1.0$	-0.1364	0.0442	0.09	-0.6503	0.5586	0.80
ζ		2.4567			7.6554	
$n = 200$						
$\gamma_1 = -2$	-0.0654	0.1870	0.06	-0.0721	0.1859	0.05
$\gamma_2 = -1.5$	0.0014	0.1776	0.05	0.0024	0.1774	0.05
$\gamma_3 = -0.1$	-0.0101	0.0495	0.10	-0.0101	0.0494	0.06
$\beta_1 = 1.5$	0.2107	0.0523	0.17	0.3992	0.1712	0.99
$\beta_2 = 2.0$	0.1975	0.0453	0.13	0.3680	0.1448	0.99
$\beta_3 = 0.2$	-0.0133	0.0006	0.10	-0.0263	0.0011	0.12
$\beta_4 = 1.0$	-0.3520	0.1371	0.13	-0.6579	0.4561	0.99
ζ		2.5793			6.7281	

Table 4.16: ZIP with complex mixing probability with large portion of zero and poorly separated Poisson component, with 1% moderate outliers having abnormal covariates, $n = 64$ and 200

$n = 64$	RES			MLE		
Parameters	Bias	MSE	Size	Bias	MSE	Size
$\gamma_1 = -1$	-0.1364	0.7116	0.06	-0.2788	0.4304	0.06
$\gamma_2 = -0.5$	0.0072	0.3818	0.05	0.0104	0.3834	0.04
$\gamma_3 = -0.1$	-0.0119	0.0925	0.04	-0.0101	0.0925	0.05
$\beta_1 = 0.5$	0.1098	0.0520	0.09	0.2494	0.1426	0.41
$\beta_2 = 1.0$	0.0972	0.0419	0.06	0.2154	0.1112	0.45
$\beta_3 = 0.2$	0.0044	0.0064	0.14	0.0106	0.0066	0.07
$\beta_4 = 1.0$	-0.2466	0.1293	0.13	-0.5326	0.4829	0.62
ζ		0.6934			1.0900	
$n = 200$						
$\gamma_1 = -1$	0.0510	0.1128	0.07	0.0696	0.1123	0.06
$\gamma_2 = -0.5$	0.0459	0.0914	0.04	0.0551	0.0910	0.04
$\gamma_3 = -0.1$	-0.0116	0.0165	0.03	-0.0115	0.0164	0.03
$\beta_1 = 0.5$	0.2506	0.0771	0.14	0.3806	0.1616	0.94
$\beta_2 = 1.0$	0.2301	0.0662	0.18	0.353	0.1395	0.95
$\beta_3 = 0.2$	-0.0033	0.0016	0.07	-0.0066	0.0016	0.11
$\beta_4 = 1.0$	-0.3441	0.2281	0.15	-0.6802	0.4902	0.95
ζ		0.4854			0.8056	

we also restrict attention to just one of the four sites involved in the study, from which 86 teachers' data were available.

A simple histogram of the insult data is presented in Figure 1 (a). From this plot it is apparent that a very large proportion of the teachers reported 0 insults. However, there are also large frequencies of insult counts that are much larger than 0 indicating possible zero inflation in these data. In addition, there are a few teachers who reported very large numbers of insults (15, 20, 30) which are clearly outlying relative to the main portion of the data and which may strongly affect inferences on the treatments. Given this data structure and experimental design, a natural model to consider here is a zero inflated Poisson analysis of covariance type model for the post-treatment insult count in which the baseline count is treated as a covariate, and which is fit with a robust methodology to account appropriately for the presence of the few extremely high observations in the data. By using the proposed robust methods, we aim to automatically downweight the extreme observations with potentially large influence on the estimates of the treatment effects.

Specifically, we assume that y_{ij}, the number of insults for the jth teacher in the ith treatment, follows a ZIP distribution with Poisson mean $\log(\mu_{ij}) = \lambda_i + \beta \log(\texttt{baseline}_{ij} + 0.01)$. In addition, we considered two models for the mixing probability p_{ij}. First, we assumed $p_{ij} = p$ for all i, j, and fit this model with the ML, MHD, and RES estimation methods. However, there is little reason a priori to assume that the presence of excess zeros is independent of the treatments and baseline response, so we also investigated a similar model to that for μ, and model $\log(\frac{p_{ij}}{1-p_{ij}}) = \gamma_1 + \gamma_2 \log(\texttt{baseline}_{ij} + 0.01)$.

Parameter estimates and standard errors for these models appear in Table 4.17. The results for the RES method here were obtained by using tuning constant $c = 1.65$ and weights $w(\boldsymbol{x}_i) = \sqrt{1 - h_i}$. From these results it can be seen that the RES and MHD methods produce similar estimates to each other in model 1 but quite different

results to ML, especially for λ. For model 2, in which there is regression structure for p, the MHD method is not available; but again, results for RES and ML differ substantially. There is also evidence here (based on Wald test of γ_2 in model 2), that the mixing probability depends at least on the baseline response so that having to restrict attention to a constant mixing probably model here, as is necessary for use of the MHD method, is an undesirable restriction. Note that downweighting the effect of outliers in the RES analysis results in qualitatively different results with respect to the treatment effect here. In particular, the Wald test of $H_0 : \lambda_1 = \cdots = \lambda_4$ gives p-values of 0.918 and 0.012 under the RES and ML fitting methods, respectively. This difference underscores the importance of robustifying the analysis so that estimates and inferences are not unduly influenced by the effects of a small number of extreme values.

To further compare the RES and ML fits of model 2, we produced plots of chi-square residuals. These plots are somewhat similar in concept to half normal plots (Atkinson, 1981; Vieira et al., 2000), and were constructed by first binning the data into b bins according to the observed values of the response. In this case, we used $b = 7$ bins corresponding to 0, 1, 2, 3, $[4, 5]$, $(5, 12]$ and $(12, \infty)$. Then we calculated residuals defined as the contribution to the chi-square goodness of fit statistic $r_j = \frac{n\left\{f_n(y) - f_{\boldsymbol{\theta},n}(y)\right\}^2}{f_{\boldsymbol{\theta},n}(y)}$ for each bin based on the fitted model (using RES or ML). The idea behind this plot is if the model is correctly specified, any collection of $b-1$ residuals should be approximately iid $\chi^2(1)$ random variables. The residual plot is a plot of ordered $r_{(j)}$ against $\chi_1^{-1}\{(j - 0.5)/b\}$, $j = 1, 2, \ldots, b$. A simulated envelope for this chi-square plot was constructed in the same way as is typically done in half normal plots. An additional 19 data sets were simulated based on the fitted model. Chi-square residuals $r_{(j)}(d)$ for data sets $d = 1, \ldots, 19$ were then calculated and ordered, where $r_{(1)}(d) < r_{(2)}(d) < \ldots < r_{(b)}(d)$ for each d. Then for each bin j, the median, minimum and maximum of $r_{(j)}(d)$ over the 19 simulated data sets were

Table 4.17: Parameters Estimates and Standard Errors (SE) of Model 1 and Model 2 for MVPP

	RES,c=1.65	MLE	MHDE
Model 1 Parameters	Estimate (SE)	Estimate (SE)	Estimate (SE)
γ	-0.7206 (0.3169)	-0.6420 (0.2939)	-0.4463 (0.3481)
λ_1	1.2497 (0.1785)	1.4126 (0.1384)	1.1466 (0.3820)
λ_2	1.2739 (0.2314)	1.1935 (0.1864)	1.2983 (0.3247)
λ_3	1.1383 (0.1683)	1.3635 (0.1218)	1.0245 (0.5030)
λ_4	1.2382 (0.2981)	1.8838 (0.1536)	1.1609 (0.7822)
β	0.1918 (0.0524)	0.2438 (0.0437)	0.1643(0.0962)
Model 2 Parameters	Estimate (SE)	Estimate (SE)	
γ_1	-0.9410 (0.3086)	-0.9061 (0.3015)	-
γ_2	-0.2696 (0.0915)	-0.2706 (0.0939)	-
λ_1	1.2989(0.1793)	1.4720 (0.1359)	-
λ_2	1.3187 (0.2316)	1.2595 (0.1866)	-
λ_3	1.1557 (0.1692)	1.3831 (0.1216)	-
λ_4	1.2841 (0.2935)	1.9283 (0.1458)	-
β	0.1477 (0.0472)	0.2019 (0.0390)	-

plotted alongside the residuals of the original data set. The plots show that most of the residuals fall within the boundaries of the envelope using RES (Figure 1(b)), but not with ML (Figure 1(c)), indicating that the former method is more appropriate for these data. The one extremely large point in Figure 1(b) is actually a desirable feature. It corresponds to the outlying values in the largest bin, which should have large residuals if the estimation is downweighting these outliers as intended.

4.6 Summary

In this thesis, we proposed two robust methods, MHD and RES estimation for ZI regression models. Simulation results were largely consistent with expectations, indicating that both the MHD and RES methods provide substantial protection against outliers and poor component separation relative to ML. However, as described above, the MHD method leads to identifiability problems for some models that are identifiable when fit with ML or the RES approach and is therefore, substantially more narrowly applicable. The Mallows class estimating equations we proposed in the RES method perform well in downweighting outliers in y and/or covariates $\boldsymbol{x}$. However, it does require specification of the tuning constant c, which does affect the efficiency of the parameter estimators. Further research is ongoing to develop methodology for optimal selection of c.

A natural extension of the RES approach would be to generalize this method to the clustered data context (e.g., longitudinal data). We are currently pursuing this goal by combining our approach with that of Hall and Zhang (2004) who recently proposed an estimation method for marginal ZI regression models for clustered data via generalized estimating equations.

Acknowledgements

The authors wish to express their sincere gratitude to the Multisite Violence Prevention Project for providing the data analyzed in Section 4.5.

Appendix

Verification of Conditions of Rosen et al.'s theory

Let (S_y, F_y) and (S_z, F_z) be σ-finite measurable spaces with a product measure space $(S_y \times S_z, F_{yz})$, $S(.|.)$ has the following two properties,

(1) For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$E_Y[S(\boldsymbol{\theta}|\boldsymbol{\theta}, Y)] = \int_{S_y \times S_z} s_i(y, z; \boldsymbol{\theta}) f(y, z|\boldsymbol{\theta}) d\mu(y) d\mu(z) = 0. \tag{4.6.19}$$

Since

$$E[Z_i - \frac{\exp(\boldsymbol{G}_i\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{G}_i\boldsymbol{\gamma})}|\boldsymbol{\theta}] = 0,$$

and

$$E[\psi_c(\frac{Y_i - \mu(\boldsymbol{B}_i\boldsymbol{\beta})}{\sqrt{\nu(\boldsymbol{B}_i\boldsymbol{\beta})}}) - a_i(\boldsymbol{\beta}, c)|\boldsymbol{\theta}] = 0.$$

Then we get

$$E_Y[s_i(Y_i, Z_i; \boldsymbol{\theta})|\boldsymbol{\theta}] = 0.$$

(2)

$$S(\boldsymbol{\phi}|\boldsymbol{\theta}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{z=0,1} s_i(y_i, z_i, \boldsymbol{\phi})\frac{p_i(y_i|z_i)p_i(z_i)}{\sum_{z=0,1} p_i(y_i|z_i)p_i(z_i)}.$$

Since s_i is continuous function for each $(y, z) \in S_y \times S_z$, so $S(.|.)$ is a bivariate continuous function on $\boldsymbol{\theta} \times \boldsymbol{\theta}$.

4.7 Reference

[1] Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.

[2] Adimari, G. and Ventura, L. (2001). Robust inference for generalized linear models with application to logistic regression. *Statistics and Probability Letters*, **55**, 413-419.

[3] Beran, R. (1977). Minimum Hellinger distance for parametric models. *The Annals of Statistics*, **5**, 445-463.

[4] Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**, 1022-1030.

[5] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* Chapman and Hall, New York.

[6] Christmann, A. (1998). Positive breakdown point estimation in categorical regression models. *Habilitation thesis*. University of Dortmund, Department of Statistics.

[7] Cutler, A., Cordero-Brana, O. (1996). Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, **91**, 1716-1724.

[8] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1-38.

[9] Hall, D. B. and Zhang, Z., (2004). Marginal models for zero inflated clustered data. *Statistical Modelling*, **4**, 161-180.

[10] Hall, D. B. and Berenhaut K. S., (2002). Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. *Canadian Journal of Statistics*, **29**, 77-97.

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

[12] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383-393.

[13] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986) *Robust Statistics.* New York: John Wiley and Sons.

[14] Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association*, **89**, 897-904.

[15] Huber, P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73-101.

[16] Huber, P. J. (2004). *Robust Statistics.* Wiley.

[17] Karlis, D. and Xekalaki, E. (2001). Robust inference for finite Poisson mixtures. *Journal of Statistical Planning and Inference*, **93**, 93-115.

[18] Künsch, H. R., Stefanski, L. A and Carroll, R. J. (1989). Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, With Applications to Generalized Linear Models. *Journal of the American Statistical Association*, **84**, 460-466.

[19] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

[20] Lindsay, B. G. and Roeder, K. (1992). Residual diagnostics for mixture models. *Journal of the American Statistical Association*, **87**, 785-794.

[21] Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, **22**, 1081-1114.

[22] Lu. Z., Hui. Y. and Lee, A. (2003) Minimum Hellinger distance estimation for finite mixtures of Poisson regression models and its applications. *Biometrics*, **59**, 1016-1026.

[23] Preisser J. S. and Qaqish, B. F. (1999). *Biometrics*, **55**, 574-579.

[24] Ridout, M., Hinde, J. and Demétrio, CGB (1998). Models for count data with many zeros. *Invited Paper, the XIXth International Biometric conference, Cape Town, South Africa*, 79-92.

[25] Rosen, O, Jiang, W.X. and Tanner, M.A (2000). Mixtures of marginal models. *Biometrika*, **87**, 391-404.

[26] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871-880.

[27] Simpson, D. G. (1987). Minimum Hellinger distance estimation for analysis of count data. *Journal of the American Statistical Association*, **82**, 802-807.

[28] Vieira A.M.C., Hinde J.P., and Demétrio C.G.B. (2000). Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, **27**, 373-389.

[29] Woodward, W. A., Whitney P. and Eslinger, P. (1994). Minimum Hellinger distance estimation of mixture proportions. *Journal of Statistical Planning and Inference*, **48**, 303-319.

[30] Xiao J, Lee AH, Vemuri SR. (1999). Mixture distribution analysis for length of hospital stay for efficient funding. *Socio-economic Planning Sciences*, **33**, 39-59.

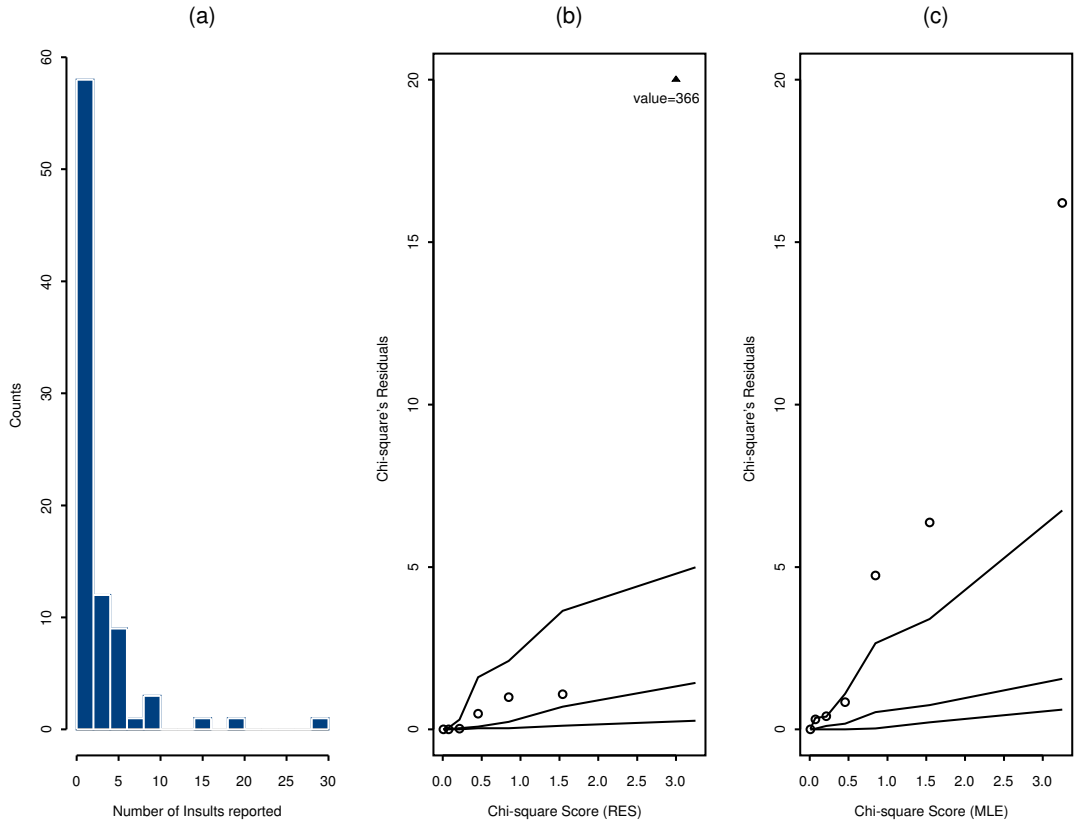

Figure 4.1: Histogram of the raw data and chi-square plots of Pearson's residuals for models fit to MVPP data.

Chapter 5

Future Work

5.1 Robust Estimation for Zero-Inflated Models for Clustered Data

Recently, ZI regression models have been extended to clustered data via adding random effects into the model by several authors (Hall, 2000; Yau and Lee, 2001). More recently, Hall and Zhang (2004) incorporate generalized estimating equations (GEEs) with EM algorithm to fit such marginal models for clustered data. However, since those methods are not designed to be resistant to potential outliers or influential data, they can be highly influenced by anomalous data points. One of the approaches described in Chapter 4, the RES approach, can easily be extended to the clustered data context.

For clustered data, the ZI model can be defined as follows. Assume the random variable

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } p_{ij}; \\ \mathrm{F}_2(y_{ij}; \theta_{ij}, \phi) & \text{with probability } 1 - p_{ij}. \end{cases}$$

where $i = 1, ..., K$ and $j = 1, ..., n_i$. In addition, we assume F_2 is in the exponential dispersion family with probability density function of $f_2(y_{ij}; \theta_{ij}, \phi)$. Also, we assume that the (conditional) mean, $\boldsymbol{\zeta}_i$, of $f_2(y_{ij}; \theta_{ij}, \phi)$ depends on covariates through some link function g, i.e. $\eta(\boldsymbol{\zeta}_i) = \boldsymbol{B}_i\boldsymbol{\beta}$, and the mixing probabilities, $\boldsymbol{p}_i$, are related to covariates, $\boldsymbol{G}_i$, via a link function too, such as the logit. Here, $\boldsymbol{B}_i$ and $\boldsymbol{G}_i$ are matrices of covariates for the i^{th} subject.

Hall and Zhang (2004) proposed an expectation solution (ES algorithm that incorporate GEEs into the S step by introducing missing values u_{ij}, where $u_{ij} = 1$ if y_{ij} is from zero state, otherwise, $u_{ij} = 0$. They also defined working correlation matrices $\boldsymbol{P}(\boldsymbol{\rho})$ and $\boldsymbol{R}(\boldsymbol{\delta})$, where $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ are unknown correlation parameters. At the E-step, the expectation of u_{ij} is calculated under the current parameter estimates $\boldsymbol{\beta}^{(l)}$, $\boldsymbol{\gamma}^{(l)}$ and $\phi^{(l)}$, which yields

$$u_{ij}^{(l)} = 1_{\{y_{ij}=0\}}[1 + \{1 - p_{ij}(\boldsymbol{\gamma}^{(l)})\} f_2(y_{ij}; \boldsymbol{\beta}^{(l)}, \phi^{(l)})/p_{ij}(\boldsymbol{\gamma}^{(l)})]^{-1}.$$

At the S-step, this approach leads to a combined estimating equation for $\boldsymbol{\beta}$, $\boldsymbol{\rho}$ and ϕ

$$\sum_{i=1}^{K} \begin{pmatrix} \frac{\partial \boldsymbol{\zeta}_i^T}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\sigma}_i^T}{\partial \tilde{\boldsymbol{\rho}}} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_{i11}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}_{i22}^{-1} \end{pmatrix} \boldsymbol{H}_i \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\zeta}_i \\ \boldsymbol{s}_i - \boldsymbol{\sigma}_i \end{pmatrix} = \mathbf{0}. \tag{5.1.1}$$

Here, $\tilde{\boldsymbol{\rho}} = (\boldsymbol{\rho}^T, \phi)^T$, $\boldsymbol{V}_{i11} = \boldsymbol{D}_i^{1/2}\{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\}\boldsymbol{P}(\boldsymbol{\rho})\boldsymbol{D}_i^{1/2}\{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\}$, $\boldsymbol{s}_i = \text{vech}\{(\boldsymbol{y}_i - \boldsymbol{\zeta}_i)(\boldsymbol{y}_i - \boldsymbol{\zeta}_i)^T\}$, $\boldsymbol{\sigma}_i = E(\boldsymbol{s}_i) = \text{vech}(\boldsymbol{V}_{i11})$, $\boldsymbol{H}_i = \text{diag}[\boldsymbol{j}_{n_i} - \boldsymbol{u}_i^{(l)}, \text{vech}\{(\boldsymbol{j}_{n_i} - \boldsymbol{u}_i^{(l)})(\boldsymbol{j}_{n_i} - \boldsymbol{u}_i^{(l)})^T\}]$, where $\boldsymbol{D}_i(\boldsymbol{\zeta}_i) = \text{diag}\{a(\phi)v(\zeta_{i1}), \ldots, a(\phi)v(\zeta_{in_i})\}$, $\boldsymbol{U}_i^{(l)} = \text{diag}(1 - u_{i1}^{(l)}, \ldots, 1 - u_{in_i}^{(l)})$, $\boldsymbol{j}_{n_i}$ is an $n_i \times 1$ vector of ones, and $\boldsymbol{V}_{i22}^{-1}$ is a weight matrix. In particular, $\boldsymbol{V}_{i22}$ has elements given by the relation $cov(s_{ijk}, s_{i\ell m}) = \sigma_{ij\ell}\sigma_{ikm} + \sigma_{ijm}\sigma_{ik\ell}$ where $s_{ijk} = (y_{ij} - \zeta_{ij})(y_{ik} - \zeta_{ik})$ and $\sigma_{ijk} = E(s_{ijk})$ are the elements of $\boldsymbol{s}_i$ and $\boldsymbol{\sigma}_i$, respectively. The vech denotes the vector-half function that stacks the columns of a matrix including only those elements on or below the diagonal. For $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, this combined estimating equation is

$$\sum_{i=1}^{K} \begin{pmatrix} \frac{\partial \boldsymbol{p}_i^T}{\partial \boldsymbol{\gamma}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\tau}_i^T}{\partial \boldsymbol{\alpha}} \end{pmatrix} \begin{pmatrix} \boldsymbol{W}_{i11}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{W}_{i22}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_i^{(l)} - \boldsymbol{p}_i \\ \boldsymbol{t}_i - \boldsymbol{\tau}_i \end{pmatrix} = \mathbf{0}, \tag{5.1.2}$$

where $\boldsymbol{W}_{i11} = \boldsymbol{A}_i^{1/2}\{\boldsymbol{p}_i(\boldsymbol{\gamma})\}\boldsymbol{R}(\boldsymbol{\delta})\boldsymbol{A}_i^{1/2}\{\boldsymbol{p}_i(\boldsymbol{\gamma})\}$, $\boldsymbol{A}_i(\boldsymbol{p}_i) = \text{diag}\{p_{i1}(1-p_{i1}), \ldots, p_{in_i}(1-p_{in_i})\}$, $\boldsymbol{t}_i = \text{vech}\{(\boldsymbol{u}_i^{(l)} - \boldsymbol{p}_i)(\boldsymbol{u}_i^{(l)} - \boldsymbol{p}_i)^T\}$, $\boldsymbol{\tau}_i = E(\boldsymbol{t}_i) = \text{vech}(\boldsymbol{W}_{i11})$, and $\boldsymbol{W}_{i22}^{-1}$ is a weight matrix. Similarly, $\boldsymbol{W}_{i22}$ has elements given by $cov(t_{ijk}, t_{i\ell m}) = \tau_{ij\ell}\tau_{ikm} + \tau_{ijm}\tau_{ik\ell}$, where $t_{ijk} = (u_{ij}^{(l)} - p_{ij})(u_{ik}^{(l)} - p_{ik})$ and $\tau_{ijk} = E(t_{ijk})$ are from $\boldsymbol{t}_i$ and $\boldsymbol{\tau}_i$.

As in Chapter 4, to create the robust approach, we propose to replace $(\boldsymbol{y}_i - \boldsymbol{\zeta}_i)$ by $(\boldsymbol{\psi}_i - \boldsymbol{a}_i)$, where $\boldsymbol{\psi}_i \equiv \Psi_c(\frac{\boldsymbol{y}_i - \boldsymbol{\zeta}_i}{\sqrt{\text{var}(\boldsymbol{y}_\text{i})}})\sqrt{\text{var}(\boldsymbol{y}_\text{i})}$, $\boldsymbol{a}_i = \text{E}(\boldsymbol{\psi}_\text{i})$ and Ψ_c is defined as (4.3.13). Accordingly, we redefine $\boldsymbol{s}_i$ as follows. $\boldsymbol{s}_i = \text{vech}(\boldsymbol{\psi}_\text{i} - \boldsymbol{a}_\text{i})(\boldsymbol{\psi}_\text{i} - \boldsymbol{a}_\text{i})^\text{T}$. We can iteratively compute the E step and S step until convergence. This approach allows the fit to downweight the outliers. The asymptotic properties of the estimator will be established using the results of Rosen et al. (2000).

5.2 Identification of Outliers and Leverage Points

Identification of outliers or highly influential data points is an important part of robust statistical estimation and inference. Checking residuals from a fit is a popular approach to detect outliers or even leverage points. The residuals from a robust fit automatically show outliers and should be more reliable than those from a classical method, for example ML, because in a non-robust estimation method outliers inflated the residual variance, which makes outlier detection more difficult. Graphical methods of inspection and more formal outlier detection techniques will be studied in the future.

5.3 References

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

[2] Hall, D. B. and Zhang, Z. (2004). Marginal models for zero inflated clustered data *Statistical Modelling*, **4**, 161-180.

[3] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986) *Robust Statistics.* New York: John Wiley and Sons.

[4] Rosen, O., Jiang W and Tanner M.A. (2000). Mixtures of marginal models. *Biometrika*, **87**, 391-404.