Robust Estimators for Finite Mixtures of Count Data Regression Models and their Applications

by

Ti-Jen Tsao

A dissertation submitted to the Graduate Faculty in Economics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2010

This manuscript has been read and accepted for the
Graduate Faculty in Economics in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Partha Deb

Date — Chair of Examining Committee

Merih Uctum

Date — Executive Officer

Professor Partha Deb

Professor Michael Grossman

Professor Wim Vijverberg

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Robust Estimators for Finite Mixtures of Count Data Regression Models and their Applications

by

Ti-Jen Tsao

Adviser: Professor Partha Deb

Finite mixtures of count data regression models have been successfully used for modeling discrete responses arising from heterogeneous populations. But the maximum likelihood (ML) estimator for such models are sensitive to data contamination and extreme values. This dissertation develops two robust estimators for finite mixtures of count data regression models. One is the minimum Hellinger distance (MHD) estimator and the other is the minimum L_2 error (L_2E) estimator, a special case of the minimum density power divergence estimator. Two Monte Carlo simulation studies show that the MHD and L_2E estimators are more robust than the ML one but come at the cost of efficiency. However, the robustness property of the MHD and L_2E estimators is deteriorated as the mixing probability approaches one.

For empirical application, this dissertation uses the data from Dionne et al. (1996), the extent of non-payments of personal loans in Spain, and from Deb and Trivedi (2002), counts of utilization from the RAND Health Insurance Experiment, respectively. The estimated results show that the two-component Poisson mixture regression model is the best fit model for the first data set and the two-component negative binomial one mixture regression model for the second data set. But both of the model specifications are preferred to be estimated by the ML estimation that could be attributed to the flexibility of the finite mixture model and data processing procedures.

Acknowledgments

I would like to express my deepest gratitude to my family, whose support paves the way to my Ph.D. degree. I thank my supervisory committee members for their advice and encouragement. I also thank Montserrat Guillén and Partha Deb for providing me their data sets so that the developed models and estimators can be empirically tested.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For more than a century, the finite mixture models (FMM) have been broadly studied in the theoretical and practical contexts to model the heterogeneity in data. Their importance in the statistical analysis is attributed to their flexibility, which is apt for modeling unknown distributional shapes of data points, e.g., overdispersion or multiple bumps. About their theories and applications of modeling, McLachlan and Peel (2000) gave a comprehensive introduction supplemented by the related literature survey in a variety of research fields. In this dissertation, the interest in the FMMs is their application in regression setting. For this purpose, the generalized linear models (GLM) (McCullagh and Nelder (1989)) are substituted for component densities of the FMMs. With the versatility of the FMMs, finite mixtures of GLMs are capable to explain continuous or discrete responses arising from a heterogeneous population that transcends the single distribution assumption of the GLMs. The spotlight in this dissertation is merely shined on the discrete case, finite mixtures of count data GLMs.

To better accommodate the count data with excess zeros, Deb and Trivedi (1997, 2002) employed the FMM variants of count data GLMs to compete the commonly used hurdle and zero-inflated regression models. Based on the definition of hurdle and zero-inflated models, excess zero counts are produced by a binary process within a two-step data generating process (DGP), which is consequently

specified as a binary choice model appended to a count data GLM. However, since empirical data are often collected over a given time period without any indication of how many subsequent responses are triggered by a primarily binary choice, modeling these characteristic data sets by a hurdle or zero-inflated model could distort the economic representation. Different from the two-step models, finite mixtures of count data GLMs let the unobserved heterogeneity endogenously split a population into homogeneous, latent subpopulations. This approach not only exploits the statistical flexibility of the FMM but also explains population heterogeneity according to the reality of data collection. Deb and Trivedi (1997, 2002) showed that finite mixtures of count data GLMs were preferred to the hurdle or zero-inflated model no matter form a perspective of adequacy-of-fit or economic interpretation.

Although model estimation of the FMMs normally depends on likelihood-based estimation methods, it is known that likelihood-based estimators become unstable when data are contaminated or have extreme values. Such data sets are frequently observed in social and economic studies. Facing this kind of problem, social scientists often subjectively drop inliers[1] or outliers, but that carries the risk of losing valuable information. Besides, it is unrealistic to do this when a sample size is considerably large. In the statistical literature, the minimum distance estimation methods have been formally proven with robustness. They are widely used in biological and engineering fields but seldom seen in economic research. This dissertation introduces two members of the minimum distance estimation family: one is the minimum Hellinger distance (MHD) estimation method and the other is the minimum L_2 error (L_2E) estimation method, a special case of the minimum density power divergence (MDPD) estimation method. Following that, the two minimum distance estimators for count data regression models are sequentially developed.

The MHD estimation method is first introduced by Beran (1977), who pre-

[1] An inlier is a data value that lies in the interior of a statistical distribution and is in error.

sented that the MHD estimator could remove the instability of estimates from perturbations but still preserved asymptotic efficiency under a specified regular parametric family of densities. Simpson (1987) derived the corresponding asymptotic properties for discrete distributions and proved that the MHD estimator was asymptotically equivalent to the maximum likelihood (ML) estimator when the model was correctly specified. Even though the MHD estimator can enjoy both robustness and asymptotic efficiency under certain circumstances, it is required to estimate a nonparametric kernel density that involves associated complications, e.g., bandwidth selection. Due to this, Basu et al. (1997, 1998) proposed another minimum distance estimator, the MDPD estimator, which minimized the divergence between the assumed parametric density and true density without using any nonparametric smoothing.

The trade-off between the robustness and asymptotic efficiency of the MDPD estimator is decided by a tuning parameter $\alpha \geq 0$. As α increases, the robustness of the MDPD estimator is improved by a relative-to-the-model downweighting for outliers, but that comes at a price of less efficient estimation. To address this issue, several examples were tested in Basu et al. (1997, 1998), and the heuristic conclusion suggested that α between 0.1 and 0.25 gave satisfactory robustness. Yet there has been no strict guideline to settle on a proper value of α while retaining efficiency. When $\alpha = 1$, the MDPD estimator is to minimize the integrated squared error between the model and true densities. From this point, Scott (1998, 2001) defined the L_2E estimation theory and presented that it received advantages of less computation time and accurate optimization results. Recently, Warwick (2005) and Warwick and Jones (2005) used the L_2E estimator as a pilot estimator to choose a proper value of α in the MDPD estimation method.

The importance of the MHD and MDPD estimators for the FMMs is underscored by the growing literature. Several key articles are as follows. In work related to the MHD estimator for the FMMs, Cutler and Cordero-Brana (1996) provided a comprehensive Monte Carlo studies on the robustness and efficiency of

the estimator. They concluded that the estimator was asymptotically efficient if data came from the parametric family and was robust to gross-error contamination. Karlis and Xekalaki (1998) extended Cutler and Cordero-Brana (1996) to a discrete case by defining the MHD estimator for the Poisson mixtures. Their experiment showed that the MHD estimator was more robust than the ML one and attained the first-order efficiency. Meanwhile, they introduced an estimation algorithm, which facilitated computation without requiring the second-order derivative. In terms of the MDPD estimation method for the FMMs, Scott (2001, 2004) presented that the L_2E criteria was not only easy to be set up for some complicated model specifications but also suitable for analyzing massive data when data clearing was unfeasible and estimation efficiency was not the first priority.

Surprisingly, there is still little attention drawn to developing the MHD and MDPD estimators for finite mixtures of count data regression models. So far, only Lu et al. (2003) presented an MHD estimation theory for the finite mixture Poisson regression model. But their theory was obtained by minimizing the Hellinger distance between the estimated unconditional marginal densities, instead of conditional ones in a regression context. Nonetheless, Lu et al. (2003) claimed that the estimates should be consistent because the unconditional marginal densities were identifiable. Considering the modeling convenience, this dissertation takes their approach to empirical application. Regarding the MDPD estimator for finite mixtures of count data regression models, there is lack of relevant research into the corresponding estimation theory. This dissertation contributes the derivation of its asymptotic properties based on a result in Basu et al. (1997, 1998), which verified that the MDPD estimator was also the M-estimator. As a result, the asymptotic properties of this estimator can be defined through the asymptotic theory of M-estimates.

One of my motivations behind the proposed approach is to pursue a better analysis of economic data consisting of heterogeneous subpopulations and extreme values. This kind of data is commonly observed in the research on demand for

retail credits or health care due to various consumer behaviors contained in private information sets. For illustration purpose, the proposed model specification and estimation methods are employed to study the data sets used in Dionne et al. (1996) and Deb and Trivedi (2002), respectively. The first data set is the extent of non-payments of personal loans in Spain and the second one is counts of utilization from the RAND Health Insurance Experiment. To the best of my knowledge, no finite mixtures of GLMs have been applied in the existing literature on retail credits. The empirical application to the first data set is hoped to provide a better understanding of heterogeneity in this subject. For the second data set, Deb and Trivedi (2002) already showed that the finite mixture count data regression model had a better fit than the hurdle model but their estimation method was the ML algorithm.

The rest of the dissertation is organized as follows. Chapter 2 introduces finite mixtures of count data regression models and two minimum distance estimators, the MHD and MDPD estimators. For model and estimator selection, the diagnostic tools are presented in Chapter 3. Two Monte Carlo studies are carried out in Chapter 4 to compare efficiency and robustness among the ML, MHD, and L_2E estimators. Chapter 5 applies the proposed approach to analyze, respectively, non-payments of personal loans in Spain and health care demand in the United States. Finally, Chapter 6 concludes the findings and provides further extensions of this dissertation.

Chapter 2

Model and Estimator Developments

Finite mixtures of count data regression models are established in Section 2.1. The density specifications includes the Poisson, negative binomial one, and negative binomial two distributions. After that, two minimum distance estimation methods, the minimum Hellinger distance estimation and the minimum density power divergence estimation, are developed in Section 2.2. The minimum L_2 error estimation is introduced as a special case of the minimum density power divergence estimation.

2.1 Finite Mixtures of Count Data Regression Models

Let the random count variable Y_i be the response and the random vector $\boldsymbol{X}_i$ the covariates, where $\{(Y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^q \mid Y_i = 0, 1, 2, \ldots, q \geq 1, \text{and}\, i = 1, \ldots, n\}$ is a set of independent random pairs. The finite mixture regression model is defined by

$$f_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i) \;=\; \sum_{j=1}^{k} \pi_j f_j(y_i \mid \boldsymbol{x}_i), \tag{2.1}$$

where $k \geq 1$, $\pi_j \in (0, 1)$, and $\sum_{j=1}^{k} \pi_j = 1$.

For count data regression models, the Poisson distribution is usually served as the benchmark distribution which not only gives basic description of a random variable but also provides easy convergence conditions. But the equidispersion property of the Poisson is a weakness for fitting the empirical data with overdispered counts. To fix this problem, the negative binomial distribution, equipped with an overdispersion parameter, is adopted.

Therefore, finite mixtures of Poisson regressions (FMP) is developed by specifying $f_j(y_i \mid \boldsymbol{x}_i)$ in (2.1) as

$$f_j\left(y_i \mid \boldsymbol{x}_i\right) = e^{-\mu_{ij}}\mu_{ij}^{y_i}/y_i!, \tag{2.2}$$

where the mean equation $\mu_{ij} = \exp\left(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_j\right)$, the conditional variance $\sigma_{ij} = \mu_{ij}$, and $\boldsymbol{\theta} = \left(\pi_1, \ldots, \pi_{k-1}, \boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_k^\intercal\right)^\intercal$. To develop finite mixtures of negative binomial regressions (FMNB), $f_j(y_i \mid \boldsymbol{x}_i)$ is defined by

$$f_j\left(y_i \mid \boldsymbol{x}_i\right) = \frac{\Gamma\left(\alpha_{ij}^{-1} + y_i\right)}{\Gamma\left(\alpha_{ij}^{-1}\right)\Gamma\left(y_i + 1\right)}\left(\frac{\alpha_{ij}^{-1}}{\alpha_{ij}^{-1} + \mu_{ij}}\right)^{\alpha_{ij}^{-1}}\left(\frac{\mu_{ij}}{\alpha_{ij}^{-1} + \mu_{ij}}\right)^{y_i}, \tag{2.3}$$

where $\mu_{ij} = \exp\left(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_j\right)$ and $\sigma_{ij} = \mu_{ij}\left(1 + \alpha_{ij}\mu_{ij}\right)$. If the precision parameter $\alpha_{ij} = \psi_j/\mu_{ij}$, $f_j(y_i \mid \boldsymbol{x}_i)$ is constructed by the NB1 density which has a linear variance form. If $\alpha_{ij} = \psi_j$, $f_j(y_i \mid \boldsymbol{x}_i)$ is constructed by the NB2 density which has a quadratic variance form. Thus $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{k-1}, \psi_1, \ldots, \psi_k, \boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_k^\intercal)^\intercal$ and ψ_j is an overdispersion parameter.

For finite mixtures of count data regression models, the ratio of conditional variance to conditional mean can be greater or less than 1 that implies over- or underdispersion can be accommodated in the data. However, their manner to cope with over- or underdispersion is different from that a hurdle regression model does.[1] Table 2.1 lists the acronyms for all model specifications applied in Chapters 4 and 5.

[1]See Appendix A for the further discussion.

Table 2.1: Model Description

Acronym	Description
SP	standard Poisson regression model
SNB1	standard NB1 regression model
SNB2	standard NB2 regression model
HP	hurdle model with a truncated Poisson regression model
HNB1	hurdle model with a truncated NB1 regression model
HNB2	hurdle model with a truncated NB2 regression model
2-FMP	2-component Poisson mixture regression model
2-FMNB1	2-component NB1 mixture regression model
2-FMNB2	2-component NB2 mixture regression model
3-FMP	3-component Poisson mixture regression model
3-FMNB1	3-component NB1 mixture regression model
3-FMNB2	3-component NB2 mixture regression model

2.2 MHD and MDPD Estimation Methods

Two minimum distance estimation methods, the MHD and the MDPD estimation methods, are introduced in Sections 2.2.1 and 2.2.2, respectively. The L_2E estimation method, a special case of the MDPD estimation method, is included in Section 2.2.2. All of these estimators were proved to be consistent and have asymptotic normalities.

2.2.1 Minimum Hellinger Distance Estimation

According to Beran (1977), the MHD estimator of $\boldsymbol{\theta}$ is that value $\hat{\boldsymbol{\theta}}_{MHD}$ in the parameter space $\boldsymbol{\Theta}$ which minimizes the Hellinger distance between a specified parametric density, $g_{\boldsymbol{\theta}}$, and a suitable nonparametric density, g_n. The corresponding joint distributions are denoted as $G_{\boldsymbol{\theta}}$ and G_n, respectively, and the objective function is

$$\hat{\boldsymbol{\theta}}_{MHD} = \arg\min_{\boldsymbol{\theta}} \parallel g_{\boldsymbol{\theta}}^{1/2} - g_n^{1/2} \parallel_2^2 . \tag{2.4}$$

To develop the MHD estimator for regression models, let $g_{\boldsymbol{\theta}}$ and g_n be the joint densities of $(Y, \boldsymbol{X})$, where $g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) = f_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})$ and $g_n(y, \boldsymbol{x}) = f_n(y \mid \boldsymbol{x}) f_{\boldsymbol{X}}$

$(\boldsymbol{x})$. Thus

$$\hat{\boldsymbol{\theta}}_{MHD} = \arg\min_{\boldsymbol{\theta}} \parallel g_{\boldsymbol{\theta}}^{1/2}(y, \boldsymbol{x}) - g_n^{1/2}(y, \boldsymbol{x}) \parallel_2^2,$$

which is equivalent to

$$\hat{\boldsymbol{\theta}}_{MHD} = \arg\min_{\boldsymbol{\theta}} \iint \left[f_{\boldsymbol{\theta}}^{1/2}(y \mid \boldsymbol{x}) - f_n^{1/2}(y \mid \boldsymbol{x}) \right]^2 f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x} dy. \tag{2.5}$$

Considering the difficulty to implement $f_n(y \mid \boldsymbol{x})$ in the objective function for count data regression models, Lu et al. (2003) minimized the Hellinger distance with respect to $\boldsymbol{\theta}$ between the unconditional densities, $f_{\boldsymbol{\theta}}(y)$ and $f_n(y)$. They approximated $f_{\boldsymbol{\theta}}(y)$ by a consistent estimator, $f_{\boldsymbol{\theta},n}(y) = n^{-1}\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(y \mid \boldsymbol{X}_i)$, and took $f_n(y)$ to be the empirical mass function $f_n(y) = N_y/n$, where $y = 0, 1, 2, \ldots, m$ and $m = \max\{Y_i \mid i = 1, \ldots, n\}$. Consequently, $\hat{\boldsymbol{\theta}}_{MHD}$ for count data regression models is defined as

$$\hat{\boldsymbol{\theta}}_{MHD} = \arg\min_{\boldsymbol{\theta}} \sum_{y=0}^{m} \left[f_{\boldsymbol{\theta},n}^{1/2}(y) - f_n^{1/2}(y) \right]^2, \tag{2.6}$$

which is the same as

$$\hat{\boldsymbol{\theta}}_{MHD} = \arg\max_{\boldsymbol{\theta}} \sum_{y=0}^{m} f_{\boldsymbol{\theta},n}^{1/2}(y)\, f_n^{1/2}(y).$$

For the asymptotic properties of $\hat{\boldsymbol{\theta}}_{MHD}$, Simpson (1987, Section 3) imposed smoothness conditions to derive the asymptotic normality of $\hat{\boldsymbol{\theta}}_{MHD}$ in discrete densities. Based on this, the asymptotic normality of $\hat{\boldsymbol{\theta}}_{MHD}$ for count data regression models is

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{MHD} - \boldsymbol{\theta}_0\right) \underset{d}{\to} N\left(\mathbf{0}, \frac{1}{4}\ddot{\boldsymbol{V}}^{-1}(\boldsymbol{\theta}_0)\, \boldsymbol{i}(\boldsymbol{\theta}_0)\, \ddot{\boldsymbol{V}}^{-1}(\boldsymbol{\theta}_0)\right), \tag{2.7}$$

where $\boldsymbol{V}(\boldsymbol{\theta}_0) = \parallel f_{\boldsymbol{\theta}_0}^{1/2}(y) - f_n^{1/2}(y) \parallel_2^2$, $\ddot{\boldsymbol{V}}(\boldsymbol{\theta}_0) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}}\boldsymbol{V}(\boldsymbol{\theta}_0)$, $\boldsymbol{i}(\boldsymbol{\theta}_0) = \sum_{y=0}^{m} \boldsymbol{l}_{\boldsymbol{\theta}_0}(y)\, \boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}$

(y) $f_{\boldsymbol{\theta}_0}(y)$, and $\boldsymbol{l}_{\boldsymbol{\theta}_0}(y) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}_0}(y)$.

The asymptotic variance of $\hat{\boldsymbol{\theta}}_{MHD}$ Avar $\left(\hat{\boldsymbol{\theta}}_{MHD}\right)$ is consistently estimated by

$$\widehat{\text{Avar}}\left(\hat{\boldsymbol{\theta}}_{MHD}\right) = \frac{1}{4}\ddot{\boldsymbol{V}}^{-1}\left(\hat{\boldsymbol{\theta}}_{MHD}\right)\boldsymbol{i}\left(\hat{\boldsymbol{\theta}}_{MHD}\right)\ddot{\boldsymbol{V}}^{-1}\left(\hat{\boldsymbol{\theta}}_{MHD}\right). \tag{2.8}$$

If $G_n \equiv G_{\boldsymbol{\theta}}$, Simpson (1987, Theorem 2) proved Avar $\left(\hat{\boldsymbol{\theta}}_{MHD}\right)$ could be simplified as

$$\text{Avar}\left(\hat{\boldsymbol{\theta}}_{MHD}\right) = \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0). \tag{2.9}$$

Instead of using the simplified Avar $\left(\hat{\boldsymbol{\theta}}_{MHD}\right)$ above, Lu et al. (2003) derived theirs on the basis of another result in Simpson (1987, Theorem 2) that $\hat{\boldsymbol{\theta}}_{MHD}$ was asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{ML}$ when $G_n \equiv G_{\boldsymbol{\theta}}$. Therefore,

$$\boldsymbol{i}(\boldsymbol{\theta}_0) = \sum_{y=0}^{m}\left\{\boldsymbol{l}_{\boldsymbol{\theta}_0}(y)\boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}(y) f_n(y) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}} f_{\boldsymbol{\theta}_0}(y)\right\}. \tag{2.10}$$

The proof of (2.10) is in Appendix B. Avar $\left(\hat{\boldsymbol{\theta}}_{MHD}\right)$ in Lu et al. (2003) is then estimated by

$$\widehat{\text{Avar}}\left(\hat{\boldsymbol{\theta}}_{MHD}\right) = \left[\sum_{y=0}^{m}\left\{\boldsymbol{l}_{\hat{\boldsymbol{\theta}}_{MHD}}(y)\boldsymbol{l}_{\hat{\boldsymbol{\theta}}_{MHD}}^{\intercal}(y) f_n(y) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}} f_{\hat{\boldsymbol{\theta}}_{MHD}}(y)\right\}\right]^{-1}. \tag{2.11}$$

Since the condition $G_n \equiv G_{\boldsymbol{\theta}}$ is uncertain for the empirical applications in Chapter 5, the computation of $\widehat{\text{Avar}}\left(\hat{\boldsymbol{\theta}}_{MHD}\right)$ is based on (2.8).

2.2.2 Minimum Density Power Divergence Estimation

Unlike the MHD estimation method, Basu et al. (1997, 1998) developed the MDPD estimation method to avoid the use of nonparametric smoothing. The MDPD estimator $\hat{\boldsymbol{\theta}}_{MDPD}$ of $\boldsymbol{\theta}$ in the parameter space $\boldsymbol{\Theta}$ minimizes the divergence between the assumed parametric density $g_{\boldsymbol{\theta}}$ and the true density g

$$\hat{\boldsymbol{\theta}}_{MDPD} = \arg\min_{\boldsymbol{\theta}} d_\alpha(g, g_{\boldsymbol{\theta}}), \tag{2.12}$$

where $d_\alpha(g, g_{\boldsymbol{\theta}}) = \int \left\{ g_{\boldsymbol{\theta}}^{1+\alpha}(\boldsymbol{z}) - (1+1/\alpha)\, g(\boldsymbol{z})\, g_{\boldsymbol{\theta}}^{\alpha}(\boldsymbol{z}) + (1/\alpha)\, g^{1+\alpha}(\boldsymbol{z}) \right\} d\boldsymbol{z}$. In this estimation method the tuning parameter $\alpha \geq 0$ decides the trade-off between robustness and efficiency. Since the integrand in $d_\alpha(g, g_{\boldsymbol{\theta}})$ is undefined when $\alpha = 0$, Basu et al. (1997, 1998) defined this case as the Kullback-Leibler divergence via an approximation, i.e., $d_0(g, g_{\boldsymbol{\theta}}) = \lim_{\alpha \to 0} d_\alpha(g, g_{\boldsymbol{\theta}}) = \int \log\left\{ g(\boldsymbol{z}) / g_{\boldsymbol{\theta}}(\boldsymbol{z}) \right\} g(\boldsymbol{z})\, d\boldsymbol{z}$. As a result, the ML estimation method is a special case of the MDPD estimation method. When $\alpha = 1$, $d_1(g, g_{\boldsymbol{\theta}}) = \int \left[g_{\boldsymbol{\theta}}(\boldsymbol{z}) - g(\boldsymbol{z}) \right]^2 d\boldsymbol{z}$ is the integrated squared error which is the same as the estimation method developed by Scott (1998, 2001), the minimum L_2 error estimation method.

To extend the MDPD estimation method from parametric distributions to regression models, Basu et al. (1997) replaced $g_{\boldsymbol{\theta}}$ and g with the conditional densities $f_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$ and $f(y \mid \boldsymbol{x})$, respectively, and assumed the marginal density $f_{\boldsymbol{X}}(\boldsymbol{x})$ is the empirical distribution of $\{\boldsymbol{X}_i \mid i = 1, \ldots, n\}$. Thus

$$\hat{\boldsymbol{\theta}}_{MDPD} = \arg\min_{\boldsymbol{\theta}} \left[n^{-1} \sum_{i=1}^{n} \int f_{\boldsymbol{\theta}}^{1+\alpha}(y \mid \boldsymbol{X}_i)\, dy - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}^{\alpha}(Y_i \mid \boldsymbol{X}_i) \right]. \tag{2.13}$$

For count data regression models, the objective function is

$$\hat{\boldsymbol{\theta}}_{MDPD} = \arg\min_{\boldsymbol{\theta}} \left[n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{m} f_{\boldsymbol{\theta}}^{1+\alpha}(y \mid \boldsymbol{X}_i) - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}^{\alpha}(Y_i \mid \boldsymbol{X}_i) \right]. \tag{2.14}$$

Because Basu et al. (1998, Section 3) showed that $\hat{\boldsymbol{\theta}}_{MDPD}$ was actually the M-estimator, the asymptotic properties of $\hat{\boldsymbol{\theta}}_{MDPD}$ can be established by the asymptotic theory of M-estimates.[2] For $\hat{\boldsymbol{\theta}}_{MDPD}$, the ρ and $\boldsymbol{\Psi}$ functions of the M-estimator are defined, respectively, as $\rho(\boldsymbol{\theta}) = \int f_{\boldsymbol{\theta}}^{1+\alpha}(y \mid \boldsymbol{X}_i)\, dy - (1+1/\alpha)\ f_{\boldsymbol{\theta}}^{\alpha}(Y_i \mid \boldsymbol{X}_i)$ and $\boldsymbol{\Psi}(\boldsymbol{\theta}) = \int \boldsymbol{u}_{\boldsymbol{\theta}}(y \mid \boldsymbol{X}_i)\, f_{\boldsymbol{\theta}}^{1+\alpha}(y \mid \boldsymbol{X}_i)\, dy - \boldsymbol{u}_{\boldsymbol{\theta}}(Y_i \mid \boldsymbol{X}_i)\, f_{\boldsymbol{\theta}}^{\alpha}(Y_i \mid \boldsymbol{X}_i)$,

[2] See Hampel et al. (1986) for more details about M-estimators.

where $\boldsymbol{u}_{\boldsymbol{\theta}}(\cdot) = \partial \log f_{\boldsymbol{\theta}}(\cdot) / \partial \boldsymbol{\theta}$. Then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{MDPD} - \boldsymbol{\theta}_0\right) \underset{d}{\rightarrow} N\left(\mathbf{0}, E\left[\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0)\right]^{-1} \boldsymbol{\Sigma} E\left[\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0)\right]^{-1}\right), \tag{2.15}$$

where $\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}^{\intercal}} \boldsymbol{\Psi}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Sigma} = E\left[\boldsymbol{\Psi}(\boldsymbol{\theta}_0)\boldsymbol{\Psi}(\boldsymbol{\theta}_0)^{\intercal}\right]$. The expansions of $\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Sigma}$ are presented in Appendix C. The asymptotic variance of $\hat{\boldsymbol{\theta}}_{MDPD}$ Avar $\left(\hat{\boldsymbol{\theta}}_{MDPD}\right)$ is consistently estimated by

$$\widehat{\text{Avar}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right) = \left[\frac{1}{n}\sum_{i=1}^{n} \widehat{\dot{\boldsymbol{\Psi}}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right)\right]^{-1} \widehat{\boldsymbol{\Sigma}} \left[\frac{1}{n}\sum_{i=1}^{n} \widehat{\dot{\boldsymbol{\Psi}}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right)\right]^{-1} \tag{2.16}$$

The expansions of $\widehat{\dot{\boldsymbol{\Psi}}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right)$ and $\widehat{\boldsymbol{\Sigma}}$ for count data regression models are showed in Appendix D.

Chapter 3

Model and Estimator Selections

To select the best fit model estimated by the ML, MHD, and L_2E estimation methods and also to decide the suitable estimator, three diagnostic tools are applied: the Pearson's chi-square statistic, root mean squared error, and cross-validation. They are sequentially introduced in Section 3.1 and 3.2.

3.1 Pearson's Chi-square Statistic and Root Mean Squared Error

The Pearson's chi-square statistic χ^2_P is given by

$$\chi^2_P = \sum_{c=1}^{C} \left[nf\left(y=c\right) - n\hat{f}\left(y=c\right) \right]^2 / \left[n\hat{f}\left(y=c\right) \right], \tag{3.1}$$

where C is the total number of mutually exclusive cells into one of which the same count y is collected, f is the sample cell frequency, $\hat{f}$ is the fitted one, and n is the sample size. The degrees of freedom of the χ^2_P is $C-1$, and $\hat{f}\left(y=c\right)$ is approximated by $n^{-1}\sum_{i=1}^{n} f_{\hat{\boldsymbol{\theta}}}\left(y=c \mid \boldsymbol{x}_i\right)$. To evaluate the adequacy-of-fit in more detail, the subsample Pearson's chi-square statistic $\chi^2_{P|x}$ is developed and denoted

as

$$\chi^2_{P|x} = \sum_{c=1}^{C} \left[nf\left(y = c \mid x\right) - n\hat{f}\left(y = c \mid x\right) \right]^2 / \left[n\hat{f}\left(y = c \mid x\right) \right], \quad (3.2)$$

where $\hat{f}\left(y = c \mid x\right) \approx n_x^{-1} \sum_{i|x} f_{\hat{\boldsymbol{\theta}}}\left(y = c \mid x_i\right)$. However, there is a deficiency in applying the Pearson's chi-square goodness-of-fit test because the estimation error of $\hat{f}$ is not controlled.[1] Hence, the statistic values are ranked as a measure of fit: the smaller the statistic value, the better fit.

The root mean squared error $RMSE$ is known as the standard error of the regression which measures the deviation of actual response values from their corresponding fitted ones. It is denoted by

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 / \left(n - q\right)}, \quad (3.3)$$

where y_i and $\hat{y}_i$ are the ith response value and its fit, respectively, n is the sample size, and q is the number of estimated parameters in a model. To assist with the evaluation, the subsample root mean squared error $RMSE_{|x}$ is also adopted and specified by

$$RMSE_{|x} = \sqrt{\sum_{i|x}^{n_x} \left(y_i - \hat{y}_i\right)^2 / \left(n_x - q\right)}. \quad (3.4)$$

3.2 Cross-validation

There are two concerns for only using the diagnostic tools introduced in Section 3.1. First, their statistic values belong to the in-sample measure that throws doubt on whether their fitting evaluation is consistent with that of the out-of-sample fit. Second, there exists a selection bias of estimator in the formula of the χ^2_P and $RMSE$, respectively, that is explained as follows.

[1] Cameron and Trivedi (1998, Chapter 5) attributed this estimation error to $\hat{f}\left(y = c\right)$, which was the arithmetic average of $f_{\hat{\boldsymbol{\theta}}}\left(y = c \mid \boldsymbol{x}_i\right)$. Although they provided another test, the conditional moment test which controlled the estimation error by taking into account the covariance matrix, it is hard to examine whether the required assumptions for this test are satisfied or not in practice. Therefore, the conditional moment test is not popularly used in empirical studies.

For the χ^2_P, the objective function of the MHD estimation method (2.6) is similar to the χ^2_P formula (3.1). Both of them measure the deviation between the sample and fitted cell frequencies. In turn, the MHD estimator is more likely picked up by this goodness-of-fit statistic than the ML and L_2E estimators. Although the L_2E estimator is in the class of the minimum distance estimators, its objective function does not directly include the sample cell frequencies.

For the $RMSE$, the objective function of the ML estimator gives each observation equal weight like the $RMSE$. Besides, the ML estimates of the Poisson regression are equivalent to the estimates that minimize the sum of squared residuals. On the other hand, the weight in the minimum distance estimation methods is assigned by the parametric density. Therefore, the $RMSE$ tends to choose the ML estimator.

To avoid these problems, the cross-validation is an appropriate technique to select the best fit model and suitable estimator. The procedure of cross-validation starts with randomly assigning observations to two subsamples, i.e., "training sample" and "hold-out sample", with 70% and 30%, respectively, of an entire sample. The training sample is used for estimation and the hold-out sample for forecast comparison based on those estimates. In this dissertation the χ^2_P and $RMSE$ are calculated as performance measures for model and estimator selection.

Chapter 4

Monte Carlo Simulation Studies

The following Monte Carlo studies are designed to compare the finite sample properties among the ML, MHD, and L_2E estimators. All of the experiments are processed in MATLAB environment and the optimization algorithm **KNITRO** of Ziena Optimization, Inc. is used as a main tool to solve the optimization problems.[1] The same optimization method is applied to the empirical application in Chapter 5.

4.1 Efficiency Comparison among the ML, MHD, and L_2E Estimators

For efficiency comparison, the DGP is modeled as the 2-FMP,

$$f_{\boldsymbol{\theta}}\left(y_i \mid \boldsymbol{x}_i\right) \;=\; \sum_{j=1}^{2} \pi_j f_j(y_i \mid \boldsymbol{x}_i),$$

where $\left\{\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^{\intercal} \mid x_{i1} = x_{i3} = 1, x_{i2} \text{ and } x_{i4} \sim U\left(0,1\right)\right\}$ and $\boldsymbol{\theta} =$

[1] Several optimization algorithms have been tested to solve the estimation problems here, including **FMINUNC** of The MathWorks, Inc., **NPSOL** of Stanford Business Software, Inc., and **UCSOLVE** of TOMLAB Optimization, Inc. According to my experience, the best outcome is achieved by **KNITRO** combined with the analytical gradient and Hessian routines. The automatic differentiation technique can give a hand with the analytical gradient and Hessian routines, but it asks for more time and computer memory to solve problems. For the introductions of optimization algorithm **KNITRO**, **FMINUNC**, **NPSOL**, and **UCSOLVE**, see Waltz and Plantenga (2009), Coleman and Zhang (2009), and Holmström et al. (2008, 2009), respectively.

$(\pi_1, \beta_1, \beta_2, \beta_3, \beta_4)^\intercal$ with two parameter settings, i.e., $\boldsymbol{\theta}_0 = (0.5, 0.6, 1, 2.5, 1.5)^\intercal$ and $(0.95, 0.6, 1, 2.5, 1.5)^\intercal$.[2] Each of the experiments generates 256 replications with 1e+2 and 1e+4 observations, respectively.[3]

The simulation results are presented by the average (Ave), standard deviation (StD), and mean squared error (MSE) of parameter estimates produced by the ML, MHD, and L_2E estimation methods.[4] At the same time, the distributional plots of parameter estimates are provided as a visual aid. To examine whether the χ^2_P and $RMSE$ are fair diagnostic tools to select the suitable estimator from the ML, MHD, and L_2E estimators, their statistic values are ranked for each replication and then the ranking frequencies are counted. It is expected that the best estimator has the most high rankings if the χ^2_P and $RMSE$ are unbiased.

Table 4.1 presents the Aves, StDs, and MSEs of the three estimators, and Figures 4.1 and 4.2 present the histograms and kernel density estimates of the parameter estimates. All of them show that the MHD and L_2E estimators are likely biased and more inefficient than the ML one, but their finite sample properties are improved as the sample size rises from 1e+2 to 1e+4.

Raising π_1 from 0.5 to 0.95, the biasedness and inefficiency of the MHD and L_2E estimates become shockingly serious at the sample size equal to 1e+2, but the ML estimates are not impacted significantly. When the sample size grows to 1e+4, these estimation problems for the MHD and L_2E estimators are less severe. There is one common phenomenon among the ML, MHD, and L_2E estimators

[2]The purpose of designing two different values for π_1, i.e., 0.5 and 0.95, is to examine whether the finite sample properties of the ML, MHD, and L_2E estimators could change as the mixing probability approaches 1. For the covariate values, Lu et al. (2003) had two kinds of settings, $(0.6, 1, 2.5, 1.5)$ and $(0.6, 1, 0.8, 1.2)$, in their Monte Carlo simulation study, but only the first, well-separated, covariate setting is used here. There are two reasons to do so: first, from computational perspective the finite mixture model does not work well when covariates values of different component densities are such close; second, from economic perspective there is no necessity of applying the finite mixture model if the heterogeneity between subpopulations is minor.

[3]The choice of the replication number 256 is based on statisticians' heuristic rule and considers the computation time. Scott (1998, 2001) also adopted this number in his experiments; on the other hand, there was only 100 replications in the experiments of Lu et al. (2003). The observation number from 1e+2 to 1e+4 provides a good indication of the convergence speeds for the ML, MHD, and L_2E estimators. For similar experiments in Lu et al. (2003), the authors only considered the sample size equal to 100 and 200, respectively.

[4]The estimated covariates of the smaller component expected value are labeled as the first component covariates, i.e., $E(y \mid \beta_1 x_1 + \beta_2 x_2) > E(y \mid \beta_3 x_3 + \beta_4 x_4)$.

Table 4.1: Averages, Standard Deviations, and Mean Squared Errors of the ML, MHD, and L_2E Estimates (2-component Poisson Mixture Regression Models)

	ML			MHD			L_2E		
	Ave	StD	MSE	Ave	StD	MSE	Ave	StD	MSE
$\boldsymbol{\theta}_0$	Sample Size = 1e+2								
$\pi_1 = 0.5$	0.500	0.005	2.2e-5	0.600	0.034	0.011	0.503	0.023	0.001
$\beta_1 = 0.6$	0.589	0.191	0.036	0.772	0.358	0.157	0.586	0.230	0.053
$\beta_2 = 1.0$	1.010	0.285	0.081	0.623	0.649	0.561	1.014	0.353	0.125
$\beta_3 = 2.5$	2.499	0.066	0.004	2.512	0.219	0.048	2.496	0.085	0.007
$\beta_4 = 1.5$	1.502	0.098	0.010	1.337	0.258	0.093	1.522	0.129	0.017
	Sample Size = 1e+4								
$\pi_1 = 0.5$	0.500	4.6e-4	2.1e-7	0.501	0.002	3.2e-6	0.500	0.003	6.2e-6
$\beta_1 = 0.6$	0.599	0.021	4.6e-4	0.601	0.038	0.001	0.599	0.024	0.001
$\beta_2 = 1.0$	1.001	0.033	0.001	0.995	0.068	0.005	1.000	0.037	0.001
$\beta_3 = 2.5$	2.501	0.006	4.2e-5	2.501	0.014	1.9e-4	2.501	0.008	6.8e-5
$\beta_4 = 1.5$	1.499	0.009	8.9e-5	1.494	0.016	2.9e-4	1.499	0.012	1.5e-4
	Sample Size = 1e+2								
$\pi_1 = 0.95$	0.949	0.004	1.8e-5	0.894	0.263	0.072	0.734	0.302	0.137
$\beta_1 = 0.6$	0.598	0.131	0.017	0.012	3.577	13.09	0.268	1.970	3.976
$\beta_2 = 1.0$	0.999	0.211	0.044	1.004	3.584	12.79	1.153	2.093	4.388
$\beta_3 = 2.5$	2.459	0.303	0.093	1.754	2.625	7.418	1.315	1.152	2.725
$\beta_4 = 1.5$	1.541	0.486	0.237	1.433	2.796	7.793	1.769	3.686	13.60
	Sample Size = 1e+4								
$\pi_1 = 0.95$	0.950	2.6e-4	6.6e-8	0.952	0.001	3.7e-6	0.950	0.003	9.1e-6
$\beta_1 = 0.6$	0.599	0.014	1.9e-4	0.600	0.023	0.001	0.599	0.016	2.6e-4
$\beta_2 = 1.0$	1.001	0.021	4.4e-4	1.000	0.039	0.001	1.001	0.025	0.001
$\beta_3 = 2.5$	2.499	0.021	4.3e-4	2.499	0.064	0.004	2.498	0.025	0.001
$\beta_4 = 1.5$	1.501	0.030	0.001	1.465	0.071	0.006	1.503	0.039	0.001

Figure 4.1: Distributions of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 1 (2-component Poisson Mixture Regression Models)

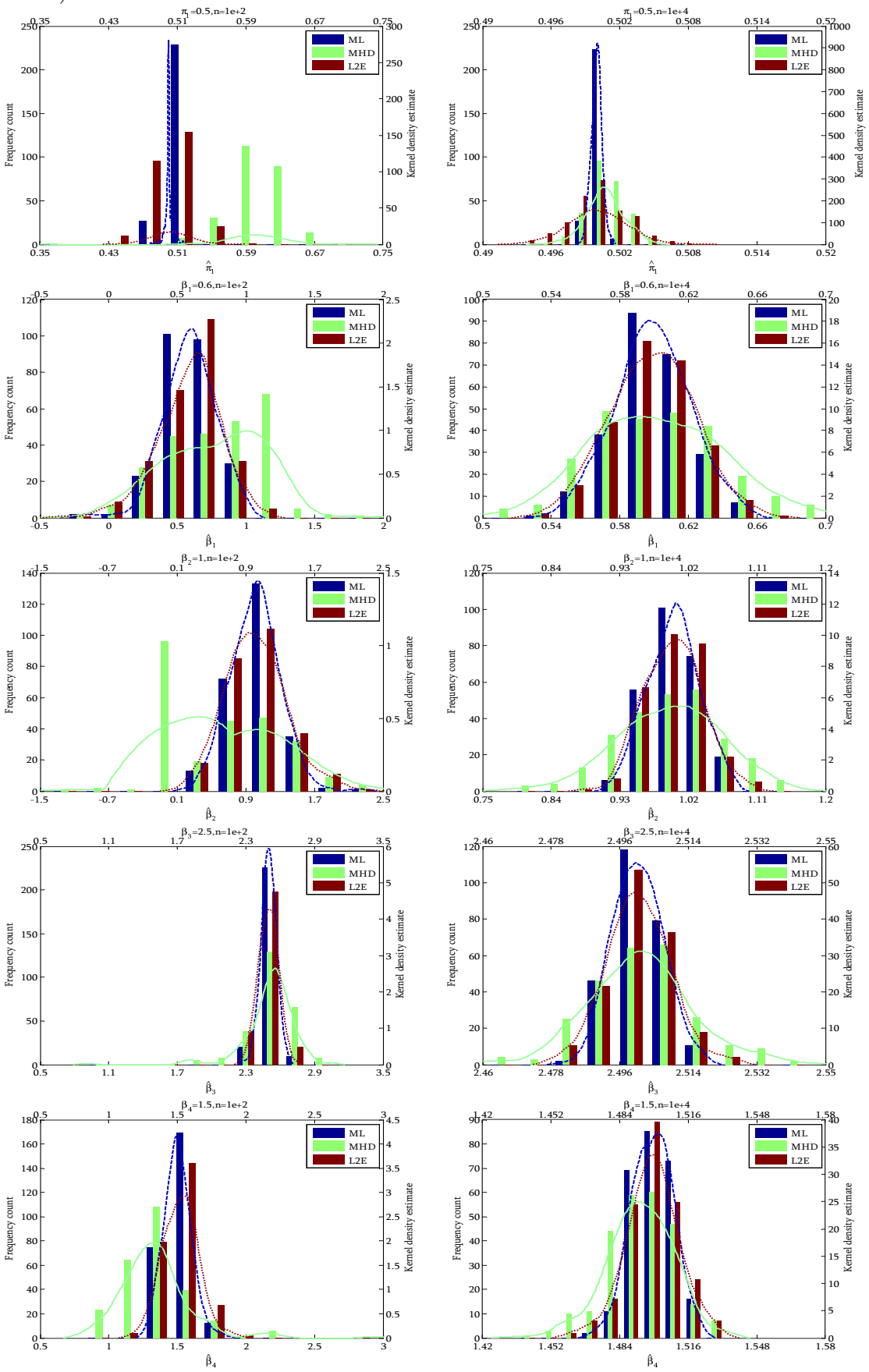
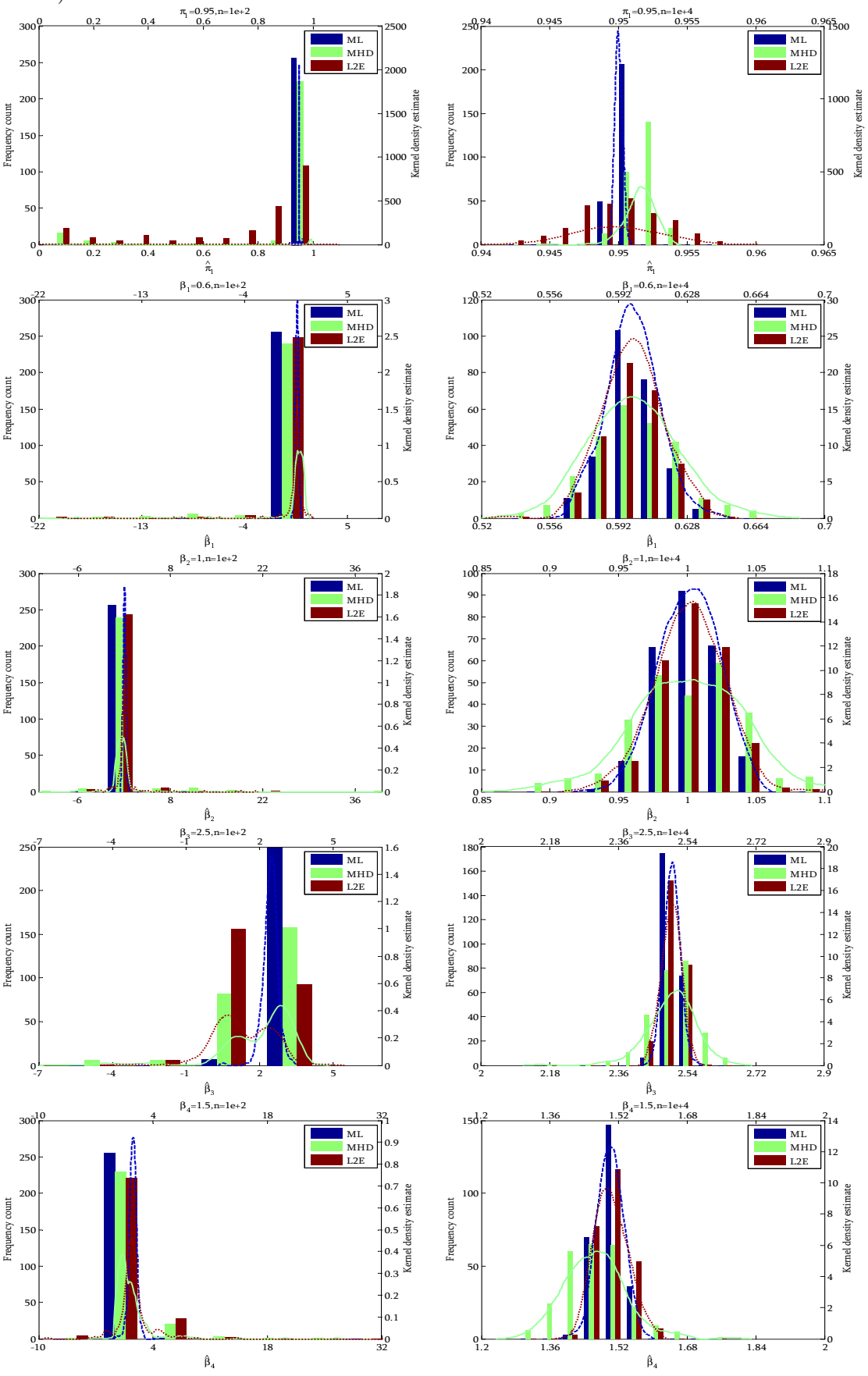


Figure 4.2: Distributions of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 2 (2-component Poisson Mixture Regression Models)



that the estimates of β_3 and β_4 have much larger StDs than the other parameter estimates at $\pi_1 = 0.95$.

Next, the ranking frequencies of the χ^2_P and $RMSE$ in Table 4.2 assess the adequacy-of-fit among the ML, MHD, and L_2E estimators. It is known that when a model is correctly specified, the ML estimator must be the best estimator. Indeed, this Monte Carlo study has the anticipated conclusion according to Table 4.1, and Figures 4.1 and 4.2. Consequently, the ML estimator should have the most high rankings of the χ^2_P and $RMSE$.

However, Table 4.2 shows that the MHD estimator has the most No. 1 rankings of the χ^2_P as the sample size rises to 1e+4 from 1e+2. On the other hand, the ranking frequencies of the $RMSE$ supports that the ML estimator is the best estimator no matter the sample size is equal to 1e+2 or 1e+4. Table 4.2 supports the argument in Section 3.2 that the MHD estimator is favored by the χ^2_P.

4.2 Robustness Comparison among the ML, MHD, and L_2E Estimators

For robustness comparison, the DGP is constructed by adding a third component, a Poisson distribution, to the 2-FMP. The third component plays a role of the "contamination" component in the DGP where it contaminates the simulated data via the contamination rate a and mean parameter z. The DGP is given by

$$f_{\boldsymbol{\theta}^c}(y_i \mid \boldsymbol{x}_i) \;=\; (1-a)\sum_{j=1}^{2}\pi_j f_j(y_i \mid \boldsymbol{x}_i) + a g_z(y_i), \tag{4.1}$$

where $\mu_{iz} = \exp(z)$ is the mean of the third component $g_z(y_i)$. Then $\boldsymbol{\theta}^c = (\boldsymbol{\theta}, a, z)^\intercal$ and $\boldsymbol{\theta} = (\pi_1, \beta_1, \beta_2, \beta_3, \beta_4)^\intercal$.

By changing the values of a and z in (4.1), two following contamination scenarios are created to examine the robustness of the three estimators. The first scenario is to alter the value of a from 0.02 to 0.2 with an incremental interval

Table 4.2: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators (2-component Poisson Mixture Regression Models)

	Parameter Setting 1						Parameter Setting 2					
	Sample Size = 1e+2			Sample Size = 1e+4			Sample Size = 1e+2			Sample Size = 1e+4		
Ranking†	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
						Pearson's Chi-square Statistic						
1	185	10	61	20	224	12	233	2	21	118	126	12
2	68	9	179	151	20	85	22	90	144	130	77	49
3	3	237	16	85	12	159	1	164	91	8	53	195
						Root Mean Squared Error						
1	179	11	66	118	99	39	200	15	41	170	40	46
2	76	10	170	107	63	86	37	64	155	67	78	111
3	1	235	20	31	94	131	19	177	60	19	138	99

Note:
† The No. 1 ranking has the smallest statistic value.

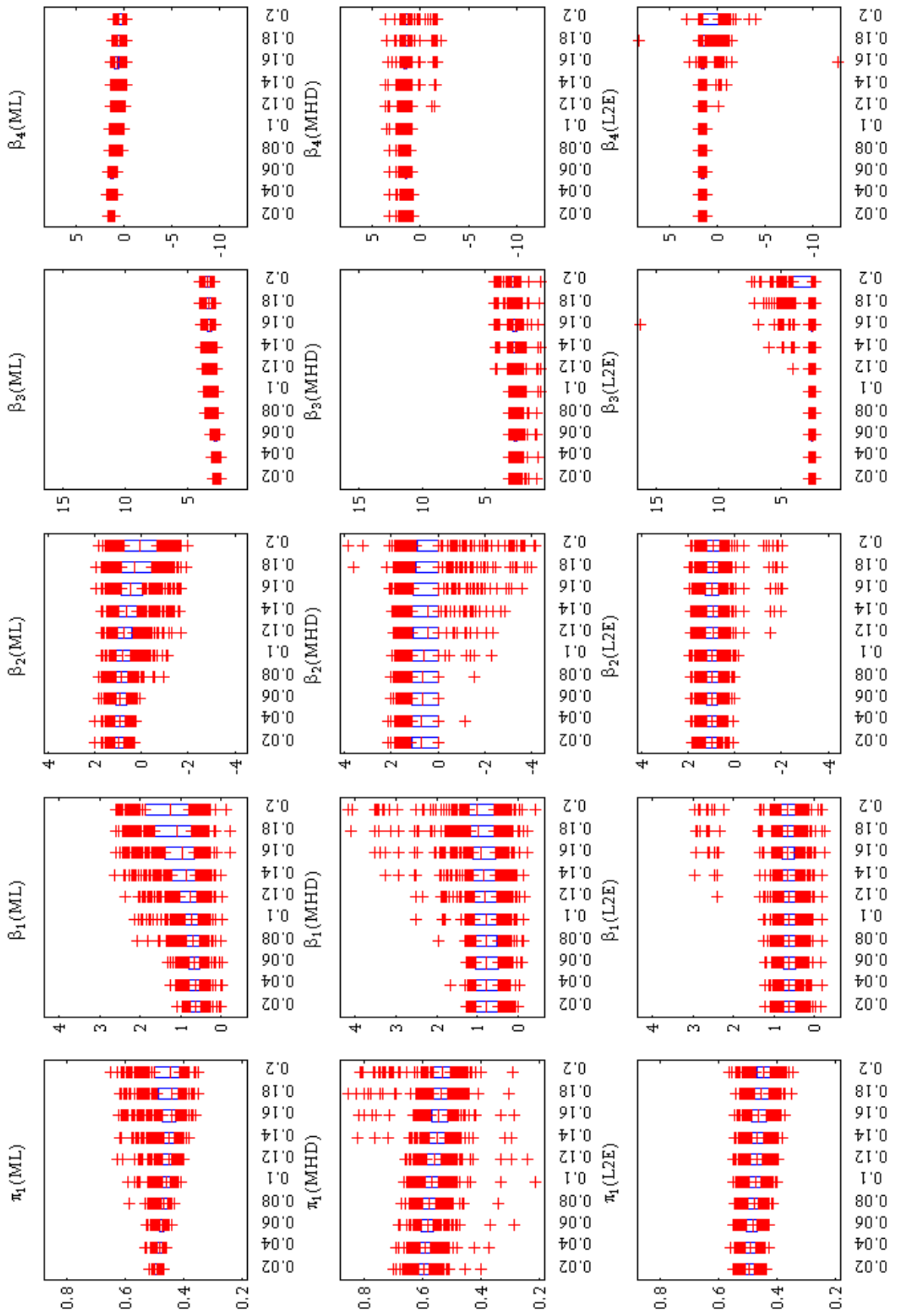


Figure 4.3: Boxplots of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 1 with Sample Size Equal to 1e+2 under Contamination Scenario 1

Notes:
To consistently compare the ML, MHD, and L_2E estimates of every parameter, the y axes are restricted to the same scales.

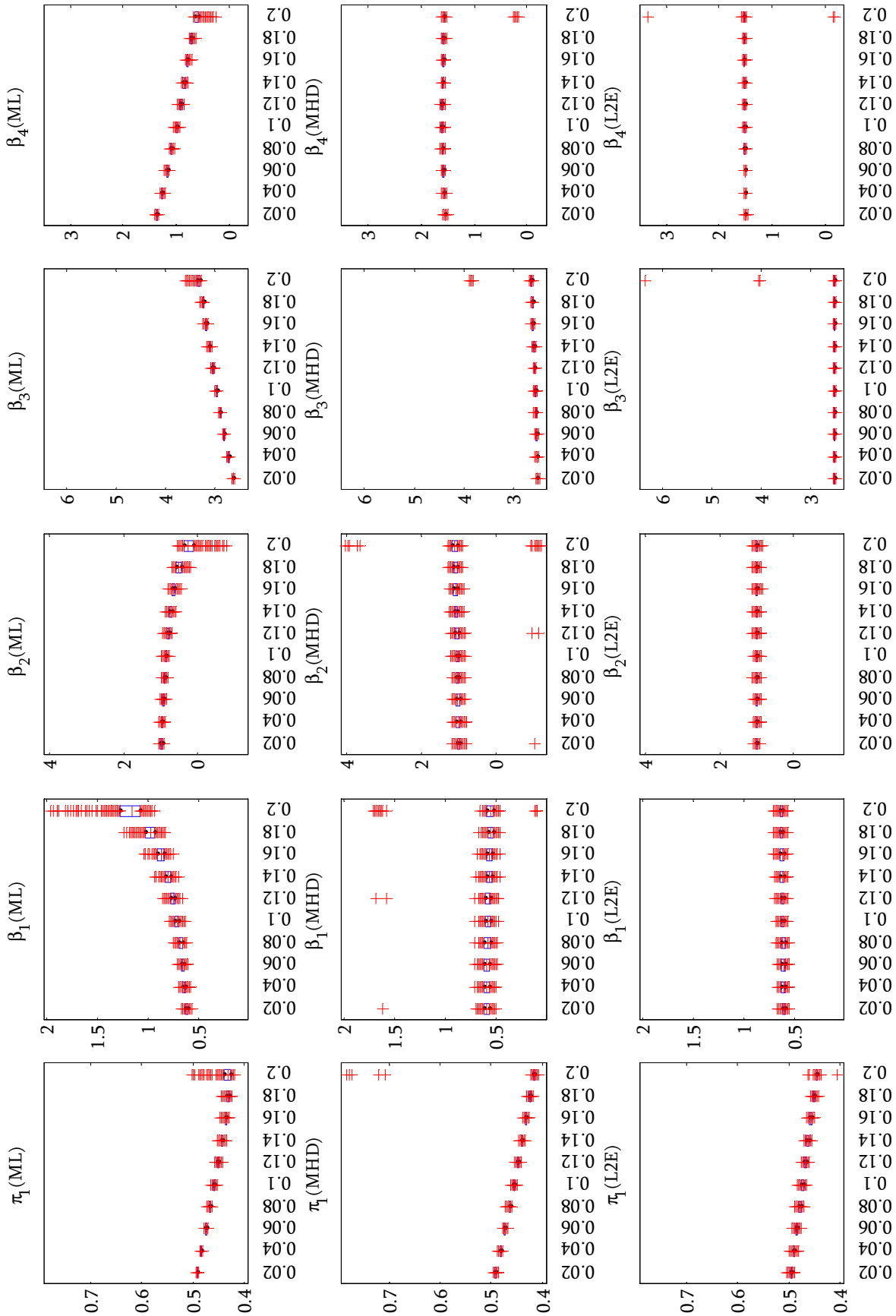


Figure 4.4: Boxplots of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 1 with Sample Size Equal to 1e+4 under Contamination Scenario 1

Notes:
To consistently compare the ML, MHD, and L_2E estimates of every parameter, the y axes are restricted to the same scales.

0.02 and fix z at 4. The second scenario is to hold a constant at 0.01 and change z from 0.5 to 5 with an incremental interval 0.5. Each contamination scenario is further considered with the effects of the sample size and the mixing probability, respectively. The sample size is set from 1e+2 to 1e+4 and the mixing probability is specified by 0.5 and 0.95, separately. The true values of the coefficient parameters are the same as those in the previous section, $\boldsymbol{\beta} = (0.6, 1, 2.5, 1.5)^{\intercal}$. There are 256 replications for every one of ten experiments under each contamination scenario.

Under the contamination scenario 1, the sample size is observed to significantly influence the robustness of the MHD and L_2E estimators. At the sample equal to 1e+2, Figures 4.3 and 4.5 present that their estimates scatter over the plot and are worse than the ML ones. When the sample size grows to 1e+4, the two minimum distance estimators demonstrate their resistance to the contaminated data according to Figures 4.4 and 4.6. However, the MHD and L_2E estimators do not have any robustness advantage of estimating π_1 compared with their robust estimates of the coefficients. Like the ML estimator, the MHD and L_2E estimates of π_1 constantly deviate from the true value as a rises and even more significantly when $\pi_1 = 0.95$.

To consider the impact of the mixing probability, the value of π_1 is raised from 0.5 to 0.95. As shown in Figures 4.5 and 4.6, the ranges of the MHD and L_2E estimates are found wider than those in Figures 4.3 and 4.4, particularly for the MHD estimator. This change shows that robustness of the two minimum distance estimators is eroded as the mixing probability approaches to 1. Concentrating on the experiments with the sample size equal to 1e+4, Figures 4.4 and 4.6 present that the breakdown points for the MHD and L_2E estimators move inversely with the value of π_1. When $\pi_1 = 0.5$, their estimates are likely away from the true values at $a = 0.2$ which plunges to 0.04 when $\pi_1 = 0.95$. Besides, the MHD estimates are more fluctuated than the L_2E ones.

Tables 4.3, 4.4, 4.5, and 4.6 give strong evidence that the χ^2_P favors the MHD

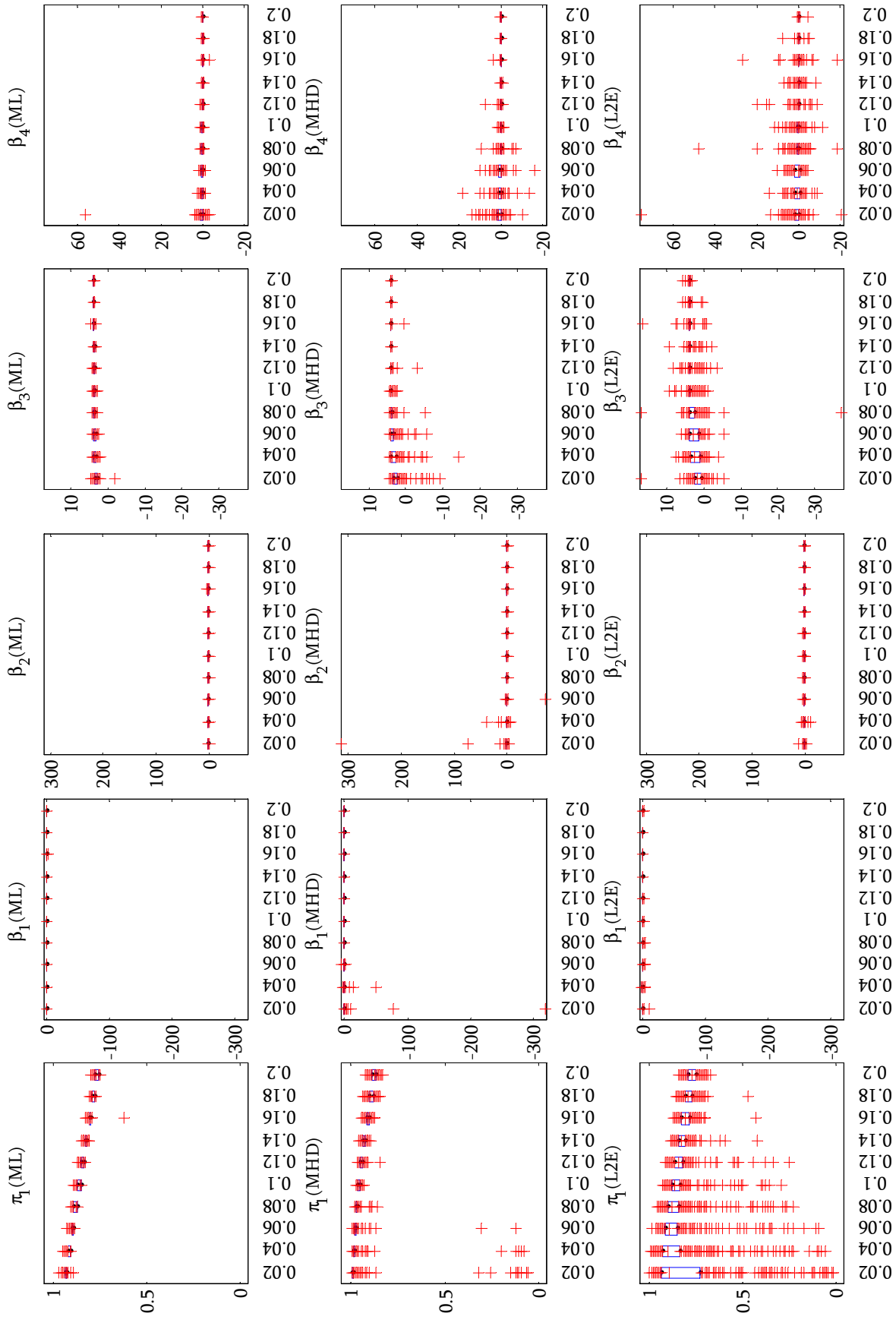
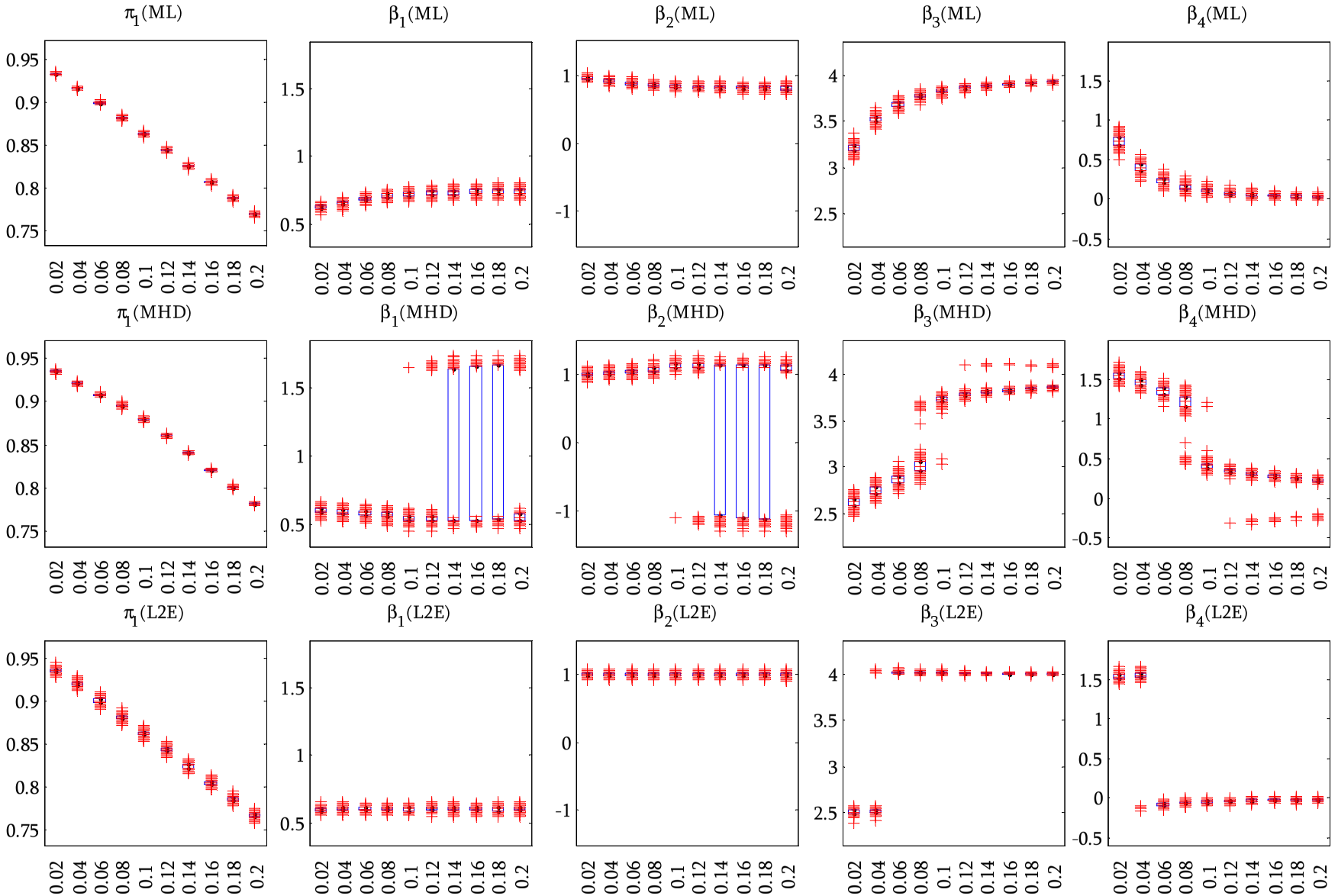


Figure 4.5: Boxplots of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 2 with Sample Size Equal to 1e+2 under Contamination Scenario 1

Notes:
To consistently compare the ML, MHD, and L_2E estimates of every parameter, the y axes are restricted to the same scales.

Figure 4.6: Boxplots of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 2 with Sample Size Equal to 1e+4 under Contamination Scenario 1



Notes:
To consistently compare the ML, MHD, and L_2E estimates of every parameter, the y axes are restricted to the same scales.

Table 4.3: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators at Parameter Setting 1 with Sample Size Equal to 1e+2 under Contamination Scenario 1

	Pearson's Chi-square Statistic										Root Mean Squared Error									
Ranking†	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2
								ML												
1	93	40	19	9	3	1	0	0	0	0	188	213	228	246	250	253	251	253	250	245
2	124	117	76	52	20	12	12	25	48	77	67	43	28	9	6	3	4	3	6	11
3	39	99	161	195	233	243	244	231	208	179	1	0	0	1	0	0	1	0	0	0
								MHD												
1	4	3	3	6	4	6	17	42	72	106	4	1	0	0	0	0	1	1	4	9
2	39	96	158	189	229	238	235	212	183	150	13	11	9	14	19	32	58	84	107	126
3	213	157	95	61	23	12	4	2	1	0	239	244	247	242	237	224	197	171	145	121
								L_2E												
1	159	213	234	241	249	249	239	214	184	150	64	42	28	10	6	3	4	2	2	2
2	93	43	22	15	7	6	9	19	25	29	176	202	219	233	231	221	194	169	143	119
3	4	0	0	0	0	1	8	23	47	77	16	12	9	13	19	32	58	85	111	135

Note:† The No. 1 ranking has the smallest statistic value.

Table 4.4: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators at Parameter Setting 1 with Sample Size Equal to 1e+4 under Contamination Scenario 1

	Pearson's Chi-square Statistic										Root Mean Squared Error									
Ranking†	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2
								ML												
1	0	0	0	0	0	0	0	0	0	0	239	255	256	256	256	256	256	256	256	256
2	0	0	0	0	0	0	0	0	0	10	17	1	0	0	0	0	0	0	0	0
3	256	256	256	256	256	256	256	256	256	246	0	0	0	0	0	0	0	0	0	0
								MHD												
1	256	256	256	256	256	256	256	256	256	256	17	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	142	202	245	256	256	256	256	256	256	252
3	0	0	0	0	0	0	0	0	0	0	97	53	11	0	0	0	0	0	0	4
								L_2E												
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	256	256	256	256	256	256	256	256	256	246	97	53	11	0	0	0	0	0	0	4
3	0	0	0	0	0	0	0	0	0	10	159	203	245	256	256	256	256	256	256	252

Note:† The No. 1 ranking has the smallest statistic value.

Table 4.5: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators at Parameter Setting 2 with Sample Size Equal to 1e+2 under Contamination Scenario 1

	Pearson's Chi-square Statistic										Root Mean Squared Error									
Ranking†	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2
	ML																			
1	212	204	203	192	186	181	163	163	161	153	231	243	236	234	234	221	220	215	213	216
2	44	52	53	64	70	74	92	91	93	100	23	12	20	22	22	35	36	40	43	40
3	0	0	0	0	0	1	1	2	2	3	2	1	0	0	0	0	0	1	0	0
	MHD																			
1	0	0	0	1	2	3	4	5	5	5	5	0	0	0	0	0	0	0	0	0
2	71	51	38	30	19	17	8	7	7	6	70	82	78	72	54	41	25	22	14	8
3	185	205	218	225	235	236	244	244	244	245	181	174	178	184	202	215	231	234	242	248
	L_2E																			
1	44	52	53	63	68	72	89	88	90	98	20	13	20	22	22	35	36	41	43	40
2	141	153	165	162	167	165	156	158	156	150	163	162	158	162	180	180	195	194	199	208
3	71	51	38	31	21	19	11	10	10	8	73	81	78	72	54	41	25	21	14	8

Note:† The No. 1 ranking has the smallest statistic value.

Table 4.6: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators at Parameter Setting 2 with Sample Size Equal to 1e+4 under Contamination Scenario 1

	Pearson's Chi-square Statistic										Root Mean Squared Error									
Ranking†	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2
	ML																			
1	62	69	13	4	0	0	0	0	0	0	255	256	256	254	165	136	141	165	174	159
2	92	85	94	177	223	147	79	34	11	7	1	0	0	2	91	120	115	91	82	97
3	102	102	149	75	33	109	177	222	245	249	0	0	0	0	0	0	0	0	0	0
	MHD																			
1	145	129	0	0	0	0	0	0	0	0	1	0	0	2	91	120	115	91	82	97
2	110	124	152	75	33	109	177	222	245	249	251	252	6	10	153	106	63	68	62	63
3	1	3	104	181	223	147	79	34	11	7	4	4	250	244	12	30	78	97	112	96
	L_2E																			
1	49	58	243	252	256	256	256	256	256	256	0	0	0	0	0	0	0	0	0	0
2	54	47	10	4	0	0	0	0	0	0	4	4	250	244	12	30	78	97	112	96
3	153	151	3	0	0	0	0	0	0	0	252	252	6	12	244	226	178	159	144	160

Note:† The No. 1 ranking has the smallest statistic value.

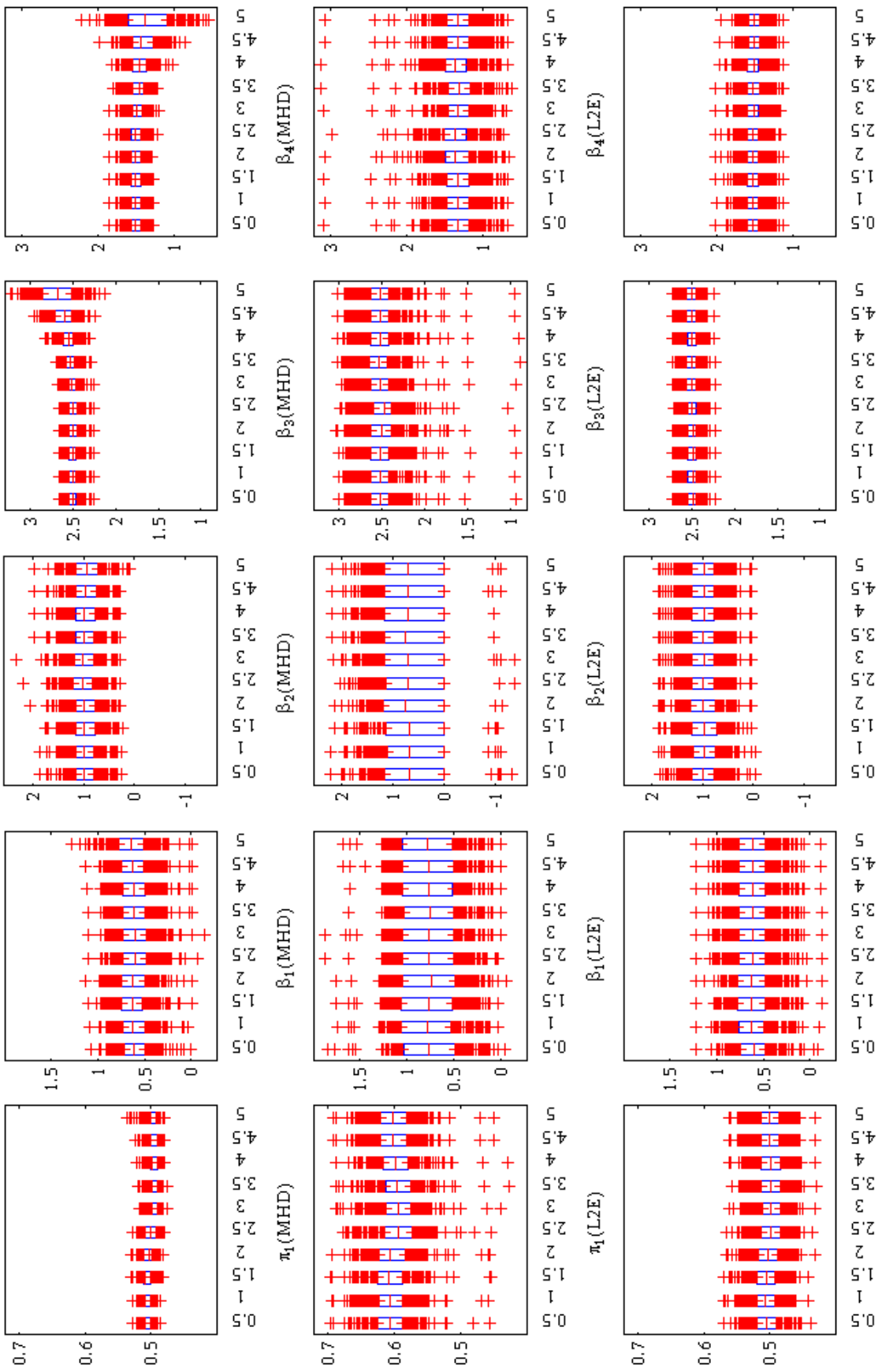


Figure 4.7: Boxplots of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 1 with Sample Size Equal to 1e+2 under Contamination Scenario 2

Notes:
To consistently compare the ML, MHD, and L_2E estimates of every parameter, the y axes are restricted to the same scales.

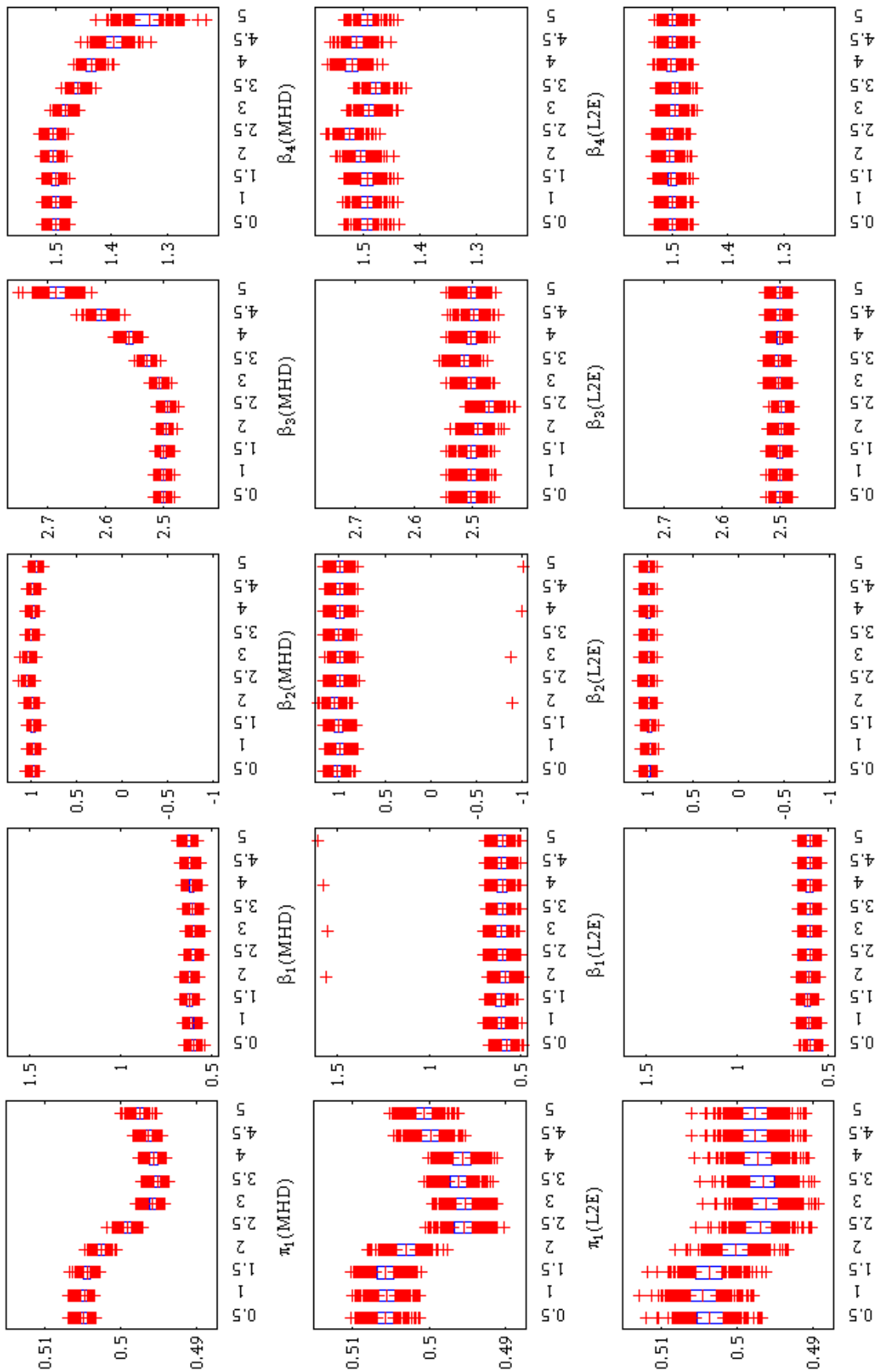


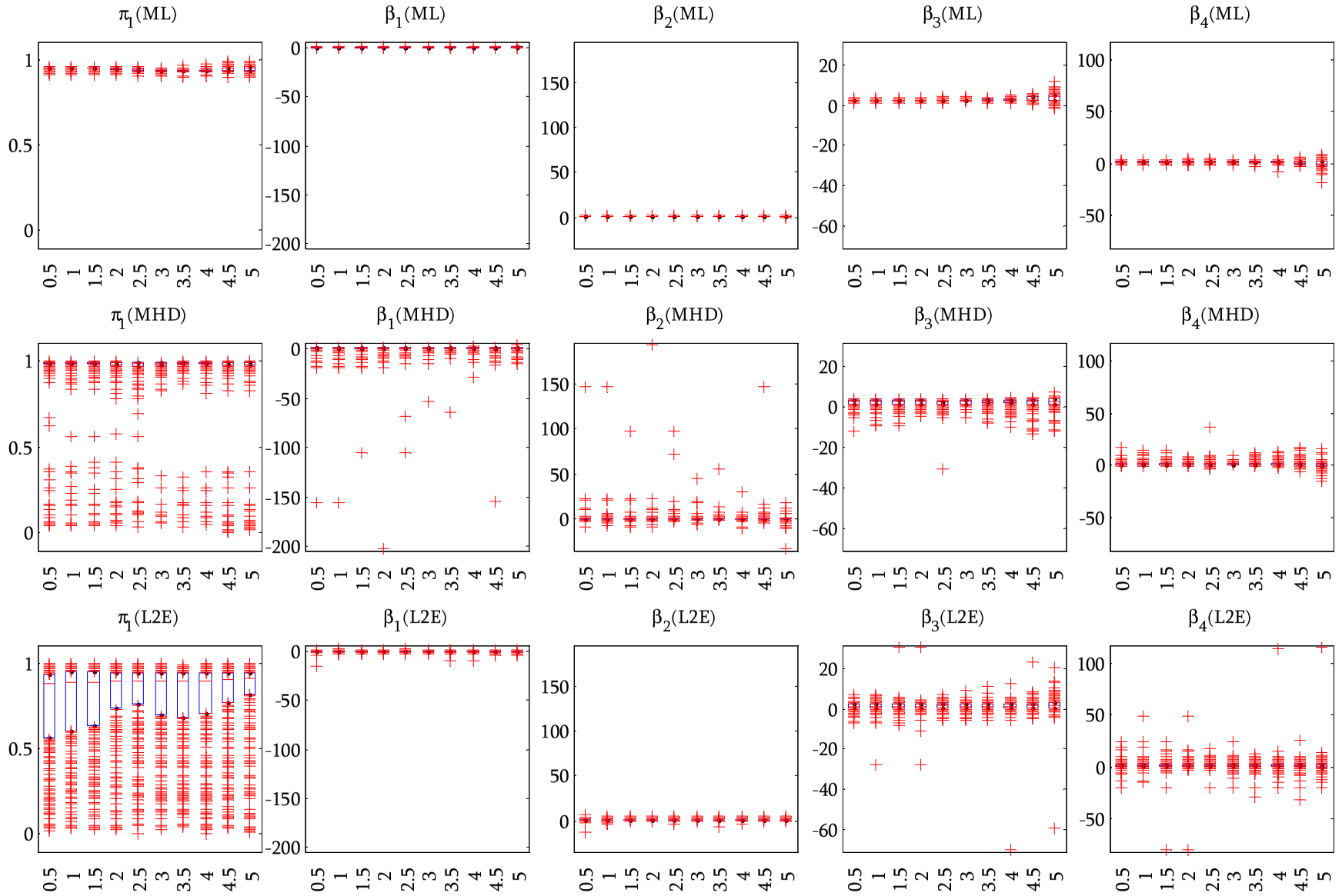
Figure 4.8: Boxplots of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 1 with Sample Size Equal to 1e+4 under Contamination Scenario 2

Notes:
To consistently compare the ML, MHD, and L_2E estimates of every parameter, the y axes are restricted to the same scales.

estimator and the $RMSE$ the ML one. According to the ranking frequencies of the χ^2_P, the MHD estimator tends to own high rankings when the sample size is equal to 1e+4. On the other hand, no matter what the sample size is, the ML estimator is the best estimator if based on the ranking frequencies of the $RMSE$.

Under the contamination scenario 2, there are a couple of common features about the robustness property of the ML, MHD, and L_2E estimators compared with the contamination scenario 1. First, as the sample size rises to 1e+4 from 1e+2, the MHD and L_2E estimates are more resistant to the contaminated data, but their robustness property deteriorates when π_1 moves to 0.95 from 0.5 as shown in Figures 4.7, 4.8, 4.9, and 4.10. Second, the χ^2_P and $RMSE$ still favor the MHD and ML estimators, respectively, according to the ranking frequencies in Tables 4.7 4.8, 4.9, and 4.10.

Figure 4.9: Boxplots of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 2 with Sample Size Equal to 1e+2 under Contamination Scenario 2



Notes:
To consistently compare the ML, MHD, and L_2E estimates of every parameter, the y axes are restricted to the same scales.

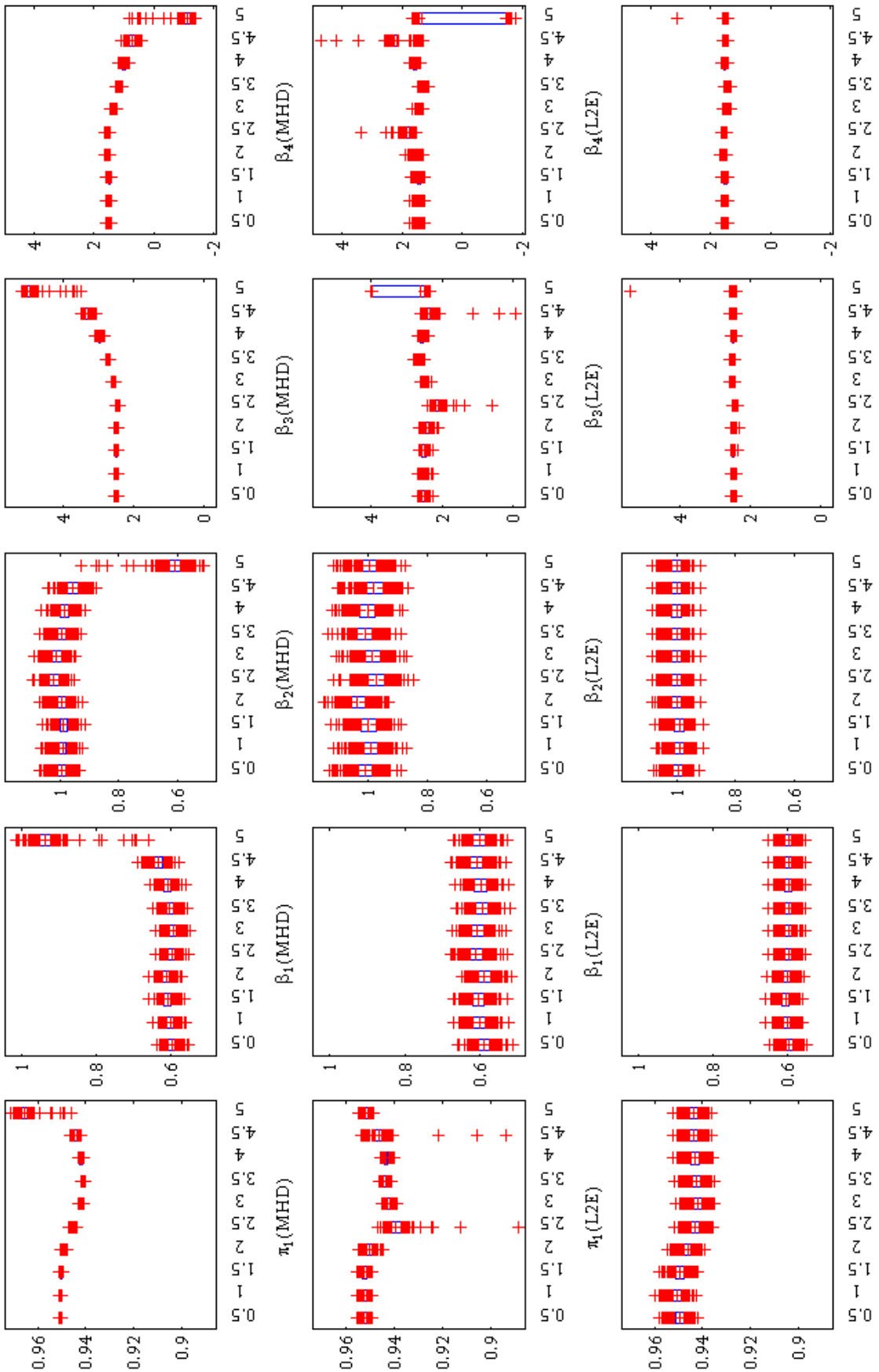


Figure 4.10: Boxplots of Parameter Estimates Produced by the ML, MHD, and L_2E Estimations at Parameter Setting 2 with Sample Size Equal to 1e+4 under Contamination Scenario 2

Notes:
To consistently compare the ML, MHD, and L_2E estimates of every parameter, the y axes are restricted to the same scales.

Table 4.7: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators at Parameter Setting 1 with Sample Size Equal to 1e+2 under Contamination Scenario 2

	Pearson's Chi-square Statistic										Root Mean Squared Error									
Ranking†	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
								ML												
1	166	166	168	162	178	173	157	120	57	23	165	168	162	169	164	168	170	171	188	205
2	89	83	85	90	75	81	95	124	156	117	89	86	92	85	90	86	85	85	64	49
3	1	7	3	4	3	2	4	12	43	116	2	2	2	2	2	2	1	0	4	2
								MHD												
1	4	5	3	6	5	7	7	5	5	7	11	10	10	11	13	12	11	9	3	0
2	10	11	12	14	17	13	8	15	42	112	12	11	11	14	13	12	8	5	4	5
3	242	240	241	236	234	236	241	236	209	137	233	235	235	231	230	232	237	242	249	251
								L_2E												
1	86	85	85	88	73	76	92	131	194	226	80	78	84	76	79	76	75	76	65	51
2	157	162	159	152	164	162	153	117	58	27	155	159	153	157	153	158	163	166	188	202
3	13	9	12	16	19	18	11	8	4	3	21	19	19	23	24	22	18	14	3	3

Note:† The No. 1 ranking has the smallest statistic value.

Table 4.8: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators at Parameter Setting 1 with Sample Size Equal to 1e+4 under Contamination Scenario 2

	Pearson's Chi-square Statistic										Root Mean Squared Error									
Ranking†	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
								ML												
1	8	16	18	2	2	17	1	0	0	0	106	118	106	110	116	78	146	189	256	256
2	148	150	155	188	121	168	75	8	0	0	120	107	119	122	113	118	64	65	0	0
3	100	90	83	66	133	71	180	248	256	256	30	31	31	24	27	60	46	2	0	0
								MHD												
1	234	230	227	249	254	236	211	251	210	104	98	95	106	109	107	139	73	66	0	0
2	15	15	16	4	1	18	45	5	46	152	70	63	56	74	93	90	148	93	120	24
3	7	11	13	3	1	2	0	0	0	0	88	98	94	73	56	27	35	97	136	232
								L_2E												
1	14	10	11	5	0	3	44	5	46	152	52	43	44	37	33	39	37	1	0	0
2	93	91	85	64	134	70	136	243	210	104	66	86	81	60	50	48	44	98	136	232
3	149	155	160	187	122	183	76	8	0	0	138	127	131	159	173	169	175	157	120	24

Note:† The No. 1 ranking has the smallest statistic value.

Table 4.9: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators at Parameter Setting 2 with Sample Size Equal to 1e+2 under Contamination Scenario 2

	Pearson's Chi-square Statistic										Root Mean Squared Error									
Ranking†	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
								ML												
1	227	227	232	221	231	228	221	218	207	166	210	206	211	207	201	217	219	215	200	145
2	29	29	24	35	25	28	35	38	49	84	34	37	31	35	39	27	28	32	31	40
3	0	0	0	0	0	0	0	0	0	6	12	13	14	14	16	12	9	9	25	71
								MHD												
1	0	0	0	0	0	0	0	0	1	3	13	11	10	11	12	11	11	10	12	27
2	89	89	85	63	76	84	84	71	60	58	67	63	66	78	70	79	87	67	66	102
3	167	167	171	193	180	172	172	185	195	195	176	182	180	167	174	166	158	179	178	127
								L_2E												
1	29	29	24	35	25	28	35	38	48	87	33	39	35	38	43	28	26	31	44	84
2	138	138	147	158	155	144	137	147	147	114	155	156	159	143	147	150	141	157	159	114
3	89	89	85	63	76	84	84	71	61	55	68	61	62	75	66	78	89	68	53	58

Note:† The No. 1 ranking has the smallest statistic value.

Table 4.10: Ranking Frequencies of the Pearson's Chi-square Statistic and Root Mean Squared Error for the ML, MHD, and L_2E Estimators at Parameter Setting 2 with Sample Size Equal to 1e+4 under Contamination Scenario 2

	Pearson's Chi-square Statistic										Root Mean Squared Error									
Ranking†	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
								ML												
1	100	110	107	39	15	79	78	57	70	1	154	167	154	183	165	131	219	231	255	8
2	141	131	133	202	133	132	83	71	121	0	87	69	84	60	73	96	16	24	1	70
3	15	15	16	15	108	45	95	128	65	255	15	20	18	13	18	29	21	1	0	178
								MHD												
1	143	131	139	210	200	145	116	152	11	2	42	46	49	37	44	105	18	25	1	1
2	73	72	71	30	34	65	117	100	81	253	84	73	73	140	112	134	197	173	75	178
3	40	53	46	16	22	46	23	4	164	1	130	137	134	79	100	17	41	58	180	77
								L_2E												
1	13	15	10	7	41	32	62	47	175	253	60	43	53	36	47	20	19	0	0	247
2	42	53	52	24	89	59	56	85	54	3	85	114	99	56	71	26	43	59	180	8
3	201	188	194	225	126	165	138	124	27	0	111	99	104	164	138	210	194	197	76	1

Note:† The No. 1 ranking has the smallest statistic value.

Chapter 5

Empirical Applications

To illustrate the proposed approach introduced in Chapter 2 and 3, two data sets are examined in Section 5.1 and 5.2, respectively. The first data set, the extent of non-payments of personal loans in Spain, was studied in Dionne et al. (1996). The second data set, counts of utilization from the RAND Health Insurance Experiment, was studied in Deb and Trivedi (2002).

5.1 Analysis of Non-payments of Personal Loans in Spain

The proposed approach applied to the data set used in Dionne et al. (1996) is to study the non-payments behavior of personal loans in Spain. This data set has typical characteristics of count data that the responses contain a large proportion of the zeros and the unconditional sample variance is bigger than the unconditional sample mean. This kind of data is often collected from a heterogeneous population with extreme values so that there is an additionally good reason to apply the proposed approach.

However, the data set used for this study has been restricted to two data-processing procedures that could impact the results. One is that Dionne et al. (1996) dropped observations with extreme covariate values beyond three standard

deviations and the other is that they excluded observations with more than twelve non-payments. Since the main purpose of developing the minimum distance estimation methods is to avoid any subjective elimination of inliers or outliers which might lose valuable information, the two procedures adopted by Dionne et al. (1996) reduce the representation of the estimates produced by the minimum distance estimation methods.

The econometric model used in Dionne et al. (1996) was an extension of the hurdle model, which merged the logit and two conditional truncated distributions to jointly estimate the default probability and non-payments of good and bad loans, respectively. To implement this extended hurdle model, Dionne et al. (1996) had to split the whole sample into two subsamples, "non-defaulters" and "defaulters" and the discrimination depended on whether or not an observed person owed four or more monthly payments. The hurdle thresholds for the two subsamples were set at the non-payments number of zero for the non-defaulters and at three for the defaulters.

By using the ML estimation method, the estimated results in Dionne et al. (1996) showed that the significant variables affecting each distribution were not the same and more importantly the two truncated distributions give different explanation of non-payment behaviors between the non-defaulters and defaulters. Their findings implied that a single count data GLM or traditional hurdle regression model was not sufficient to evaluate credit scoring systems.

There are a couple of concerns about the approach applied in Dionne et al. (1996): one is that the defaulters subsample only had 336 observations of the total sample 2,446 observations; the other is that among their eighteen covariates, excluding the intercept, only one covariate was not a dummy variable. Both of them raise a question as to whether the current optimization algorithm could find a feasibly global optimum in their estimation problem.[1] For this reason, the hurdle model is substituted for their approach in the following subsection of model

[1]I use Stata and MATLAB to try to replicate their results but only have the same estimates as theirs for the logit part.

comparison that also returns to conventional model comparison. The best fit model will be selected from the standard, hurdle, and finite mixture regression models, whose density specification include the Poisson, NB1, and NB2.[2]

5.1.1 Data and Summary Statistics

The data set used in Dionne et al. (1996) covers personal loans for consumption granted by a Spanish bank, and was retrieved in May 1989. This type of personal loans was usually borrowed for a short period and required to pay monthly constant amounts before the loan was terminated. Dionne et al. (1996) eliminated records from the raw data (4,691) by the following criteria: first, they dropped incomplete records (2,010); second, observations with extreme covariate values, beyond three standard deviations, were deleted (18); third, they excluded observations with more than twelve non-payments (217). The final sample contained the remaining 2,446 loans.

Definitions and summary statistics for the variables are presented in Table 5.1. For the explanatory variables, there are three categories, i.e., (1) personal variables (age, education, marital status), (2) socio-economic variables (income, housing ownership, geographical location), and (3) financial variables (loan duration, loan age, collateral, bank relationship, store credit).[3]

5.1.2 Results

The results, including model specification, estimation, and evaluation, are given in the following manner: first, by applying the ML estimation method, the standard, hurdle, and finite mixture regression models are compared in Section 5.1.2.1; second, the final decision on the best fit model and suitable estimator is made in Section 5.1.2.2. Based on the decision, further analysis of the estimated result is presented in Section 5.1.2.3.

[2]Here the standard count data GLM is referred to as the standard regression model.
[3]For more details of data processing, see Dionne et al. (1996, Section 4).

Table 5.1: Variable Definitions and Summary Statistics

Variable	Definition	Mean	SD
Y	number of non-payments†	1.11	2.20
DT6	1 if total contract duration of return period is more than four years	0.33	0.47
DUREEA	number of months from the beginning of the contract at the sampling date‡	1.88	1.08
AGE1	1 if the age group is 18-24 years	0.09	0.28
AGE2	1 if the age group is 25-39 years	0.48	0.50
AGE3	1 if the age group is 40 years or more	0.43	0.50
DESTIN	1 if the credit is used to purchase a good with a collateral	0.45	0.50
ETU1	1 if the client has not completed primary education	0.04	0.20
ETU2	1 if the client has completed primary education	0.49	0.50
ETU3	1 if the client has completed higher education	0.32	0.47
ETU4	1 if the client has a university degree	0.15	0.36
RECSAL	1 if the client receives the salary through the bank	0.40	0.49
M1	1 if married, non-owner, salary under $3,000	0.21	0.40
M2	1 if married, non-owner, salary higher than or equal to $3,000	0.02	0.14
M3	1 if married, owner, salary under $3,000	0.08	0.27
M4	1 if married, owner, salary higher than or equal to $3,000	0.69	0.46
NM1	1 if not married, non-owner	0.22	0.41
NW2	1 if not married, owner	0.78	0.41
CENTRE	1 if the credit is granted by a store	0.16	0.36
RESID	1 if resident in the city for at least four years	0.74	0.44
Z1	1 if south (Andalucla, Canarias, Castilla-La Mancha, Extremadura, Murcia)	0.28	0.45
Z2	1 if north (Aragon, Asturias, Cantabria, Castilla-Le6n, Galicia, Navarra, Pals Vasco)	0.29	0.45
Z3	1 if east (Baleares, Catalunya, Valencia)	0.30	0.46
Z4	1 if center (Madrid)	0.13	0.34

Notes:

† The frequency of zero counts is 68.07%.

‡ The variable values are divided by 10 to avoid an error might caused by floating-point calculation.

5.1.2.1 Model Comparison among the Standard, Hurdle, and Finite Mixture Regression Models

The competing models applied here include every model specification listed in Table 2.1. To determine the best fit, the comparison process utilizes the following diagnostic tools sequentially: the AIC, BIC, LR test[4], Pearson's chi-square statistic, root mean squared error, and fitted descriptive statistics.

Table 5.2 reports the values of the information criteria and LR ratios. Based on the information criteria, it is observed that adding an overdispersion parameter to a standard regression model or applying a finite mixture regression model can give a much better fit than the SP that implies significantly unobserved heterogeneity existing in the data. For the AIC values, every density specification prefers the 3-component mixture regression model and the 3-FMP has the smallest value, 6095. As anticipated, the BIC is more likely to select a parsimonious model and the SNB1 with the BIC value equal to 6333 is the top choice. For the LR tests, all the null hypotheses are rejected in favor of complicated models, and the preferred model within each density specification is like that of the AIC, the 3-component mixture regression model.

Table 5.3 compares the sample cell frequencies with the fitted ones and also reports the values of the χ^2_P and $RMSE$. For the cell fitted frequencies, the finite mixture regression models overwhelmingly have better fits than the standard and hurdle regression models, except at the count zero where the hurdle regression models can fit perfectly due to their model specifications incorporating a binary choice model. Using the χ^2_P to summarize the cell fitted frequencies, its statistic values demonstrate a descending sequence from a standard regression model to a

[4]Deb and Trivedi (1997, Section 3.2) drew a road map to evaluate models based on the AIC, BIC, and LR test. They referred the proof of the adequacy of the information criteria to Leroux (1992). They gave a warning of using the LR test to decide whether an extra component density was necessary in the finite mixture regression model. Because the true asymptotic distribution of the LR ratio is not $\chi^2(1)$, using the traditional critical value could mislead a user to under-reject the false null hypothesis and then mistakenly choose a simpler model; see Böhning et al. (1994) for further details. To avoid this problem when conducting the LR tests in the first empirical application, the difference of estimated parameter numbers between two competing models is used as the degrees of freedom to compute a stricter χ^2 critical value.

Table 5.2: Information Criteria and Likelihood Ratio Tests

	Information Criteria											
	SP	HP	2-FMP	3-FMP	SNB1	HNB1	2-FMNB1	3-FMNB1	SNB2	HNB2	2-FMNB2	3-FMNB2
AIC	9075	6399	6190	6095[a,c]	6217	6130	6123	6097[a]	6226	6148	6125	6099[a]
BIC	9185	6620	6416[b]	6437	6333[b,d]	6356	6361	6456	6342[b]	6374	6363	6459
$q^{\dagger}$	19	38	39	59	20	39	41	62	20	39	41	62
	Likelihood Ratio Tests[‡]											
	H_0:SP	2713	2924	3060	H_0:SNB1	124.9	136.0	204.1	H_0:SNB2	116.6	143.2	211.3
			H_0:2-FMP	135.2			H_0:2-FMNB1	68.14			H_0:2-FMNB2	68.10

Notes:

[†] q is the number of estimated parameters.

[a] The model is preferred by the AIC within a single density specification.

[b] The model is preferred by the BIC within a single density specification.

[c] The model is preferred by the AIC among all density specifications.

[d] The model is preferred by the BIC among all density specifications.

[‡] The 5% critical value for the hypothesis tests, SP vs. HP and SNB vs. HNB, is $\chi^2(5\%, 19) = 30.14$. For non-standard nested hypothesis tests, the 5% conservatively critical values, respectively, are $\chi^2(5\%, 20) = 31.41$ for SP vs. 2-FMP and 2-FMP vs. 3-FMP, $\chi^2(5\%, 21) = 32.67$ for SNB vs. 2-FMNB and 2-FMNB vs. 3-FMNB, $\chi^2(5\%, 40) = 55.76$ for SP vs. 3-FMP, and $\chi^2(5\%, 42) = 58.12$ for SNB vs. 3-FMNB.

Table 5.3: Fitted Frequencies, Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values of All Models Estimated by the ML Estimation

		Poisson Density Specification				NB1 Density Specification				NB2 Density Specification			
		SP	HP	2-FMP	3-FMP	SNB1	HNB1	2-FMNB1	3-FMNB1	SNB2	HNB2	2-FMNB2	3-FMNB2
Count	Sample	Fitted Frequencies											
0	68.07	39.02	68.07	67.37	67.86	67.10	68.07	67.09	67.72	67.12	68.07	67.71	67.87
1	11.08	31.74	5.149	11.77	10.93	12.23	9.076	9.935	10.90	13.26	8.932	10.95	10.58
2	4.129	16.72	6.751	4.505	5.198	6.389	6.461	6.333	5.419	6.385	6.744	5.455	5.579
3	2.984	7.466	6.456	4.134	3.879	3.990	4.648	4.432	3.969	3.745	4.807	3.985	3.983
4	4.334	3.091	5.076	3.744	3.200	2.692	3.345	3.171	3.214	2.416	3.360	3.212	3.186
5	2.944	1.226	3.499	2.976	2.594	1.896	2.403	2.285	2.576	1.652	2.337	2.541	2.548
6+	6.460	0.739	4.999	5.507	6.335	5.700	5.997	5.946	6.200	5.429	5.749	6.139	6.254
		Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values											
	χ^2_P †	2310	252.9	16.05[a,c]	21.52	69.88	56.95	49.70	24.69[a]	99.08	67.23	25.67[a]	27.70
	$RMSE$	2.126[b,d]	2.140	2.160	2.157	2.129[b]	2.148	2.148	2.147	2.148	2.147	2.145[b]	2.148
	obj. ‡	-4518	-3162	-3056	-2988	-3088	-3026	-3020	-2986	-3093	-3035	-3022	-2987

Notes:

[a] The model is preferred by the χ^2_P within a single density specification.

[b] The model is preferred by the $RMSE$ within a single density specification.

[c] The model is preferred by the χ^2_P among all density specifications.

[d] The model is preferred by the $RMSE$ among all density specifications.

† The 5% critical value of the Pearson's chi-square test is $\chi^2(5\%, 6) = 12.592$.

‡ The objective function value is the log likelihood.

hurdle regression model and then to finite mixture regression models regardless of the density specification. The smallest two χ^2_P values occur to the 2 and 3-FMPs, i.e., 16.05 and 21.52, respectively. Hence, the finite mixture regression models fit the data better than the standard and hurdle regression models from this perspective.

For the $RMSE$, it should not be surprising to see that the SP has the smallest value, 2.126. As mentioned in Section 3.2, the ML estimates of the SP are equal to the estimates which minimize the sum of squared residuals. Together with the second smallest $RMSE$ value, 2.129, for the SNB1, the two smallest values are confirmed by the "Mean" result in Table 5.4, where the Mean values of the SP and SNB1 are almost equal to the sample mean, 1.109.[5] Compared with the previous outcome based on the χ^2_P values, the $RMSE$ and "Mean" values of the 2 and 3-FMPs clearly do not support that they are the best choices. To solve this dilemma, the following techniques are applied to decide which model fits actually better than the others. The first technique is to analyze the movement of the χ^2_P by changing the number of cells, and the second one is to examine the $\chi^2_{P|x}$ and $RMSE_{|x}$.

The reason to analyze the movement of the χ^2_P is that some of the models might not fit the data very well in the far right tail. Therefore, when the number of cells increases, the changing statistic value can reveal more information about the right tail fitting.[6] Table 5.5 shows that the χ^2_P value grows with the number of cell but is relatively small for finite mixture regression models. This supports that the finite mixture regression model does a better job than the standard and hurdle regression models. For the 2-FMP, the trend of the χ^2_P values tells that this model does not fit very well in the far right tail. The finding is consistent with its relatively small "Var." and "Max." values in Table 5.4 compared with the sample

[5]The Law of Total Expectations states that the expected value of the conditional mean equals the unconditional mean.

[6]Normally, the minimum allowed expected count for a bin is 5, but this is not a must. Besides, the series of χ^2_P values here are only compared on a cross-model basis.

Table 5.4: Fitted Descriptive Statistics of All Models Estimated by the ML Estimation

	Poisson Density Specification				NB1 Density Specification				NB2 Density Specification			
	SP	HP	2-FMP	3-FMP	SNB1	HNB1	2-FMNB1	3-FMNB1	SNB2	HNB2	2-FMNB2	3-FMNB2
Mean	1.109	1.117	1.038	1.088	1.109	1.120	1.112	1.082	1.137	1.123	1.094	1.089
Var.	1.109	3.526	3.989	4.504	6.130	4.878	5.061	4.437	8.370	5.211	4.865	4.527
Min.	0.167	0.119	0.404	0.292	0.194	0.132	0.165	0.279	0.125	0.115	0.267	0.274
Max.	4.580	5.044	2.833	5.248	4.025	5.548	5.370	4.869	6.484	5.921	5.539	5.258

Notes:
"Mean" is the sample average of the fitted means, "Var." is the sample average of the fitted variances, "Min." is the minimum of the fitted means, and "Max." is the maximum of the fitted means.
The sample mean and variance are 1.109 and 4.861, respectively.

Table 5.5: Movement of Pearson's Chi-square Statistics of All Models Estimated by the ML Estimation

	Poisson Density Specification				NB1 Density Specification				NB2 Density Specification			
Cells	SP	HP	2-FMP	3-FMP	SNB1	HNB1	2-FMNB1	3-FMNB1	SNB2	HNB2	2-FMNB2	3-FMNB2
5	1857	242.7	14.45	20.93	63.76	56.03	48.06	24.23	88.23	66.48	25.18	27.04
6	2310	252.9	16.05	21.52	69.88	56.95	49.70	24.69	99.08	67.23	25.67	27.70
7	3098	276.3	26.22	22.95	70.83	57.14	49.73	26.29	102.5	67.34	26.92	28.95
8	4357	305.0	40.03	25.53	71.70	57.36	49.76	29.07	105.9	67.38	28.85	31.20
9	5505	314.3	42.77	25.61	81.38	60.27	53.68	29.13	123.1	72.05	29.33	31.38
10	6159	314.6	42.87	28.36	97.13	68.73	63.36	31.82	146.6	83.60	33.98	34.57
11	6160	324.2	52.01	38.63	116.8	83.91	79.29	42.16	171.9	102.0	47.22	45.41

variance, 4.861.[7]

To examine the $\chi^2_{P|x}$ and $RMSE_{|x}$, Table 5.6 presents the values referred to the two representative covariates, **DESTIN** and **RESID,** which have the largest subsamples when they equal 1.[8] For the $\chi^2_{P|x}$ values, the finite mixture regression model generally has the smaller value than the standard and hurdle regression models that is consistent to the earlier result obtained by the χ^2_P values. But for the $RMSE_{|x}$ values, the result is different from that obtained by the $RMSE$ values, except for the SP. The SNB1 does not have the second smallest $RMSE_{|x}$ values, and the other models also have different rankings of the $RMSE_{|x}$ values compared with their rankings of the $RMSE$ values. Consequently, the χ^2_P is a more reliable goodness-of-fit statistic here than the $RMSE$.

To conclude the diagnoses of model comparison which have been done so far, the finite mixture regression model is the top choice of model specification for the data used in Dionne et al. (1996), contrasting with the standard and hurdle regression models. This result strongly supports that the data should be divided into latent classes for a better economic interpretation. Referring to the density specification in finite mixture regression models, even though there is no absolute answer to prefer a particular one, the Poisson density specification performs relatively better than the negative binomial ones. The possible explanation for this phenomenon is that Dionne et al. (1996) excluded observations with more than twelve non-payments so that it is not as necessary to incorporate an overdispersion parameter in the finite mixture regression model.

5.1.2.2 Model and Estimator Choices for Finite Mixture Regression Models

After recognizing the goodness-of-fit of finite mixture regression models, the next step is to select the best fit model from them. Meanwhile, the MHD and L_2E

[7]The Analysis of Variance formula states that the unconditional variance is the sum of the variance of the conditional mean plus the expected value of the conditional variance.

[8]The subsample size of **DESTIN** and **RESID** equal to 1 are 1,109 and 1,805, respectively.

Table 5.6: Subsample Pearson's Chi-square Statistics and Root Mean Squared Errors of All Models Estimated by the ML Estimation

	Poisson Density Specification				NB1 Density Specification				NB2 Density Specification			
	SP	HP	2-FMP	3-FMP	SNB1	HNB1	2-FMNB1	3-FMNB1	SNB2	HNB2	2-FMNB2	3-FMNB2
	Subsample of **DESTIN**=1											
$\chi^2_{P\|x}$	850.5	125.4	22.65	23.96	44.45	41.84	40.28	40.99	58.36	46.41	26.17	26.23
$RMSE_{\|x}$	1.811	1.827	1.874	1.877	1.827	1.821	1.898	3.010	1.823	1.830	1.844	1.870
	Subsample of **RESID**=1											
$\chi^2_{P\|x}$	1854	202.8	8.062	25.62	42.23	48.41	42.37	37.52	76.10	54.79	24.77	27.42
$RMSE_{\|x}$	2.049	2.066	2.084	2.074	2.051	2.073	2.079	2.087	2.068	2.073	2.067	2.066

Table 5.7: Overdispersion Parameters of the FMNBs Produced by the ML, MHD, and L_2E Estimations

	FMNB1			FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E
	2-component Mixture Regression Models					
$\hat{\psi}_1$	3.201	8e-21	9.913	3.254	6.3e-7	0.286
$\hat{\psi}_2$	1.160	4e-13	5e-13	1.5e-7	5e-17	7.823
	3-component Mixture Regression Models					
$\hat{\psi}_1$	5.7e-5	0.796	7.2e-6	1.038	3.254	0.346
$\hat{\psi}_2$	2.6e-5	6e-32	8.9e-7	0.165	0.002	1.7e-6
$\hat{\psi}_3$	3.5e-5	2e-26	2.2e-8	5.3e-6	1e-17	1.5e-9

estimators are both applied to model estimation in order to decide whether the two minimum distance estimators are more suitable than the ML one for this data set. Two more diagnostic procedures are required to accomplish these tasks. One is using the cross-validation to examine whether the out-of-sample fit is consistent with the in-sample one, and the other is interpreting parameter estimates to verify whether the estimates make economic sense.

Because Section 5.1.2.1 concludes that the FMNBs do not necessarily fit the data set better than the FMPs, more analyses are performed here to investigate whether the FMPs should be selected rather than the FMNBs. Table 5.7 summarizes overdispersion parameter estimates of the FMNBs produced by the ML, MHD, and L_2E estimations. For the 3-FMNBs, almost all of the estimates are close to zero when the density specification is the NB1 and at lease one of them for each estimator is close to zero when the density specification is the NB2. For the 2-FMNBs, all the MHD estimates are near zero, and ψ_2 of the L_2E and ML estimates based on the NB1 and NB2 density specifications, respectively, are close to zero. This result is verified by Tables 5.8 and 5.9 where the values of the Mean and Var. are almost identical for the corresponding component densities in the FMNBs.[9]

While solving the FMNB estimation problems, the optimization process for the MHD and L_2E estimators becomes unsmooth. Convergence rate is slow and the ML estimates are not always good starting values. Tables 5.10 and 5.11 show that the trouble causes awful fitted cell frequencies, large goodness-of-fit statistics, and abnormal objective function values of the FMNBs produced by the MHD and L_2E estimations.

For the subsample fitting performance, Tables 5.12 and 5.13 present the $\chi^2_{P|x}$

[9]In Tables 5.7, 5.8, and 5.9 some inconsistencies might cause a curiosity between overdispersion parameter estimates and the values of the Mean and Var. First, $\hat{\psi}_2$ of the MHD in the 2-FMNB2 and $\hat{\psi}_3$ of the L_2E in the 3-FMNB2 are near zero, but the Mean and Var. are apart that is due to very large fitted means. Second, $\hat{\psi}_1$ of the MHD in the 3-FMNB2 is 3.254 but the Mean and Var. are almost equivalent. This is because the fitted means are extremely small. A quick way to have an approximated proof of these inconsistencies is to plug those numerical values into the conditional variance formula.

Table 5.8: Fitted Descriptive Statistics of the ML, MHD, and L_2E Estimations (2-component Mixture Regression Models)

	2-FMP			2-FMNB1			2-FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
	Mixture density								
Mean	1.038	1.073	1.020	1.112	2.497	1.565	1.094	2e+30	3.824
Var.	3.989	4.479	5.539	5.061	13.67	17.20	4.865	1e+63	4.7e+3
Min.	0.404	0.647	0.031	0.165	0.045	0.073	0.267	0.416	0.016
Max.	2.833	1.440	14.08	5.370	106.2	24.27	5.539	3e+33	540.6
	Component density 1								
Mean	0.153	0.154	0.096	0.336	1.799	0.995	0.479	5.190	3.345
Var.	0.153	0.154	0.096	1.411	1.799	10.86	2.003	5.190	17.96
Min.	0.009	0.003	2e-08	0.006	1.9e-4	0.045	0.012	1.783	0.033
Max.	1.103	0.267	1.530	3.668	144.2	6.074	4.873	15.95	111.8
	Component density 2								
Mean	4.042	4.437	2.806	2.960	4.337	3.202	4.682	2e+30	4.091
Var.	4.042	4.437	2.806	6.392	4.337	3.202	4.682	4e+47	7.0e+3
Min.	1.493	3.006	0.087	0.434	0.105	0.024	1.461	8e-54	2e-11
Max.	10.66	5.735	41.24	16.28	11.92	92.40	14.54	4e+33	841.3

Notes:
"Mean" is the sample average of the fitted means, "Var." is the sample average of the fitted variances, "Min." is the minimum of the fitted means, and "Max." is the maximum of the fitted means. The component densities are put in order with the smallest "Mean" at the top. The sample mean and variance are 1.109 and 4.861, respectively.

Table 5.9: Fitted Descriptive Statistics of the ML, MHD, and L_2E Estimations (3-component Mixture Regression Models)

	3-FMP			3-FMNB1			3-FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
	Mixture density								
Mean	1.088	1.074	1.2e+5	1.082	2e+10	26.40	1.089	2.594	2e+18
Var.	4.504	4.490	3e+13	4.437	3e+22	3.1e+5	4.527	14.94	1e+40
Min.	0.292	0.717	0.025	0.279	7e-10	0.019	0.274	0.190	0.024
Max.	5.248	1.342	1.2e+8	4.869	6e+12	1.6e+4	5.257	99.27	3e+21
	Component density 1								
Mean	0.073	6.6e-5	0.063	0.110	0.014	2.270	0.098	5.7e-5	2.632
Var.	0.073	6.6e-5	0.063	0.110	0.025	2.270	0.234	5.7e-5	10.91
Min.	1.9e-9	1.4e-8	2e-17	1.5e-7	8.5e-5	0.050	7e-14	1.4e-6	0.052
Max.	1.494	0.004	26.32	4.571	0.312	43.03	5.851	5.8e-4	64.78
	Component density 2								
Mean	1.369	0.257	2.587	1.143	1.641	35.66	1.152	1.605	6.010
Var.	1.369	0.257	2.587	1.143	1.641	35.66	1.649	1.688	6.079
Min.	2.2e-7	0.209	0.072	3.2e-7	2e-59	3e-26	0.003	9.7e-5	5e-24
Max.	13.37	0.302	54.27	11.71	8.647	4.5e+4	11.46	131.7	9.1e+3
	Component density 3								
Mean	4.939	4.437	5.3e+5	4.940	3e+10	47.68	4.986	5.908	7e+18
Var.	4.939	4.437	5.3e+5	4.940	3e+10	47.68	4.986	5.908	1e+32
Min.	1.686	2.873	5e-11	1.619	3e-11	6e-10	1.686	0.736	5e-40
Max.	13.97	5.615	5.6e+8	14.46	1e+13	1.4e+4	14.27	28.15	1e+22

Notes:
"Mean" is the sample average of the fitted means, "Var." is the sample average of the fitted variances, "Min." is the minimum of the fitted means, and "Max." is the maximum of the fitted means. The component densities are put in order with the smallest "Mean" at the top. The sample mean and variance are 1.109 and 4.861, respectively.

Table 5.10: Fitted Frequencies, Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values for the ML, MHD, and L_2E Estimators (2-component Mixture Regression Models)

		2-FMP			2-FMNB1			2-FMNB2		
		ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
Count	Sample	Fitted Frequencies								
0	68.07	67.37	68.38	69.26	67.09	75.13	68.02	67.71	49.47	68.06
1	11.08	11.77	11.13	12.95	9.935	3.417	10.23	10.95	1.332	10.42
2	4.129	4.505	4.074	5.625	6.333	6.7e+13	5.649	5.455	2.307	5.477
3	2.984	4.134	3.531	3.390	4.432	2.0e+49	3.597	3.985	2.6e+67	3.376
4	4.334	3.744	3.363	2.245	3.171	2e+140	2.485	3.212	6.6e+52	2.280
5	2.944	2.976	2.912	1.561	2.285	2e+120	1.821	2.541	1.8e+38	1.636
6+	6.460	5.507	6.611	4.967	5.946	-2e+140	8.193	6.139	-2.6e+67	8.752
		Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values								
	χ^2_P †	16.05	9.076[a]	106.4	49.70	0	73.82	25.67	-2e+64	95.73
	$RMSE$	2.160	2.201	2.315	2.148	5.906	2.565	2.145[b]	7e+31	19.47
	$obj.$ ‡	-3056	4.432e-3	-0.4967	-3020	-5.1e+139	-0.4992	-3022	-1.19e+38	-0.4994

Notes:

[a] The estimation method is preferred by the χ^2_P among all 2-component mixture regression models.

[b] The estimation method is preferred by the $RMSE$ among all 2-component mixture regression models.

† The 5% critical value of the Pearson's chi-square test is $\chi^2(5\%, 6) = 12.592$.

‡ The objective function values are the log likelihood, Hellinger distance, and L_2 error for the ML, MHD, and L_2E estimators, respectively.

Table 5.11: Fitted Frequencies, Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values for the ML, MHD, and L_2E Estimators (3-component Mixture Regression Models)

		3-FMP			3-FMNB1			3-FMNB2		
		ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
Count	Sample	Fitted Frequencies								
0	68.07	67.86	68.37	68.90	67.72	100.0	69.05	67.87	77.82	67.85
1	11.08	10.93	11.14	10.37	10.90	3.180	10.26	10.58	9.164	9.468
2	4.129	5.198	4.066	5.737	5.419	9.2e-21	5.903	5.579	3.692	5.277
3	2.984	3.879	3.483	3.502	3.969	4.62e+4	3.626	3.983	2.070	3.230
4	4.334	3.200	3.337	2.281	3.214	8e+117	2.370	3.186	1.343	2.130
5	2.944	2.594	2.927	1.564	2.576	1.6e+92	1.629	2.548	0.947	1.486
6+	6.460	6.335	6.671	7.646	6.200	-8e+117	7.157	6.254	4.968	10.55
		Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values								
	χ^2_P †	21.52	9.263[a]	93.74	24.69	0	85.11	27.70	327.7	142.9
	$RMSE$	2.157	2.237	2.9e+6	2.147[b]	2e+11	410.6	2.148	5.252	6e+19
	$obj.$‡	-2988	4.453e-3	-0.5025	-2986	-2.1e+104	-0.5045	-2988	-8.14e+46	-0.5039

Notes:

[a] The estimation method is preferred by the χ^2_P among all 3-component mixture regression models.

[b] The estimation method is preferred by the $RMSE$ among all 3-component mixture regression models.

† The 5% critical value of the Pearson's chi-square test is $\chi^2(5\%, 6) = 12.592$.

‡ The objective function values are the log likelihood, Hellinger distance, and L_2 error for the ML, MHD, and L_2E estimators, respectively.

and $RMSE_{|x}$ values of the ML, MHD, and L_2E estimators. For each estimator, even though the FMPs do not always have the smallest values of the $\chi^2_{P|x}$ and $RMSE_{|x}$, their overall fitting performance is constantly better than that of the FMNB1s and FMNB2s no mater for a 2 or 3-component mixture model. One example is that for the ML estimator of 2-component mixture models the 2-FMNB1 and 2 have smaller $RMSE_{|x}$ values than the 2-FMP in the subsample of **RESID**, but their $\chi^2_{P|x}$values are much bigger than that of the 2-FMP, even in the subsample of **DESTIN**. The other example is that for the L_2E estimator of 3-component mixture models the 3-FMNB1 is the best fit model in the subsample of **DESTIN**, but it fits poorly in the subsample of **RESID**.

Back to Table 5.7, although the ML estimates of the 2-FMNB1 and the L_2E estimates of the 2-FMNB2 are not trivial, their χ^2_P and $RMSE$ values in Table 5.10 do not support that they fit better than the 2-FMP. Each of them has one statistic value considerably larger than the corresponding value of the 2-FMP. This phenomenon also happens to their $\chi^2_{P|x}$ and $RMSE_{|x}$ values in Table 5.12.

In view of this evidence, the FMNBs are proved to be overparameterized by overdispersion parameters no matter for which estimator. Furthermore, this overparameterization creates a floating-point calculation problem for the asymptotic variances of parameter estimates whose values become unreliable when an overdispersion parameter estimate is extremely small. As a result, the NB1 and 2 are excluded from the density specification, and that leaves only the 2 and 3-FMPs to be the final candidates to fit the data.

To finalize the model specification and select an appropriate estimator for it, the cross-validation is applied. As proven in Chapter 4, the χ^2_P shows favor to the MHD estimator and the $RMSE$ the ML one. These statistics are not allowed to assess the adequacy-of-fit among the ML, MHD, and L_2E estimators directly. But via the cross-validation, not only can the better fit of the 2 or 3-FMP be decided, but also the suitability of an estimator is verified.

Figure 5.1 presents the χ^2_P and $RMSE$ values of the 2 and 3-FMPs for the

Table 5.12: Subsample Pearson's Chi-square Statistics and Root Mean Squared Errors for the ML, MHD, and L_2E Estimators (2-component Mixture Regression Models)

	2-FMP			2-FMNB1			2-FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
	Subsample of **DESTIN**=1								
$\chi^2_{P\|x}$	22.65	15.18	137.1	40.28	3e+37	46.95	26.17	1396	91.93
$RMSE_{\|x}$	1.874	1.923	4.654	1.898	2.617	2651	1.844	40.17	5723
	Subsample of **RESID**=1								
$\chi^2_{P\|x}$	8.062	6.363	85.05	42.37	NaN	91.44	24.77	0	111.0
$RMSE_{\|x}$	2.084	2.122	2.286	2.079	978.4	18.29	2.067	695.5	10.49

Table 5.13: Subsample Pearson's Chi-square Statistics and Root Mean Squared Errors for the ML, MHD, and L_2E Estimators (3-component Mixture Regression Models)

	3-FMP			3-FMNB1			3-FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
	Subsample of **DESTIN**=1								
$\chi^2_{P\|x}$	23.96	15.08	109.1	40.99	15.18	41.84	26.23	15.08	266.3
$RMSE_{\|x}$	1.877	1.955	132.8	3.010	2.252	36.65	1.870	1.971	4E+19
	Subsample of **RESID**=1								
$\chi^2_{P\|x}$	25.62	5.364	81.65	37.52	6.326	85.98	27.42	6.440	110.2
$RMSE_{\|x}$	2.074	2.181	143.1	2.087	2.210	5159	2.066	2.167	76.44

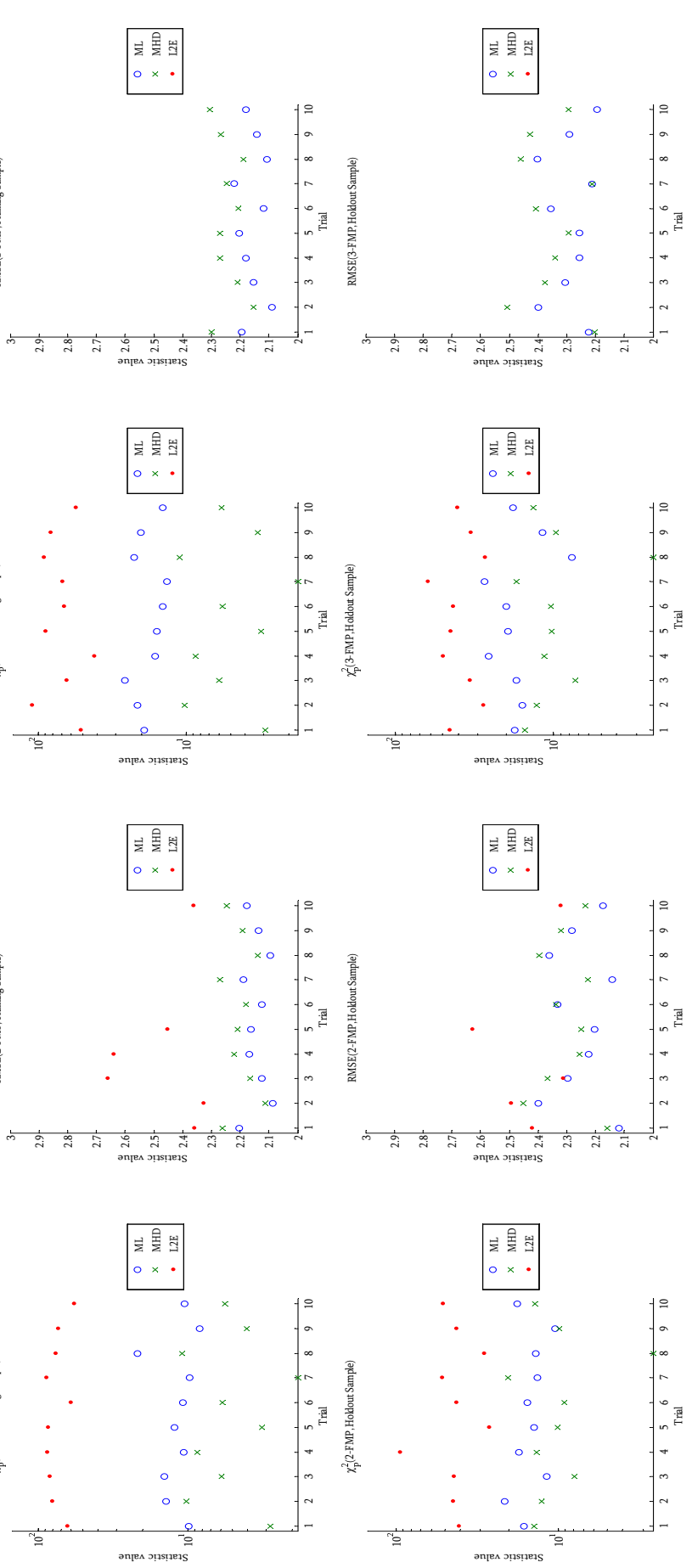


Figure 5.1: Pearson's Chi-square Statistics and Root Mean Squared Errors for the ML, MHD, and L_2E Estimators from the Cross-validation (2 and 3-component Poisson Mixture Regression Models)

Notes:
To consistently compare the statistic values among the ML, MHD, and L_2E estimators, the y axes are restricted to the same scales in all plots that makes some L_2E points lie beyond the upper bound.

ML, MHD, and L_2E estimators from the cross-validation.[10] Apparently, the L_2E estimator is the worst among the three because its two goodness-of-fit statistics almost have the largest values in both the training and holdout samples for either the 2 or 3-FMP. Figure 5.2 verifies the discrepancy between the ML and MHD estimators and the L_2E estimator. The component densities of the L_2E estimator are different from those of the other two estimators in the 2 and 3-FMPs. To take a close look at the fitting performance of the ML and MHD estimators, Figure 5.3 filters out the goodness-of-fit statistics of the L_2E estimator in Figure 5.1.

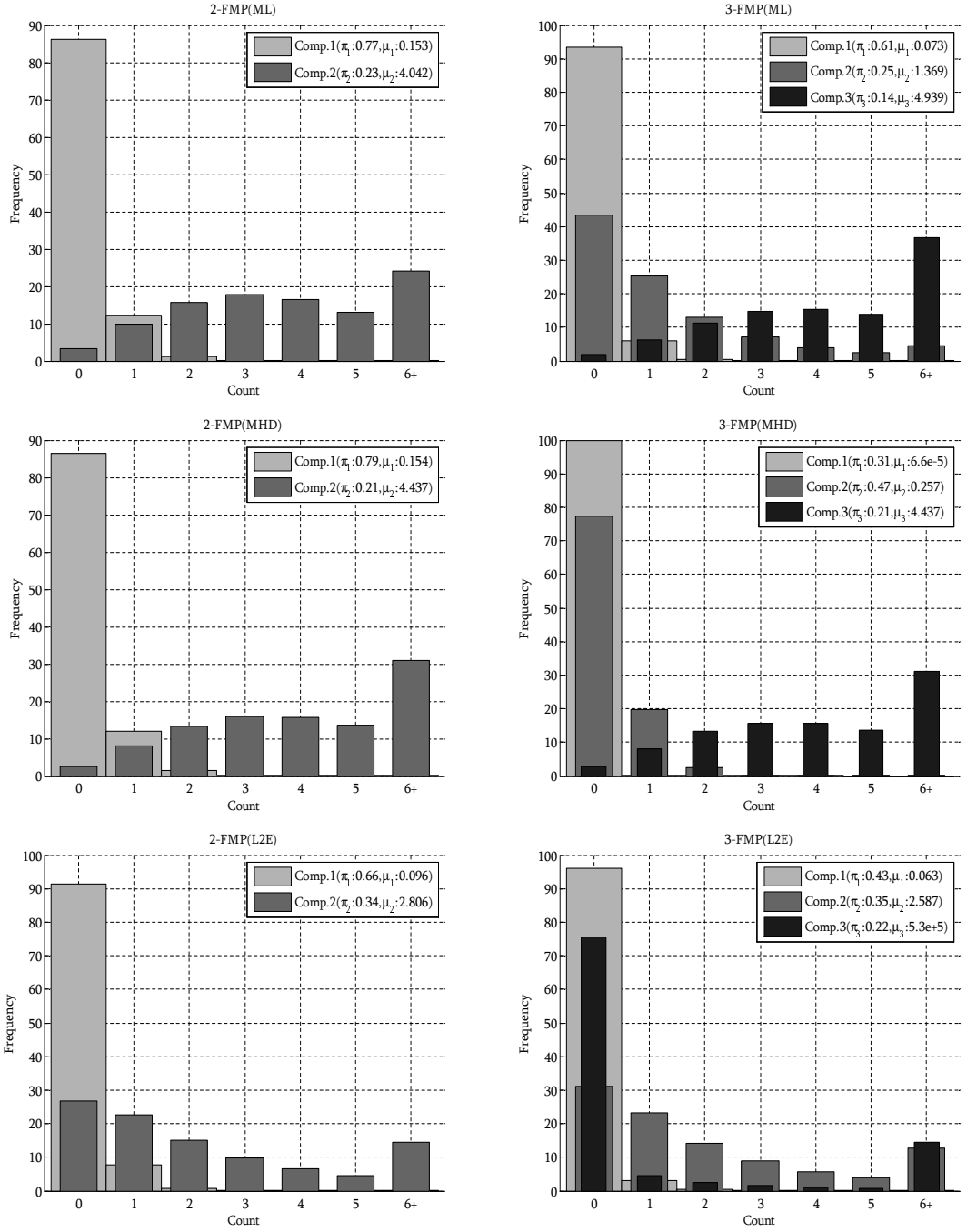
For the MHD estimator, it is hard to decide whether the 2 or 3-FMP has a better fit in either a training or holdout samples if only based on the χ^2_P, but the $RMSE$ helps distinguish their fitting performance. In both the training and holdout samples, the 2-FMP constantly has smaller $RMSE$ values than the 3-FMP. For the ML estimator, the 2-FMP clearly fits better than the 3-FMP no matter in the training or holdout sample on the basis of the χ^2_P and $RMSE$. At this stage, the 2-FMP can be concluded as the preferred model for the data set.

On the other hand, the cross-validation gives an ambiguous answer to choose the best estimator for the 2-FMP. Based on the $RMSE$ values, the ML estimator continually maintains its best fitting performance in the training and holdout samples, but the χ^2_P values of the ML estimator are higher than those of the MHD one in either the training or holdout sample, except in Trail 7 of the holdout samples. This conundrum is consistent with the similar component densities of the 2-FMP for the ML and MHD estimators in Figure 5.2. To solve this, studying their parameter estimates from an economic perspective can give a hand.

For this task, two covariates are picked in Table 5.14. One is **DESTIN** which equals 1 if the credit is used to purchase a good with a collateral, and the other is **RECSAL** which equals 1 if the client receives the salary through the bank. The reason to select these two covariates is that they definitely have negative impact on the non-payments number. For **DESTIN**, when a client can purchase a good

[10]To balance lengthy computation against proper sampling, there are ten trials conducted in this experiment.

Figure 5.2: Component Densities of the ML, MHD, and L_2E Estimations (2 and 3-component Poisson Mixture Regression Models)



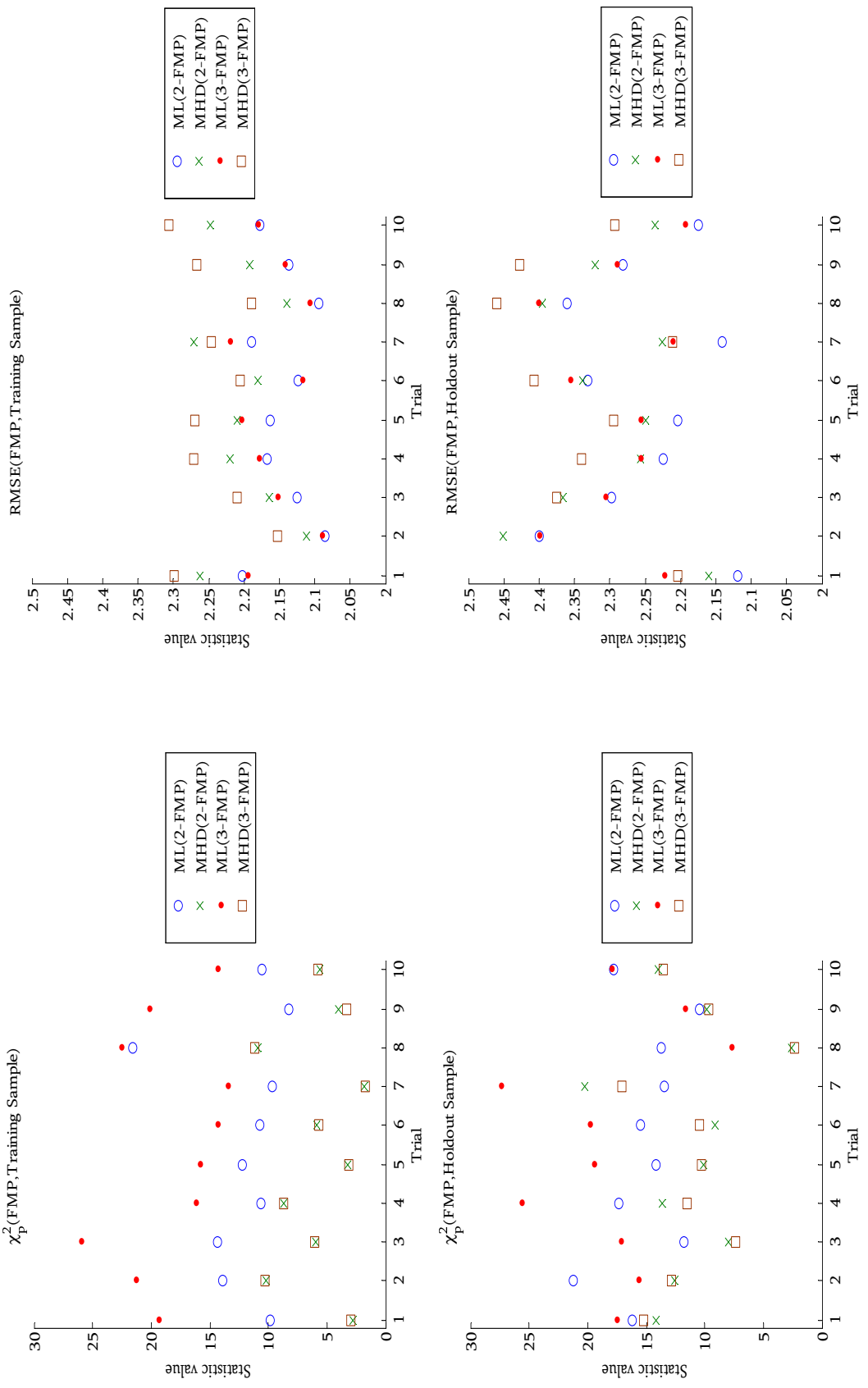


Figure 5.3: Pearson's Chi-square Statistics and Root Mean Squared Errors for the ML and MHD Estimators from the Cross-validation (2 and 3-component Poisson Mixture Regression Models)

with a collateral, that usually implies he has certain amount of wealth. Moreover, he will avoid the collateral to be liquidated because most of time the selling price is underestimated and that is not his best interest to let it happen. For **RECSAL**, the bank not only collects more this client's credit information but also has a deeper relationship with him through his salary account. This gives the bank an edge to analyze his credit quality in more detail. Once the bank offers him a loan, that means the probability of non-payments is relatively low. Besides, this client is more likely to pay on time and avoids going into default because he wants to keep a good relationship with this type of bank for future borrowing needs.

In Table 5.14 only the ML estimates of these two covariates meet the expectation. Both of them have anticipated signs in the component 1 and 2, and reach economic and statistical significance as well. For the MHD estimates, only the estimate of **RECSAL** in the component 2 is statistically significant, but it does not reach economic significance and also contradicts the anticipated sign. Based on the findings, the ML estimator is the best estimator for the 2-FMP.

5.1.2.3 Further Analysis of the Preferred Model and Estimator

While applying the t-test to examine the significance of the $\hat{\pi}$ and $\hat{\psi}$, Deb and Trivedi (1997) emphasized that the distributions of the test statistics were not the conventional ones because both $\hat{\pi}$ and $\hat{\psi}$ had boundary values. Since the 2-FMP is chosen as the best fit model for this data set, only the significance of $\hat{\pi}$ needs to be considered here. In Table 5.14, the value of π is very precisely estimated under the ML estimation method, which again strengthens the hypothesis that more than one latent class exists.

Table 5.8 indicates that the Mean values are diverse for the two component densities in the 2-FMP estimated by the ML estimation method, i.e., 0.153 vs. 4.042. This discrepancy can be explained by the corresponding chart in Figure 5.2 where the component density 1 has a large proportion of the count zero and the component density 2 has a substantial fraction of the count greater than or equal

Table 5.14: 2-FMP Parameter Estimates of the ML, MHD, and L_2E Estimations

	ML				MHD				L_2E			
Variables	Component 1		Component 2		Component 1		Component 2		Component 1		Component 2	
INTERCEPT	-2.46***	(0.47)	1.25***	(0.16)	-1.37***	(0.32)	1.10***	(0.13)	-3.80***	(0.80)	-0.02	(0.35)
DT6	0.48***	(0.19)	0.05	(0.06)	-4e-3	(0.01)	3e-7**	(1e-7)	0.86**	(0.35)	0.12	(0.15)
DUREEA	0.06	(0.08)	0.13***	(0.03)	-3e-3	(4e-3)	2e-6*	(1e-6)	0.18	(0.12)	0.41***	(0.09)
AGE1	0.35	(0.39)	0.25*	(0.15)	0.02	(0.02)	9e-6	(7e-6)	-0.04	(0.70)	0.29	(0.24)
AGE2	0.39*	(0.22)	0.19***	(0.07)	5e-3	(0.01)	-4e-6*	(2e-6)	0.29	(0.43)	0.35**	(0.17)
DESTIN	-0.46**	(0.20)	-0.25***	(0.07)	2e-3	(3e-3)	-1e-6	(1e-6)	-0.48	(0.37)	-1.13***	(0.20)
ETU1	-0.43	(0.84)	0.19	(0.13)	0.02	(0.03)	8e-6	(5e-6)	-0.14	(1.10)	1.37***	(0.30)
ETU2	0.45*	(0.27)	0.10	(0.10)	0.02	(0.03)	9e-6	(6e-6)	0.65	(0.51)	0.65**	(0.28)
ETU3	0.36	(0.32)	-0.05	(0.11)	0.03	(0.04)	1e-5	(7e-6)	0.65	(0.58)	0.38	(0.30)
RECSAL	-0.49**	(0.22)	-0.34***	(0.09)	-4.44	(21.5)	6e-7**	(3e-7)	-0.65	(0.55)	-0.90***	(0.15)
M1	-0.42	(0.41)	0.13*	(0.08)	3e-3	(4e-3)	0.65***	(0.12)‡	-0.40	(0.67)	0.80***	(0.17)
M2	0.04	(0.48)	0.34***	(0.11)	-9e-4	(9e-4)	0.65***	(0.12)	0.63	(0.69)	1.25***	(0.25)
M3	0.77***	(0.29)	0.33***	(0.11)	-0.01	(0.01)	0.65***	(0.12)	1.18***	(0.45)	0.42*	(0.23)
NM1	0.04	(0.26)	0.10	(0.09)	-3e-3	(3e-3)	0.65***	(0.12)	0.03	(0.45)	0.62***	(0.21)
CENTRE	-0.29	(0.24)	0.12	(0.09)	-0.04	(0.05)	2e-6	(1e-6)	0.14	(0.54)	-0.56*	(0.31)
RESID	-0.12	(0.25)	-0.03	(0.06)	0.01	(0.01)	1e-7**	(7e-8)	-0.14	(0.39)	-0.19	(0.13)
Z2	0.37	(0.27)	-0.17**	(0.07)	-4e-3	(0.01)	-3e-6*	(2e-6)	0.79**	(0.40)	-0.32*	(0.18)
Z3	0.29	(0.24)	-0.30***	(0.07)	-0.01	(0.01)	-1e-5*	(5e-6)	0.69	(0.45)	-0.35**	(0.16)
Z4	-0.55	(0.42)	-0.14	(0.15)	-0.01	(0.02)	-1e-5*	(5e-6)	-12.9***	(1.60)	-0.34	(0.36)
$\hat{\pi}$	0.77	(0.01)			0.79	(0.02)			0.66	(0.02)		
obj.†	-3056.055				4.43231e-3				-0.496691			

Notes:

***, **, and * mean significant at the 1%, 5%, and 10% levels, respectively, but those does not apply to $\hat{\pi}$ because it has a boundary value.

† The objective function values are the log likelihood, Hellinger distance, and L_2 error for the ML, MHD, and L_2E estimators, respectively.

‡ The MHD estimates and standard errors of M1, M2, M3, and NM1 are (0.646102, 0.646093, 0.646103, 0.646089) and (0.124450, 0.124448, 0.124450, 0.124448), respectively.

to six. Moreover, depending on the "Mean" which represents the non-payments number for the average person in each latent class, the component density 1 and 2 can be denoted as the "low-risk" and "high-risk" groups, respectively, for behavior analysis. However, the group names do not literally mean that the low-risk group never has a high number of non-payments and the high-risk group would not pay on time. These two groups have their own distributions just like those in Figure 5.2.

5.2 Analysis of Health Care Demand in the United States

Deb and Trivedi (2002) used the data on counts of utilization from the RAND Health Insurance Experiment to contrast the hurdle model with the finite mixture model. Based on model properties, they defined that the hurdle model distinguished between users and non-users of health care and the finite mixture model distinguished between infrequent and frequent users. Their result provided strong evidence in favor of the finite mixture model which gave different estimates of policy-relevant measures when calculated for hypothetical individuals with specific characteristics.

Since Deb and Trivedi (1997, 2002) had proved that the finite mixture regression model fitted better than the standard and hurdle regression models, the model comparison in the following subsection only focuses on 2 and 3-component mixture regression models which are estimated by the ML, MHD and L_2E estimations, respectively. The same diagnostic techniques used in Section 5.1 are employed to compare the adequacy-of-fit among the models and estimators.

Referring to the sample size, this data set with 20,186 observations is much larger than the data set of Dionne et al. (1996). The simulation studies in Chapter 4 show that the MHD and L_2E estimates are more convincing if the sample size is large enough. However, there is a deficiency to directly apply the proposed

approach to the data. The data is composed of a panel of families, with all members of a given family in the sample. It is ideal that the objective functions take into account the correlation, but considering the computational difficulty the estimators are still specified as the original functional forms for this empirical application, which is also consistent with the manner in the previous literature.

5.2.1 Data and Summary Statistics

The data set studied in Deb and Trivedi (2002) was collected from the Rand Health Insurance Experiment which recorded the enrollee's use of medical care services and health status throughout the randomly assigned period. This experiment was conducted from 1974 to 1982 with around 8,000 enrollees in 2,823 families from 6 sites across the U.S. The final sample used in their paper only consisted of individuals in the fee-for-service plans, which contained 20,186 observations.

Definitions and summary statistics for the variables are presented in Table 5.15. For application purpose, only one of the two utilization measures is picked as a response variable, i.e., the number of contacts with a physician (**MDU**). The explanatory variables can be classified into three main groups: (1) insurance plan variables (coinsurance rate, deductibles, maximum dollar-expenditure function, participation-incentive payment function), (2) health status variables (chronic conditions, existence of a physical limitation, self-reported health status), and (3) socio-economic variables (family income, family size, age, gender, race, education).[11]

5.2.2 Results

Since Deb and Trivedi (1997, 2002) found that the finite mixture regression model gave a better fit and interpretation than the conventional models, these findings are used as a starting point to evaluate the proposed approach. The data set is only modeled by 2 and 3-component mixture regression models, which are in

[11]For more details of data processing, see Deb and Trivedi (2002, Section 3).

Table 5.15: Variable Definitions and Summary Statistics

Variable	Definition	Mean	SD
MDU	number of outpatient visits to an MD†	2.86	4.50
LC	ln(coinsurance + 1), $0 \leq$ coinsurance ≤ 100	1.71	1.96
IDP	1 if individual deductible plan	0.22	0.41
LPI	ln(max(1, annual participation incentive payment))	4.71	2.70
FMDE	0 if IDP = 1; otherwise ln(max(1, MDE/(0.01 coinsurance)))	3.15	3.64
LINC	ln(family income)	8.71	1.23
LFAM	ln(family size)	1.25	0.54
AGE	Age in years	25.7	16.8
FEMALE	1 if the person is female:	0.52	0.50
CHILD	1 if age is less than 18	0.40	0.49
FEMCHILD	FEMALE $*$ CHILD	0.19	0.40
BLACK	1 if race of household head is black	0.18	0.38
EDUCDEC	education of the household head in years	12.0	2.81
PHYSLIM	1 if the person has a physical limitation	0.12	0.32
DISEASE	index of chronic diseases	11.2	6.74
HLTHG	1 if self-rated health is good	0.36	0.48
HLTHF	1 if self-rated health is fair	0.08	0.27
HLTHP	1 if self-rated health is poor	0.01	0.12

Notes:† The frequency of zero counts is 31.25%.

Table 5.16: Overdispersion Parameters of the FMNB Estimated by the ML, MHD, and L_2E Estimations

	FMNB1			FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E
	2-component Mixture Regression Models					
$\hat{\psi}_1$	1.684	0.002	0.111	0.388	2.0e-4	0.213
$\hat{\psi}_2$	7.070	0.003	2.319	1.007	0.107	0.677
	3-component Mixture Regression Models					
$\hat{\psi}_1$	1.067	0.023	0.084	0.355	1e-17	0.290
$\hat{\psi}_2$	1.442	0.002	0.363	0.480	7e-13	0.459
$\hat{\psi}_3$	7.359	0.002	2.288	0.984	1.7e-6	0.145

combination with the Poisson, NB1, and NB2 density specifications, respectively. All of the models are estimated by the ML, MHD, and L_2E estimations. For model and estimator selection, the required diagnostic tools are as follows: the Pearson's chi-square statistic, root mean squared error, fitted descriptive statistics, and cross-validation.

Before assessing the three estimators, the task starts with model selection for which the first priority is to figure out whether the negative binomial density specification is a better choice than the Poisson one for this data set. Table 5.16 presents the ML, MHD, and L_2E estimates of overdispersion parameters for the 2 and 3-component mixture regression models based on the NB1 and 2 density specifications. Clearly, the ML and L_2E estimates are different from zero but that is not necessary for the MHD ones. The MHD estimates of overdispersion parameters are all close to zero, except $\hat{\psi}_1$ in the 3-FMNB1 and $\hat{\psi}_2$ in the 2-FMNB2. This result is verified by Tables 5.17 and 5.18 where the "Var." values are greater than the "Mean" values in the FMNBs for the ML and L_2E estimators but are almost equal for the MHD one.

From the adequacy-of-fit perspectives, Tables 5.19 and 5.20 present the fitted frequencies and the χ^2_P and $RMSE$ values of the 2 and 3-component mixture regression models, and Tables 5.21 and 5.22 report the $\chi^2_{P|x}$ and $RMSE_{|x}$ values for the corresponding models and estimators conditional on **FEMALE** and **CHILD** equal to 1, respectively.[12]

For 2-component mixture regression models, Table 5.19 presents that the ML estimator of the 2-FMNB1 has both the smallest χ^2_P and $RMSE$ values, 10.49 and 4.283, respectively, compared with that of the 2-FMP and 2-FMNB2. Based on the comparison of sample and fitted frequencies, the ML estimator of the 2-FMNB1 also shows a better fit over the entire range of the distribution than that of the other two density specifications. However, for the MHD and L_2E estimators, each of them only has one smallest goodness-of-fit statistic value to support the

[12]The subsamples of **FEMALE** and **CHILD** equal to 1 have more observations than others, i,e., 10,435 and 8,103, respectively.

Table 5.17: Fitted Descriptive Statistics of the ML, MHD, and L_2E Estimations (2-component Mixture Regression Models)

	2-FMP			2-FMNB1			2-FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
	Mixture density								
Mean	2.858	2.803	2.250	2.860	2.803	2.583	2.875	2.803	2.877
Var.	12.25	7.296	5.345	17.44	7.160	9.129	16.55	8.827	16.66
Min.	0.329	0.153	0.218	0.312	0.202	0.447	0.372	0.019	0.221
Max.	20.78	44.85	26.99	25.65	34.64	21.03	34.67	62.66	50.83
	Component density 1								
Mean	1.548	0.405	0.940	2.137	2.600	1.085	1.634	2.384	1.171
Var.	1.548	0.405	0.940	5.736	2.606	1.206	3.698	2.387	1.979
Min.	0.058	1.1e-6	0.010	0.224	6.0e-4	2.8e-4	0.010	6.0e-5	0.001
Max.	15.76	69.35	20.01	13.79	68.98	9.332	41.52	66.16	47.21
	Component density 2								
Mean	9.233	3.343	4.216	5.792	2.928	3.084	4.384	9.140	3.928
Var.	9.233	3.343	4.216	46.74	2.936	10.24	27.42	22.59	20.94
Min.	1.632	0.187	0.529	0.672	0.316	0.596	0.811	0.296	0.357
Max.	45.22	54.94	37.47	73.76	34.51	25.06	26.34	35.74	53.06

Notes:
"Mean" is the sample average of the fitted means, "Var." is the sample average of the fitted variances, "Min." is the minimum of the fitted means, and "Max." is the maximum of the fitted means. The component densities are put in order with the smallest "Mean" at the top. The sample mean and variance are 2.861 and 20.293, respectively.

Table 5.18: Fitted Descriptive Statistics of the ML, MHD, and L_2E Estimations (3-component Mixture Regression Models)

	3-FMP			3-FMNB1			3-FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
	Mixture density								
Mean	2.863	2.804	2.598	2.837	2.803	2.669	2.863	2e+15	2.736
Var.	16.91	10.45	8.978	17.44	9.826	10.45	18.33	7e+33	14.24
Min.	0.398	0.055	0.238	0.209	0.267	0.328	0.452	6e-13	0.418
Max.	19.48	27.04	38.81	25.50	40.66	26.73	42.16	2e+19	50.45
	Component density 1								
Mean	1.125	1.9e-6	0.710	1.850	0.432	1.165	1.718	0.088	1.778
Var.	1.125	1.9e-6	0.710	3.825	0.442	1.262	4.816	0.088	4.922
Min.	0.026	7e-24	0.002	0.040	3.0e-9	0.006	1.4e-4	5e-105	3.9e-4
Max.	12.93	0.002	17.46	14.61	67.77	26.23	83.25	1.2e+3	83.00
	Component density 2								
Mean	5.445	2.348	2.533	2.384	2.274	1.340	2.526	2.4e+9	2.817
Var.	5.445	2.348	2.533	5.822	2.277	1.826	7.015	1e+10	12.52
Min.	0.942	0.037	0.140	0.091	0.266	1.1e-4	0.439	2e-10	0.134
Max.	32.74	19.53	47.62	19.67	31.97	16.72	21.26	1e+13	71.99
	Component density 3								
Mean	20.81	4.528	6.934	6.251	3.669	3.950	6.780	3e+15	4.262
Var.	20.81	4.528	6.934	52.26	3.676	12.99	58.96	7e+28	7.284
Min.	3.577	0.081	0.972	0.532	0.136	0.613	1.232	3e-25	1.528
Max.	63.69	69.29	67.45	78.95	67.18	41.03	32.17	2e+19	21.23

Notes:
"Mean" is the sample average of the fitted means, "Var." is the sample average of the fitted variances, "Min." is the minimum of the fitted means, and "Max." is the maximum of the fitted means. The component densities are put in order with the smallest "Mean" at the top. The sample mean and variance are 2.861 and 20.293, respectively.

Table 5.19: Fitted Frequencies, Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values for the ML, MHD, and L_2E Estimators (2-component Mixture Regression Models)

		2-FMP			2-FMNB1			2-FMNB2		
		ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
Count	Sample	Fitted Frequencies								
0	31.25	24.54	31.29	31.39	31.61	31.29	31.76	31.02	31.30	31.64
1	18.90	24.25	18.94	21.21	18.95	18.93	19.13	20.37	18.94	20.12
2	13.85	16.75	13.79	13.94	13.28	13.82	13.32	13.37	13.81	13.28
3	9.333	9.898	9.463	9.778	9.444	9.435	9.477	8.992	9.457	8.982
4	6.663	5.630	6.603	7.165	6.746	6.602	6.818	6.227	6.592	6.249
5	4.795	3.468	4.760	5.228	4.839	4.775	4.957	4.445	4.764	4.465
6	3.413	2.506	3.507	3.702	3.491	3.513	3.642	3.262	3.515	3.264
7	2.631	2.089	2.610	2.534	2.539	2.606	2.699	2.451	2.613	2.435
8	2.021	1.866	1.951	1.687	1.867	1.945	2.013	1.878	1.949	1.847
9	1.422	1.682	1.464	1.103	1.390	1.461	1.509	1.463	1.460	1.422
10	1.021	1.484	1.103	0.717	1.050	1.105	1.135	1.155	1.101	1.109
11+	4.706	5.834	4.509	1.545	4.793	4.512	3.533	5.369	4.510	5.188
		Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values								
	χ^2_P †	1036	4.874	1439	10.49	4.808[a]	93.74	65.68	4.844	52.43
	$RMSE$	4.286	5.009	4.341	4.283[b]	5.151	4.299	4.301	5.020	4.449
	$obj.$ ‡	-45499	1.374e-3	-0.2050	-42037	1.370e-3	-0.2074	-42155	1.376e-3	-0.2065

Notes:

[a] The estimation method is preferred by the χ^2_P among all 2-component mixture regression models.

[b] The estimation method is preferred by the $RMSE$ among all 2-component mixture regression models.

† The 5% critical value of the Pearson's chi-square test is $\chi^2(5\%, 11) = 19.675$.

‡ The objective function values are the log likelihood, Hellinger distance, and L_2 error for the ML, MHD, and L_2E estimators, respectively.

Table 5.20: Fitted Frequencies, Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values for the ML, MHD, and L_2E Estimators (3-component Mixture Regression Models)

		3-FMP			3-FMNB1			3-FMNB2		
		ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
Count	Sample	Fitted Frequencies								
0	31.25	28.83	31.29	31.43	32.02	31.29	31.79	30.97	63.66	31.67
1	18.90	22.07	18.94	20.25	18.59	18.93	19.24	20.05	1.965	19.26
2	13.85	13.93	13.82	13.33	13.29	13.82	13.35	13.75	1.041	13.13
3	9.333	8.812	9.427	9.103	9.543	9.431	9.310	9.383	0.690	9.430
4	6.663	6.152	6.607	6.463	6.826	6.603	6.567	6.466	0.508	6.886
5	4.795	4.692	4.779	4.755	4.875	4.776	4.730	4.534	0.399	5.026
6	3.413	3.675	3.513	3.586	3.490	3.513	3.485	3.246	0.329	3.652
7	2.631	2.825	2.603	2.729	2.514	2.605	2.617	2.374	0.281	2.647
8	2.021	2.097	1.944	2.071	1.830	1.944	1.991	1.773	0.247	1.921
9	1.422	1.508	1.461	1.557	1.351	1.461	1.529	1.350	1.5e+84	1.403
10	1.021	1.061	1.105	1.159	1.013	1.106	1.180	1.046	1.4e+69	1.034
11+	4.706	4.352	4.513	3.575	4.654	4.513	4.211	5.052	-1.5e+84	3.935
		Pearson's Chi-square Statistics, Root Mean Squared Errors, and Objective Function Values								
	χ^2_P †	162.3	4.773[a]	105.5	17.76	4.819	25.27	38.28	0	49.05
	$RMSE$	4.291	4.771	4.339	4.270[b]	4.860	4.297	4.305	2e+17	4.439
	$obj.$ ‡	-42731	1.370e-3	-0.2066	-41910	1.369e-3	-0.2085	-41980	-2.9e+40	-0.2077

Notes:

[a] The estimation method is preferred by the χ^2_P among all 3-component mixture regression models.

[b] The estimation method is preferred by the $RMSE$ among all 3-component mixture regression models.

† The 5% critical value of the Pearson's chi-square test is $\chi^2(5\%, 11) = 19.675$.

‡ The objective function values are the log likelihood, Hellinger distance, and L_2 error for the ML, MHD, and L_2E estimators, respectively.

2-FMNB1, i.e., $\chi^2_P = 4.808$ for the MHD estimator and $RMSE = 4.299$ for the L_2E estimator.

Referring to fitted frequencies, the MHD estimator actually have very similar results across the three density specifications that is consistent with their small overdispersion parameter estimates in Table 5.16. For the L_2E estimator, the sample and fitted frequencies are much closer to each other based on the negative binomial density specification than the Poisson one. Between the 2-FMNB1 and 2-FMNB2, although the χ^2_P prefers the latter, its smaller χ^2_P value, 52.43, is attributed to the better fitted frequencies of the count nine and higher. The 2-FMNB2 does not fit the cell frequencies well on the lower counts compared with the 2-FMNB1.

Regarding the $\chi^2_{P|x}$ and $RMSE_{|x}$ values, Table 5.21 presents that these two subsample goodness-of-fit statistics give the same conclusion of the density selection for each estimator as that in Table 5.19. The ML estimator of the 2-FMNB1 has the smallest statistic values for either **FEMALE** or **CHILD** subsample. The negative binomial density specification instead causes an unstable MHD estimator that generates abnormal statistic values in both of the subsamples. For the L_2E estimator, the $\chi^2_{P|x}$ and $RMSE_{|x}$ values of the 2-FMNB1 are smaller than those of the other models, except the $RMSE_{|x}$ value in the subsample of **CHILD**.

For 3-component mixture regression models, Table 5.20 presents that for the ML and L_2E estimators the 3-FMNB1 is the best fit model based on its smallest χ^2_P and $RMSE$ values. In terms of the MHD estimator, the estimation algorithm fails in the 3-FMNB2, and the χ^2_P and $RMSE$ values of 3-FMNB1 are greater than those of the 3-FMP. But for the fitted frequencies of the MHD estimator, the 3-FMP and 3-FMNB1 have very close values across the whole range of counts. Again, this result verifies the small estimates of overdispersion parameters in Table 5.16.

Table 5.22 reports the corresponding $\chi^2_{P|x}$ and $RMSE_{|x}$ values but the result is only consistent with that in Table 5.20 for the ML estimator. Among the three

Table 5.21: Subsample Pearson's Chi-square Statistics and Root Mean Squared Errors for the ML, MHD, and L_2E Estimators (2-component Mixture Regression Models)

	2-FMP			2-FMNB1			2-FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
	Subsample of **FEMALE**=1								
$\chi^2_{P\|x}$	501.3	2.924	548.3	14.62	2.595	23.39	30.14	5e+34	44.24
$RMSE_{\|x}$	4.624	5.209	4.679	4.616	5.382	4.642	4.648	2.8e+5	4.812
	Subsample of **CHILD**=1								
$\chi^2_{P\|x}$	316.2	13.04	492.9	25.56	0	31.67	27.70	16.35	346.0
$RMSE_{\|x}$	3.873	4.132	3.920	3.860	1e+75	4.7e+7	3.874	4.080	3e+11

Table 5.22: Subsample Pearson's Chi-square Statistics and Root Mean Squared Errors for the ML, MHD, and L_2E Estimators (3-component Mixture Regression Models)

	3-FMP			3-FMNB1			3-FMNB2		
	ML	MHD	L_2E	ML	MHD	L_2E	ML	MHD	L_2E
	Subsample of **FEMALE**=1								
$\chi^2_{P\|x}$	102.1	2.399	34.57	5.936	2.627	73.04	18.79	2.601	271.0
$RMSE_{\|x}$	4.627	5.146	4.675	4.631	5.043	3.2e+7	4.666	4.835	1e+89
	Subsample of **CHILD**=1								
$\chi^2_{P\|x}$	55.09	10.43	199.7	34.27	9.629	218.2	23.63	8.624	16.17
$RMSE_{\|x}$	3.886	4.122	3.882	3.844	4.482	4.1e+4	4.051	4.410	3.949

density specifications of the ML estimator, the 3-FMNB1 has relatively small values for both of the $\chi^2_{P|x}$ and $RMSE_{|x}$. On the other hand, the two subsample goodness-of-fit statistics of the MHD estimator do not specifically indicate which density specification has a better fit. Each of them picks different density specification and then reverts when the subsample is changed. The contradiction for the L_2E estimator is that the 3-FMNB1 does not fit well in the two subsamples; instead, the 3-FMP fits better than the other two models based on the the $\chi^2_{P|x}$ and $RMSE_{|x}$ values, except the $\chi^2_{P|x}$ value in the subsample of **CHILD**.

Based on these results above, the ML estimator is the only estimator among the three explicitly has a preferred density specification for the 2 and 3-component mixture regression models and coincidentally both of them are the NB1 density. Using this preferred density specification of the ML estimator as a benchmark, the cross-validation computes the χ^2_P and $RMSE$ values for the ML, MHD, and L_2E estimators in combination with the 2-FMNB1 and 3-FMNB1, respectively, all of which are presented in Figure 5.4.

For the 2-FMNB1, although the χ^2_P favors the MHD estimator in the training sample, this overwhelming preference does not continue in the holdout sample where the ML estimator has five smallest χ^2_P values in the ten trials. Moreover, the $RMSE$ completely prefers the ML estimator to the MHD one no matter in the training or holdout sample, but surprisingly the ML and L_2E estimator have very close $RMSE$ values. For the 3-FMNB1, the same outcome is produced by the χ^2_P that the MHD estimator owns all the smallest statistic values in the training sample but in the holdout sample the ML has four smallest values in eight trials, excluding the trial 1 and 4 where the MHD estimation algorithm fails. Based on the $RMSE$, the ML estimator is obviously the best choice, and this time it does not have a close value to that of the L_2E estimator for every trial.

At this point, the ML estimator has been confirmed as the best estimator among the three based on the cross-validation result in Figure 5.4, but it is not clear whether the 2 or 3-FMNB1 is the preferred model. To assess the adequacy

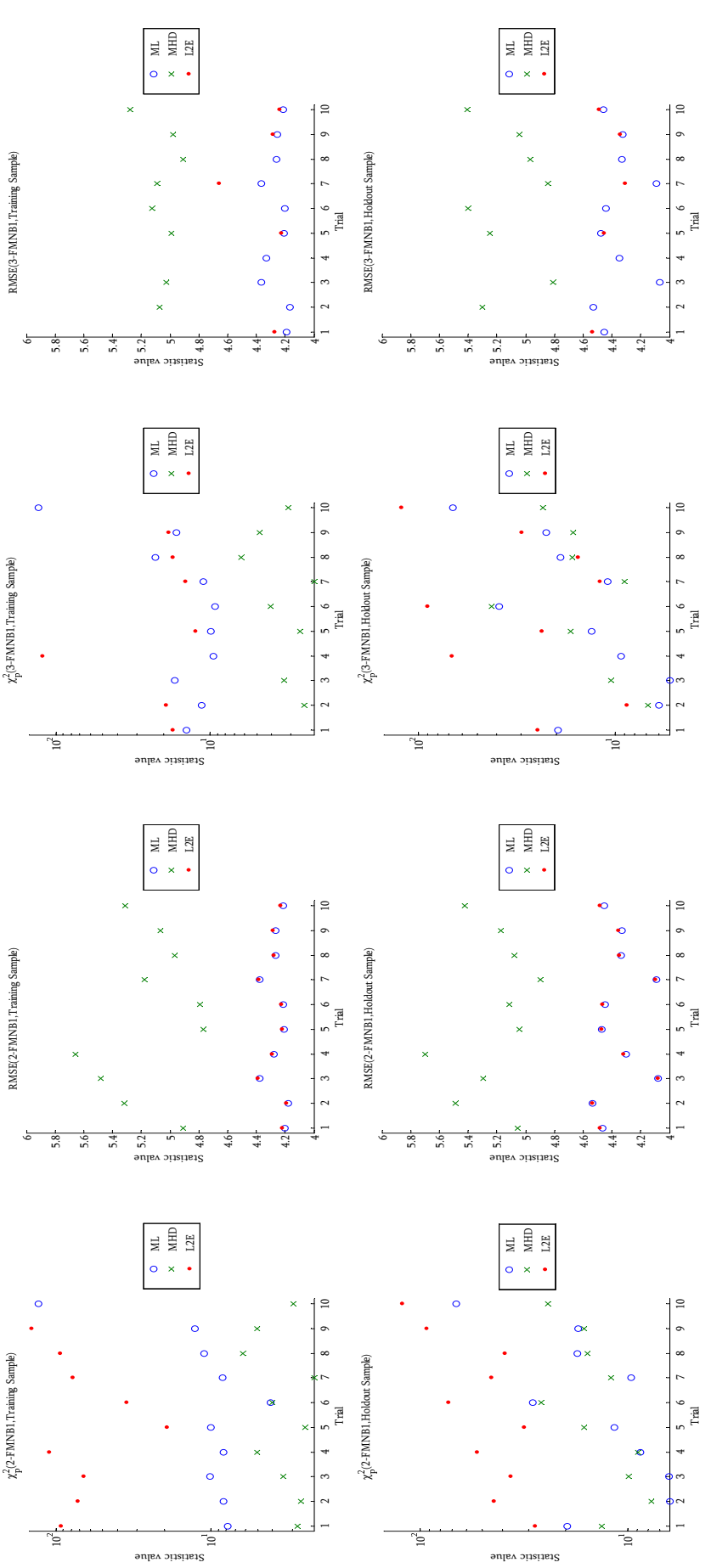


Figure 5.4: Pearson's Chi-square Statistics and Root Mean Squared Errors for the ML, MHD, and L_2E Estimators from the Cross-validation (2 and 3-component Negative Binomial 1 Mixture Regression Models)

Notes:
To consistently compare the statistic values among the ML, MHD, and L_2E estimators, the y axes are restricted to the same scales so that some points are beyond the maximum limit of y axes.

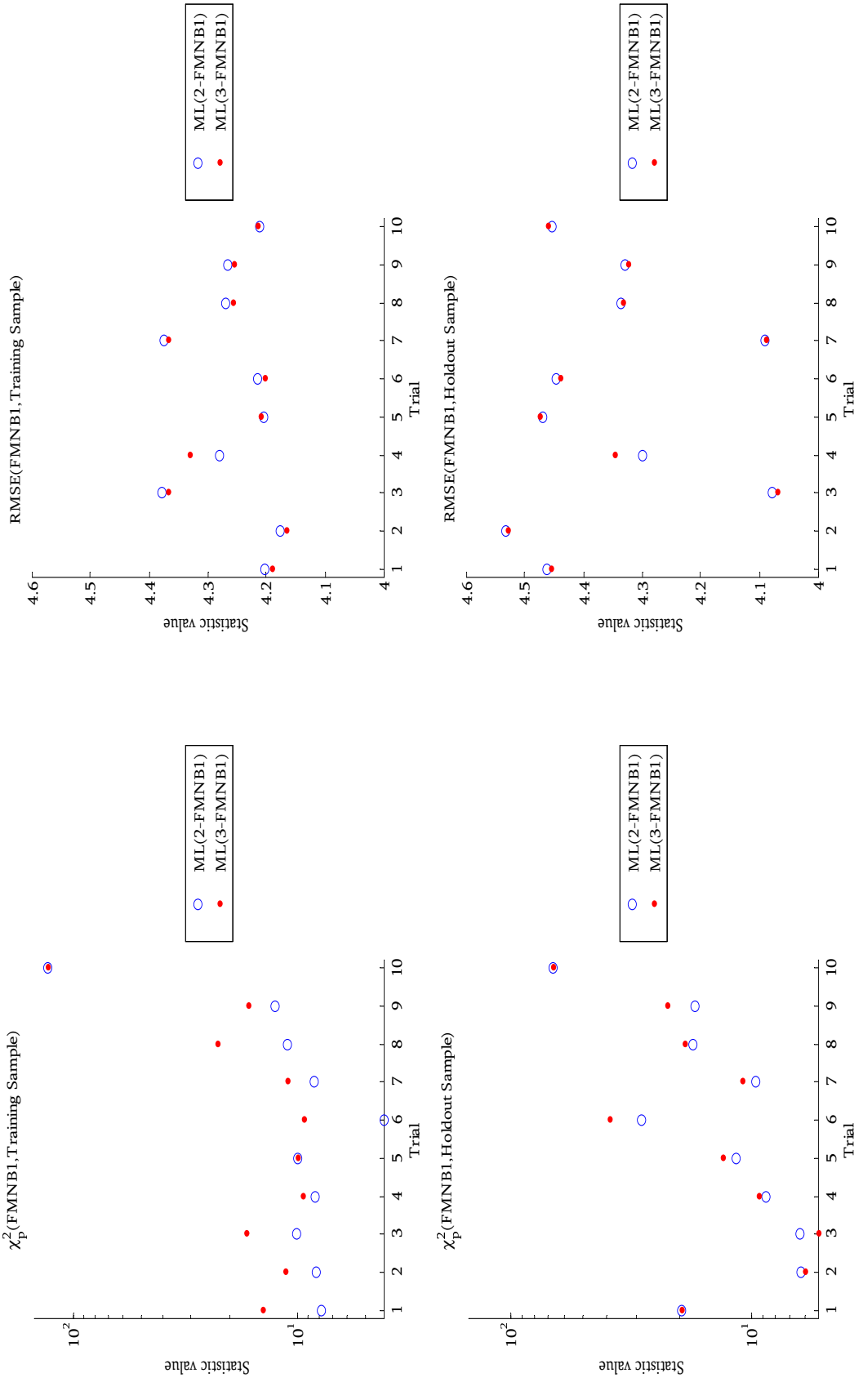


Figure 5.5: Pearson's Chi-square Statistics and Root Mean Squared Errors for the ML Estimator from the Cross-validation (2 and 3-component Negative Binomial 1 Mixture Regression Models)

of fit between these two models, Figure 5.5 focuses on the ML estimators of the 2 and 3-FMNB1. For the χ^2_P, the 2-FMNB1 has more lower points in the plots than the 3-FMNB1 does, i.e., nine and seven in the training and holdout samples, respectively. However, the $RMSE$ does not provide strong evidence to support a particular model. In both the training and holdout samples, although the $RMSE$ value of the 3-FMNB1 is smaller than that of the 2-FMNB1 in seven out of ten trials, the difference is actually tiny. To sum up, the 2-FMNB1 is selected as the best fit model and this is consistent with the result in Deb and Trivedi (2002).

5.3 Discussion of the Best Estimator in Empirical Applications

In the two empirical applications, the choice of the best estimator is the ML estimator. There are two reasons to explain this choice. First, it can be attributed to flexibility of the finite mixture model. Based on the estimated results in Deb and Trivedi (1997, 2002) and this dissertation, the finite mixture regression model is shown to enormously improve the goodness-of-fit compared with the standard and hurdle regression models. That leaves less room for the MHD and L_2E estimators to down weight the exotic observations, especially for observations in the far right tail.

The other reason to explain why the ML estimator is selected is data processing procedure. The data set studied in the first empirical application already excluded extreme values when it was provided by the corresponding author of Dionne et al. (1996). For the data set used in the second empirical application, Deb and Trivedi (2002) claimed that careful data collection minimized the contaminated sample. As a result, there is less necessity of using the MHD and L_2E estimators.

To verify this argument, an experiment is carried out to examine whether or not there exists a certain amount of contaminated sample in each data set. There are 100 replications in this experiment and for each replication the simulated sam-

ple is constructed by randomly drawing 95 percent of the original observations. According to the experiment design, it is expected that the contaminated data can be largely eliminated in some replications if the original data sets are indeed contaminated. Then by applying the same model specification and the ML estimation to those simulated samples, the "experimental" estimates should be significantly different from the original estimates of the entire sample.

Based on the result in Table 5.23, the range of 100 experimental estimates is acceptably narrow for each statistically significant covariate, except for the component 1 intercept in the second empirical application.[13] Therefore, it is safe to say that the contamination problem is restricted for the two original data sets. In this situation, the ML estimator is preferred to the MHD and L_2E ones.

[13]In fact, there are only two abnormal values, 0.1953 and 0.1755, among the 100 replications. The third largest value is -0.2155.

Table 5.23: Ranges of Experimental Estimates for the Two Empirical Applications

Empirical Application 1 (ML Estimates of the 2-FMP)							Empirical Application 2 (ML Estimates of the 2-FMNB1)						
Variables	Component 1†			Component 2			Variables	Component 1			Component 2		
INTERCEPT	-2.46***	-2.73	-2.20	1.25***	1.11	1.35	INTERCEPT	-0.44**	-0.54	0.20	0.57	-1.87	0.74
DT6	0.48***	0.41	0.61	0.05	0.01	0.08	LC	-0.20***	-0.21	-0.19	-0.02	-0.04	-0.01
DURER	0.06	-0.01	0.11	0.13***	0.11	0.15	DIP	-0.52***	-0.54	-0.50	-0.33***	-0.37	-0.28
AGE1	0.35	0.12	0.50	0.25*	0.14	0.33	LPI	0.02***	0.02	0.03	0.02	0.01	0.02
AGE2	0.39*	0.28	0.51	0.19***	0.15	0.25	FMDE	0.04***	0.03	0.04	-0.03**	-0.04	-0.02
DESTIN	-0.46**	-0.56	-0.35	-0.25***	-0.29	-0.22	LINC	0.08***	0.00	0.09	0.06	0.04	0.35
ETU1	-0.43	-1.22	-0.13	0.19	0.06	0.28	LFAM	-0.08***	-0.09	-0.05	-0.18***	-0.27	-0.14
ETU2	0.45*	0.25	0.61	0.10	0.01	0.15	AGE	6e-4	0.00	0.00	4e-3	0.00	0.01
ETU3	0.36	0.12	0.50	-0.05	-0.16	0.02	FEMALE	0.41***	0.39	0.43	0.31***	0.27	0.36
RECSAL	-0.49**	-0.62	-0.36	-0.34***	-0.42	-0.31	CHILD	0.33***	0.29	0.37	0.32***	0.25	0.40
M1	-0.42	-0.76	-0.20	0.13*	0.07	0.19	FEMCHILD	-0.41***	-0.45	-0.39	-0.36***	-0.43	-0.26
M2	0.04	-0.22	0.35	0.34***	0.25	0.41	BLACK	-0.80***	-0.89	-0.76	-0.58***	-0.66	-0.39
M3	0.77***	0.57	0.93	0.33***	0.24	0.39	EDUCDEC	0.03***	0.02	0.03	0.01	0.00	0.02
NM1	0.04	-0.10	0.18	0.10	0.04	0.14	PHYSLIM	0.14***	0.12	0.16	0.39***	0.34	0.44
CENTRE	-0.29	-0.45	-0.18	0.12	0.05	0.16	DISEASE	0.02***	0.02	0.02	0.03***	0.03	0.03
RESID	-0.12	-0.25	0.13	-0.03	-0.07	0.01	HLTHG	0.04*	0.03	0.05	0.04	0.01	0.08
Z2	0.37	0.27	0.57	-0.17**	-0.21	-0.09	HLTHF	0.19***	0.17	0.22	0.27***	0.22	0.33
Z3	0.29	0.18	0.41	-0.30***	-0.35	-0.25	HLTHP	0.52***	0.46	0.56	0.39***	0.28	0.57
Z4	-0.55	-0.78	-0.33	-0.14	-0.25	-0.05	$\hat{\psi}$	1.68***	1.57	1.74	7.07***	6.50	7.52
$\hat{\pi}$	0.77	0.76	0.78				$\hat{\pi}$	0.80	0.77	0.82			

Notes:

† For every component density, there are three columns under it, which present the original estimates of the entire sample, minimum experimental estimates, and maximum experimental estimates, respectively.

***, **, and * mean significant at the 1%, 5%, and 10% levels, respectively, but those does not apply to $\hat{\pi}$ because it has a boundary value.

Chapter 6

Conclusion

Two robust estimators for finite mixtures of count data regression models are developed in this dissertation: the MHD estimator and the L_2E estimator, which is a special case of the MDPD estimator. The Monte Carlo simulation studies show that the MHD and L_2E estimators are more robust than the ML one by sacrificing efficiency but the robustness of these two minimum distance estimators is deteriorated as the mixing probability approaches one.

In the first empirical application, the estimated results show that the heterogeneity of the data set is not simply explained by an overdispersion parameter of a negative binomial distribution built in either a single or a two-step count data regression model. Instead, the 2-FMP which categorizes the sample into two latent classes, low-risk and high-risk, fits the heterogeneity better than the traditional models. In the second empirical application, the preferred model specification, the 2-FMNB1, is the same as that concluded by Deb and Trivedi (2002). For robust estimates, there is no convincing evidence to replace the ML estimator with the MHD or L_2E one in the two empirical applications. This outcome is attributable to the flexibility of the finite mixture model and data processing procedures.

From a computational perspective, there are two concerns about the two minimum distance estimators. First, high dimension of a covariate vector not only slows their convergence rate, but also reduces their estimation accuracy. Recent

research about dimension reduction in regression could be a remedy for this nuisance. Second, according to my experience, the ML estimates suggested by previous literature are not necessarily good starting values for the MHD and L_2E estimation problems in the case of finite mixtures of count data regression models. This will require more simulation studies to design a better searching algorithm for the two minimum distance estimation problems.

For the MHD estimation method and estimator selection tools, some improvements are still required. The MHD objective function in Lu et al. (2003) was constructed in an unconditional way. A better approach is to implement a conditional nonparametric density in the MHD estimation method, which returns to the regression definition and may also avoid bizarre results in the empirical application. Regarding estimator selection, the χ^2_P and $RMSE$ are proved to favor the MHD and ML estimators, respectively. Although the cross-validation cures this problem, it is worth pursuing a fair diagnostic statistic to give a direct and fast evaluation.

An interesting empirical application for future research is to evaluate banks' credit score methodologies by the proposed approach. Banks like to grant consumers loans on the basis of their credit scores, e.g., FICO, but there are some discussions about loan brokers or agents who might boost their business by polishing consumers' credit reports, especially for those with the credit scores just below the minimum requirement. The proposed approach can be used to estimate default probabilities or non-payments numbers and then compare the predicted power with that of the existing credit score methodologies, particularly for those controversial loans.

Appendices

Appendix A: Accommodation of Over- and Under-dispersion by Hurdle and Finite Mixture Models

First, let the hurdle regression model be specified by

$$f_{\boldsymbol{\theta}}\left(y_i \mid \boldsymbol{x}_i\right) = \begin{cases} f_1\left(0 \mid \boldsymbol{x}_i\right) & \text{if } y_i = 0, \\ \left\{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)\right\} \frac{f_2(y_i|\boldsymbol{x}_i)}{1-f_2(0|\boldsymbol{x}_i)} & \text{if } y_i > 0, \end{cases}$$

where $f_1(\cdot)$, the first step, is a binary choice model, and $f_2(\cdot)/\{1 - f_2(\cdot)\}$, the second step, is a zero-truncated model. It is elastic to use different distributions to define $f_1(\cdot)$, but the logit distribution is commonly adopted. For the empirical application in Chapter 5, $f_1(\cdot)$ is specified as the logit regression, i.e., $f_1\left(0 \mid \boldsymbol{x}_i\right) = 1/\exp\left(\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}_1\right)$, and $f_2(\cdot)$ is specified as the Poisson and negative binomial regressions, respectively.

The hurdle model with a truncated Poisson regression (HP) assumes that $f_2(\cdot)$ is the Poisson regression whose functional form is like (2.2) with $\mu_{i2} = \exp\left(\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}_2\right)$ so that $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^{\intercal}, \boldsymbol{\beta}_2^{\intercal})^{\intercal}$. The conditional mean and variance of the HP are, respectively,

$$E\left(Y_i \mid \boldsymbol{X}_i\right) = \frac{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)}{1 - f_2\left(0 \mid \boldsymbol{x}_i\right)}\mu_{i2}$$

and

$$Var\left(Y_i \mid \boldsymbol{X}_i\right) = \frac{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)}{1 - f_2\left(0 \mid \boldsymbol{x}_i\right)} \left[\mu_{i2} + \left\{1 - \frac{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)}{1 - f_2\left(0 \mid \boldsymbol{x}_i\right)}\right\} \mu_{i2}^2\right].$$

Since

$$Var\left(Y_i \mid \boldsymbol{X}_i\right) / E\left(Y_i \mid \boldsymbol{X}_i\right) = 1 + \left\{1 - \frac{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)}{1 - f_2\left(0 \mid \boldsymbol{x}_i\right)}\right\} \mu_{i2},$$

where $Var\left(Y_i \mid \boldsymbol{X}_i\right) / E\left(Y_i \mid \boldsymbol{X}_i\right)$ can be either greater or less than 1, the HP allows for over- or underdispersion in data.

For the hurdle model with a truncated negative binomial regression (HNB), let $f_2\left(\cdot\right)$ denote the negative binomial regression and the functional form is analogous to (2.3) with $\mu_{i2} = \exp\left(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_2\right)$; thus $\boldsymbol{\theta} = (\psi, \boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_2^\intercal)^\intercal$. The conditional mean and variance of the HNB are, respectively,

$$E\left(Y_i \mid \boldsymbol{X}_i\right) = \frac{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)}{1 - f_2\left(0 \mid \boldsymbol{x}_i\right)} \mu_{i2}$$

and

$$Var\left(Y_i \mid \boldsymbol{X}_i\right) = \frac{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)}{1 - f_2\left(0 \mid \boldsymbol{x}_i\right)} \left[\mu_{i2}\left(1 + \alpha_i \mu_{i2}\right) + \left\{1 - \frac{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)}{1 - f_2\left(0 \mid \boldsymbol{x}_i\right)}\right\} \mu_{i2}^2\right].$$

Here

$$Var\left(Y_i \mid \boldsymbol{X}_i\right) / E\left(Y_i \mid \boldsymbol{X}_i\right) = 1 + \left\{\alpha_i + 1 - \frac{1 - f_1\left(0 \mid \boldsymbol{x}_i\right)}{1 - f_2\left(0 \mid \boldsymbol{x}_i\right)}\right\} \mu_{i2}$$

shows that the HNB not only can handle overdispersion in data but also underdispersion.

In terms of finite mixtures of count data regression models, the conditional mean is

$$E\left(Y_i \mid \boldsymbol{X}_i\right) = \sum_{j=1}^{k} \pi_j \mu_{ij}.$$

The conditional variance of the FMP and FMNB are given, respectively, by

$$Var\left(Y_i \mid \boldsymbol{X}_i\right) = \sum_{j=1}^{k} \pi_j \left(\mu_{ij} + \mu_{ij}^2\right) - \left(\sum_{j=1}^{k} \pi_j \mu_{ij}\right)^2$$

and

$$Var\left(Y_i \mid \boldsymbol{X}_i\right) = \sum_{j=1}^{k} \pi_j \left\{\mu_{ij}\left(1 + \alpha_{ij}\mu_{ij}\right) + \mu_{ij}^2\right\} - \left(\sum_{j=1}^{k} \pi_j \mu_{ij}\right)^2 .$$

Because $Var\left(Y_i \mid \boldsymbol{X}_i\right)/E\left(Y_i \mid \boldsymbol{X}_i\right)$ can be greater or less than 1, the FMP and FMNB admit over- or underdispersion in data as well.

Appendix B: Proof of Avar $\left(\hat{\boldsymbol{\theta}}_{MHD}\right)$ in Lu et al. (2003)

When the model is correctly specified, $G_n \equiv G_{\boldsymbol{\theta}}$, Simpson (1987) proved that Avar $\left(\hat{\boldsymbol{\theta}}_{MHD}\right) = \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0)$ and $\hat{\boldsymbol{\theta}}_{MHD}$ was asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{ML}$. Based on the second result, Lu et al. (2003) further specified $\boldsymbol{i}(\boldsymbol{\theta}_0)$ by the following manner.

$$
\begin{aligned}
i(\boldsymbol{\theta}_0) &= \sum_{y=0}^{m} \boldsymbol{l}_{\boldsymbol{\theta}_0}(y)\, \boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}(y)\, f_{\boldsymbol{\theta}_0}(y) \\
&= E\left[\boldsymbol{l}_{\boldsymbol{\theta}_0}(y)\, \boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}(y)\right] \\
&= E\left[\boldsymbol{l}_{\boldsymbol{\theta}_0}(y_i)\, \boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}(y_i)\right] \\
&\overset{a}{=} -E\left[\boldsymbol{H}_{\boldsymbol{\theta}_0}(y_i)\right] \ (\boldsymbol{H}_{\boldsymbol{\theta}_0}(\cdot) \text{ is the Hessian}) \\
&= -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}} \log f_{\boldsymbol{\theta}_0}(y_i)\right] \\
&= -E\left[-\frac{\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(y_i)\, \frac{\partial}{\partial\boldsymbol{\theta}^{\intercal}} f_{\boldsymbol{\theta}_0}(y_i)}{f_{\boldsymbol{\theta}_0}^2(y_i)} + \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}} f_{\boldsymbol{\theta}_0}(y_i)}{f_{\boldsymbol{\theta}_0}(y_i)}\right] \\
&= E\left[\boldsymbol{l}_{\boldsymbol{\theta}_0}(y_i)\, \boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}(y_i) - \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}} f_{\boldsymbol{\theta}_0}(y_i)}{f_{\boldsymbol{\theta}_0}(y_i)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{l}_{\boldsymbol{\theta}_0}(y_i)\, \boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}(y_i) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}} f_{\boldsymbol{\theta}_0}(y_i)\right] \\
&= \sum_{y=0}^{m}\left[\boldsymbol{l}_{\boldsymbol{\theta}_0}(y)\, \boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}(y)\, f_{\boldsymbol{\theta}_0}(y) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}} f_{\boldsymbol{\theta}_0}(y)\right] \\
&= \sum_{y=0}^{m}\left[\boldsymbol{l}_{\boldsymbol{\theta}_0}(y)\, \boldsymbol{l}_{\boldsymbol{\theta}_0}^{\intercal}(y)\, f_n(y) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}} f_{\boldsymbol{\theta}_0}(y)\right].
\end{aligned}
$$

Appendix C: Expansions of $\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Sigma}$

Avar $\left(\hat{\boldsymbol{\theta}}_{MDPD}\right) = E\left[\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0)\right]^{-1} \boldsymbol{\Sigma} E\left[\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0)\right]^{-1}$, where $\boldsymbol{\Psi}(\boldsymbol{\theta}_0) = \int \boldsymbol{u}_{\boldsymbol{\theta}_0}(y \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{1+\alpha}(y \mid \boldsymbol{X}_i)\, dy - \boldsymbol{u}_{\boldsymbol{\theta}_0}(Y_i \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{\alpha}(Y_i \mid \boldsymbol{X}_i)$, $\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}^{\intercal}} \boldsymbol{\Psi}(\boldsymbol{\theta}_0)$, and $\boldsymbol{\Sigma} = E\Big[\boldsymbol{\Psi}(\boldsymbol{\theta}_0) \boldsymbol{\Psi}(\boldsymbol{\theta}_0)^{\intercal}\Big]$. $\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Sigma}$ are expanded as follows.

$$
\begin{aligned}
\dot{\boldsymbol{\Psi}}(\boldsymbol{\theta}_0) &= \frac{\partial}{\partial \boldsymbol{\theta}^{\intercal}} \boldsymbol{\Psi}(\boldsymbol{\theta}_0) \\
&= \int \Big\{ \dot{\boldsymbol{u}}_{\boldsymbol{\theta}_0}(y \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{1+\alpha}(y \mid \boldsymbol{X}_i) + (1+\alpha) \boldsymbol{u}_{\boldsymbol{\theta}_0}(y \mid \boldsymbol{X}_i) \boldsymbol{u}_{\boldsymbol{\theta}_0}^{\intercal}(y \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{1+\alpha} \\
&\quad (y \mid \boldsymbol{X}_i)\, dy \Big\} - \Big\{ \dot{\boldsymbol{u}}_{\boldsymbol{\theta}_0}(Y_i \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{\alpha}(Y_i \mid \boldsymbol{X}_i) + \alpha \boldsymbol{u}_{\boldsymbol{\theta}_0}(Y_i \mid \boldsymbol{X}_i) \boldsymbol{u}_{\boldsymbol{\theta}_0}^{\intercal}(Y_i \mid \boldsymbol{X}_i) \\
&\quad f_{\boldsymbol{\theta}_0}^{\alpha}(Y_i \mid \boldsymbol{X}_i) \Big\}, \\
\boldsymbol{\Sigma} &= E\left[\boldsymbol{\Psi}(\boldsymbol{\theta}_0) \boldsymbol{\Psi}(\boldsymbol{\theta}_0)^{\intercal}\right] \\
&= E\Bigg[\left\{ \int \boldsymbol{u}_{\boldsymbol{\theta}_0}(y \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{1+\alpha}(y \mid \boldsymbol{X}_i)\, dy \right\} \cdot \left\{ \int \boldsymbol{u}_{\boldsymbol{\theta}_0}(y \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{1+\alpha}(y \mid \boldsymbol{X}_i)\, dy \right\}^{\intercal} \\
&\quad -2 \left\{ \int \boldsymbol{u}_{\boldsymbol{\theta}_0}(y \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{1+\alpha}(y \mid \boldsymbol{X}_i)\, dy \right\} \cdot \left\{ \boldsymbol{u}_{\boldsymbol{\theta}_0}(Y_i \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{\alpha}(Y_i \mid \boldsymbol{X}_i) \right\}^{\intercal} \\
&\quad + \boldsymbol{u}_{\boldsymbol{\theta}_0}(Y_i \mid \boldsymbol{X}_i) \boldsymbol{u}_{\boldsymbol{\theta}_0}^{\intercal}(Y_i \mid \boldsymbol{X}_i) f_{\boldsymbol{\theta}_0}^{2\alpha}(Y_i \mid \boldsymbol{X}_i) \Bigg], \\
\boldsymbol{u}_{\boldsymbol{\theta}_0}(\cdot) &= \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}_0}(\cdot), \text{ and} \\
\dot{\boldsymbol{u}}_{\boldsymbol{\theta}_0}(\cdot) &= \frac{\partial}{\partial \boldsymbol{\theta}^{\intercal}} \boldsymbol{u}_{\boldsymbol{\theta}_0}(\cdot).
\end{aligned}
$$

Appendix D: Expansions of $\dot{\boldsymbol{\Psi}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right)$ and $\widehat{\boldsymbol{\Sigma}}$ for Count Data Regression Models

$\widehat{\text{Avar}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right) = \left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\dot{\boldsymbol{\Psi}}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right)\right]^{-1}\widehat{\boldsymbol{\Sigma}}\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\dot{\boldsymbol{\Psi}}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right)\right]^{-1}$. For count data regression models, $\widehat{\dot{\boldsymbol{\Psi}}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right)$ and $\widehat{\boldsymbol{\Sigma}}$ have the following expansion forms, respectively.

$$
\begin{aligned}
\widehat{\dot{\boldsymbol{\Psi}}}\left(\hat{\boldsymbol{\theta}}_{MDPD}\right) &= \sum_{y=0}^{m}\Big\{\dot{\boldsymbol{u}}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)\,f^{1+\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)+(1+\alpha)\,\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)\\
&\quad \boldsymbol{u}^{\intercal}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)\,f^{1+\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)\Big\}-\dot{\boldsymbol{u}}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i)\,f^{\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}\\
&\quad (Y_i\mid\boldsymbol{X}_i)-\alpha\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i)\,\boldsymbol{u}^{\intercal}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i)\,f^{\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i),\\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}\Bigg[\Bigg\{\sum_{y=0}^{m}\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)\,f^{1+\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)\Bigg\}\cdot\Bigg\{\sum_{y=0}^{m}\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}\\
&\quad (y\mid\boldsymbol{X}_i)\,f^{1+\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)\Bigg\}^{\intercal}-2\Bigg\{\sum_{y=0}^{m}\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}(y\mid\boldsymbol{X}_i)\,f^{1+\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}\\
&\quad (y\mid\boldsymbol{X}_i)\Bigg\}\cdot\Bigg\{\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i)\,f^{\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i)\Bigg\}^{\intercal}\\
&\quad +\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i)\,\boldsymbol{u}^{\intercal}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i)\,f^{2\alpha}_{\hat{\boldsymbol{\theta}}_{MDPD}}(Y_i\mid\boldsymbol{X}_i)\Bigg],\\
\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}(\cdot) &= \frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\hat{\boldsymbol{\theta}}_{MDPD}}(\cdot),\ \text{and}\\
\dot{\boldsymbol{u}}_{\hat{\boldsymbol{\theta}}_{MDPD}}(\cdot) &= \frac{\partial}{\partial\boldsymbol{\theta}^{\intercal}}\boldsymbol{u}_{\hat{\boldsymbol{\theta}}_{MDPD}}(\cdot).
\end{aligned}
$$

Bibliography

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1997). Robust and efficiency estimation by minimising a density power divergence. Technical Report 7, Department of Mathematics, University of Olso.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46:373–388.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge, New York.

Coleman, T. F. and Zhang, Y. (2009). *Optimization Toolbox User's Guide*. The MathWorks, Inc., 4 edition.

Cutler, A. and Cordero-Brana, O. I. (1996). Minimum hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91(436):1716–1723.

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

Deb, P. and Trivedi, P. K. (2002). The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics*, 21(4):601 – 625.

Dionne, G., Artís, M., and Guillén, M. (1996). Count data models for a credit scoring system. *Journal of Empirical Finance*, 3(3):303 – 325.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics : The Approach Based on Influence Functions*. Wiley, New York.

Holmström, K., Göran, A. O., and Edvall, M. M. (2008). *User's Guide for TOMLAB /NPSOL*. TOMLAB Optimization, Inc.

Holmström, K., Göran, A. O., and Edvall, M. M. (2009). *User's Guide for TOMLAB*. TOMLAB Optimization, Inc., 7 edition.

Karlis, D. and Xekalaki, E. (1998). Minimum hellinger distance estimation for poisson mixtures. *Computational Statistics & Data Analysis*, 29(1):81 – 103.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360.

Lu, Z., Hui, Y. V., and Lee, A. H. (2003). Minimum hellinger distance estimation for finite mixtures of poisson regression models and its applications. *Biometrics*, 59(4):1016–1026.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience.

Scott, D. W. (1998). Parametric modeling by minimum l2 error. Technical Report 98-3, Rice University, Dept. of Statistics.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43(3):274–285.

Scott, D. W. (2004). Outlier detection and clustering by partial mixture modeling. In *COMPSTAT Symposium*. Physica-Verlag/Springer.

Simpson, D. G. (1987). Minimum hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, 82(399):802–807.

Waltz, R. A. and Plantenga, T. D. (2009). *KNITRO User's Manual*. Ziena Optimization, Inc., 6 edition.

Warwick, J. (2005). A data-based method for selecting tuning parameters in minimum distance estimators. *Computational Statistics & Data Analysis*, 48(3):571 – 585.

Warwick, J. and Jones, M. C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation & Simulation*, 75(7):p581 – 588.