Abstract

Finite mixture models provide a concrete way to capture unobserved heterogeneity and thereby create clusters of data. In a regression setting, unobserved heterogeneity occurs when responses come from more than one subpopulation or when extreme values are present in a data. Finite mixture of regression models are widely applied in many diverse areas such as economics, genetics, medicine and psychology, among others. With the advent of modern computers and softwares, the method of maximum likelihood became the most widely applied estimation technique in finite mixture regression models. Even so, it is well known that maximum likelihood estimation for mixture models has many drawbacks, including their sensitivity to extreme values and data contamination.

In this dissertation, we introduce a robust estimation method for finite mixture of regression models based on a minimum integrated $L_2$ distance for count response. It is shown that our robust estimator, called $L_2E$, is consistent and asymptotically normal. Furthermore, our $L_2E$ estimator of the number of mixture components, the so called mixture complexity, is also shown to be consistent. Through Monte Carlo simulations, we compare the performance of our $L_2E$ with those of the maximum likelihood (ML) estimator and the minimum
Hellinger distance (MHD) estimator. It is shown that our $L_2E$ is highly competitive to the ML estimator and a better alternative to the MHD estimator.

The $L_2E$ estimator is also shown to be more robust than the MHD estimator, when samples are generated from gross-error contaminated (three-component) mixture of Poisson regression models. For the poorly-separated model with the small fraction of extreme values, we propose a modified $L_2E$ criterion, called $L_2E_p$, with a penalty function. The $L_2E_p$ estimator is also shown to be consistent and asymptotically normal for the case of two-component mixture regression models. In addition, the $L_2E_p$ estimator of mixture complexity is also shown to be consistent. Finally, the performance of these two methods are illustrated for two real data sets.

Next, we introduce a modified $L_2E$ estimator of mixture complexity in finite mixture models for continuous data. We show via simulations that the modified $L_2E$ estimator of mixture complexity helps detect a small fraction of extreme values and improve the efficiency when the true finite mixture model contains a small mixing proportion and are poorly-separated. We also consider non-normal mixture models and study the performance of our $L_2E$ estimator in gamma and lognormal mixtures via Monte Carlo simulations and real data analyses.

**Index Words:** Asymptotic normality, Consistency, Finite mixtures, Finite mixtures of regression models, $L_2E$ estimation, Mixture complexity, Penalty function, Robustness
$L_2E$ Estimation for Finite Mixture of Regression Models with Applications

and

$L_2E$ with Penalty and Non-normal Mixtures

by

Jaejun Lee

M.S., Virginia Polytechnic Institute and State University, 2006

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2010
Estimation for Finite Mixture of Regression Models with Applications

and

$L_2 E$ with Penalty and Non-normal Mixtures

by

Jaejun Lee

Approved:

Major Professor: T. N. Sriram

Committee: Gauri Datta
Nicole Lazar
William P. McCormick
Lynne Seymour

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
July 2010
DEDICATION

To my parents
Acknowledgments

First of all, I would like to express my sincere appreciation to my advisor, Professor T. N. Sriram, for all of his advice, guidance, and encouragement. He has demonstrated that he is a masterly statistician, an amazing teacher, and a great communicator during my research years. He has given me encouragement and confidence to decompose and annihilate insuperable problems and obstructions that I have encountered in my research.

I also want to thank the other members of my committee: Professors Ishwar Basawa, Gauri Datta, Nicole Lazar, William P. McCormick, and Lynne Seymour for their direct and fundamental contributions to this dissertation. Moreover, I would like to especially thank Professor Lazar for her constructive and insightful comments that improved this work. I also wish to thank all the faculty and staff in the department for their kind assistance and guidance.

Finally and most importantly, a special thanks is extended to my parents. Without their invaluable love and constant support, I could never have accomplished the most important goal in my life.
# Table of Contents

<table>
<thead>
<tr>
<th>Acknowledgments</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
</tbody>
</table>

## Chapter

1. **Introduction and Literature Review**
   - 1.1 An Overview of Finite Mixture Models
   - 1.2 References

2. **$L_2E$ Estimation for Finite Mixture of Regression Models with Applications**
   - 2.1 Introduction
   - 2.2 Finite Mixture of Regression Models
   - 2.3 $L_2E$ Criterion
   - 2.4 $L_2E$ Estimator of Mixture Complexity
   - 2.5 Theoretical Properties of $L_2E$
   - 2.6 Computational Details
   - 2.7 Simulation Studies
   - 2.8 Data Analysis
   - 2.9 Discussion
   - 2.10 Appendix
   - 2.11 References
3  \textit{L}_2\textit{E} \textit{with Penalty and Non-normal Mixtures} \hspace{1cm} 73

3.1 \textbf{Introduction} \hspace{1cm} 75

3.2 \textbf{L}_2\textit{E Estimator} \hspace{1cm} 77

3.3 \textbf{Computational Details} \hspace{1cm} 80

3.4 \textbf{Simulation Studies} \hspace{1cm} 82

3.5 \textbf{Data Analysis} \hspace{1cm} 85

3.6 \textbf{Discussion} \hspace{1cm} 90

3.7 \textbf{References} \hspace{1cm} 91

4 \textbf{Conclusions} \hspace{1cm} 94
## List of Figures

2.1  The Penalty function 1  .................................................. 32
2.2  The Penalty function 2 .................................................. 33
2.3  Maternity length of stay ................................................. 45
2.4  Histograms of LOS and LOS by clusters  .......................... 51
2.5  Histograms of crashes and crashes by clusters  ..................... 57
2.6  Histograms of LOS and LOS by clusters ($L_2E_p$ vs MLE)  .... 68
3.1  Frequency histogram of the data simulated from (3.12) ......... 82
3.2  Generating density functions for two different gamma mixtures 84
3.3  Plots of fitted gamma and lognormal mixtures: Italian surgical procedure times 89
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Comparison between MLE, MHDE, and $L_2E$ for Cases 1-4: Relative Efficiencies</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>Comparison between MLE, MHDE, and $L_2E_p$ for Case 4: Relative Efficiencies</td>
<td>36</td>
</tr>
<tr>
<td>2.3</td>
<td>The relative efficiency of the $L_2E$ vs the MLE (well-separated models)</td>
<td>37</td>
</tr>
<tr>
<td>2.4</td>
<td>The relative efficiency of the $L_2E$ vs the MLE (poorly-separated models)</td>
<td>38</td>
</tr>
<tr>
<td>2.5</td>
<td>The relative efficiency of the $L_2E$ vs the MHDE (well-separated models)</td>
<td>39</td>
</tr>
<tr>
<td>2.6</td>
<td>The relative efficiency of the $L_2E$ vs the MHDE (poorly-separated models)</td>
<td>40</td>
</tr>
<tr>
<td>2.7</td>
<td>Mixture complexity estimation results</td>
<td>42</td>
</tr>
<tr>
<td>2.8</td>
<td>Results of fitting Poisson mixture regression models to maternity length of stay data</td>
<td>47</td>
</tr>
<tr>
<td>2.9</td>
<td>Model-based clusters using the $L_2E_p$</td>
<td>48</td>
</tr>
<tr>
<td>2.10</td>
<td>Model-based clusters using the $L_2E$</td>
<td>49</td>
</tr>
<tr>
<td>2.11</td>
<td>Model-based clusters using the MHDE</td>
<td>50</td>
</tr>
<tr>
<td>2.12</td>
<td>Summary statistics for Toronto motor vehicle crashes data</td>
<td>54</td>
</tr>
<tr>
<td>2.13</td>
<td>$L_2E$ estimation of mixture complexity using $\hat{m}_{L_2E}$: Toronto crashes data</td>
<td>54</td>
</tr>
<tr>
<td>2.14</td>
<td>Results of fitting negative binomial mixture regression models to Toronto motor vehicle crashes data</td>
<td>55</td>
</tr>
<tr>
<td>2.15</td>
<td>Model-based clusters using our $L_2E$ and Bayesian method of Park and Lord (2009)</td>
<td>56</td>
</tr>
<tr>
<td>2.16</td>
<td>$L_2E$ and $L_2E_p$ estimation of mixture complexity using $\hat{m}<em>{L_2E}$ and $\hat{m}</em>{L_2E_p}$, respectively, in Section 2.8.1</td>
<td>65</td>
</tr>
<tr>
<td>2.17</td>
<td>Two-component Poisson mixture regression fit using $L_2E_p$ vs. MLE: LOS data</td>
<td>66</td>
</tr>
<tr>
<td>2.18</td>
<td>Model-based clusters using the MLE</td>
<td>67</td>
</tr>
</tbody>
</table>
3.1 Mixture complexity estimation results for four-component normal mixture in (3.12) ................................................................. 83
3.2 SLC data parameter estimates .............................................. 86
3.3 Conformity to the additive model ....................................... 87
3.4 Estimates of gamma and lognormal mixtures: Italian surgical procedure times 89
Chapter 1

Introduction and Literature Review

1.1 An Overview of Finite Mixture Models

Finite mixture models have been historically used to model heterogeneous populations that arise in biological, physical, and social sciences. The earliest application of finite mixture models can be traced back to the work of Pearson (1894), who proposed a mixture of two normal distributions to fit Weldon’s crabs data set via the method of moments. Since then, they have been widely applied to many fields such as economics, genetics, medicine, physics, psychology, and other disciplines (for example, Tanner, 1962; Broadben, 1966; Clark, Chapman, Coulson, and Hasselblad, 1968; Quandt and Ramsey, 1978; McManus, 1983; Schork, Allison, and Thiel, 1996; Ott, 1999). Also, they have been extensively used for cluster analysis, latent class analysis, discriminant analysis, image analysis, survival analysis, disease mapping, meta analysis, and more (for example, Schlattmann and Böhning, 1993; Böhning, 1999). Due to their flexibility, finite mixture models have received extensive attention over several decades, from both theoretical and practical perspectives. There is a vast literature on theory and applications of mixture models; see, Everitt and Hand (1981), Titterington, Smith, and Makov (1985), Lindsay (1995), Böhning (1999), McLachlan and Peel (2000), Böhning and Seidel (2003) and references therein.

Maximum likelihood estimation (MLE) using the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977) is a popular way to estimate the parameters in a finite mixture model. The EM is a general-purpose algorithm for iterative computation of maximum likelihood (ML) estimates, when the observations can be viewed as the incomplete data. While the MLE has desirable properties such as consistency and efficiency when there
is no contamination in the data, it is highly unstable when there are extreme values/outliers in the data or when the components of the underlying mixture model are essentially indistinguishable (Aitkin and Wilson, 1980). There are also other well-known disadvantages with the EM algorithm such as its sensitivity to initial values, slow convergence, and inability to provide estimation of the asymptotic variance-covariance matrix. Although modifications of EM algorithm customized for special cases have been proposed in the literature, there is no single algorithm that can be used in every case.

When there is contamination in the data, robust methods such as M-estimation and minimum distance estimation are feasible alternatives to likelihood-based methods. M-estimators, introduced by Huber (1964), provide a flexible class of estimators which generalize ML estimators and are consistent and asymptotically normal under certain regularity conditions. However, M-estimators usually sacrifice first-order efficiency at the model to achieve their robustness and they are not easily adapted for mixtures.

Minimum distance estimation methods yield yet another class of robust estimators. Some of the distance-based estimation methods include the Wolfowitz distance, the Levy distance, the Kolmogorov-Smirnov distance, the Cramér-von Mises distance, the Hellinger distance, and the squared $L_2$ distance. Many articles have investigated the behavior of these distance measures for mixtures; for example, Choi and Bulgren (1968) and Choi (1969) used the Wolfowitz distance; Yakowitz (1969) used the Levy distance; Deeley and Kruse (1968) and Blum and Susarla (1977) used the Kolmogorov-Smirnov distance; Macdonald (1971) and Woodward, Parr, Schucany, and Lindsay (1984) used the Cramér-von Mises distance; Lindsay (1994), Cutler and Cordero-Braña (1996), and Karlis and Xekalaki (1998) used the Hellinger distance; and Clarke (1989), Clarke and Heathcote (1994), and Scott (2001) used the squared $L_2$ distance. For a comprehensive review of the properties of minimum distance estimators for finite mixture models, see Titterington et al. (1985).

Within the class of minimum distance methods, a density-based minimum Hellinger distance estimation (MHDE) introduced by Beran (1977) has been shown to attain first-order
efficiency at the model density and simultaneously possess strong robustness properties. Beran (1977), Stather (1981), and Simpson (1987, 1989) provided several basic properties of the minimum Hellinger distance (MHD) method and illustrated its usefulness in statistical inference. Cutler and Cordero-Braña (1996) provided a MHD estimator in the context of finite mixture models for continuous data when the exact form of the component densities are unknown but are thought to be close to members of some parametric family. In addition, they showed that their estimator is efficient at the parametric model and robust under gross-error contaminations. Karlis and Xekalaki (1998) developed a MHD estimator for finite Poisson mixture models and also established that their MHD estimators are robust and efficient. Recently, Woo and Sriram (2006, 2007) developed a MHD based estimator of the number of mixture components, known as the *mixture complexity*. They showed that their estimator is strongly consistent for any parametric family of mixtures. Furthermore, they illustrated through extensive simulations that their estimator determines the true mixture complexity, whether or not the postulated mixture model is correctly specified.

The MHDE approach is an attractive alternative to ML estimation because the former yields estimators that are fully efficient at the parametric model and simultaneously robust against model misspecification. However, as noted by many authors including Basu, Harris, Hjort, and Jones (1998), Scott (1999, 2001), and Markatou (2000, 2001), the MHDE method involves some practical challenges such as the selection of an appropriate nonparametric kernel density estimator and optimal bandwidth that are intrinsic to kernel smoothing methods.

Motivated by the work of Rudemo (1982), Bowman (1984), and Terrell (1990), Scott (1998, 1999, 2001, and 2004) proposed an estimator for parametric modeling based on an integrated squared error criterion, called the $L_2E$, which does not require any nonparametric smoothing but has many computational advantages. Scott illustrated that the $L_2E$ is especially suited for parameter-rich models such as mixture models. He also showed that the $L_2E$ approach is computationally feasible and yields robust estimators like all other min-
imum distance methods. He pointed out that $L_2E$ is a special class of robust estimator like a median-based estimator as well as a particular example of M-estimator, both of which sacrifice some asymptotic efficiency but have substantial computational benefits in difficult estimation problems. Moreover, he demonstrated that the $L_2E$ estimates are superior to the MHD estimates under gross-error contamination, especially for smaller samples.

The $L_2E$ is a special case of a density-based power divergence measure of Basu et al. (1998), who developed a whole continuum of divergence estimators, called minimum density power divergence (MDPD) estimators, indexed by a tuning constant $\alpha$ that begin with the MLE ($\alpha = 0$) and interpolate to an $L_2E$ ($\alpha = 1$) and beyond, where $\alpha$ controls the trade-off between efficiency and robustness. Through empirical studies, Basu et al. (1998) showed that the efficiency of MDPD decreases as $\alpha$ increases and it is moderately lower for larger values of $\alpha$, including $\alpha = 1$. While the $L_2E$ incurs moderate loss of efficiency at the parametric model, within the family of density-based power divergence measures, the approach has distinct computational advantages, especially for the continuous case (Scott, 1999, 2001).

Ever since Pearson’s pioneering work, normal mixtures have provided the most popular model-based approach to analyze continuous data coming from populations consisting of a finite number of subclasses. However, in many circumstances the component normal distribution assumption is clearly inappropriate, especially when the data originate from distributions that are heavily skewed, heavy-tailed, or concentrated on nonnegative values. In such scenarios, inference based on the normality assumption can be misleading (Lin, Lee, and Yen, 2007); thus, it is natural to consider mixture models with non-normal components such as lognormal or gamma. A comprehensive review of applications of non-normal mixture models appears in Titterington et al. (1985, Chapter 2). More recently, Slud (1997) proposed a two-component exponential mixture model to test imperfect debugging in software reliability. Gruet, Philippe, and Robert (1999) also considered a mixture of exponential distributions to analyze data that are skewed and multi-modal. Wiper, Insua, and Ruggeri (2001) proposed a Bayesian density estimation method based on a gamma mixture model. Wijaya, Harada,
and Horton (2008) compared normal, lognormal, and gamma mixture models for analyzing the marginal distributions of expression values for individual genes over a broad range of normal tissues.

In a regression setting, unobserved heterogeneity occurs when the responses come from more than one subpopulation or when there exists extreme values in the data. Examples of data sets with extreme values are plentiful. For example, data sets such as the length of stay in hospital (Lu, Hui, and Lee, 2003), crash occurrence around and within intersections (Park and Lord, 2009), and baseball salaries (Khalili and Chen, 2007) all come from more than one subpopulation and contain extreme values. Once again, the finite mixture of regression model framework provides a natural way of modeling data that exhibit unobserved population heterogeneity in a regression setting. Examples of such data sets also arise in many applications, particularly in biology (Wang, Puterman, Cockburn, and Le, 1996), economics (Quandt, 1972), and genetics (Zhang and Merikangas, 2000).

The literature on finite mixture regression models is vast; see for example, Follmann and Lambert (1989), Wang et al. (1996), Wang and Puterman (1998), Aitkin (1999), Lu et al. (2003), Khalili and Chen (2007), and Park and Lord (2009). Wang et al. (1996) studied mixed Poisson regression models with covariate dependent rates and showed that the MLE based on the EM algorithm is efficient but not robust. Lu et al. (2003) proposed a MHDE for finite mixtures of Poisson regression models.

When covariates are present, Poisson regression models are widely used to analyze the effect of covariates on count data. However, observed count data analyzed under such models often exhibit overdispersion so that, conditional on covariates, the Poisson assumption of mean-variance equality is no longer valid. Consequently, many authors have studied the effects of overdispersion on inferences made under a Poisson model (for example, Cox, 1983; McCullagh and Nelder, 1989; Breslow, 1990; Cameron and Trivedi, 1990). Wang et al. (1996) regarded overdispersion as an intrinsic aspect of data which requires modeling. By using finite mixture models to explain overdispersion, they adopted the viewpoint that there are a finite
number of unobservable categories of observations which may be characterized by different values of regression coefficients. With this motivation, Wang et al. (1996) proposed a class of Poisson mixture models that relates covariates to component Poisson rates.

Finite mixtures of Poisson regression models and negative binomial regression models have been used in modeling overdispersed count data arising in behavioral health-care demand, utilization and cost (Mullahy, 1997; Deb and Trivedi, 1997, 2002; Deb and Holmes, 2000), health planning and management (Xiao, Lee, and Vemuri, 1999), disease mapping, genetics and risk assessment (Schlattmann and Böhning, 1993; Beyers and Shenton, 1999; Böhning, 1999; Lawson, Biggeri, Böhning, Lesaffre, Viel, and Bertollini, 1999), to name a few. Recently, Deb and Holmes (2000) and Deb and Burgess (2007) used finite mixtures of lognormal regression models and gamma regression models, respectively, to study the expenditure (cost) of behavioral healthcare.

The estimation literature for finite mixture of regression models reviewed above, with the exception of Lu et al. (2003), does not provide methodologies that are inherently robust against extreme values and model misspecifications. Chapter 2 focuses squarely on developing a robust estimation method that does not require nonparametric smoothing but has computational advantages. More specifically, in Chapter 2 we develop a robust estimation method for finite mixture of regression models for count response based on the $L_2$ distance between parametric conditional and true conditional mass functions, without imposing any specific distributional assumptions. Our proposed $L_2E$ estimator is shown to be consistent and asymptotically normal under certain regularity conditions. Under the same setting, we also develop a $L_2E$ based estimation of mixture complexity in finite mixture of regression models. Once again, the resulting $L_2E$ estimator of mixture complexity is shown to be strongly consistent, under certain regularity conditions. Monte Carlo simulations for finite mixture of Poisson regression models are carried out in order to assess the performance of our $L_2E$ and compare our results with those for MHDE and MLE methods. Additionally, we investigate the robustness of $L_2E$ when samples come from gross-error contaminated mix-
ture of Poisson regression models. Finally, we illustrate the use of our $L_2E$ method through analysis of two real data sets and compare the conclusions with existing analyses in the literature. We argue convincingly that our $L_2E$ provides a competitive alternative to other methods in the literature.

In Chapter 3, we revisit $L_2E$ estimation of mixture complexity in finite mixture models when covariates are not present. This problem was considered in a recent doctoral dissertation of Umashanger (2009). More specifically, we provide a more comprehensive $L_2E$ estimation procedure which can handle non-normal mixture models and cases where a small fraction of plausible extreme values are present. More specifically, we introduce a modified $L_2E$ approach with penalty and illustrate its use in the robust estimation of mixture complexity for normal mixture models with small mixture components. The modification, however, is not restricted to any specific distribution; it is useful more generally. Here, two real data sets are analyzed to illustrate the performance of the $L_2E$ of Umashanger (2009) and the modified $L_2E$.

1.2 References


Count Data,” Computational and Statistical Data Analysis, 51, 4379–4392.

ison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion,”


Chapter 2

$L_2 E$ Estimation for Finite Mixture of Regression Models with Applications

†

For count response, we propose a robust estimation method for finite mixture of regression models based on minimum integrated $L_2$ distance between parametric conditional and true conditional mass functions. The estimator, called $L_2 E$, is shown to be consistent and asymptotically normal. Monte Carlo simulations for finite mixture of Poisson regression models show that the $L_2 E$ is highly competitive to the maximum likelihood (ML) estimator and a better, yet viable, alternative to a minimum Hellinger distance (MHD) estimator, when the model is correctly specified. When samples come from gross-error contaminated mixture of Poisson regression models, it is shown that the $L_2 E$ is considerably more robust than the MHD and ML estimators. We also develop a $L_2 E$ based estimation of mixture complexity in finite mixture of regression models. Once again, the resulting $L_2 E$ estimator of mixture complexity is shown to be strongly consistent, under certain regularity conditions. For the poorly-separated model with a small fraction of extreme values, we propose a modified $L_2 E$ criterion, called $L_2 E_p$, with a penalty function. It is also shown that the $L_2 E_p$ estimator is consistent and asymptotically normal for the case of two-component mixture regression models. Additionally, we also show that the $L_2 E_p$ estimator of mixture complexity is consistent. For real data sets on Maternity Length of Hospital Stay and Toronto Motor Vehicle Crashes each containing a fraction of extreme values, we show that a two-component Poisson mixture regression model and a two-component negative binomial mixture regression model, respectively, based on the $L_2 E$ criterion provide a better overall fit in terms of capturing the heterogeneity in the count responses and providing interpretability of component profiles via their respective covariates. We believe that our $L_2 E$ criterion developed for the finite mixture regression framework will be useful for analyzing count data sets arising in other contexts as well; particularly in health economics.

Key words and Phrases: Asymptotic normality; Consistency; Extreme values; Finite mixture of regression models; $L_2$ distance; $L_2 E$ functional; Penalty function; Robustness.
2.1 Introduction

Minimum distance estimation methods are attractive alternatives to the maximum likelihood estimation (MLE) method because they provide estimates that are intrinsically robust against extreme values, but possibly incur some loss of efficiency at the correct parametric model. For finite mixtures, the minimum Hellinger distance estimation (MHDE) method (Beran, 1977) yields estimators that are fully efficient at the parametric model and simultaneously robust against gross-error contaminations; see e.g., Cutler and Cordero-Braña (1996), Karlis and Xekalaki (1998, 1999), Woo and Sriram (2006, 2007) and references therein. For finite mixture of Poisson regression models, Lu, Hui, and Lee (2003) developed a MHDE approach, which minimizes the Hellinger distance between the parametric unconditional (marginal) mass function and the empirical mass function of the count response. They showed that their approach provides a viable alternative to the MLE method considered in Wang, Puterman, Cockburn, and Le (1996).

A more natural approach in a regression setting, however, is to develop minimum distance estimation methods based on conditional density or mass functions. In this chapter, we develop a robust estimation method for finite mixture of regression models for count response based on the \(L_2\) distance between parametric conditional and true conditional mass functions, without imposing any specific distributional assumptions. Extension of our \(L_2\) estimation method to the continuous response case with covariates is straightforward, but a full treatment of it is deferred to future work.

Whereas the MHDE approach leads to efficiency and robustness in general, as noted by many authors (e.g., Basu, Harris, Hjort, and Jones, 1998; Scott, 1999, 2001; Markatou, 2000, 2001), the method involves practical challenges such as requiring a nonparametric estimator of the true conditional density or mass functions and selection of appropriate bandwidths. Lu et al. (2003) attempted to overcome this by adopting an unconditional MHDE approach for finite mixture of Poisson regression models, but the consistency of their MHD estimates
depends on whether or not their unconditional marginal mass function is identifiable; see Remarks 1 and 2 in Lu et al. (2003) for a specific discussion.

In a series of articles, Scott (1998, 1999, 2001, and 2004) studied an estimator for parametric modeling based on an integrated squared error criterion, called the $L_2E$, which does not require any nonparametric smoothing but has many computational advantages. In fact, the $L_2E$ is a special case of a density-based power divergence measure of Basu et al. (1998), who developed a whole continuum of divergence estimators indexed by a tuning constant $\alpha$ that begins with the MLE ($\alpha = 0$) and interpolates to an $L_2E$ ($\alpha = 1$) and beyond, where $\alpha$ controls the trade-off between efficiency and robustness. While the $L_2E$ incurs moderate loss of efficiency at the parametric model, within the family of density-based power divergence measures, the approach has distinct computational advantages, especially when the response is continuous (Scott, 1999, 2001).

The rest of the chapter is organized as follows. Section 2.2 gives the definition of a finite mixture regression model. In Section 2.3, we introduce the $L_2E$ methodology for finite mixture of regression models. Section 2.4 proposes an estimator of mixture complexity based on an $L_2$ model selection criterion. A theorem establishing the consistency and asymptotic normality of $L_2E$ estimator is stated in Section 2.5 but proved in the Appendix. Computational details are given in Section 2.6. Monte Carlo simulations for finite mixture of Poisson regression models given in Section 2.7.1 show that our $L_2E$ is highly competitive to the MLE and a better, yet viable, alternative to the MHDE of Lu et al. (2003), when the model is correctly specified. When there is gross-error contamination in mixture of Poisson regression models, it is shown in Section 2.7.4 that the $L_2E$ is considerably more robust than the MHDE and MLE. To further illustrate the robustness of $L_2E$ against extreme values, we apply our methodology to the following two real count data sets, each of which contains a fraction of extreme values: (i) 1998/1999 Maternity Length of Hospital Stay data from a Western Australia hospital (Lu et al., 2003) and (ii) 1995 Motor Vehicle Crashes data at urban four-legged signalized intersections in Toronto, Canada (Park and Lord, 2009); see
Sections 2.8.1 and 2.8.2, respectively, for details. We make a compelling argument that a two-component Poisson mixture regression model for the data in (i) and a two-component negative binomial mixture regression model for the data in (ii) combined with our \( L_2E \) approach provide a better overall fit, not only in terms of capturing the heterogeneity in the count response, but also in terms of providing interpretability of component profiles via their respective covariates. We begin by introducing some basic notations which will be used throughout this chapter.

2.2 Finite Mixture of Regression Models

Suppose \( \{(Y_i, X_i), i = 1, ..., n\} \) denotes a set of independent random pairs where, for each \( i \), \( Y_i \) is a count response variable and \( X_i = (1, X_{i1}, \ldots, X_{ip})^T \) is an \( \mathbb{R}^{(p+1)} \)-valued vector of covariates having density \( f_X \). Let \((Y, X)\) denote a generic member of the sample and \( G \) denote the class of all joint densities \( g(y, x) = f(y|x)f_X(x) \) of \((Y, X)\). From now on we will not use the subscript \((Y, X)\) for members of \( G \).

Let \( f_0(y|x) \) denote the true conditional probability mass function (p.m.f.) of \( Y \) given \( X = x \) and let \( g_0(y, x) = f_0(y|x)f_X(x) \) denote the corresponding joint density. Suppose we postulate that the conditional p.m.f. of \( Y \) given \( X = x \) belongs to a specified parametric family \( F_m = \{f_{\theta_m}(y|x) : \theta_m \in \Theta_m \subseteq \mathbb{R}^k\} \) for each fixed positive integer \( m < \infty \) and for some \( k \)-dimensional space \( \mathbb{R}^k \), such that \( f_{\theta_m} \) can be represented as a finite mixture regression model of the form

\[
f_{\theta_m}(y_i|x_i) = \sum_{j=1}^{m} \pi_j f_j(y_i|\lambda_{ij}), \quad y_i \in \mathcal{Y} = \{0, 1, 2, \ldots\},
\]

(2.1)

where \( f_j(y_i|\lambda_{ij}) \) is a parametric p.m.f. with a link function \( \lambda_{ij} = h(x_i^T \beta_j) \), and mixing proportions \( \pi_j \geq 0, \sum_{j=1}^{m} \pi_j = 1 \), for \( j = 1, \ldots, m \). Here, \( \theta_m = (\pi, \beta)^T \in \mathbb{R}^k \) with \( \pi = (\pi_1, \ldots, \pi_{m-1}), \beta = (\beta_1^T, \ldots, \beta_m^T) \), and \( k = (p+2)m-1 \). For finite mixture of Poisson regression model, \( f_j(y_i|\lambda_{ij}) = e^{-\lambda_{ij}} \lambda_{ij}^{y_i}/y_i! \) with mean \( \lambda_{ij} = \exp(x_i^T \beta_j) \); see Section 2.8.2 for a negative binomial mixture regression model.
2.3 \(L_2E\) Criterion

Let \(G\) denote the class of all joint distributions \(G_{\text{Y,X}}\) with corresponding joint density \(g_{\text{Y,X}}\) belonging to \(G\) defined in Section 2.2. Following Basu et al. (1998) [also see Scott (2001)], define an \(x\)-conditional version of the squared \(L_2\) distance between the postulated conditional p.m.f. \(f_{\theta_m}\) in (2.1) and the true conditional p.m.f. \(f_0\) as

\[
l_2(f_{\theta_m}(\cdot|x), f_0(\cdot|x)) = \sum_{y=0}^{\infty} [f_{\theta_m}(y|x) - f_0(y|x)]^2.
\] (2.2)

Now, define the overall \(L_2\) distance between the postulated joint density \(g_{\theta_m}(y,x) = f_{\theta_m}(y|x)f_X(x)\) and the true joint density \(g_0(y,x)\) as

\[
L_2(g_{\theta_m}, g_0) = \int l_2(f_{\theta_m}(\cdot|x), f_0(\cdot|x))f_X(x)dx = \int \sum_{y=0}^{\infty} [f_{\theta_m}(y|x) - f_0(y|x)]^2f_X(x)dx. \tag{2.3}
\]

For each fixed \(m\), we define an \(L_2E\) functional \(T_{m}^{L_2E}\) on \(G\) by the requirement that for every \(g \in G\),

\[
T_{m}^{L_2E}(g) = \{ \theta_m \in \Theta_m : L_2(g_{\theta_m}, g) = \min_{\delta_m \in \Theta_m} L_2(g_{\delta_m}, g) \}. \tag{2.4}
\]

Expand the right side of (2.3) and note that the last term \(\int \sum_{y=0}^{\infty} f_0^2(y|x)f_X(x)dx\) does not depend on \(\theta_m\); therefore, the \(L_2E\) functional is

\[
T_{m}^{L_2E}(g) = \arg \min_{\theta_m} \left[ \int \sum_{y=0}^{\infty} f_{\theta_m}^2(y|x)f_X(x)dx - 2 \int \sum_{y=0}^{\infty} f_{\theta_m}(y|x)f_0(y|x)f_X(x)dx \right]. \tag{2.5}
\]

In (2.5), note that \(\int \sum_{y=0}^{\infty} f_{\theta_m}(y|x)f_0(y|x)f_X(x)dx = E f_{\theta_m}(y|X)\), which can be estimated by \(n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i|X_i)\) and the first term, \(\int \sum_{y=0}^{\infty} f_{\theta_m}^2(y|x)f_X(x)dx\), can be estimated by \(n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y|X_i)\) (see Basu et al., 1997; Scott, 2001). Then, we define the \(L_2E\) estimator of \(\theta_m\) as

\[
\hat{\theta}_{m}^{L_2E} = \arg \min_{\theta_m} \left[ n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y|X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i|X_i) \right]. \tag{2.6}
\]

As mentioned in Section 2.1, the most natural approach to minimum distance estimation in a (count) regression setting is to minimize the distance between the conditional densities defined in (2.3). Instead, Lu et al. (2003) developed an MHD estimator of \(\theta_m\) by minimizing
the distance between the estimated unconditional marginal densities. More specifically, their MHDE is defined as
\[
\hat{\theta}_{n,m}^{\text{MHDE}} = \arg\min_{\theta_m} \left[ \sum_{y=0}^{\infty} \left( \frac{f_1^{1/2}(y)}{n} - \frac{f_{\theta_m,n}^{1/2}(y)}{n} \right)^2 \right],
\]
where \( f_n(y) = N_y/n \) with \( N_y \) the frequency of \( y \) among \( y_1, \ldots, y_n \), and \( f_{\theta_m,n}(y) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_j f_j(y_j | \lambda_{ij}) \) with the Poisson probability mass function \( f_j \) and \( \lambda_{ij} = \exp(x_i^T \beta_j) \).

Lu et al. (2003) themselves remark that “It is not obvious that this approach estimates the parameters corresponding to the conditional model.” Also, note that their MHDE algorithm can only be applied to Poisson mixture regression models. Furthermore, their method cannot be directly extended to the continuous response case.

Unlike Lu et al. (2003)’s MHDE approach for Poisson mixture regression models, clearly our \( L_2E \) is based on minimizing distance between conditional mass functions and it is not (component) distribution-specific. Also, our approach can be extended easily to the continuous response case by replacing \( \sum_{y=0}^{\infty} \) in (2.2) by \( \int_{-\infty}^{\infty} \) and \( \sum_{y=0}^{\infty} f_{\theta_m}^2(y | X_i) \) in (2.6) by \( \int_{-\infty}^{\infty} f_{\theta_m}^2(y | X_i) \) \( dy \), which has a closed form expression for normal (Scott, 2001), gamma, and lognormal mixtures.

2.4 \( L_2E \) Estimator of Mixture Complexity

In Sections 2.2 and 2.3, we assumed that the number of mixture components \( m \), the so called mixture complexity, in (2.1) is known and developed an \( L_2E \) method to estimate only the parameter \( \theta_m \). In many practical applications, however, the mixture complexity may be unknown. In this section, we propose an estimator of the mixture complexity based on \( L_2E \).

For a parametric family \( F_m \) defined in Section 2.2, the class \( F_m \subseteq F_{m+1} \) for all \( m \) and we denote \( F = \bigcup_{m=1}^{\infty} F_m \). For an unknown conditional p.m.f. \( f_0 \), define the index of the economical representation of \( f_0 \), relative to the family of mixtures \( F_m \), as
\[
m_0 = m(f_0) = \min\{m : f_0 \in F_m\}. \quad (2.7)
\]
If indeed $f_0$ is a finite mixture then $m_0$ is finite and denotes the true mixture complexity; otherwise $m_0 = \infty$. Note that $m_0$ represents the most parsimonious mixture model representation for $f_0$. We also assume that, for an unknown joint density $g_0$

$$m_0 = m(f_0) = m(g_0) = \min\{m : g_0 \in \mathcal{G}_m\}, \quad (2.8)$$

with $\mathcal{G}_m \subseteq \mathcal{G}_{m+1}$ for all $m$ and $\mathcal{G} = \bigcup_{m=1}^{\infty} \mathcal{G}_m$, where a semi-parametric family $\mathcal{G}_m = \{g_{\theta_m}(y, x) = f_{\theta_m}(y|x)f_X(x) : \theta_m \in \Theta_m \subseteq \mathbb{R}^k\}$. Thus, our goal is to find a semi-parametric conditional density estimator of the form

$$\hat{f}_n^*(y_i|x_i) = \sum_{j=1}^{\hat{m}_n} \hat{\pi}_j f_j(y_i|\hat{\lambda}_{ij}), \quad (2.9)$$

with the property that $\hat{m}_n \to m_0$ almost surely (a.s.) as $n \to \infty$. Accordingly, if $f_0 \in \mathcal{F}_m$ for some $m$, then $\hat{f}_n^* \to f_0$. If $f_0 \notin \mathcal{F}_m$ for any $m$, then $\hat{m}_n \to \infty$ a.s.; nevertheless, $\hat{f}_n^* \to f_0$.

In order to develop a robust estimation procedure of $m_0$, we proceed as in Woo and Sriram (2006, 2007) and Umashanger and Sriram (2009), and introduce a model selection criterion based on $L_2(g_{\hat{\theta}_{n,m}^L,E}, \hat{g}_n)$ defined by

$$LIC = L_2(g_{\hat{\theta}_{n,m}^L,E}, \hat{g}_n) + n^{-1} \log q(m),$$

where $\hat{g}_n$ is the empirical distribution of $\{(Y_i, X_i), i = 1, \ldots, n\}$ and $q(m)$ is a penalizing function depending on $m$. In the AIC/BIC type of model selection criteria, generally the penalty term is a function of the number of parameters in the model. However, for mixture complexity estimation, Poland and Shachter (1994) suggested using a logarithmic function of the number of mixture components $m$ in order to create a slowly increasing penalty function. In fact, they suggested using $q(m) = am^\kappa$ (see Poland and Shachter (1994), Sections 4 and 5) for an appropriate choice of $a$ and $\kappa$ to be determined based on sample size in a data set. Motivated by this, we use such a penalty function and this works well in our simulation studies and data analysis, as also noted in Umashanger and Sriram (2009).

Here, the value of $m$ yielding the minimum $LIC$ specifies the best model. Since $\mathcal{G}_m \subseteq \mathcal{G}_{m+1}$, we have $L_2(g_{\hat{\theta}_{n,m}^L,E}, \hat{g}_n) \geq L_2(g_{\hat{\theta}_{n,m+1}^L,E}, \hat{g}_n)$. Therefore, we penalize the first term in $LIC$
with a slowly increasing function of $m$. A simple heuristic to search for the best model from a sequence of nested models is to try successive models, starting with the smallest, and stop with model $m$ when its LIC value is lesser than that for model $(m+1)$. That is, this heuristic stops when

$$L_2(g_{\theta_{n,m}^{L_2E}}^\ast, \hat{g}_n) - L_2(g_{\theta_{n,m+1}^{L_2E}}^\ast, \hat{g}_n) \leq n^{-1} \log[q(m + 1)/q(m)].$$

Motivated by Poland and Shachter (1994), setting $\zeta_{n,m} = n^{-1} \log[q(m + 1)/q(m)] = n^{-1} \log[(m + 1)/m]$ leads to the following estimator of $m_0$ defined by (See Sections 2.7.4 and 2.8.1 for the choice of $\kappa$)

$$\hat{m}_{n}^{L_2E} = \min\{m : L_2(g_{\theta_{n,m}^{L_2E}}^\ast, \hat{g}_n) \leq L_2(g_{\theta_{n,m+1}^{L_2E}}^\ast, \hat{g}_n) + \zeta_{n,m}\}, \quad (2.10)$$

where $\{\zeta_{n,j} ; j \geq 1\}$ are positive sequences of threshold values chosen in such a way they tend to zero as $n \to \infty$.

### 2.5 Theoretical Properties of $L_2E$

In this section, we first state a Proposition which establishes basic properties concerning the $L_2E$ functional $T_{m}^{L_2E}(g)$ defined in Section 2.3. In addition, we state a set of regularity conditions and a theorem to establish the consistency and asymptotic normality of $\hat{\theta}_{n,m}^{L_2E}$ defined in Section 2.3. Finally, a consistency Theorem for $\hat{m}_{n}^{L_2E}$ defined in Section 2.4 is stated here. The proofs of the Proposition and Theorems are given in Appendices 2.10.1, 2.10.2, and 2.10.3.

Using the notation in Sections 2.2 and 2.3, let $\tilde{G} \subset G$ denote the sub-class of joint distributions for which the following condition holds: For each $m$, there is a compact set $C_m \subseteq \Theta_m$ such that for every $g \in \tilde{G}$,

$$\inf_{\delta_m \in \Theta_m - C_m} L_2(g_{\delta_m}, g) > L_2(g_{\theta_m}, g),$$

for some $\theta_m^* \in C_m$, where $g_{\delta_m}$ and $g_{\theta_m}$ are members of the class $G_m = \{g_{\theta_m}(y, x) = f_{\theta_m}(y|x)f_X(x) : \theta_m \in \Theta_m \subseteq \mathbb{R}^k\}$. If $\Theta_m$ is compact, then $C_m = \Theta_m$. We will assume
that \( g_{\theta_m} \) is continuous in \( \theta_m \), the class \( \mathcal{G}_m \) is identifiable, and \( G_{\theta_m} \) denotes the distribution corresponding to \( g_{\theta_m} \).

**Proposition.** The following hold for the functional \( T_{L^2E}^m(g) \) defined in (2.4) for \( g \in \mathcal{G} \):

(i) \( T_{L^2E}^m(g) \) satisfying (2.4) exists.

(ii) If \( T_{L^2E}^m(g) \) is unique, then the functional \( T_{L^2E}^m \) is continuous at \( g \) in \( L_2 \) defined in (2.3).

(iii) \( T_{L^2E}^m(g_{\theta_m}) = \theta_m \) uniquely for every \( \theta_m \in \Theta_m \).

We now state the following regularity conditions required to establish the consistency and asymptotic normality of the \( L_2E \) estimator defined in (2.6).

(C.1) The distributions \( G_{\theta_m} \) and \( G \) have common support \( \mathcal{Y} \), which does not depend on the parameter \( \theta_m \).

(C.2) The \( \hat{\theta}_{L^2E}^m \) lies in an open subset \( \mathfrak{S} \) of the parameter space \( \Theta_m \). For almost all \( y \in \mathcal{Y} \), \( f_{\theta_m}(y|x) \) is three times differentiable with respect to \( \theta_m \), and the third partial derivative is continuous with respect to \( \theta_m \).

(C.3) The integral \( \int f_{\theta_m}^2(y|x)f_X(x)dx \) can be differentiated three times with respect to \( \theta_m \), and the derivative can be taken under the integral sign.

(C.4) The matrix \( J = J(\theta_m) \), defined by

\[
J(\theta_m) = \int \left[ \sum_{y=0}^{\infty} u_{\theta_m}(y|x)u_{\theta_m}^T(y|x)f_{\theta_m}^2(y|x) \right] f_X(x)dx \\
+ \int \left[ \sum_{y=0}^{\infty} \{ i_{\theta_m}(y|x) - u_{\theta_m}(y|x)u_{\theta_m}^T(y|x) \} \{ f_0(y|x) - f_{\theta_m}(y|x) \} f_{\theta_m}(y|x) \right] f_X(x)dx,
\]

is positive definite almost everywhere in \( \mathfrak{S} \), where the model score function \( u_{\theta_m}(y|x) = \partial \log f_{\theta_m}(y|x)/\partial \theta_m \) and the model information function \( i_{\theta_m}(y|x) = -\partial^2 \log f_{\theta_m}(y|x)/\partial \theta_m \partial \theta_m^T \).
and the matrix $K = K(\theta_m)$ is defined as

$$K(\theta_m) = \int \left[ \sum_{y=0}^{\infty} u_{\theta_m}(y|x)u_{\theta_m}^T(y|x)f_0(y|x)f_{\theta_m}^2(y|x) \right] f_X(x)dx - \int \xi(x)\xi^T(x)f_X(x)dx,$$

with $\xi(x) = \sum_{y=0}^{\infty} u_{\theta_m}(y|x)f_0(y|x)f_{\theta_m}(y|x)$.

(C.5) For $V_{\theta_m}(Y, X) = \sum_{y=0}^{\infty} f_{\theta_m}^2(y|x) - 2f_{\theta_m}(Y|X)$, there exist functions $M_{rst}(Y, X)$ such that $|\partial^3 V_{\theta_m}(Y, X)/\partial \theta_{m,r}\partial \theta_{m,s}\partial \theta_{m,t}| \leq M_{rst}(Y, X)$ and $E_G[M_{rst}(Y, X)] < \infty$ for all $r, s,$ and $t$, where the expected value is with respect to $G$. Here, $\theta_{m,r}, \theta_{m,s}$ and $\theta_{m,t}$ are the $r$-th, $s$-th and $t$-th components of $\theta_m$.

**Theorem 1.** For each integer $m \geq 1$, under the conditions (C.1) to (C.5),

(i) $\tilde{\theta}_{n,m}^{L_2E} \overset{P}{\longrightarrow} \theta_m$, and

(ii) $\sqrt{n}(\tilde{\theta}_{n,m}^{L_2E} - \theta_m) \overset{D}{\longrightarrow} N(0, J^{-1}KJ^{-1})$,

as $n \to \infty$, where $\overset{P}{\longrightarrow}$ and $\overset{D}{\longrightarrow}$ denote convergence in probability and distribution, respectively.

**Remark 1.** Suppose that the true conditional p.m.f $f_0$ is equal to $f_{\theta_m}$, where $\theta_m$ is the true value of the parameter. Then $J$, $K$, and $\xi$ in (C.4) reduce to

$$J = J(\theta_m) = \int \left[ \sum_{y=0}^{\infty} u_{\theta_m}(y|x)u_{\theta_m}^T(y|x)f_{\theta_m}^2(y|x) \right] f_X(x)dx,$$

and

$$K = K(\theta_m) = \int \left[ \sum_{y=0}^{\infty} u_{\theta_m}(y|x)u_{\theta_m}^T(y|x)f_{\theta_m}^3(y|x) \right] f_X(x)dx - \int \xi(x)\xi^T(x)f_X(x)dx,$$

with $\xi(x) = \sum_{y=0}^{\infty} u_{\theta_m}(y|x)f_{\theta_m}^2(y|x)$. Thus, given a parametric model such as Poisson or negative binomial mixture regressions, the asymptotic covariance matrix of $\tilde{\theta}_{n,m}^{L_2E}$ (See part (ii) of Theorem 1) can be estimated by

$$\hat{J} = \hat{J}(\tilde{\theta}_m) = n^{-1}\sum_{i=1}^{n} \sum_{y=0}^{\infty} u_{\tilde{\theta}_m}(y|X_i)u_{\tilde{\theta}_m}^T(y|X_i)f_{\tilde{\theta}_m}^2(y|X_i), \quad (2.12)$$
and
\[
\hat{K} = \hat{K}(\hat{\theta}_m) = n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} u_{\hat{\theta}_m}^T(y|X_i) u_{\hat{\theta}_m}(y|X_i) f_{\hat{\theta}_m}^3(y|X_i) - n^{-1} \sum_{i=1}^{n} \left[ \sum_{y=0}^{\infty} u_{\hat{\theta}_m}(y|X_i) f_{\hat{\theta}_m}^2(y|X_i) \right] \left[ \sum_{y=0}^{\infty} u_{\hat{\theta}_m}(y|X_i) f_{\hat{\theta}_m}^2(y|X_i) \right]^T. \tag{2.13}
\]

Note that in the data analyses described in Sections 2.8.1 and 2.8.2, we compute the standard error of the \(L_2E\) estimates of \(\theta_m\) using the above formulas.

**Theorem 2.** Suppose all the assumptions of the Proposition stated above are satisfied.

If \(g_0\) is a finite mixture with mixture complexity \(m_0 < \infty\), then for any sequence \(\zeta_{n,m} \to 0\)
\[\hat{m}_n^{L_2E} \to m_0 \quad \text{a.s.}\]
as \(n \to \infty\), where \(\hat{m}_n^{L_2E}\) and \(m_0\) are as defined in (2.10) and (2.8), respectively. If \(g_0\) is not a finite mixture, then \(\hat{m}_n^{L_2E} \to \infty \) a.s.

### 2.6 Computational Details

Here, we give some computational details about \(\hat{\theta}_{n,m}^{L_2E}\) in (2.6) and \(\hat{m}_n^{L_2E}\) in (2.10). The \(L_2E\) can be performed using any standard nonlinear optimization method; we used a quasi-Newton algorithm called the Broyden-Fletcher-Goldfarb-Shanno (BFGS) for our computations. We used R software for Linux on an AMD Opteron dual-processor workstation, invoking the *optim* routine to calculate the \(L_2E\) estimates. As for computation time, it took on the average about 17 seconds CPU time to compute one \(L_2E\) estimate of \(\theta_2\) based on a simulated data of size \(n = 500\) from a two-component Poisson mixture regression model considered in Table 2.1 in Section 2.7.1, for instance. Also, the number of iterations required for convergence was usually about between 50 and 90.

There are distinct computational advantages that make our \(L_2E\) a more attractive, yet viable, alternative to the MHDE method of Lu et al. (2003). First, computation of the \(L_2E\) objective function in (2.6) is much simpler than that of the MHDE, which is an extension of
the HELMIX algorithm of Karlis and Xekalaki (1998) to finite mixture of Poisson regression models. Secondly, compared to our $L_2E$, the HELMIX algorithm is computationally slower since the HELMIX is an EM type algorithm and highly sensitive to the choice of initial values (Karlis and Xekalaki, 1998). Thirdly, as mentioned at the end of Section 2.3, our $L_2E$ can be easily extended to the continuous response case, while retaining its computational conveniences for some known mixture models.

As for estimation of mixture complexity, our estimator of $m_0$ defined in (2.10) can be rewritten as

$$\hat{m}_n^{L_2E} = \min \{m : L_2(\hat{\theta}_{n,m}) \leq L_2(\hat{\theta}_{n,m+1}) + \zeta_{n,m}\}, \quad (2.14)$$

where

$$L_2(\theta_m) = \left[ n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y|X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i|X_i) \right] \quad (2.15)$$

(see equation (2.6)). Computation of an estimate of mixture complexity begins with a single density model $f_{\theta_1}$ and proceeds iteratively through the following four steps.

- **Step 1:** Compute an estimate $\hat{\theta}_{n,1}^{L_2E}$ which minimizes $L_2(\theta_1)$, and compute $L_2(\hat{\theta}_{n,1}^{L_2E})$.

- **Step 2:** Now add another component yielding a mixture with two components ($m = 2$) and compute $L_2(\hat{\theta}_{n,2}^{L_2E})$ as in **Step 1**.

- **Step 3:** Calculate the difference $L_2(\hat{\theta}_{n,1}^{L_2E}) - L_2(\hat{\theta}_{n,2}^{L_2E})$ and compare it with the threshold value $\zeta_{n,m}$ in (2.10). If $\zeta_{n,m}$ is greater than this difference, then stop and report $\hat{m}_n^{L_2E} = 1$. Otherwise go to **Step 4**.

- **Step 4:** Repeat **Step 2** and **Step 3** by adding one more component to the previous mixture and comparing the difference until the first value $m = m^*$ for which the difference $L_2(\hat{\theta}_{n,m^*}^{L_2E}) - L_2(\hat{\theta}_{n,m^*+1}^{L_2E})$ falls below the threshold value $\zeta_{n,m^*}$. At this point, the procedure terminates and declares $m^*$ as an estimate of the mixture complexity. Note that, at this stage, our procedure automatically provides the best parametric fit determined by $\hat{\theta}_{n,m^*}^{L_2E}$.
2.7 Simulation Studies

In this section, we carry out two different simulation studies in order to assess the performance of our estimator $\hat{\theta}_{n,m}^{L_2E}$. In the first study, data come from four different two-component Poisson mixture regression models. The second study examines the robustness of our estimator under gross-error contaminations; that is, data are generated from four different three-component Poisson mixture regressions, where the third component represents three different levels of contamination. These are described in detail in the following two subsections, respectively. In the last subsection, we numerically assess the performance of $\hat{m}_{n}^{L_2E}$ for two different two-component Poisson mixture regression models.

2.7.1 Comparison with MLE and MHDE

In order to assess the performance of $\hat{\theta}_{n,m}^{L_2E}$, we generated data from the two-component Poisson mixture regression model

$$f_{\theta_2}(y_i|\mathbf{x}_i) = \pi_1 f_1(y_i|\lambda_{i1}) + (1 - \pi_1) f_2(y_i|\lambda_{i2}),$$

(2.16)

where $\lambda_{i1} = \exp(\beta_{10} + \beta_{11}x_i)$, $\lambda_{i2} = \exp(\beta_{20} + \beta_{21}x_i)$, for $i = 1, \ldots, n$ and the covariate $\mathbf{x}$ is a uniform $[0,1]$ random variable. As in Lu et al. (2003), we considered four different values for the parameter vector $\theta_2 = (\pi_1, \beta_{10}, \beta_{11}, \beta_{20}, \beta_{21})$ given by

- **Case 1**: $(0.5, 0.6, 1.0, 2.5, 1.5);$  
  **Case 2**: $(0.95, 0.6, 1.0, 2.5, 1.5);$ 
- **Case 3**: $(0.5, 0.6, 1.0, 0.8, 1.2);$  
  **Case 4**: $(0.95, 0.6, 1.0, 0.8, 1.2).$

These four cases cover a variety of scenarios, in the first two of which the component regression parameters are well-separated, whereas in the last two cases they are poorly-separated. Also, in Cases 2 and 4 the mixing proportion of the second component is much smaller compared to the first; that is, $\pi_2 = 0.05$. 

For each of the four cases above, we implemented our $L_2E$ computational algorithm described in Section 2.6 for sample sizes $n = 200$ and 500. For each sample size, we performed 200 Monte Carlo replications, each yielding an estimate of $\theta_2$. In order to avoid any sampling bias and ensure fair comparison between our estimate and the MLE and MHDE, we computed each of the three estimates (MLE, MHDE and $L_2E$) using the same generated data set (in each replication) with identical initial values. Realizing that the MHDE and MLE are very sensitive to the choice of initial values, we used the actual values in Cases 1 to 4 above as initial values for each replication; this adds another layer of fairness in comparison. Incidentally, we used the publicly available R code to compute the MLE but used the published code of Lu et al. (2003) to compute the MHDE, which is available from the Biometrics website [http://www.biometrics.tibs.org/datasets/020128.html](http://www.biometrics.tibs.org/datasets/020128.html).

We report in Table 2.1 below the relative efficiency (R.E.) of the MHDE compared to the MLE; $L_2E$ compared to the MLE; and $L_2E$ compared to the MHDE, for each component of vector $\theta_2$ defined above. Here, R.E. of the $L_2E$ (or the MHDE) is defined as the mean squared error (MSE) of the MLE divided by that of the $L_2E$ (or the MHDE). For ease of comparison, in Table 2.1 R.E. > 1 are identified in bold.

For the well-separated, correctly specified case considered in Case 1, Table 2.1 shows that, as expected, the MLE is more efficient than the $L_2E$ and the MHDE for all sample sizes. Nevertheless, the relative efficiencies of $L_2E$ and MHDE increase for some parameters as sample size increases. It is important to note from the column $L_2E$ vs MHDE that $L_2E$ is significantly more efficient than the MHDE for all parameters and sample sizes.

For the well-separated, correctly specified case considered in Case 2, the MLE once again is more efficient than the $L_2E$ and MHDE for all parameters. Especially, note that for the second component parameters, the MHDE performs poorly against the MLE and its performance worsens even when $n$ increases. On the contrary, the performance of $L_2E$ slightly improves for the second component parameters as $n$ increases. Once again, the $L_2E$ is much
more efficient than the MHDE for every parameter and sample size and its efficiency improves as \( n \) increases.

Table 2.1: Comparison between MLE, MHDE, and \( L_2E \) for Cases 1-4: Relative Efficiencies

<table>
<thead>
<tr>
<th>Case 1</th>
<th>( x \sim \text{U}[0,1] ) parameters</th>
<th>Sample Size = 200</th>
<th></th>
<th>Sample Size = 500</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE vs ( L_2E ) vs MHDE ( L_2E )</td>
<td>MLE vs MHDE ( L_2E )</td>
<td>MLE vs MHDE ( L_2E )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \pi_1 = 0.5 )</td>
<td>( \beta_{10} = 0.6 )</td>
<td>( \beta_{11} = 1.0 )</td>
<td>( \beta_{20} = 2.5 )</td>
<td>( \beta_{21} = 1.5 )</td>
<td></td>
</tr>
<tr>
<td>0.326772</td>
<td>0.191262</td>
<td>0.121576</td>
<td>0.164791</td>
<td>0.129507</td>
<td>0.616576</td>
</tr>
<tr>
<td>0.814837</td>
<td>0.625721</td>
<td>0.617223</td>
<td>0.637569</td>
<td>0.63673</td>
<td>0.743595</td>
</tr>
<tr>
<td>( \beta_{10} = 0.6 )</td>
<td>( \beta_{11} = 1.0 )</td>
<td>( \beta_{20} = 2.5 )</td>
<td>( \beta_{21} = 1.5 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.326772</td>
<td>0.191262</td>
<td>0.121576</td>
<td>0.164791</td>
<td>0.129507</td>
<td>0.616576</td>
</tr>
<tr>
<td>0.814837</td>
<td>0.625721</td>
<td>0.617223</td>
<td>0.637569</td>
<td>0.63673</td>
<td>0.743595</td>
</tr>
</tbody>
</table>

For the poorly-separated scenario considered in Case 3 with \( n = 200 \) and 500, Table 2.1 shows that the \( L_2E \) is more efficient compared to the MLE for every parameter. While the MHDE is also more efficient than the MLE for the mixing proportion and second component parameters, its performance is significantly poor compared to the MLE for the first component parameters. Note that for all sample sizes, the \( L_2E \) is more efficient than the MHDE for all parameters, except the mixing proportion. Nevertheless, as noted earlier, the
$L_2E$ is more efficient than the MLE in this case for the mixing proportion. Interestingly, however, the MHDE performs significantly poorly against the $L_2E$ for the first component parameters.

For the poorly-separated scenario with a small (second component) mixing proportion considered in Case 4, $L_2E$ is more efficient than the MHDE for all parameters and sample sizes. However, when $n = 200$, the $L_2E$ is not as efficient as the MLE for the mixing proportion and first component parameters, but it is more efficient than the MLE for the second component parameters. When $n = 500$, the $L_2E$ is more efficient than the MLE for the mixing proportion but it still remains inefficient compared to the MLE for the first component parameters.

In summary, $L_2E$ is more efficient than the MHDE for all sample sizes and all the four cases, except for the mixing proportion in Case 3. For the two difficult cases (Cases 3 and 4), the $L_2E$ is more efficient than the MLE for Case 3, but it is only more efficient than MLE for second component parameters (when $n = 200$) and the mixing proportion in Case 4 (when $n = 500$). On the contrary, the MHDE performs poorly against the MLE for Case 4, but is more efficient than the MLE for some parameters in Case 3. It should be mentioned that for Cases 3 and 4 (these are difficult cases), the R.E. values for MHDE (relative to the MLE) we obtained using Lu et al. (2003)’s published code were drastically different from those obtained using the numbers reported in Tables 3 and 4 of Lu et al. (2003; see Section 6). This is another reason why we recomputed the MHDE estimates using Lu et al. (2003)’s published code in all our simulations above.

2.7.2 $L_2E$ with Penalty

Whereas the $L_2E$ is more efficient than the MLE in the poorly-separated model in Case 3, it is not as efficient as the MLE when estimating a small mixing proportion and first component parameters in the poorly-separated model in Case 4. This is not entirely surprising.
because, when there is a small mixing proportion in a two-component finite mixture regression model—this is possible if there is a small fraction of extreme values in a real data—the inherent robustness of $\hat{\theta}_{L^2E}^{n,m}$ automatically down weighs the extreme values, which in turn impedes accurate estimation of the small mixing proportion and first component parameters.

To improve the R.E. of $L_2E$ in this case, we introduce the following modified $L_2E$ criterion with a penalty function:

$$L_{2p}(\theta_m) = n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y | X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i | X_i) + p(\pi)$$

$$= L_2(\theta_m) + p(\pi)$$ (2.17)

where $L_2(\theta_m)$ is the (original) estimating equation defined in (2.15) and the penalty function

$$p(\pi) = \min \{ \tau \tanh[\eta(\gamma - (0.5 - \pi_1))], \ldots, \tau \tanh[\eta(\gamma - (0.5 - \pi_m))] \},$$

(2.18)

for some positive $\tau$, $\eta$, and $\gamma$, to be specified below. Henceforth, we will refer to the modified $L_2$ criterion with the penalty as $L_{2E}p$. The resulting estimator of $\theta_m$ is given by

$$\hat{\theta}_{L^2E}^{n,m} = \arg \min_{\theta_m} \left[ n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y | X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i | X_i) + p(\pi) \right].$$

(2.19)

![Figure 2.1: The Penalty function 1](image)
When $m = 2$, the penalty function reduces to $p(\pi_1) = \tau \tanh[\eta(\gamma - |0.5 - \pi_1|)]$. Figure 2.1 above shows the graph of a generic penalty function, where the values, $\tau$, $\eta$, and $\gamma$ correspond to the height, kurtosis, and width of the function, respectively. In particular, the height tends to $2\tau$ as $\eta\gamma$ increases, whereas the height goes to zero as $\eta\gamma \to 0$. Moreover, the penalty function is *platykurtic* when $\eta$ is large, while it is *leptokurtic* when $\eta$ is small.

![Figure 2.1: The Penalty function 1](image1)

![Figure 2.2: The Penalty function 2](image2)

Figure 2.2 shows the graph of the penalty function for two different sets of values of $\tau$, $\eta$, and $\gamma$. Note from Figure 2.2 that the penalty function takes a (positive) constant value around a large neighborhood of 0.5 and then decreases slowly, becoming negative when $\pi_1$ is roughly near 0.8 and finally settles at a negative value when $\pi_1$ is close to 1. Intuitively, addition of
such a penalty function to our original estimating function $L_2(\theta_m)$ should improve estimation of $\pi_1$ when the true value of $\pi_1$ is close to 1 (or 0). We also checked through simulations (not given here) that when the true value of $\pi_1$ is not very large (or small), the penalty function does not affect the efficiency of $\hat{\theta}_{n,m}$. In the literature similar penalty function have been added to estimating equations to improve estimation; see, for instance, Li, Chen, and Marriott (2009) for other penalty functions added to the loglikelihood.

It is reasonable to ask when one should use $L_2E_p$ for real data. The decision to use $L_2E_p$ on real data would be largely dictated by the presence of small subpopulation(s) in the data, which would be apparent from the histogram and/or examination of percentile values. When in doubt, it is always advisable to fit the data using both $L_2E$ and $L_2E_p$, and then carefully examine the adequacy of each fit and interpretations given by covariates to determine the final model. As to the choice of values for $\tau$, $\eta$, and $\gamma$ when one decides to use a penalty function, we will show below that the estimator, $\hat{\theta}_{n,m}$, is less sensitive to different choices of values.

Before we numerically assess the performance of $\hat{\theta}_{n,2}^{L_2E_p}$, we show that for two-component mixture of regression models $\hat{\theta}_{n,2}^{L_2E_p}$ continues to enjoy the same theoretical properties as $\hat{\theta}_{n,2}^{L_2E}$ stated in Theorem 1 in Section 2.5 above. This result is stated as a Corollary below, whose proof is given in Appendix 2.10.4.

**Corollary 1.** For two-component mixtures of regression models, under the same regularity conditions as in Section 2.5 with $\pi_1 \neq 0.5$ as $n \to \infty$,

1. $\hat{\theta}_{n,2}^{L_2E_p} \overset{P}{\to} \theta_2$, and
2. $\sqrt{n}(\hat{\theta}_{n,2}^{L_2E_p} - \theta_2) \overset{D}{\to} N(0, J^{-1}K^*J^{-1})$,

where $J = J(\theta_2)$ is defined in condition (C.4) in Section 2.5 and $K^* = K^*(\theta_2)$ is indicated in Appendix 2.10.4.
Remark 2. In order to estimate the asymptotic covariance matrix of $\hat{\theta}_{n,2}^{L_2E_p}$ (See part (ii) of Corollary 1), we also suppose (as done in Remark 1) that the true conditional p.m.f $f_0$ is equal to $f_{\theta_m}$, where $\theta_m$ is the true value of the parameter. In the data analysis described in Section 2.8.1 we compute the standard error of the $L_2E_p$ estimates of $\theta_2$ using $\hat{J}$ (see (2.12) in Remark 1) and $\hat{K}^*$, where $\hat{K}^*$ is similar to $\hat{K}$ in (2.13) with modification as indicated in Appendix 2.10.4.

Corollary 2. Suppose all the assumptions of the Proposition stated in Section 2.5 are satisfied. If $g_0$ is a finite mixture with mixture complexity $m_0 < \infty$, then for any sequence $\zeta_{n,m} \to 0$

$$\hat{m}^{L_2E_p}_n \longrightarrow m_0 \text{ a.s.}$$

as $n \to \infty$, where

$$\hat{m}^{L_2E_p}_n = \min \left\{ m : L_{2p}(\hat{\theta}^{L_2E_p}_{n,m}) \leq L_{2p}(\hat{\theta}^{L_2E_p}_{n,m+1}) + \zeta_{n,m} \right\},$$

(2.20)

with $L_{2p}(\theta_m)$ is as defined in (2.17) with the penalty function $p(\pi)$ as in (2.18), and $m_0$ is as defined in (2.8). If $g_0$ is not a finite mixture, then $\hat{m}^{L_2E_p}_n \to \infty$ a.s. The proof is almost identical to that of Theorem 2; hence is omitted.

Returning to Case 4 of our simulation study, we used the penalty $p(\pi_1)$ with two different sets of threshold values, $\tau = 0.001$, $\eta = 4$, and $\gamma = 0.3$, and $\tau = 0.48$, $\eta = 15$, and $\gamma = 0.3$ (Figure 2.2) and computed the estimates $\hat{\theta}^{\text{MLE}}_{n,m}$, $\hat{\theta}^{\text{MHDE}}_{n,m}$, $\hat{\theta}^{L_2E_p}_{n,m}$, and their R.E.s. Once again, all the three estimates were computed using the same data sets and initial values. Table 2.2 shows that the use of the penalty function substantially improves the performance of the $L_2E_p$ for Case 4 in comparison with Table 2.1. More precisely, the $L_2E_p$ outperforms the MLE (and continues to outperform MHDE) for all parameters and sample sizes. Note that the performance of the $L_2E_p$ is not too sensitive to different choice of values for $\tau$, $\eta$, and $\gamma$. In conclusion, the performance of the $L_2E_p$ constructed using the penalty function (with
\( \tau = 0.001, \eta = 4, \) and \( \gamma = 0.3, \) or \( \tau = 0.48, \eta = 15, \) and \( \gamma = 0.3 \) is better than that of the \( L_2E, \) MHDE, and MLE for all parameters in Case 4, for which the modification was intended.

Table 2.2: Comparison between MLE, MHDE, and \( L_2E_p \) for Case 4: Relative Efficiencies

<table>
<thead>
<tr>
<th>Parameters ( x \sim U[0,1] )</th>
<th>( \tau = 0.001, \eta = 4, ) and ( \gamma = 0.3 )</th>
<th>( \tau = 0.48, \eta = 15, ) and ( \gamma = 0.3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi_1 = 0.95 )</td>
<td>0.681786</td>
<td>0.683319</td>
</tr>
<tr>
<td>( \beta_{10} = 0.6 )</td>
<td>0.114905</td>
<td>0.112983</td>
</tr>
<tr>
<td>( \beta_{11} = 1.0 )</td>
<td>0.103443</td>
<td>0.101265</td>
</tr>
<tr>
<td>( \beta_{20} = 0.8 )</td>
<td>0.001794</td>
<td>0.000982</td>
</tr>
<tr>
<td>( \beta_{21} = 1.2 )</td>
<td>0.001954</td>
<td>0.001064</td>
</tr>
</tbody>
</table>

2.7.3 Robustness against Gross-error Contaminations

Here, we investigate the robustness of \( L_2E \) against 100\( \varepsilon \)% gross-error contaminated mixture Poisson regression models defined by

\[
f_{\theta, \varepsilon, \lambda_{i3}}(y_i | x_i) = (1 - \varepsilon)f_{\theta_2}(y_i | x_i) + \varepsilon f_{\lambda_3}(y_i | \lambda_{i3}),
\]

where \( f_{\theta_2} \) is as in (2.16), \( \varepsilon \) is the proportion of contamination, and the contaminating value \( \lambda_{i3} = \exp(\beta_{30} + \beta_{31}x_i) \) is large compared to those of \( \lambda_{i1} \) and \( \lambda_{i2}. \) More precisely, we set \( \varepsilon = 0.01, 0.05, \) and \( 0.10, \) and the regression coefficients of the third (contaminating) component are set at \( \beta_3 = (3.7, 0) \) for well-separated models (Case 1 and 2) and \( \beta_3 = (2.8, 0) \) for poorly-separated models (Case 3 and 4). Thus, the third component means are approximately 40.4 and 16.4 for well- and poorly-separated models, respectively; these are relatively large compared to each of the respective second component means: \( \lambda_{i2} \approx 28.3 \) and 4.3.
We perform simulation studies based on 200 replications with sample sizes \( n = 200 \) and 500 when \( \theta_2 \) is as in Cases 1-4. We generated samples from \( f_{\theta_2, \varepsilon, \lambda_3}(y_i|\mathbf{x}_i) \) and computed the estimates \( \hat{\theta}_{L^2E}^{n,2}, \hat{\theta}_{MLE}^{n,2}, \) and \( \hat{\theta}_{MHDE}^{n,2} \). We also used the published code in Lu et al. (2003) to compute the MHDE, which is available from the *Biometrics* website.

We assessed the robustness of \( L^2E \) to the contaminations above through R.E. of the \( L^2E \) compared to the MLE and MHDE, respectively. The R.E. of \( L^2E \) compared to the MLE is reported in Tables 2.3 and 2.4, and the R.E. of \( L^2E \) compared to the MHDE is reported in Tables 2.5 and 2.6. As before, in Tables 2.3, 2.4, 2.5, and 2.6, R.E. > 1 is identified in bold and R.E. > 0.9 is identified in italic for ease of comparison.

Table 2.3: The relative efficiency of the \( L^2E \) vs the MLE (well-separated models)

<table>
<thead>
<tr>
<th>Case 1: ( \theta_2 = (0.5, 0.6, 1.0, 2.5, 1.5) ) with ( \beta_3 = (3.7, 0) )</th>
<th>( \pi_1 )</th>
<th>( \beta_{10} )</th>
<th>( \beta_{11} )</th>
<th>( \beta_{20} )</th>
<th>( \beta_{21} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>( \varepsilon = 0.01 )</td>
<td>( \varepsilon = 0.05 )</td>
<td>( \varepsilon = 0.10 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.737952</td>
<td>0.608151</td>
<td>0.619237</td>
<td>1.579440</td>
<td>1.377853</td>
</tr>
<tr>
<td>500</td>
<td>0.881512</td>
<td>0.854848</td>
<td>0.767001</td>
<td>2.210057</td>
<td>1.761970</td>
</tr>
<tr>
<td>200</td>
<td>1.080872</td>
<td>0.844260</td>
<td>0.798431</td>
<td>9.596001</td>
<td>7.747073</td>
</tr>
<tr>
<td>500</td>
<td>1.231113</td>
<td>0.940577</td>
<td>0.823489</td>
<td>21.97846</td>
<td>16.30218</td>
</tr>
<tr>
<td>200</td>
<td>1.427806</td>
<td>1.178158</td>
<td>0.966605</td>
<td>21.63284</td>
<td>15.74818</td>
</tr>
<tr>
<td>500</td>
<td>1.722218</td>
<td>1.276738</td>
<td>1.020011</td>
<td>51.66155</td>
<td>42.62790</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 2: ( \theta_2 = (0.95, 0.6, 1.0, 2.5, 1.5) ) with ( \beta_3 = (3.7, 0) )</th>
<th>( \pi_1 )</th>
<th>( \beta_{10} )</th>
<th>( \beta_{11} )</th>
<th>( \beta_{20} )</th>
<th>( \beta_{21} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>( \varepsilon = 0.01 )</td>
<td>( \varepsilon = 0.05 )</td>
<td>( \varepsilon = 0.10 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.485216</td>
<td>0.703264</td>
<td>0.628402</td>
<td>2.620229</td>
<td>1.439340</td>
</tr>
<tr>
<td>500</td>
<td>0.608732</td>
<td>0.694026</td>
<td>0.698692</td>
<td>6.502045</td>
<td>5.125489</td>
</tr>
<tr>
<td>200</td>
<td>0.974810</td>
<td>0.797510</td>
<td>0.750726</td>
<td>3.333506</td>
<td>3.250667</td>
</tr>
<tr>
<td>500</td>
<td>1.191335</td>
<td>0.914877</td>
<td>0.863363</td>
<td>5.018670</td>
<td>4.809153</td>
</tr>
<tr>
<td>200</td>
<td>1.118773</td>
<td>0.964406</td>
<td>0.905058</td>
<td>1.575239</td>
<td>1.559673</td>
</tr>
<tr>
<td>500</td>
<td>1.170962</td>
<td>1.171522</td>
<td>1.054599</td>
<td>1.300150</td>
<td>1.286163</td>
</tr>
</tbody>
</table>

Table 2.3 shows that the \( L^2E \) is more efficient than the MLE for the second component parameters for all sample sizes and whether or not the second component is relatively small.
This shows that the efficiency of MLE is severely affected when the contamination is far away from the components. Particularly, as contamination levels become severe, the performance of $L_2E$ for the mixing proportion and the first component regression coefficients become better than that of MLE. More precisely, when the level of contamination is small ($\varepsilon = 0.01$), the MLE is more efficient than the $L_2E$ for the mixing proportion and the first component parameters, for all sample sizes. However, when the level of contamination is severe ($\varepsilon = 0.10$), the $L_2E$ performs well for the mixing proportion and is almost as efficient as the MLE for the first component regression coefficients. In fact, in Case 1 the R.E. of $L_2E$ increases as $n$ becomes large for all parameters, but in Case 2 $L_2E$ for the second component when $\varepsilon = 0.10$ is somewhat less efficient than when $\varepsilon = 0.05$.

Table 2.4: The relative efficiency of the $L_2E$ vs the MLE (poorly-separated models)

<table>
<thead>
<tr>
<th>Case 3: $\theta_2 = (0.5, 0.6, 1.0, 0.8, 1.2)$ with $\beta_3 = (2.8, 0)$</th>
<th>$\pi_1$</th>
<th>$\beta_{10}$</th>
<th>$\beta_{11}$</th>
<th>$\beta_{20}$</th>
<th>$\beta_{21}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$\varepsilon = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>4.517429</td>
<td>0.448477</td>
<td>0.494971</td>
<td>12.18322</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>4.603598</td>
<td>0.272792</td>
<td>0.198393</td>
<td>34.64517</td>
</tr>
<tr>
<td>$n$</td>
<td>$\varepsilon = 0.05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>4.392321</td>
<td>0.279952</td>
<td>0.279210</td>
<td>26.80257</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>3.004578</td>
<td>0.330749</td>
<td>0.213445</td>
<td>23.09939</td>
</tr>
<tr>
<td>$n$</td>
<td>$\varepsilon = 0.10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>2.632669</td>
<td>0.308430</td>
<td>0.300645</td>
<td>10.31775</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>2.017586</td>
<td>0.376643</td>
<td>0.323677</td>
<td>4.976262</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 4: $\theta_2 = (0.95, 0.6, 1.0, 0.8, 1.2)$ with $\beta_3 = (2.8, 0)$</th>
<th>$\pi_1$</th>
<th>$\beta_{10}$</th>
<th>$\beta_{11}$</th>
<th>$\beta_{20}$</th>
<th>$\beta_{21}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$\varepsilon = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.056166</td>
<td>0.215350</td>
<td>0.245812</td>
<td>4.776708</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.072936</td>
<td>0.191845</td>
<td>0.220873</td>
<td>6.102565</td>
</tr>
<tr>
<td>$n$</td>
<td>$\varepsilon = 0.05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.015779</td>
<td>0.364749</td>
<td>0.335484</td>
<td>2.828692</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.011233</td>
<td>0.281297</td>
<td>0.296801</td>
<td>2.488701</td>
</tr>
<tr>
<td>$n$</td>
<td>$\varepsilon = 0.10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0.328451</td>
<td>0.451107</td>
<td>0.438692</td>
<td>1.741227</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.696215</td>
<td>0.635812</td>
<td>0.576968</td>
<td>1.154113</td>
</tr>
</tbody>
</table>
For the poorly-separated model considered in Case 3, Table 2.4 shows that the $L^2E$ for the mixing proportion as well as the second component parameters is more efficient than the MLE for all levels of contamination and sample sizes. However, the MLE is more efficient than the $L^2E$ for the first component parameters, for all levels of contamination and sample sizes. In Case 4, the $L^2E$ for the second component parameters performs well for almost all levels of contamination and sample sizes. However, the R.E. of $L^2E$ is less than that of MHE for the mixing proportion and the first component parameters even though it increases to some extent as contaminations become severe.

Table 2.5: The relative efficiency of the $L^2E$ vs the MHDE (well-separated models)

<table>
<thead>
<tr>
<th>Case 1: $\theta_2 = (0.5, 0.6, 1.0, 2.5, 1.5)$ with $\beta_3 = (3.7, 0)$</th>
<th>$\pi_1$</th>
<th>$\beta_{10}$</th>
<th>$\beta_{11}$</th>
<th>$\beta_{20}$</th>
<th>$\beta_{21}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>2.211064</td>
<td>3.788419</td>
<td>5.413387</td>
<td>3.233910</td>
<td>3.738408</td>
</tr>
<tr>
<td>500</td>
<td>1.204132</td>
<td>3.374628</td>
<td>4.449664</td>
<td>2.995921</td>
<td>3.706434</td>
</tr>
<tr>
<td>$\varepsilon = 0.05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1.461698</td>
<td>2.906491</td>
<td>4.465220</td>
<td>4.196069</td>
<td>4.816903</td>
</tr>
<tr>
<td>500</td>
<td>0.839290</td>
<td>3.764244</td>
<td>5.130940</td>
<td>4.840694</td>
<td>5.255434</td>
</tr>
<tr>
<td>$\varepsilon = 0.10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.836985</td>
<td>2.877944</td>
<td>3.793446</td>
<td>5.741952</td>
<td>5.015456</td>
</tr>
<tr>
<td>500</td>
<td>0.938615</td>
<td>3.730368</td>
<td>5.083854</td>
<td>8.159027</td>
<td>7.599006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 2: $\theta_2 = (0.95, 0.6, 1.0, 2.5, 1.5)$ with $\beta_3 = (3.7, 0)$</th>
<th>$\pi_1$</th>
<th>$\beta_{10}$</th>
<th>$\beta_{11}$</th>
<th>$\beta_{20}$</th>
<th>$\beta_{21}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1.410025</td>
<td>2.637362</td>
<td>3.339077</td>
<td>9.435256</td>
<td>3.553203</td>
</tr>
<tr>
<td>500</td>
<td>1.167526</td>
<td>2.231918</td>
<td>3.060181</td>
<td>17.55445</td>
<td>9.748067</td>
</tr>
<tr>
<td>$\varepsilon = 0.05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.146686</td>
<td>2.782098</td>
<td>4.158173</td>
<td>3.986203</td>
<td>3.580838</td>
</tr>
<tr>
<td>500</td>
<td>0.358659</td>
<td>2.091982</td>
<td>2.854258</td>
<td>6.540490</td>
<td>5.411358</td>
</tr>
<tr>
<td>$\varepsilon = 0.10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.242866</td>
<td>2.965674</td>
<td>4.423789</td>
<td>2.006965</td>
<td>1.912535</td>
</tr>
<tr>
<td>500</td>
<td>0.496513</td>
<td>1.960772</td>
<td>2.523431</td>
<td>1.504467</td>
<td>1.425561</td>
</tr>
</tbody>
</table>

As can be seen from Table 2.5, which gives the R.E. of the $L^2E$ compared to the MHDE, for Cases 1 and 2 the $L^2E$ is clearly more efficient for all regression coefficients. However, for the mixing proportion, when the contamination is moderate or severe ($\varepsilon = 0.05$ or 0.10),
the MHDE is better than $L_2E$, but our $L_2E$ is still competitive to the MHDE. As for Cases 3 and 4, the performance of the $L_2E$ is better than the MHDE for almost all parameters. However, in Case 3 the $L_2E$ becomes less efficient compared to the MHDE for $\pi_1$ and $\beta_{20}$ when the level contamination is severe and $n = 500$, and in Case 4 $L_2E$ is worse than the MHDE for the parameters when $\varepsilon = 0.05$ with $n = 200$ and $\varepsilon = 0.10$ with $n = 200$ and 500, respectively. Since the contamination parameters are the same for the cases considered in Tables 2.3 and 2.5, in the situations where the MLE is as efficient or more efficient than the $L_2E$ but the latter is more efficient than the MHDE, we can conclude that the performance of the MHDE is much worse than the MLE.

Table 2.6: The relative efficiency of the $L_2E$ vs the MHDE (poorly-separated models)

| Case 3: $\theta_2 = (0.5, 0.6, 1.0, 0.8, 1.2)$ with $\beta_3 = (2.8, 0)$ | $\pi_1$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{20}$ | $\beta_{21}$ |
|---|---|---|---|---|
| $\varepsilon = 0.01$ | | | | | |
| $n$ | $\varepsilon$ | 200 | 1.283822 | 2.177727 | 3.136509 | 2.933183 | 1.726719 |
| 200 | 1.597331 | 1.876266 | 3.122642 | 3.059138 | 2.259245 |
| $n$ | $\varepsilon$ | 200 | 2.713668 | 1.491362 | 1.518433 | 2.850503 | 3.583477 |
| 500 | 1.658705 | 1.237762 | 1.104381 | 3.094385 | 5.436454 |
| $n$ | $\varepsilon$ | 200 | 1.127083 | 2.869497 | 3.186586 | 1.871035 | 2.483163 |
| 500 | 0.514167 | 5.607890 | 6.772185 | 0.832411 | 3.801424 |

| Case 4: $\theta_2 = (0.95, 0.6, 1.0, 0.8, 1.2)$ with $\beta_3 = (2.8, 0)$ | $\pi_1$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{20}$ | $\beta_{21}$ |
|---|---|---|---|---|
| $\varepsilon = 0.01$ | | | | | |
| $n$ | $\varepsilon$ | 200 | 1.124379 | 4.894744 | 6.251754 | 1910.096 | 1016.815 |
| 500 | 0.693797 | 3.929831 | 5.094624 | 3866.875 | 1621.695 |
| $n$ | $\varepsilon$ | 200 | 0.447041 | 5.160351 | 6.362797 | 10.57424 | 6.331524 |
| 500 | 1.119443 | 7.882234 | 11.42163 | 1.057486 | 1.823140 |
| $n$ | $\varepsilon$ | 200 | 3.077861 | 7.133456 | 9.961060 | 0.713484 | 2.142887 |
| 500 | 9.696371 | 21.40473 | 29.86705 | 0.439754 | 1.574259 |

In conclusion, for the well-separated case, the $L_2E$ performs better than (or as good as) the MLE when the contamination is moderate to severe. However, when the contamination
is small, the MLE performs better than the $L_2E$ for the mixing proportion and the first component parameters. The performance of $L_2E$ is much superior compared to the MHDE for the well-separated case, but the $L_2E$ for the mixing proportion is less efficient than the MHDE when there is a severe contamination. For the poorly-separated model in Case 3, the $L_2E$ is more efficient than the MLE for the mixing proportion and the second component and is also more efficient than the MHDE for most parameters. However, in Case 4 $L_2E$ for only the second component coefficients performs better than the MLE. Nonetheless the $L_2E$ does better than the MHDE as in Case 3. Overall, our $L_2E$ is inclined to protect the second component against the contamination; thus, the MLE is competitive in the first component, but not in the second one. In particular, in the poorly-separated case, there is no clear winner among the three estimators, but the $L_2E$ and MHDE seem to be competitive with the MLE.

2.7.4 Evaluation of $\hat{m}_n^{L_2E}$

In order to assess the performance of $\hat{m}_n^{L_2E}$, we generated data from two-component Poisson mixture regression model (2.16) in Section 2.7.1 and we considered two cases with respective parameter vectors:

**Case 1**: (0.5, 0.6, 1.0, 2.5, 1.5); **Case 4**: (0.95, 0.6, 1.0, 0.8, 1.2).

Note again that the component regression parameters of Case 1 are well-separated, whereas Case 4 is poorly-separated and its mixing proportion of the second component is small (0.05). Our preliminary simulation studies, where we carried out the simulations for sample size $n = 100$ showed that in Case 1 our $L_2E$ correctly detected the true mixture complexity 98% of the time. However, in Case 4 our procedure failed to perform well; that is, $\hat{m}_n^{L_2E}$ is often underestimated in the case of a poorly-separated model with a small mixing proportion. As we mentioned in Section 2.7.1, this is a general characteristic of $L_2E$. Therefore, we revisited the $L_2E_p$ methodology defined in Section 2.7.1 to improve the performance of $\hat{m}_n^{L_2E}$. In Sections 2.4 and 2.6, we introduced an estimator of $m$ defined by

$$\hat{m}_n^{L_2E} = \min \left\{ m : L_2(\hat{\theta}_{n,m}^{L_2E}) \leq L_2(\theta_{n,m+1}^{L_2E}) + \zeta_{n,m} \right\},$$
where $L_2(\theta_m) = \left[ n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2 (y|X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m} (Y_i|X_i) \right]$ and the threshold $
abla_{m,n} = (\kappa/n) \ln[(m+1)/m]$ for $\kappa > 0$. Our $L_2 E_p$ estimator of $m$ is that

$$\hat{m}_{n}^{L_2 E_p} = \min \left\{ m : L_2 p(\hat{\theta}_{n,m}^{L_2 E_p}) \leq L_2 p(\hat{\theta}_{n,m+1}^{L_2 E_p}) + \nabla_{n,m} \right\},$$

where $L_2 p(\theta_m)$ is as defined in (2.17) with the penalty function $p(\pi)$ as in (2.18). For three threshold values, we used the same numbers as in Section 2.7.1; that is, $\tau = 0.001$, $\eta = 4$, and $\gamma = 0.3$.

We performed 100 Monte Carlo replications of our algorithm for sample size $n = 1000$, each yielding an estimate of mixture complexity with the LIC threshold value $\nabla_{n,m} = n^{-1} \kappa \log((m+1)/m)$, where we used $\kappa = 0.6$ as in Umashanger and Sriram (2009). We then tally the estimated number of components (out of 100 replications). These counts are reported for each case in Table 2.7. The simulation results in Table 2.7 show that for the well-separated model considered in Case 1 our $\hat{m}_{n}^{L_2 E}$ correctly estimates the true mixture complexity 99% of the time. In the poorly-separated case with a small mixing proportion (Case 4), $\hat{m}_{n}^{L_2 E_p}$ also correctly identifies a large percentage of times, which indicates that $L_2 E_p$ identifies a small second component.

Table 2.7: Mixture complexity estimation results

<table>
<thead>
<tr>
<th>$m_0 = 2$</th>
<th>Estimated number of components</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 1000$</td>
<td></td>
</tr>
<tr>
<td>$L_2 E$ (Case 1)</td>
<td>99 1</td>
</tr>
<tr>
<td>$L_2 E_p$ (Case 4)</td>
<td>90 10</td>
</tr>
</tbody>
</table>

2.8 Data Analysis

We assess the performance of our $L_2 E$ estimator on real data sets by revisiting two examples discussed in the literature, both of which have some fraction of extreme values. The first is the Maternity Length of Stay data set analyzed in Lu et al. (2003) using MHD.
and the second is the *Toronto Motor Vehicle Crashes* data set analyzed in Park and Lord (2009) using Bayesian methodology.

2.8.1 **Maternity Length of Stay**

A major concern of policymakers worldwide is the rapidly increasing cost of healthcare. Ideally one would measure healthcare utilization by expenditures, but discrete measures of units of healthcare use, such as the number of visits to a physician and the length of stay (LOS) in a hospital (measured in days/weeks), are often more easily available from national health surveys than data on expenditures (Cameron and Trivedi, 1998). Although considerable research has been conducted into possible causes and consequences of increasing costs, a basic prerequisite of good policy, which is the ability to predict healthcare use reasonably, remains elusive. Consequently, construction of meaningful statistical models for the estimation of healthcare utilization continues to be an active area of research.

In modeling counts of medical care utilization, finite mixture regression model framework (McLachlan and Peel, 2000) is a preferred choice over two-part models (Mullahy, 1986) because the former provides a more tenable distinction for typical cross-sectional data, distinguishing between the populations of, e.g., “infrequent users” and “frequent users” of medical care (Deb and Trivedi, 1997, 2002) or “normal-stayers” and “long-stayers” in hospitals (Lu et al., 2003), the difference being determined by the available covariate information. While finite mixture models allow greater flexibility of functional form, the fact that healthcare utilization data sets are often bedevilled by the presence of extreme values severely impacts likelihood-based estimation methods; hence, conclusions drawn from these may not be completely reliable. In fact, if the estimates do not reflect the underlying behavioral structures that drive healthcare use, then predictions of actual use and policy recommendations made on their basis may have unintended consequences.

Examples of count data sets with extreme values are plentiful, especially in the healthcare literature. For example, the count data on the number of office-based physician visits
considered in Deb and Trivedi (1997) has approximately 31 percent of individuals that do
not have any visits in a year while 90 percent have 13 or fewer visits. In addition, the 95th
and 99th percentiles of the data are 20 and 44, respectively, with observations in the 50s,
100s and even 200s. These extreme values cannot be immediately judged as outliers because
one plausibly can have those numbers of doctor visits in a year, for example, if one is on
dialysis. Similarly, the LOS data considered in Lu et al. (2003), Singh and Ladusingh (2009)
and Kulinskaya, Kornbrot, and Gao (2005) all contain a fraction of extreme values (unusually
long stays).

Ever increasing patient-care costs demand that healthcare institutions improve their
resource-utilization effectiveness and efficiency. The length of an inpatient’s stay has direct
impact on patient care costs, service quality, and outcomes. Despite attempts to manage the
LOS for frequently performed medical procedures, many service providers cannot achieve the
target range allowed by the managed care system. Consequently, as an important component
in the consumption of hospital resources, LOS in hospital has drawn considerable attention
in the healthcare literature (Lave and Leinhard, 1976). LOS also provides important inform-
ation for health planning and formation of payment policy (Xiao, Lee, and Vemuri, 1999;
Wang, Yau, and Lee, 2002; Polverejan, Gardiner, Bradley, Holmes, and Rovner, 2003); hence,
it is a key performance indicator for hospital management and a key measure of hospital
efficiency.

Here, we revisit the Maternity Length of Stay count data previously analyzed in Lu et al.
(2003) using MHDE and a two-component Poisson regression model, and illustrate the use
of our $L_2E$ robust estimation methodology in conjunction with a suitable finite mixture of
regression model. This data set, extracted from the 1998/1999 Western Australian hospital
morbidity database, is available for download from the supplementary materials section
of Lu et al. (2003); see http://www.biometrics.tibs.org/datasets/020128.html. The
data consists of $n = 865$ vaginal delivery patients with multiple complicating diagnoses.
For each patient, the data contains a response variable $y$, which is defined as the number
of days stay at the hospital after delivery to discharge, and information on six covariates that are marital status (unmarried/married), admission type (emergency/planned), payment class (private/public), location (rural/urban), occupation (professional/home duties), and indigenous status (aboriginal/non-aboriginal). Each covariate takes 1/0 value according to the categories indicated in the parentheses. It can be observed that 91.5% of patients in the data set sustained 11 or fewer days of hospitalization with the average LOS of 6.24 days and median of 5 days. However, the 95th and 99th percentiles of the data are 14 and 29, respectively, with observations in the 30s and even 40s, indicating the presence of a small fraction of extreme values (see Figure 2.3). This clear presence of subpopulations within the data suggest that a robust estimation criterion that is less sensitive to extreme values and computationally feasible in conjunction with a finite mixture of regression model would more accurately capture the overall heterogeneity in the data.

![Figure 2.3: Maternity length of stay](image)

Our data analysis begins with estimation of the number of mixture components and simultaneous determination of suitable parametric distribution for components in our finite mixture of regression model. We estimated $m$ using the $L_2E$ estimator $\hat{m}_n^{L_2E}$ defined in
Section 2.4. Since the data is count, we set the $j$-th component distribution in $f_{\theta_m}(y|x)$ to be Poisson [see below (2.1)] for $j = 1, \ldots, m$.

Using the iterative algorithm described in Section 2.6, we computed $\hat{m}_{nL_2E}$ in (2.10) and (2.14) with the threshold value $\zeta_{n,m} = (2/865) \ln[(m+1)/m]$, that is, with $\kappa = 2$. Here, we used a slightly higher value of $\kappa (= 2)$ than the one suggested in Umashanger and Sriram (2009) because $n$ is large; see Sections 2 and 7 of Umashanger and Sriram (2009). This yielded $\hat{m}_{nL_2E} = 2$; see Appendix 2.10.5 for details. Therefore, we decided to fit a two-component Poisson mixture regression model to the LOS data.

We pointed out in the beginning of this section that the Maternity LOS data has a small fraction of extreme values. This data structure and the improved performance of $L_2E_p$ for models with small mixing proportion (see Table 2.2) motivated us to use $L_2E_p$ with the penalty function $p(\pi) = \min\{\tau \tanh[\eta(\gamma - (0.5 - \pi_1))], \ldots, \tau \tanh[\eta(\gamma - (0.5 - \pi_m))]\}$, where $\tau = 0.001$, $\eta = 4$, and $\gamma = 0.3$ for robust estimation of parameters in the Poisson mixture regression model. We also estimated $m$ using the $L_2E_p$ estimator $\hat{m}_{nL_2E_p}$ defined in Section 2.7.4 and computed $\hat{m}_{nL_2E_p}$ in (2.33) with the same threshold value $\zeta_{n,m}$. Incidentally, we also determined using $L_2E_p$ that $\hat{m}_{nL_2E_p} = 2$ (See Appendix 2.10.5).

Table 2.8 gives the results of our $L_2E$ and $L_2E_p$ and that of the MHDE taken from Table 7 of Lu et al. (2003). The component LOS means ($\lambda$) and standard errors (s.e.) are also given in Table 2.8. See Remark 1 in Section 2.5 and Remark 2 in Section 2.7.2 for the standard error and see Appendix 2.10.6 for a comparison of the two-component Poisson mixture regression fit using $L_2E_p$ and MLE. Table 2.8 shows that the parameter estimates obtained using $L_2E_p$ ($L_2E$) are different from the MHDE values reported in Table 7 of Lu et al. (2003). Obviously, the $L_2E_p$ fits are almost same as $L_2E$ fits except for the mixing proportion. Based on the estimated LOS mean ($\lambda$) for the second component, the $L_2E_p$ ($L_2E$) is less influenced by the extreme LOS values compared to the MHDE. As for covariate influence, our $L_2E_p$ ($L_2E$) fit shows that, among the small proportion of patients whose longer-stays are modeled by the second component, the maternity LOS is positively and significantly
Table 2.8: Results of fitting Poisson mixture regression models to maternity length of stay data

<table>
<thead>
<tr>
<th>Estimation methods</th>
<th>Component</th>
<th>$L_2E_p$</th>
<th>$L_1$ GoF</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mixing proportion</td>
<td></td>
<td></td>
<td>$0.8133$ ($0.0320$)</td>
<td>$0.9316$ ($0.0152$)</td>
</tr>
<tr>
<td>Covariate</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.e.)</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.5747 (0.0476)</td>
<td>1.3115 (0.1857)</td>
<td>1.5689 (0.0194)</td>
<td>1.8862 (0.0567)</td>
</tr>
<tr>
<td>Marital status</td>
<td>0.0247 (0.0504)</td>
<td>$0.5964$ ($0.1313$)</td>
<td>$-0.0020$ (0.0716)</td>
<td>$1.3186$ (0.0571)</td>
</tr>
<tr>
<td>Admission type</td>
<td>0.0860 (0.0531)</td>
<td>$1.0059$ (0.1503)</td>
<td>$-0.0027$ (0.0581)</td>
<td>$0.0020$ (0.1252)</td>
</tr>
<tr>
<td>Payment class</td>
<td>0.2552 (0.0857)</td>
<td>0.4491 (0.2587)</td>
<td>0.0059 (0.3704)</td>
<td>0.0013 (0.8641)</td>
</tr>
<tr>
<td>Location</td>
<td>0.2004 (0.0672)</td>
<td>$-0.2020$ (0.2066)</td>
<td>$0.8131$ (0.0443)</td>
<td>$0.0084$ (0.3022)</td>
</tr>
<tr>
<td>Occupation</td>
<td>$-0.0972$ (0.0489)</td>
<td>0.3075 (0.1414)</td>
<td>0.0054 (0.0460)</td>
<td>$-0.4920$ (0.3454)</td>
</tr>
<tr>
<td>Indigenous status</td>
<td>$-0.1523$ (0.0711)</td>
<td>$1.0967$ (0.1228)</td>
<td>$-0.0049$ (0.1606)</td>
<td>$0.4907$ (0.0795)</td>
</tr>
<tr>
<td>$\lambda$ estimate</td>
<td>4.9667</td>
<td>10.1938</td>
<td>5.6411</td>
<td>10.7952</td>
</tr>
</tbody>
</table>

related to the admission type and the indigenous status, indicating that aboriginal women admitted through emergency care tend to sustain a longer period of hospitalization. Note that our conclusion about the significance of indigenous status in the second component agrees not only with that of Lu et al. (2003), but also with similar observations made in the literature. For example, according to Lee, Wang, Yau, McLachlan, and Ng (2007), “aboriginal
mothers tend to have additional co-morbidities in conjunction with a delayed admission due to transportation difficulties, which in turn may lead to further complications and consequently a late discharge.” Usually, any delay in admission makes patients deteriorate in health, which requires them to be transported directly to the emergency care. This may also explain why admission type is significant in our second component (Kulinskaya et al., 2005; Lee et al., 2007). Furthermore, our $L_2E_p$ ($L_2E$) fit shows that unmarried women may marginally sustain longer stay among the small subpopulation of long-stayers, but rural patients and those with private insurance coverage may marginally sustain longer stay among the main subpopulation of normal-stayers.

Table 2.9: Model-based clusters using the $L_2E_p$

<table>
<thead>
<tr>
<th>Cluster 1: 92.02% (796 obs.)</th>
<th>LOS</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>45</td>
<td>5.3329</td>
<td>3.4075</td>
</tr>
<tr>
<td>Percentage of 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td>26.88% (214 obs.)</td>
<td>73.12% (582 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ad*ission type</td>
<td></td>
<td>30.40% (242 obs.)</td>
<td>69.60% (554 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td></td>
<td>5.28% (42 obs.)</td>
<td>94.72% (754 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td>14.32% (114 obs.)</td>
<td>85.68% (682 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
<td>37.44% (298 obs.)</td>
<td>62.56% (498 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td></td>
<td>10.80% (86 obs.)</td>
<td>89.20% (710 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 2: 7.98% (69 obs.)</td>
<td>LOS</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>S.D.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>45</td>
<td>16.7536</td>
<td>8.2469</td>
</tr>
<tr>
<td>Percentage of 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td>27.54% (19 obs.)</td>
<td>72.46% (50 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ad*ission type</td>
<td></td>
<td>65.22% (45 obs.)</td>
<td>34.78% (24 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td></td>
<td>5.80% (4 obs.)</td>
<td>94.20% (65 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td>8.70% (6 obs.)</td>
<td>91.30% (63 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
<td>42.03% (29 obs.)</td>
<td>57.97% (40 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td></td>
<td>20.29% (14 obs.)</td>
<td>79.71% (55 obs.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We also classified the entire sample into two clusters based on our $L_2E_p$ Poisson mixture regression fit, by assigning each individual to the component with the highest posterior
probability computed using the following:

$$
\text{Pr}(y_i \in \text{component } j) = \frac{\hat{\pi}_j f_j(y_i|\hat{\lambda}_{ij})}{\sum_{j=1}^{2} \hat{\pi}_j f_j(y_i|\lambda_{ij})},
$$

where \( f_j(y_i|\lambda_{ij}) \) is the Poisson mass function with mean \( \lambda_{ij} = \exp(x_i^T \hat{\beta}_j) \) for \( i = 1, \ldots, 865 \), and \( j = 1, 2 \). Posterior probabilities along with cluster sizes for our \( L_2E_p \) are given in Table 2.9; this profiles the two clusters using the predictor variables. It is interesting to note from Table 2.9 that the admission type, which was found to be significant in the second component by our \( L_2E_p \), helps profile the two clusters in that cluster 1 (normal-stayers) predominantly consists of planned patients, whereas emergency patients dominate cluster 2 (long-stayers)[Kulinskaya et al., 2005]. Incidentally, the clustering allocation for \( L_2E \) fit in Table 2.10 is almost same as that for \( L_2E_p \). On the contrary, the post-profiling analysis for the MHDE fit (Table 2.11) shows that occupation helps profile the two clusters, which does not substantiate the results of the MHDE fits in Table 2.8.

### Table 2.10: Model-based clusters using the \( L_2E \)

<table>
<thead>
<tr>
<th>Cluster 1: 91.56% (792 obs.)</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOS</td>
<td>1</td>
<td>45</td>
<td>5.2992</td>
<td>3.3818</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage of 1</td>
<td>26.89% (213 obs.)</td>
<td>73.11% (579 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission type</td>
<td>30.30% (240 obs.)</td>
<td>69.70% (552 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td>5.30% (42 obs.)</td>
<td>94.70% (750 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>14.27% (113 obs.)</td>
<td>85.73% (679 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td>37.12% (294 obs.)</td>
<td>62.88% (498 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td>10.86% (86 obs.)</td>
<td>89.14% (706 obs.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 2: 8.44% (73 obs.)</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOS</td>
<td>2</td>
<td>45</td>
<td>16.4932</td>
<td>8.0934</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage of 1</td>
<td>27.40% (20 obs.)</td>
<td>72.60% (53 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission type</td>
<td>64.38% (47 obs.)</td>
<td>35.62% (26 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td>5.48% (4 obs.)</td>
<td>94.52% (69 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>9.59% (7 obs.)</td>
<td>90.41% (66 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td>45.21% (33 obs.)</td>
<td>54.79% (40 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td>19.18% (14 obs.)</td>
<td>80.82% (59 obs.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.11: Model-based clusters using the MHDE

<table>
<thead>
<tr>
<th>Cluster 1: 91.21% (789 obs.)</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOS</td>
<td>1</td>
<td>31</td>
<td>5.6033</td>
<td>3.2299</td>
</tr>
<tr>
<td>Percentage of 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td>28.01% (221 obs.)</td>
<td>71.99% (568 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission type</td>
<td>32.57% (257 obs.)</td>
<td>67.43% (532 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td>5.32% (42 obs.)</td>
<td>94.68% (747 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>10.77% (85 obs.)</td>
<td>89.23% (704 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Occupation</strong></td>
<td><strong>35.99% (284 obs.)</strong></td>
<td><strong>64.01% (505 obs.)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td>10.65% (84 obs.)</td>
<td>89.35% (705 obs.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 2: 8.79% (76 obs.)</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOS</td>
<td>1</td>
<td>45</td>
<td>12.8947</td>
<td>11.6797</td>
</tr>
<tr>
<td>Percentage of 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td>15.79% (12 obs.)</td>
<td>84.21% (64 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission type</td>
<td>39.47% (30 obs.)</td>
<td>60.53% (46 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td>5.26% (4 obs.)</td>
<td>94.74% (72 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>46.05% (35 obs.)</td>
<td>53.95% (41 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Occupation</strong></td>
<td><strong>56.58% (43 obs.)</strong></td>
<td><strong>43.42% (33 obs.)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td>21.05% (16 obs.)</td>
<td>78.95% (60 obs.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Additionally, Figure 2.4 provides the frequency histogram of LOS overlayed with those belonging to clusters 1 and 2 (see Table 2.8) for our $L_2E_p \ (L_2E)$ and the MHDE of Lu et al. (2003). It illustrates that component 1 models LOS values $y < 10$, while component 2 models LOS $y \geq 10$ for all methods. However, the histograms based on Lu et al’s MHDE method (see upper right panel of Figure 2.4) show that the second component does not adequately model higher LOS compared to those for our $L_2E_p$ or $L_2E$ method.

Furthermore, to assess the adequacy of the two-component Poisson mixture regression fit provided by $L_2E_p$ and compare it to that using MHDE of Lu et al. (2003), we introduce a goodness-of-fit (GoF) criterion based on $L_1$ distance, which we believe is appropriate for the count regression setting with binary covariates and is neutral for the comparison between two
Figure 2.4: Histograms of LOS and LOS by clusters

different distance measures. The GoF criterion based on $L_1$ compares the fitted conditional mass function with a nonparametric estimate of the true conditional mass function. More
specifically, our GoF criterion is:

\[ L_1 \text{ GoF} =: \sum_{y=0}^{\infty} \sum_{i=1}^{n^*} |f_{\hat{\theta}_{n,2}}(y|x_i^*) - \hat{f}_{0,n}(y|x_i^*)|, \]

where \( x_i^* \) is a distinct row covariate-vector value in \( x \); \( n^* \) is the total number of such values; \( f_{\hat{\theta}_{n,2}}(y|x_i^*) \) is a fitted two-component Poisson mixture regression model, where \( \hat{\theta}_{n,2} \) is either \( L_2 E_p \) or MHDE, and \( \hat{f}_{0,n}(y|x_i^*) = N(y|x_i^*)/N(x_i^*) \), where \( N(y|x_i^*) \) is the frequency of \( y \) among \( y_1, \ldots, y_n \) with covariate \( x_i^* \), and \( N(x_i^*) \) is the frequency of \( x_i^* \) in \( x \).

In addition, we consider a chi-square statistic \( \chi^2 \) to examine the adequacy of each fit, which is defined as

\[ \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2}, \]

where \( \hat{\mu}_i \) and \( \hat{\sigma}_i^2 \) are estimates of the conditional mean (\( \mu_i \)) and variance (\( \sigma_i^2 \)), respectively, where \( \mu_i = E(y_i|x_i) = \sum_{j=1}^{m} \pi_j \lambda_{ij} \) and \( \sigma_i^2 = Var(y_i|x_i) = \sum_{j=1}^{m} \pi_j \lambda_{ij}^2 + \mu_i(1 - \mu_i) \), with \( \lambda_{ij} = \exp(x_i^T \beta_j) \). Table 2.8 gives the \( L_1 \) GoF value and chi-square statistic for our \( L_2 E_p \) fit, \( L_2 E \) fit, and the MHDE fit of Lu et al. (2003).

In terms of \( L_1 \) GoF and chi-square statistic criterion, we conclude that our \( L_2 E_p \) fit is much better than the MHDE fit of Lu et al. (2003). We think that MHDE’s poor performance might be due to the fact that MHDE in Lu et al. (2003) was obtained by minimizing the Hellinger distance between the estimated unconditional marginal mass functions (see (3.6) and Remark 1 in Lu et al., 2003) instead of the natural approach of minimizing the Hellinger distance between conditional mass functions. Moreover, the \( L_2 E_p \) fit is slightly better than the \( L_2 E \) fit, which supports our simulation studies of Table 2.2.

Overall, while there are some agreements between the conclusions of our \( L_2 E_p \) analysis and those of Lu et al. (2003) using MHDE regarding the influence of covariates on LOS, there is also an important difference in that we detect the admission type as a significant predictor of LOS in the small subpopulation of long stayers and, more importantly, it helps profile our two clusters of normal-stayers and long-stayers; see Table 2.9. In terms of the fitted model using a robust estimator and better interpretability of factors that influence the
two subpopulations of LOS values, we conclude that our two-component Poisson mixture regression fit based on $L_2E_p$ is superior to the MHDE fit of Lu et al. (2003).

2.8.2 Toronto Motor Vehicle Crashes

In order to study the causes of traffic accidents and reduce the number of traffic accidents and severity of collisions, it is important to develop sound and reliable statistical models for motor vehicle crashes. A crash data analysis is most important in assessing safety of transportation facilities such as intersections or freeways. A measure of traffic conditions used commonly in traffic safety research is the Annual Average Daily Traffic (AADT), which is the total volume of vehicle traffic on a highway or road for a year divided by 365 days. AADT is a useful and simple measurement of how busy the road is. Crash frequency analysis based on AADT is an aggregate way to identify hazardous sites, predict motor vehicle collisions, and develop accident modification factors (Lord and Park, 2008).

Recently, Park and Lord (2009) provided a Bayesian approach that could be used for capturing heterogeneity in a Toronto motor vehicle crash count data through the use of finite mixture regression models. In this section, we reanalyze the crash data at urban four-legged signalized intersections in Toronto, Canada in 1995. Here, the response variable ($y$) is the number of crashes in each of the 868 intersections with each intersection experiencing 0 to 54 crashes. The two covariates involved here are entering flow in vehicles/day from the major-approaches and minor-approaches at each intersection, respectively. These covariates will be referred to as major-approach AADT and minor-approach AADT, respectively. Entering traffic volumes in the data set vary widely from intersections to intersections: from 5469 to 72,178 vehicles/day for major approaches and from 53 to 42,644 vehicles/day for minor approaches. Table 2.12 provides the summary statistics for this data set; for more details see Lord (2000) and Park and Lord (2009).

Park and Lord (2009) claimed that the un-modeled heterogeneity across sites might be caused by the existence of several different latent subgroups. To support the claim, they fitted
Poisson and negative binomial (NB) regression models with different numbers of components (ranging from 1 to 4) using a Bayesian approach and finally concluded that a two-component NB mixture regression model provides a better overall fit than all other models; see Park and Lord (2009) for more details.

Table 2.12: Summary statistics for Toronto motor vehicle crashes data

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crashes</td>
<td>0</td>
<td>54</td>
<td>11.56</td>
<td>10.02</td>
</tr>
<tr>
<td>Major-approach AADT</td>
<td>5469</td>
<td>72178</td>
<td>28044.81</td>
<td>10660.39</td>
</tr>
<tr>
<td>Minor-approach AADT</td>
<td>53</td>
<td>42644</td>
<td>11010.18</td>
<td>8599.40</td>
</tr>
</tbody>
</table>

Table 2.13: $L_2E$ estimation of mixture complexity using $\hat{m}_n^{L_2E}$: Toronto crashes data

<table>
<thead>
<tr>
<th>$m$</th>
<th>$L_2(\theta_m)$</th>
<th>Difference</th>
<th>$\zeta_{n,m}$</th>
<th>$\hat{m}_n^{L_2E}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.070958</td>
<td>0.001701</td>
<td>0.001597</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.072659</td>
<td>$\textbf{0.000346}$</td>
<td>$\textbf{0.000934}$</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>-0.073005</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

First, we computed $\hat{m}_n^{L_2E}$ in (2.14) with the same threshold value as done in Section 2.8.1, which yields $\hat{m}_n^{L_2E} = 2$ (see Table 2.13). Accordingly, we decided to fit a two-component NB mixture regression model using our $L_2E$ approach; that is, $f_j(y_i|\mu_{ij}, \phi_j) = \frac{\Gamma(y_i+\phi_j)}{\Gamma(y_i+1)\Gamma(\phi_j)} \left( \frac{\phi_j}{\phi_j+\mu_{ij}} \right)^{\phi_j} \left( \frac{\mu_{ij}}{\phi_j+\mu_{ij}} \right)^{y_i}$, where $\mu_{ij} = \exp(x_i^T\beta_j)$ and $\phi_j \geq 0$ is an overdispersion parameter. Additionally, the $j$-th component mean function is $\mu_{ij} = \exp[\log(\beta_{j0}) + \beta_{j1}\log(x_{i1}) + \beta_{j2}\log(x_{i2})]$ for $i = 1, \ldots, 868$ and $j = 1, 2$ with $x_{i1}$ and $x_{i2}$ as the major-approach AADT and minor-approach AADT, respectively. Here, $\mu_{ij}$ is the $j$-th component mean number of crashes for intersection $i$. In Table 2.14, we provide a two-component NB mixture regression fit based on our $L_2E$, the Bayesian approach of Park and Lord (2009), and the MLE. We refer to Park and Lord (2009)’s fit as Bayesian Method(NB) in Table 2.14. The Bayesian two-component NB mixture regression fit given in Table 2.14 is a revised
fit that was provided to us by Park and Lord through personal communication. We will compare our $L_2E$ fit with the revised fit of Park and Lord.

Table 2.14: Results of fitting negative binomial mixture regression models to Toronto motor vehicle crashes data

<table>
<thead>
<tr>
<th>Estimation methods</th>
<th>$L_2E$</th>
<th></th>
<th>Bayesian Method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Component</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mixing proportion</td>
<td>0.6191 (0.1761)</td>
<td>0.3809 (0.1761)</td>
<td>0.5875 (0.1715)</td>
<td>0.4125 (0.1715)</td>
</tr>
<tr>
<td>Covariate</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.d.)*</td>
<td>coefficient(s.d.)*</td>
</tr>
<tr>
<td>Intercept</td>
<td>-11.3168 (1.4126)</td>
<td>-9.9411 (1.5146)</td>
<td>-9.8455 (1.1283)</td>
<td>-10.3935 (1.9191)</td>
</tr>
<tr>
<td>Major-approach AADT</td>
<td>0.9191 (0.1713)</td>
<td>0.3229 (0.1732)</td>
<td>0.4237 (0.1497)</td>
<td>0.8176 (0.1850)</td>
</tr>
<tr>
<td>Minor-approach AADT</td>
<td>0.4706 (0.1101)</td>
<td>0.9797 (0.1134)</td>
<td>0.8526 (0.0850)</td>
<td>0.4862 (0.1068)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>6.4858 (1.7511)</td>
<td>29.0378 (32.3970)</td>
<td>11.8048 (8.2181)</td>
<td>9.5775 (8.3046)</td>
</tr>
<tr>
<td>$\mu$ estimate</td>
<td>11.3227</td>
<td>12.0648</td>
<td>11.0918</td>
<td>11.5688</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimation methods</th>
<th>MLE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Component</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mixing proportion</td>
<td>0.7850 (0.1452)</td>
<td>0.2150 (0.1452)</td>
</tr>
<tr>
<td>Covariate</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.e.)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-10.0801 (0.8258)</td>
<td>-10.9368 (2.2594)</td>
</tr>
<tr>
<td>Major-approach AADT</td>
<td>0.5023 (0.0948)</td>
<td>0.9185 (0.2681)</td>
</tr>
<tr>
<td>Minor-approach AADT</td>
<td>0.7897 (0.0536)</td>
<td>0.4550 (0.1011)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>7.3332 (0.9628)</td>
<td>28.4089 (25.2929)</td>
</tr>
<tr>
<td>$\mu$ estimate</td>
<td>10.8164</td>
<td>14.2228</td>
</tr>
</tbody>
</table>

* Park and Lord (2009) provided posterior means and standard deviations for their estimated coefficients.

Our $L_2E$ estimates in Table 2.14 indicate that the mean number of crashes ($\mu$) for component 2 is larger than that of component 1. This is also supported by the average number of crashes for each cluster given in Table 2.15, which classifies each observation into one of the two clusters based on the highest posterior probability calculated using our $L_2E$ fit. Furthermore, in Figure 2.5, we give the frequency histogram of crashes for $i = 1, \ldots, 868$ overlayed with those belonging to clusters 1 and 2 (see Table 2.15) for our $L_2E$ method and the Bayesian method of Park and Lord (2009). For our $L_2E$ method, we see from the left panel of Figure 2.5 that component 1 essentially models crashes $y < 30$, whereas component 2 is more skewed to the right, modeling a high proportion of $y \geq 30$. This once again supports our higher second component mean number of crashes. Nevertheless, the mean of our second component is not severely affected by a small percentage of extreme number of crashes. In
contrast, the histograms based on the Bayesian method (see right panel of Figure 2.5) of Park and Lord (2009) are drastically different from ours in that the second component does not seem to adequately model higher number of crashes.

Table 2.15: Model-based clusters using our $L_2E$ and Bayesian method of Park and Lord (2009)

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$L_2E$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 1: 69.12% (600 obs.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crashes</td>
<td>0</td>
<td>53</td>
<td>10.17</td>
<td>7.75</td>
</tr>
<tr>
<td>Major-approach AADT</td>
<td>5469</td>
<td>56697</td>
<td>26935.90</td>
<td>10107.81</td>
</tr>
<tr>
<td>Minor-approach AADT</td>
<td>53</td>
<td>41029</td>
<td>10314.43</td>
<td>7890.71</td>
</tr>
<tr>
<td>Cluster 2: 30.88% (268 obs.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crashes</td>
<td>0</td>
<td>54</td>
<td>14.67</td>
<td>13.30</td>
</tr>
<tr>
<td>Major-approach AADT</td>
<td>9506</td>
<td>72178</td>
<td>30527.44</td>
<td>11439.66</td>
</tr>
<tr>
<td>Minor-approach AADT</td>
<td>877</td>
<td>42644</td>
<td>12567.84</td>
<td>9844.85</td>
</tr>
<tr>
<td><strong>Bayesian Method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 1: 78.92% (685 obs.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crashes</td>
<td>0</td>
<td>54</td>
<td>11.67</td>
<td>10.57</td>
</tr>
<tr>
<td>Major-approach AADT</td>
<td>5469</td>
<td>72178</td>
<td>27844.29</td>
<td>10782.36</td>
</tr>
<tr>
<td>Minor-approach AADT</td>
<td>546</td>
<td>42644</td>
<td>11473.15</td>
<td>8638.34</td>
</tr>
<tr>
<td>Cluster 2: 21.08% (183 obs.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crashes</td>
<td>1</td>
<td>48</td>
<td>11.14</td>
<td>7.60</td>
</tr>
<tr>
<td>Major-approach AADT</td>
<td>5967</td>
<td>56623</td>
<td>28795.38</td>
<td>10184.37</td>
</tr>
<tr>
<td>Minor-approach AADT</td>
<td>53</td>
<td>36002</td>
<td>9277.19</td>
<td>8246.62</td>
</tr>
</tbody>
</table>

As for the influence of covariates, our $L_2E$ estimates in Table 2.14 show that the major-approach AADT increases the number of crashes in component 1, whereas the minor-approach AADT increases the number of crashes in component 2. In addition, the cluster associations and summary statistics given in Table 2.15 show a striking difference in the average values of major- and minor-approach AADT between clusters 1 and 2. That is, most of the low major- and minor-approach AADT values seem to be assigned to component 1, resulting in lower average values, and many of the high major- and minor-approach AADT values seem to be assigned to component 2. This suggests that the variabilities in major- and minor-approach AADT values are sources of overdispersion. Hence, we may conclude that as
the entering flow in vehicles/day from the major- and minor-approaches at each intersection increases so does the number of crashes.

![Histograms of Crashes and Clusters](image)

**Figure 2.5**: Histograms of crashes and crashes by clusters

### 2.9 Discussion

We have introduced an estimator, $\hat{\theta}^{L_2E}_{n,m}$, for finite mixtures of regression models, which minimizes the integrated $L_2$ distance between *conditional* mass functions. This estimator is easy to compute and less sensitive to the choice of initial values as compared to a MHD estimator obtained by minimizing the Hellinger distance between *unconditional* mass functions of the count response. We have shown that $\hat{\theta}^{L_2E}_{n,m}$ is consistent and asymptotically normal, under regularity conditions for any parametric family of finite mixture regression models.
We have also introduced an $L_2E$ estimator, $\hat{m}_{n_2}^{L_2E}$, of the unknown number of components in a finite mixture of regression model. This estimator is constructed by an information criterion approach based on $L_2$ distance. In addition, it is shown to be strongly consistent under certain regularity conditions.

We have shown using Monte Carlo simulations for finite mixture of Poisson regression models that $\hat{\theta}_{n,m}^{L_2E}$ is highly competitive to the MLE and a better, yet viable, alternative to a MHD estimator, when the model is correctly specified. We have also introduced a penalized $L_2E$, called $L_2E_p$, by adding a penalty function to improve the efficiency of $\hat{\theta}_{n,m}^{L_2E}$ when the true models contain a small mixing proportion and are poorly-separated. In this case, the $L_2E_p$ outperforms the MHDE and MLE. We have also shown that $\hat{\theta}_{n,2}^{L_2E_p}$ is consistent and asymptotically normal, under regularity conditions. Additionally, the $L_2E_p$ estimator, $\hat{m}_{n_2}^{L_2E_p}$, of mixture complexity has been shown to be consistent. When there is gross-error contamination in mixture of Poisson regression models, it has been shown via simulations that the $L_2E$ is considerably more robust than the MHDE and MLE.

Robustness of $L_2E$ against extreme values has been further illustrated using two real data sets, each of which contains a fraction of extreme values. For the 1998/1999 Maternity Length of Hospital Stay data from a Western Australia hospital morbidity database we have used the $L_2E$ and $L_2E_p$ and fitted a two-component Poisson mixture regression model. Here, we have used the $L_2E_p$ because previous studies indicate possible presence of two components with a small mixing proportion for the second component. For the 1995 Motor Vehicle Crash data at urban four-legged signalized intersections in Toronto, Canada, we have fitted a two-component negative binomial mixture regression model, but using $L_2E$. For the maternity LOS and the motor vehicle crash data sets, our conclusions based on $L_2E_p$ and $L_2E$, respectively, have been shown to be quite different from the corresponding ones in the literature. Nevertheless, we do make a compelling case that our two-component Poisson/negative binomial mixture regression models provide a better overall fit, not only in terms of capturing
the heterogeneity in the count responses, but also in terms of interpretability of component profiles via their respective covariates.

Finally, the $L_2E$ approach introduced here for finite mixture of regression models can be easily extended to the continuous response case, but we do not present the details in this dissertation. Since our $L_2E$ approach is not distribution-specific, the method can be applied to a continuous inpatient LOS data as the one considered in Lee et al. (2007). Data sets and computational codes used in this chapter may be obtained from the author.

2.10 Appendix

2.10.1 Proof of Proposition

(i) For every $g \in \tilde{G}$, let $h(\delta_m) = \|g_{\delta_m} - g\|_2$ for $\delta_m \in \Theta_m$. Suppose $\{\delta_{n,m}\} \in \Theta_m$ is any sequence such that $\delta_{n,m} \to \delta_m$ as $n \to \infty$. Then, by the Minkowski’s inequality,

$$|h(\delta_{n,m}) - h(\delta_m)|^2 = \|g_{\delta_{n,m}} - g\|_2 - \|g_{\delta_m} - g\|_2|^2 \leq \|g_{\delta_{n,m}} - g_{\delta_m}\|^2_2.$$ 

Since $g_{\theta_m}(y, x)$ is assumed to be continuous in $\theta_m$ for each pair $(y, x)$, we have that

$$\|g_{\delta_{n,m}} - g_{\delta_m}\|^2_2 = \int \sum_{y=0}^{\infty} |g_{\theta_{n,m}}(y, x) - g_{\theta_m}(y, x)|^2 dx$$

$$\leq \int \sum_{y=0}^{\infty} |f_{\theta_{n,m}}(y|x) - f_{\theta_m}(y|x)|^2 f_X(x) dx$$

$$\leq 2 \int \sum_{y=0}^{\infty} |f_{\theta_{n,m}}(y|x) - f_{\theta_m}(y|x)| f_X(x) dx.$$ 

Let define $W_n(x) = \sum_{y=0}^{\infty} |f_{\theta_{n,m}}(y|x) - f_{\theta_m}(y|x)|$. Then, by the Glick’s theorem (Devroye and Györfi, 1985, p.10) and the dominated convergence theorem, $W_n(x)f_X(x) \to 0$, and then $|h(\delta_{n,m}) - h(\delta_m)|^2 \to 0$. Thus, $h(\delta_m)$ is continuous in $\delta_m$. This, together with the assumptions made in Proposition, implies that $T^{L_2E}_m(g)$ exists. Proofs of parts (ii) and (iii) are similar to those in Theorem 1 of Beran (1977). Hence, we omit the details. □
2.10.2 Proof of Theorem 1

(i) The proof of part (i) of the Theorem follows along the lines of proof of Theorem 5.1 of Lehmann and Casella, 1998; Chapter 6. Therefore, we omit the proof of part (i).

As for (ii), let \( H_n(\theta_m) \) (henceforth denoted as \( H(\theta_m) \)) be the objective function in (2.6); that is,

\[
H(\theta_m) = n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y | X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i | X_i) = n^{-1} \sum_{i=1}^{n} V_{\theta_m}(Y_i, X_i).
\]

Now we expand \( H_r'(\theta_m) = \frac{\partial H(\theta_m)}{\partial \theta_{m,r}} \) about the true value \( \delta_m \) to get

\[
H_r'(\theta_m) = H_r'(\delta_m) + \frac{1}{2} \sum_{s=1}^{k} \sum_{t=1}^{k} (\theta_{m,s} - \delta_{m,s})(\theta_{m,t} - \delta_{m,t}) H_r''(\delta_m),
\]

where \( H_r', H_r'', \) and \( H_r''' \) indicate the first, second, and third derivatives of \( H \), and \( \delta_m^* \) is between \( \theta_m \) and \( \delta_m \), and \( k = (p + 2)m - 1 \). Replacing \( \theta_m \) by \( \hat{\theta}_{L^2E}^{n,m} \) in (2.22) and noting that \( H_r'(\hat{\theta}_{L^2E}^{n,m}) = 0 \), we obtain

\[
-n^{1/2}H_r'(\delta_m) = n^{1/2} \sum_{s=1}^{k} (\hat{\theta}_{L^2E}^{n,m,s} - \delta_{m,s}) \left[ H_r''(\delta_m) + \frac{1}{2} \sum_{t=1}^{k} (\hat{\theta}_{L^2E}^{n,m,t} - \delta_{m,t}) H_r'''(\delta_m^*) \right].
\]

This is of the form

\[
\sum_{s=1}^{k} A_{rsn} \Psi_{sn} = \Upsilon_{rn},
\]

with \( A_{rsn} = \left[ H_r''(\delta_m) + \frac{1}{2} \sum_{t=1}^{k} (\hat{\theta}_{L^2E}^{n,m,t} - \delta_{m,t}) H_r'''(\delta_m^*) \right] \), \( \Psi_{sn} = n^{1/2}(\hat{\theta}_{L^2E}^{n,m,s} - \delta_{m,s}) \), and \( \Upsilon_{rn} = -n^{1/2}H_r'(\delta_m) \). By Lemma 5.2 of Lehmann and Casella (1998); Chapter 6, the solutions \( (\Psi_1, \ldots, \Psi_k) \) of (2.23) tend in law to the solutions of

\[
\sum_{s=1}^{k} a_{rs} \Psi_s = \Upsilon_r, \text{ for } r = 1, \ldots, k,
\]

where \( (\Upsilon_1, \ldots, \Upsilon_k) \) tends weakly to \( (\Upsilon_1, \ldots, \Upsilon_k) \), and for each fixed \( r \) and \( s \), \( A_{rsn} \) tends in probability to constants \( a_{rs} \) for which the matrix \( A = ((a_{rs})) \) is nonsingular. The solution \( (\Psi_1, \ldots, \Psi_k) \) is given by

\[
\Psi_r = \sum_{s=1}^{k} b_{rs} \Upsilon_r,
\]

where \( b_{rs} = \frac{1}{a_{rs}} \).
where $A^{-1} = ((b_{rs}))$. In this case, $(\Upsilon_{1n}, \ldots, \Upsilon_{kn})$ has an asymptotic multivariate normal distribution. Also, since the true parameter $\delta_m$ must minimize the expression on the right side of (2.5); [see Proposition (iii) above], we have that

$$2 \int \sum_{y=0}^{\infty} f_{\delta_m}^2 (y|x) u_{\delta_m}^{(r)} (y|x) f_X(x) dx - 2 \int \sum_{y=0}^{\infty} f_{\delta_m} (y|x) u_{\delta_m}^{(r)} (y|x) f_0(y|x) f_X(x) dx = 0,$$

where $u_{\delta_m}^{(r)} (y|x_i) = \partial \{ \log f_{\theta_m} (y|x_i) \} / \partial \theta_m, r = 1, \ldots, k$. Then from (2.21)

$$E \left( V_{\delta_m}^{(r)} (Y, X) \right) = 2 \int \sum_{y=0}^{\infty} f_{\delta_m}^2 (y|x) u_{\delta_m}^{(r)} (y|x) f_X(x) dx$$

$$- 2 \int \sum_{y=0}^{\infty} f_{\delta_m} (y|x) u_{\delta_m}^{(r)} (y|x) f_0(y|x) f_X(x) dx = 0,$$

where $V_{\delta_m}^{(r)}$ is the derivative of $V_{\theta_m}$ with respect to $\theta_m, r$ evaluated at $\theta_m = \delta_m$, and therefore from (2.23) and (2.24), the limiting distribution of $(\Upsilon_{1n}, \ldots, \Upsilon_{kn})$ has a zero mean vector.

Now the covariance matrix of $(\Upsilon_1, \ldots, \Upsilon_k)$ is the covariance matrix of the vector formed by the elements $-2 f_{\delta_m} (Y|X) u_{\delta_m}^{(r)} (Y|X), r = 1, \ldots, k$, and it can be easily shown that this covariance matrix is $4 K(\delta_m)$, where $K(\delta_m)$ is defined in condition (C.4) above. Thus, from (2.25) the limiting distribution of $(\Psi_{1n}, \ldots, \Psi_{kn})$ is multivariate normal with mean vector zero and covariance matrix $4 A^{-1} K(\delta_m) A^{-1}$.

Finally, we will show that $A = 2 J(\delta_m)$. From condition (C.5) and the consistency of $\hat{\theta}_{n,m}^{L_2} \to \sum_{t=1}^{k} (\hat{\theta}_{n,m,t} - \delta_m, t) H_{rs}''(\delta_m^*) \xrightarrow{p} 0$, so that in this case $a_{rs}$ in (2.24) is simply the limit of $H_{rs}''(\delta_m)$. Additionally, $H_{rs}''(\theta_m) / 2$ converges to

$$J_{rs} (\theta_m) = \int \left[ \sum_{y=0}^{\infty} u_{\theta_m}^{(r)} (y|x) u_{\theta_m}^{(s)} (y|x) f_{\theta_m}^2 (y|x) \right] f_X(x) dx$$

$$+ \int \left[ \sum_{y=0}^{\infty} \{ u_{\theta_m}^{(r)} (y|x) - u_{\theta_m}^{(r)} (y|x) u_{\theta_m}^{(s)} (y|x) \} \{ f_0(y|x) - f_{\theta_m} (y|x) \} f_{\theta_m} (y|x) \right] f_X(x) dx.$$

Note that $J(\theta_m) = ((J_{rs} (\theta_m)))$ is the same as the matrix in (2.11). Thus $H_{rs}''(\delta_m)$ converges to $2 J(\delta_m)$ in probability, which completes the proof.
2.10.3 Proof of Theorem 2

Since \( \hat{g}_n(y, x) \to g_0(y, x) \) a.s. for each \((y, x)\), from the Glick’s theorem and the dominated convergence theorem

\[
\| \hat{g}_n - g_0 \|_2^2 \to 0 \text{ a.s.} \tag{2.26}
\]

Henceforth, we will suppress “a.s.”, as it will be clear from the context. Define \( h_n(\delta) = \| g_\delta - \hat{g}_n \|_2 \) and \( h_0(\delta) = \| g_\delta - g_0 \|_2 \). Then, by the Minkowski’s inequality we have \( |h_n(\delta) - h_0(\delta)| \leq \| \hat{g}_n - g_0 \|_2 \). Hence, by (2.26)

\[
\sup_{\delta} |h_n(\delta) - h_0(\delta)| \leq \| \hat{g}_n - g_0 \|_2 \to 0.
\]

Let \( \theta_{0,m} = T^{L_2E}(g_0) \) and \( \hat{\theta}_{n,m}^{L_2E} = T^{L_2E}(\hat{g}_n) \). Then, it is possible to show that

\[
\left| \min_{\delta \in \Theta_m} h_n(\delta) - \min_{\delta \in \Theta_m} h_0(\delta) \right| \to 0. \tag{2.27}
\]

That is, \( \| \hat{g}_n - g_{\hat{\theta}_{n,m}^{L_2E}} \|_2 \to \| g_0 - g_{\theta_{0,m}} \|_2 \). Therefore,

\[
L_2(g_{\hat{\theta}_{n,m}^{L_2E}}, \hat{g}_n) - L_2(g_{\hat{\theta}_{n,m+1}^{L_2E}}, \hat{g}_n) \to L_2(g_{\theta_{0,m}}, g_0) - L_2(g_{\theta_{0,m+1}}, g_0). \tag{2.28}
\]

Note from (2.8) and (2.28) that

\[
m_0 = \min\{ m : d_m = L_2(g_{\theta_{0,m}}, g_0) - L_2(g_{\theta_{0,m+1}}, g_0) \leq 0 \}.
\]

If \( g_0 \) is not a finite mixture, then \( m_0 = \infty \), which implies that \( d_m > 0 \) for all \( m > 0 \). Thus, \( \hat{m}_n \to \infty \) by (2.10) and (2.28). When \( g_0 \) is finite; that is, \( g_0 = g_{\theta_{m_0}} \), we now show that \( d_m > 0 \) for \( m < m_0 \) and \( d_m = 0 \) for \( m \geq m_0 \).

Suppose that \( m \geq m_0 \). Then since \( g_0 \in \mathcal{G}_{m_0} \subseteq \mathcal{G}_j \) (that is, \( f_0 \in \mathcal{F}_{m_0} \subseteq \mathcal{F}_j \)), \( j \geq m_0 \)

\[
L_2(g_{\theta_{m,j}^{L_2E}}, \hat{g}_n) \leq L_2(g_{\theta_{m_0}}, \hat{g}_n) \to 0, \tag{2.29}
\]

by (2.26). Thus, by (2.28) we have that \( d_m = 0 \) for \( m \geq m_0 \).

Suppose that \( m < m_0 \). Then, \( g_0 \in \mathcal{G}_{m_0} \) (\( f_0 \in \mathcal{F}_{m_0} \), but \( g_0 \notin \mathcal{G}_m \) (\( f_0 \notin \mathcal{F}_m \)). Suppose, on the contrary, \( d_m = 0 \) for some \( m < m_0 \); that is, \( L_2(g_{\theta_{0,m}}, g_0) = L_2(g_{\theta_{0,m+1}}, g_0) \). Then, for all \( \delta_{m+1} \in \Theta_{m+1} \)

\[
L_2(g_{\theta_{0,m}}, g_0) \leq L_2(g_{\theta_{m+1}}, g_0). \tag{2.30}
\]
For an arbitrary $\epsilon \in (0,1)$ and $\lambda$, let $g_{\delta_{m+1}}(y,x) = (1 - \epsilon)g_{\theta_{0,m}}(y,x) + \epsilon g_{\lambda}(y,x)$, where $g_{\lambda}(y,x) = f_{\lambda}(y|x)f_x(x)$. Then, $g_{\delta_{m+1}} \in \mathcal{G}_{m+1}$ ($f_{\delta_{m+1}} \in \mathcal{F}_{m+1}$) and from (2.30)
\[ \int \sum_{y=0}^{\infty} \left| g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right|^2 dx - \int \sum_{y=0}^{\infty} \left| (1 - \epsilon)g_{\theta_{0,m}}(y,x) + \epsilon g_{\lambda}(y,x) - g_{\lambda}(y,x) \right|^2 dx \leq 0. \]

After some algebra, we have
\[ \epsilon \int \sum_{y=0}^{\infty} \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] \left[ 2 \left\{ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right\} + \epsilon \left\{ g_{\lambda}(y,x) - g_{\theta_{0,m}}(y,x) \right\} \right] dx \leq 0, \]
which implies
\[ 2\epsilon \int \sum_{y=0}^{\infty} \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] dx \leq \epsilon^2 \int \sum_{y=0}^{\infty} \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right]^2 dx. \quad (2.31) \]

Dividing both sides of (2.31) by $\epsilon$ and letting $\epsilon \to 0$ we get
\[ \int \sum_{y=0}^{\infty} \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] dx \leq 0, \]
which implies
\[ \int \sum_{y=0}^{\infty} g_{\theta_{0,m}}(y,x) \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] dx \leq \int \sum_{y=0}^{\infty} g_{\lambda}(y,x) \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] dx. \quad (2.32) \]

Since $g_0 \in \mathcal{G}_{m_0}$ ($f_0 \in \mathcal{F}_{m_0}$), $g_0(y,x) = \int \sum_{y=0}^{\infty} f_0(y|x)f_x(x)dx$, where $f_0(y|x) = \sum_{j=1}^{m_0} \pi_{0j} f_j(y|x^0)$, and (2.32) holds for each $\lambda = \lambda_{ij}^0$, $i = 1, \ldots, n$ and $j = 1, \ldots, m_0$. Since $\sum_{j=1}^{m_0} \pi_{0j}^i = 1$, from (2.32)
\[ \int \sum_{y=0}^{\infty} g_{\theta_{0,m}}(y,x) \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] dx \leq \int \sum_{y=0}^{\infty} g_0(y,x) \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right] dx, \]
which implies that $\int \sum_{y=0}^{\infty} \left[ g_{\theta_{0,m}}(y,x) - g_{\lambda}(y,x) \right]^2 dx = 0$. This contradicts the fact that $g_0 \notin \mathcal{G}_m$ ($f_0 \notin \mathcal{F}_m$) for $m < m_0$. Therefore, $d_m = 0$ for $m < m_0$. This completes the proof. □
2.10.4 Proof of Corollary 1

(i) The proof of part (i) of the Corollary is identical to that of Theorem 1; hence is omitted.

(ii) The proof of part (ii) of the Corollary is almost identical to that of Theorem 1, but

\[
H^*(\theta_m) = n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y|X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i|X_i) + p(\pi)
\]

\[
= n^{-1} \sum_{i=1}^{n} V_{\theta_m}(Y_i|X_i) + p(\pi),
\]

and \(4K^* = 4K^*(\theta_2)\) is the covariance matrix of the vector formed by the elements

\[-2f_{\delta_2}(Y|X)u_{\delta_2}(r)(Y|X) + p^{(r=1)}(\delta_2,(r=1)), \quad r = 1, \ldots, k,\]

with \(p^{(r=1)}(\delta_2,(r=1)) = \partial p(\delta_2,(r=1))/\partial \delta_2,(r=1)\), where \(\delta_2,(r=1)\) is a true \(\pi_1\). The rest of the proof is analogous to that of Theorem 1. \(\square\)

2.10.5 \(L_2E\) Estimation of Mixture Components, \(m\)

In Sections 2.4 and 2.6, we introduced an estimator of \(m\) defined by

\[
\hat{m}_n^{L_2E} = \min\{m : L_2(\hat{\theta}_{n,m}^{L_2E}) \leq L_2(\hat{\theta}_{n,m+1}^{L_2E}) + \zeta_{n,m}\},
\]

where \(L_2(\theta_m) = \left[ n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y|X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i|X_i) \right] \) and the threshold \(\zeta_{n,m} = (\kappa/865)\ln[(m + 1)/m]\) is a decreasing function of \(m\) for \(\kappa > 0\). In Section 2.7.4, we also defined the \(L_2E_p\) estimator of \(m\) as

\[
\hat{m}_n^{L_2E_p} = \min\{m : L_{2p}(\hat{\theta}_{n,m}^{L_2E_p}) \leq L_{2p}(\hat{\theta}_{n,m+1}^{L_2E_p}) + \zeta_{n,m}\}, \quad (2.33)
\]

where \(L_{2p}(\theta_m) = \left[ n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^2(y|X_i) - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(Y_i|X_i) + p(\pi) \right] \)

with the penalty function \(p(\pi) = \min\{\tau\tanh[\eta\{\gamma - (0.5 - \pi_1)\}], \ldots, \tau\tanh[\eta\{\gamma - (0.5 - \pi_m)\}]\}\)

for some suitable constants \(\tau, \eta,\) and \(\gamma\). For these threshold values, we used same numbers as in Section 2.7.1; that is, \(\tau = 0.001, \eta = 4,\) and \(\gamma = 0.3.\)
The computation of $\hat{m}_n^{L_2E}$ and $\hat{m}_n^{L_2E_p}$ are iterative (see Section 2.6) and Table 2.16 gives all the steps leading to $\hat{m}_n^{L_2E} = \hat{m}_n^{L_2E_p} = 2$, when the threshold value $\zeta_{n,m} = (2/865) \ln[(m + 1)/m]$. To compute $\hat{m}_n^{L_2E}$, for example, we first calculate $L_2(\hat{\theta}_n^{L_2E}, \hat{\theta}_n^{L_2E}) = -0.099076$ and $L_2(\hat{\theta}_n^{L_2E}) = -0.102208$, and compute the Difference $= L_2(\hat{\theta}_n^{L_2E}, \hat{\theta}_n^{L_2E}) = 0.003132$, which is then compared to $\zeta_{n,1} = 0.001603$. Since $0.003132 > 0.001603$, we fit a 3-component Poisson mixture regression for which $L_2(\hat{\theta}_n^{L_2E}, \hat{\theta}_n^{L_2E}) = -0.103130$. Once again, the Difference $= L_2(\hat{\theta}_n^{L_2E}, \hat{\theta}_n^{L_2E}) = 0.000922 < \zeta_{n,2} = 0.000937$. Thus, the procedure stops at $\hat{m}_n^{L_2E} = 2$.

Similarly, we conclude using the values under $L_2E_p$ reported in Table 2.16 that $\hat{m}_n^{L_2E_p} = 2$.

Table 2.16: $L_2E$ and $L_2E_p$ estimation of mixture complexity using $\hat{m}_n^{L_2E}$ and $\hat{m}_n^{L_2E_p}$, respectively, in Section 2.8.1

<table>
<thead>
<tr>
<th>$m$</th>
<th>$L_2(\theta_m)$</th>
<th>Difference</th>
<th>$\zeta_{n,m}$</th>
<th>$\hat{m}_n^{L_2E}$</th>
<th>$L_2E_p(\theta_m)$</th>
<th>Difference</th>
<th>$\zeta_{n,m}$</th>
<th>$\hat{m}_n^{L_2E_p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.099076</td>
<td>0.003132</td>
<td>0.001603</td>
<td>-0.099076</td>
<td>0.003104</td>
<td>0.001603</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.102208</td>
<td><strong>0.000922</strong></td>
<td><strong>0.000937</strong></td>
<td>2</td>
<td>-0.102180</td>
<td><strong>0.000665</strong></td>
<td><strong>0.000937</strong></td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>-0.103130</td>
<td>-0.102845</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.10.6 Comparison of Two-Component Poisson Mixture Regression Fits using $L_2E_p$ and MLE

For the Maternity LOS data, we used the MLE fit provided by Lu et al. (2003) to classify the entire sample into two clusters, as done for $L_2E_p$ in Table 2.9. Table 2.17 compares the two-component Poisson mixture regression fit using $L_2E_p$ (from Table 2.8) and the MLE, while Table 2.18 gives the model-based clusters using the MLE.

In terms of the two summary measures (LOS component means and the $L_1$ Goodness-of-fit (GoF) value), the MLE is not only sensitive to extreme LOS values (the estimated second component LOS mean of MLE is 13.2729 while that of $L_2E_p$ is 10.1938 and $L_2E$ is 9.659), but also worse than the $L_2E_p$ (or $L_2E$) in terms of the $L_1$ GoF value. On the contrary, in terms of the chi-square statistic value, $\chi^2$, the MLE performs much better than the $L_2E_p$. 
and $L_2E$, which is not surprising because the chi-square criterion favors likelihood methods. The mixing proportion estimates of the MLE and $L_2Ep$ are also slightly different.

Table 2.17: Two-component Poisson mixture regression fit using $L_2Ep$ vs. MLE: LOS data

<table>
<thead>
<tr>
<th>Estimation methods</th>
<th>$L_2Ep$</th>
<th></th>
<th>MLE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Component</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mixing proportion</td>
<td></td>
<td>0.1867 (0.0320)</td>
<td>0.8569 (0.0201)</td>
<td>0.1431 (0.0201)</td>
</tr>
<tr>
<td>Covariate</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.e.)</td>
<td>coefficient(s.e.)</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.5115 (0.0318)</td>
<td>2.0000 (0.0178)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td>0.0445 (0.0361)</td>
<td>0.2126 (0.0518)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission type</td>
<td>0.1657 (0.0336)</td>
<td>0.7873 (0.0273)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td>0.2073 (0.0655)</td>
<td>0.1576 (0.1330)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>0.1074 (0.0463)</td>
<td>0.0924 (0.0638)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td>−0.0171 (0.0359)</td>
<td>0.1545 (0.0465)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td>0.0490 (0.0426)</td>
<td>0.6404 (0.0603)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$ estimate</td>
<td>4.9667</td>
<td>10.1938</td>
<td>4.9973</td>
<td>13.2729</td>
</tr>
<tr>
<td>$L_1$ GoF</td>
<td>42.1963</td>
<td>42.4517</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>2385.72</td>
<td>1695.337</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In other words, the ML estimate of the second component mixing proportion is relatively smaller than that of the $L_2Ep$. Nevertheless, comparison of the sizes of cluster 2 in Table 2.9 and Table 2.18, respectively, shows that the subpopulation of long-stayers for the MLE fit is about 0.8 times as large as that of the $L_2Ep$ fit, which implies the $L_2Ep$ properly protects the second component. Interestingly, both the $L_2Ep$ fit and the MLE fit of the second component indicate that the maternity LOS is positively and significantly related to the admission type and the indigenous status. However, in Table 2.18, none of the covariates help profile the two clusters obtained using the MLE fit. In contrast, we noted from Table 2.9 that the admission type, which was found to be highly significant in the second component by our $L_2Ep$, helps profile the two clusters in that cluster 1 (normal-stayers) predominantly consists of planned patients, whereas emergency patients dominate cluster 2 (long-stayers). In addition, the right panel of Figure 2.6 shows that the MLE definitely splits a whole LOS into two components compared with the left panel for $L_2Ep$. From these, we conclude that our $L_2Ep$ criterion also
plays a crucial role in elucidating the factors that influence the predominant subpopulation of normal-stayers and the small subpopulation of long-stayers.

Table 2.18: Model-based clusters using the MLE

<table>
<thead>
<tr>
<th>Cluster 1: 90.29% (781 obs.)</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOS</td>
<td>1</td>
<td>16</td>
<td>5.0435</td>
<td>2.3962</td>
</tr>
<tr>
<td>Percentage of 1</td>
<td>Percentage of 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td>27.40% (214 obs.)</td>
<td>72.60% (567 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission type</td>
<td>32.52% (254 obs.)</td>
<td>67.48% (527 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td>5.38% (42 obs.)</td>
<td>94.62% (739 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>14.08% (110 obs.)</td>
<td>85.92% (671 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td>38.16% (298 obs.)</td>
<td>61.84% (483 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td>11.14% (87 obs.)</td>
<td>88.86% (694 obs.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 2: 9.71% (84 obs.)</th>
<th>Min</th>
<th>Max</th>
<th>Ave</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOS</td>
<td>10</td>
<td>45</td>
<td>17.4048</td>
<td>8.5519</td>
</tr>
<tr>
<td>Percentage of 1</td>
<td>Percentage of 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td>22.62% (19 obs.)</td>
<td>77.38% (65 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission type</td>
<td>39.29% (33 obs.)</td>
<td>60.71% (51 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Payment class</td>
<td>4.76% (4 obs.)</td>
<td>95.24% (80 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>11.90% (10 obs.)</td>
<td>88.10% (74 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td>34.52% (29 obs.)</td>
<td>65.48% (55 obs.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td>15.48% (13 obs.)</td>
<td>84.52% (71 obs.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.6: Histograms of LOS and LOS by clusters ($L_2E_p$ vs MLE)
2.11 References


Chapter 3

$L_2E$ with Penalty and Non-normal Mixtures

Abstract

For continuous data, we revisit the problem of robust estimation of mixture complexity using $L_2E$. It was illustrated via extensive simulations in Umashanger (2009) that the $L_2E$ approach yields an estimator, $\hat{m}_{\text{L}_2E}$, of mixture complexity which correctly detects the true mixture complexity even when the underlying model is misspecified. However, when there is a small fraction of plausible extreme values in the data, the inherent robustness feature of $\hat{m}_{\text{L}_2E}$ often hinders correct detection of mixture complexity. In order to overcome this limitation, we propose a modified $L_2$ criterion with a penalty function, which yields another estimator of mixture complexity. Monte Carlo simulations show that our estimator based on a modified $L_2$ method correctly detects true mixture complexity even when there is a small component in finite mixtures. In addition, we also illustrate the usefulness of $\hat{m}_{\text{L}_2E}$ for correct detection of mixture complexity in non-normal mixture models. Real data sets on Red blood cell sodium-lithium countertransport (SLC) activity and Italian surgical procedure times are analyzed to illustrate the performance of $L_2E$ for normal and non-normal mixtures, respectively. We show that the use of the modified $L_2E$ estimator to analyze SLC data yields a three-component normal mixture that most closely resembles an additive (genetic) model amongst all other models proposed thus far for this data and for Italian surgical procedure times data two-component gamma/lognormal mixture model based on $L_2E$ method provides a good overall fit.

Key words and Phrases: Finite mixtures; Information criterion; Mixture complexity; Penalty; Robustness; Threshold.
3.1 Introduction

Finite mixture models provide a natural way of modeling data that come from populations consisting of a finite, but possibly unknown, number of homogeneous subpopulations. Examples of such data sets are abundant in many applications such as biological, physical and social sciences. For this reason, finite mixture models have been used in a variety of applications for more than a hundred years. The literature surrounding these models is vast and traces back to 1894 when Karl Pearson first published an article on the decomposition of two-component normal mixtures using the method of moments. Ever since, normal mixtures have been the most widely used models in many disciplines. For example, Cassie (1954) studied the length distribution of a certain type of fish using normal mixture models and found that it is reasonable to divide observations into age categories. Gregor (1969) applied normal mixtures to data arising from measuring the content of DNA in the nuclei of liver cells of rats. Mundry (1972) applied mixtures of normal distributions to the analysis of pollen grains data. Holgersson and Jorner (1976) studied normal mixtures on fiber size distributions of myelinated lumbosacral ventral root fibers from kittens of various ages. Venkataraman (1997) applied normal mixture models to construct the Value at Risk measure and showed that his model provides a reasonable fit for the data via empirical studies. More recently, mixture models of normal distributions have been applied in bioinformatics and genetics; for example, Delmar, Robin, Tronik-Le Roux, and Daudin (2005) and Tadesse, Sha, and Vannucci (2005).

As mentioned above, data from normal mixtures occur frequently in a wide range of areas such as biology, economics, marketing, medicine, and physics. However, there are circumstances when the normal mixture is clearly unsuitable. Commonly, these are situations where the intrinsic range of the variable is the nonnegative real line. Moreover, the data being modeled may exhibit skewness and multimodality, and in addition, there may be presence of heavy tail in the observed data. For example, health service researchers frequently study the length of stay (LOS) in hospital as a continuous health outcome. The LOS data
usually originates from heavily skewed distributions as we saw in the count LOS data case (see Figure 2.3 in Section 2.8.1). Failure data analysis, hazard rates modelings, and analysis in duration setups are also other popular application contexts where finite mixture of non-normal distributions arise naturally. In these instances, it may be preferable to consider alternative mixture distributions such as exponential, gamma, or lognormal densities for the data.

Literature on non-normal mixtures is also vast. For instance, Davis (1952) and Mendenhall and Hader (1958) fitted exponential components to the failure distribution of electronic valves. Dey, Kuo, and Sahu (1995) compared mixtures of normal and gamma distributions to describe their Bayesian mixture modeling. Wiper, Insua, and Ruggeri (2001) studied Bayesian inference using gamma mixture models. Liu, Pasarica, and Shao (2003) developed the likelihood ratio test statistic for testing homogeneity against gamma mixture alternatives. Recently, Wijaya, Harada, and Horton (2008) used gamma and lognormal mixtures to model skewed microarray data.

As has been pointed out in the literature and in this thesis, the $L_2E$ approach attempts to find estimates which provide an appropriate fit for the largest fraction of data possible (Scott, 1999). In particular, this phenomenon becomes more pronounced when there is a small fraction of extreme values in a data set. In order to improve the performance of $L_2E$ in estimating a small fraction of extreme values in univariate continuous data, we resort to the modified $L_2E$ criterion introduced in Section 2.7.1 which adds a penalty function to the basic $L_2E$ objective function.

In Section 3.2, we revisit the problem of mixture complexity estimation considered recently by Umashanger (2009); also see Umashanger and Sriram (2009). For continuous data, we modify the $L_2E$ criterion considered in Umashanger (2009) and propose a modified estimator of mixture complexity. Specifically, we propose a criterion, called $L_2E_p$, by introducing a penalty function to the $L_2E$ criterion considered in Umashanger (2009) for the
purpose of detecting small components. Computational details concerning our estimator are
given in Section 3.3.

In Section 3.4, we numerically assess the performance of $L_2 E$ and our $L_2 E_p$ estimators of
mixture complexity in detecting the number of components in gamma mixtures. Section 3.5
illustrates that the $L_2 E_p$ not only detects a small fraction of extreme values in a red blood
cell sodium-lithium countertransport (SLC) activity data, but also yields the best fitting
three-component mixture that satisfies a genetic (additive) model (Roeder, 1994). Moreover,
for an Italian surgical procedure times data collected at a hospital in Genoa, Italy, containing
extreme values we determine a $L_2 E$ estimate of the mixture complexity and provide fitted
two-component gamma mixture and lognormal mixture densities. We conclude that our
two-component gamma mixture provides a slightly better fit. An overall summary is given
in Section 3.6.

3.2 $L_2 E$ Estimator

For each $1 \leq m < \infty$, consider a parametric family $\mathcal{F}_m = \{F_{\theta_m} : \theta_m \in \Theta_m \subseteq \mathbb{R}^p\}$ of
distribution functions such that $F_{\theta_m}$ can be represented as a finite mixture of the form

$$
F_{\theta_m}(x) = \sum_{j=1}^{m} \pi_j F(x|\phi_j), \quad x \in \mathcal{X} \subseteq \mathbb{R},
$$

and $\theta_m = (\pi_1, \ldots, \pi_{m-1}, \phi_1^T, \ldots, \phi_m^T)^T$. The class $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$ for all $m$ and we denote
$\mathcal{F} = \bigcup_{m=1}^{\infty} \mathcal{F}_m$. For each $m$, let $f_{\theta_m}(x)$ denote the mixture density function corresponding to
$F_{\theta_m}(x)$ with component densities denoted by $f(x|\phi_j)$, for $j = 1, \ldots, m$. That is,

$$
f_{\theta_m}(x) = \sum_{j=1}^{m} \pi_j f(x|\phi_j). \quad (3.1)
$$

Note that, in this chapter we once again use notation introduced in Chapter 2; for example,
$f_{\theta_m}$ once again denotes the parametric mixture density, but there are no covariates present
here.

Let $X_1, \ldots, X_n$ be independent and identically distributed continuous random variables
with an unknown distribution $F_0$ with corresponding density function $f_0$. For an arbitrary
distribution $G$, define the index of the economical representation of $G$, relative to the family of mixtures $\mathcal{F}_m$, as $m(G) = \min\{m : G \in \mathcal{F}_m\}$. If indeed $G$ is a finite mixture then $m(G)$ is finite and denotes the true mixture complexity; otherwise $m(G) = \infty$. Note that $m(G)$ represents the most parsimonious mixture model representation for $G$. Henceforth, we let $m_0 = m(F_0)$. Our goal is to construct an estimator of $m_0$, which is a viable, robust alternative to the MHDE of Woo and Sriram (2006) and computationally much simpler than the latter.

Given the true density $f_0$ and the finite mixture $f_{\theta_m}$, Scott (1998) considered the squared $L_2$ distance between $f_{\theta_m}$ and $f_0$ and derived an estimate of $\theta_m$ as described below. Let

$$L_2(f_{\theta_m}, f_0) = \int_{-\infty}^{\infty} [f_{\theta_m}(x) - f_0(x)]^2 dx. \quad (3.2)$$

Expanding this gives

$$L_2(f_{\theta_m}, f_0) = \int_{-\infty}^{\infty} f_{\theta_m}^2(x) dx - 2 \int_{-\infty}^{\infty} f_{\theta_m}(x) f_0(x) dx + \int_{-\infty}^{\infty} f_0^2(x) dx, \quad (3.3)$$

where the last integral is a constant with respect to $\theta_m$ and therefore can be ignored. The first integral in (3.3) is often available as a closed form expression that may be evaluated for any specified value of $\theta_m$. The second integral in (3.3) is $-2 \int_{-\infty}^{\infty} f_{\theta_m}(x) dF_0(x)$, which can be estimated by $-2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(X_i)$. Combining, the $L_2E$ estimator of $\theta_m$ is given by

$$\hat{\theta}_{m,n}^{L_2E} = \arg \min_{\theta_m} \left[ \int_{-\infty}^{\infty} f_{\theta_m}^2(x) dx - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(X_i) \right], \quad (3.4)$$

Note that, for $t_m \in \Theta_m$, if we let

$$L(t_m, F) = \left[ \int_{-\infty}^{\infty} f_{t_m}^2(x) dx - 2 \int_{-\infty}^{\infty} f_{t_m}(x) dF(x) \right] \quad (3.5)$$

and define a $L_2E$ functional $T_m^{L_2E}$ on $\mathcal{F}$ by the requirement that for every $F \in \mathcal{F}$

$$T_m^{L_2E}(F) = \{ \theta_m \in \Theta_m : L(\theta_m, F) = \min_{t_m \in \Theta_m} L(t_m, F) \}. \quad (3.6)$$

Then,

$$L(\theta_m, \hat{F}_n) = \left[ \int_{-\infty}^{\infty} f_{\theta_m}^2(x) dx - 2n^{-1} \sum_{i=1}^{n} f_{\theta_m}(X_i) \right], \quad (3.7)$$
where \( \hat{F}_n \) is the empirical distribution of \( \{X_i, i = 1, \ldots, n\} \), and hence

\[
\hat{\theta}_{n,m}^{L_2E} = T_m^{L_2E}(\hat{F}_n) = \arg \min_{\theta_m} L(\theta_m, \hat{F}_n),
\] (3.8)

with \( L(\hat{\theta}_{n,m}, \hat{F}_n) = \min_{\theta_m} L(\theta_m, \hat{F}_n) \).

Now, to construct an estimator of \( m_0 \), we proceed as in Woo and Sriram (2006, Section 2) and introduce a model selection criterion based on \( L(\hat{\theta}_{n,m}^{L_2E}, \hat{F}_n) \) defined by

\[
LIC = L(\hat{\theta}_{n,m}^{L_2E}, \hat{F}_n) + n^{-1}b(n)\nu(m),
\]

where \( b(n) \) depends only on \( n \) and \( \nu(m) \) is the number of parameters in the mixture model.

Here, the value of \( m \) yielding the minimum \( LIC \) specifies the best model. Since \( \mathcal{F}_m \subseteq \mathcal{F}_{m+1} \), we have \( L(\hat{\theta}_{n,m}^{L_2E}, \hat{F}_n) \geq L(\hat{\theta}_{n,m+1}^{L_2E}, \hat{F}_n) \). Therefore, we penalize the goodness-of-fit statistic by a term proportional to the number of parameters in the mixture model. A simple heuristic to search for the best model from a sequence of nested models is to try successive models, starting with the smallest, and stop with model \( m \) when the \( LIC \) value for model \( m \) is less than that for model \( (m + 1) \). That is, this heuristic stops

\[
L(\hat{\theta}_{n,m}^{L_2E}, \hat{F}_n) + n^{-1}b(n)\nu(m) \leq L(\hat{\theta}_{n,m+1}^{L_2E}, \hat{F}_n) + n^{-1}b(n)\nu(m + 1)
\]

or, equivalently,

\[
L(\hat{\theta}_{n,m}^{L_2E}, \hat{F}_n) - L(\hat{\theta}_{n,m+1}^{L_2E}, \hat{F}_n) \leq n^{-1}b(n)[\nu(m + 1) - \nu(m)].
\]

Setting \( \zeta_{n,m} = n^{-1}b(n)[\nu(m + 1) - \nu(m)] \) naturally leads to the following estimator of \( m_0 \) defined by

\[
\hat{m}^{L_2E}_n = \min\{m : L(\hat{\theta}_{n,m}^{L_2E}, \hat{F}_n) \leq L(\hat{\theta}_{n,m+1}^{L_2E}, \hat{F}_n) + \zeta_{n,m}\}. \tag{3.9}
\]

Note that in equation (3.9) the threshold value \( \zeta_{n,m} \) has not been specified yet. It can be seen easily that threshold values directly impact the \( \hat{m}^{L_2E}_n \) values, which increase as \( \zeta_{n,m} \) values decrease. Since an \( \hat{m}^{L_2E}_n \) value determines the mixture complexity of the final mixture model, choice of \( \zeta_{n,m} \) may be viewed as model selection. Following the suggestions in Woo
and Sriram (2006, Section 4), we use the Akaike Information Criterion (AIC) threshold value \( \zeta_{n,m} = 3/n \) to numerically study the performance of \( \hat{m}_{n}^{L_2E} \) throughout this chapter.

In many instances, the robustness feature of \( L_2E \) down weights a small fraction of extreme values present in a data set (Scott, 2001). This clearly impacts any robust estimator of mixture complexity, such as \( \hat{m}_{n}^{L_2E} \) introduced in Umashanger (2009) and \( \hat{m}_{n}^{MHDE} \) introduced in Woo and Sriram (2006). To circumvent this limitation, we reintroduce a modified \( L_2E \) criterion, \( L_2E_p \), similar to the one introduced in Section 2.7.1. Our modified criterion adds a penalty function \( p(\pi) \) to \( L(\theta_m, \hat{F}_n) \) and minimizes

\[
L_p(\theta_m, \hat{F}_n) = L(\theta_m, \hat{F}_n) + p(\pi)
\]

\[
= \left[ \int_{-\infty}^{\infty} f_{\theta_m}^2(x)dx - 2n^{-1}\sum_{i=1}^{n} f_{\theta_m}(X_i) \right] + p(\pi),
\]

where \( p(\pi) = \min \{ \tau \tanh[\eta(\gamma - (0.5 - \pi_1))], \ldots, \tau \tanh[\eta(\gamma - (0.5 - \pi_m))] \} \). We define the corresponding estimator of mixture complexity by

\[
\hat{m}_{n}^{L_2E_p} = \min \{ m : L_p(\theta_{n,m}^{L_2E_p}, \hat{F}_n) \leq L_p(\theta_{n,m+1}^{L_2E_p}, \hat{F}_n) + \zeta_{n,m} \},
\]

(3.10)

where \( \zeta_{n,m} \) is as in (3.9).

### 3.3 Computational Details

Here, we give some computational details about \( \hat{m}_{n}^{L_2E} \) in (3.9) and \( L_2E \) for non-normal mixtures (e.g., gamma and lognormal mixtures). Computation of an estimate of mixture complexity using (3.9) and \( L_2E \) for non-normal mixtures can be performed using standard built-in quasi-Newton method algorithms such as \textit{nlm} or \textit{optim} routine in R software. As mentioned in Chapter 2, choice of initial values is critical in numerical optimizations involving mixture models. Extensive preliminary simulations indicated that our \( L_2E \) estimation procedure is not very sensitive to the choice of initial values. In fact, we conducted our numerical studies using the initial values based on \textit{K}-means or hierarchical clustering methods.

Computation of \( \hat{m}_{n}^{L_2E} \) is clearly an iterative procedure which can be used for any mixture density. It is outlined as follows:
• Step 1: Start with $m = 1$ (single density model), and compute $L(\hat{\theta}_{n,1}^{L_2E}, \hat{F}_n)$ from (3.7).

• Step 2: Now set $m = 2$ and compute $L(\hat{\theta}_{n,2}^{L_2E}, \hat{F}_n)$.

• Step 3: Calculate the difference $L(\hat{\theta}_{n,1}^{L_2E}, \hat{F}_n) - L(\hat{\theta}_{n,2}^{L_2E}, \hat{F}_n)$ and compare it with the threshold value $\zeta_{n,m} (= 3/n$ in all simulations and data analysis) in (3.9). If $\zeta_{n,m}$ is greater than this difference, then stop and report $\hat{m}_{n}^{L_2E} = 1$. Otherwise go to Step 4.

• Step 4: Repeat Step 2 and Step 3 by adding one more component to the previous mixture and comparing the difference until the first value $m = m^*$ for which the difference $L(\hat{\theta}_{n,m^*}^{L_2E}, \hat{F}_n) - L(\hat{\theta}_{n,m^*+1}^{L_2E}, \hat{F}_n)$ falls below the threshold value $\zeta_{n,m^*}$. At this point, the procedure terminates and declares $m^*$ as an estimate of the mixture complexity. Note that, at this stage, our procedure automatically provides the best parametric fit determined by $\hat{\theta}_{n,m^*}^{L_2E}$.

As noted in Scott (2001), the integral term in (3.7) can be computed in a closed form for normal mixtures. In addition, even for finite mixture of well-known non-normal densities such as gamma and lognormal, the integral term in (3.7) has a closed form expression. For gamma and lognormal components,

\[
G(x|\alpha_j, \beta_j) = \frac{x^{\alpha_j-1}e^{-x/\beta_j}}{\Gamma(\alpha_j)\beta_j^{\alpha_j}}, \quad \text{and} \quad L(x|\mu_j, \sigma_j^2) = \frac{\exp\left[-\left\{\log(x) - \mu_j\right\}^2/2\sigma_j^2\right]}{(x\sigma_j\sqrt{2\pi})},
\]

for $j = 1, \ldots, m$, the key integrals are

\[
\int_{-\infty}^{\infty} f_{\hat{\theta}_m}^2(x)dx = \sum_{i=1}^{m} \sum_{j=1}^{m} \pi_i \pi_j \frac{\Gamma(\alpha_i + \alpha_j - 1)}{\Gamma(\alpha_i)\Gamma(\alpha_j)} \frac{\beta_i^{\alpha_i-1}\beta_j^{\alpha_j-1}}{\beta_i + \beta_j}^{\alpha_i + \alpha_j - 1},
\]

and

\[
\int_{-\infty}^{\infty} f_{\hat{\theta}_m}^2(x)dx = \sum_{i=1}^{m} \sum_{j=1}^{m} \pi_i \pi_j \exp\left[-\left(\frac{(\mu_i - \mu_j)^2 + 2(\mu_i\sigma_i^2 + \mu_j\sigma_j^2) - \sigma_i^2\sigma_j^2}{2(\sigma_i^2 + \sigma_j^2)}\right)\right] \sqrt{2\pi(\sigma_i^2 + \sigma_j^2)},
\]

respectively, making the integral tractable and thereby significantly reducing the computation time involved in minimizing $L(\theta_m, \hat{F}_n)$. 

3.4 Simulation Studies

When there is a small fraction of extreme values in a data set, the tendency of \( \hat{\theta}_{n,m}^{L_2E} \) to automatically down weigh those values can sometimes hinder the performance of \( \hat{m}_{n}^{L_2E} \). Incidentally, a small fraction of extreme values cannot be immediately judged as outliers because they may be plausible values of an observed variable (see examples in Sections 3.5.1 and 3.5.2). Hence, it is important to detect and model that small component accurately in many practical situations.

The first simulation assesses the performance of \( \hat{m}_{n}^{L_2E} \) in (3.9) when samples are drawn from the four-component normal mixture density given by

\[
f_{\theta}(x) = 0.4N(x|(-2.5,0.36)) + 0.3N(x|(-0.5,1)) \\
+ 0.28N(x|2,0.09) + 0.02N(x|6,0.0001),
\]

where \( N \) denotes the normal density with mean and variance identified inside the parentheses. Here the fourth normal component located at 6 is tiny; see Figure 3.1.

![Figure 3.1: Frequency histogram of the data simulated from (3.12)](image-url)
We implemented our computational algorithm for sample sizes $n = 250$ drawn from (3.12) and performed 100 Monte Carlo replications of our algorithm, each yielding an estimate of mixture complexity. We then tallied the estimated number of components (out of 100 replications). These counts are reported in Table 3.1 below, where $L_2E$ and $L_2E_p$ correspond to the estimates given by (3.9) and (3.10), respectively.

Table 3.1: Mixture complexity estimation results for four-component normal mixture in (3.12)

<table>
<thead>
<tr>
<th>Estimated number of components</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 250$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_2E$</td>
<td>16</td>
<td>22</td>
<td>25</td>
<td>34</td>
<td>3</td>
</tr>
<tr>
<td>$L_2E_p$</td>
<td>38</td>
<td>61</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For this mixture, $\hat{m}^{L_2E}_{n}$ severely underestimates the true mixture complexity ($m_0 = 4$) and is able to correctly detect $m_0$ only 34% of the time. To improve the performance of $\hat{m}^{L_2E}_{n}$, we used the $L_2E_p$ criterion for estimation of mixture complexity (see 3.10) with $\tau = 0.48$, $\eta = 15$, and $\gamma = 0.3$ in the penalty function, $p(\pi)$, and $\zeta_{n,m} = 3/n$. This yielded $\hat{m}^{L_2E_p}_{n} = 3$ about 38% of the time, but $\hat{m}^{L_2E_p}_{n} = 4$ about 61% of the time, indicating that the modified criterion $L_2E_p$ helps detect the small component. Note that $L_2E_p$, in addition to correctly detecting $m_0$, also simultaneously provides accurate estimates of $\theta_{m_0}$. In Section 3.5.1, we will illustrate the use of $\hat{m}^{L_2E_p}_{n}$ for a real data.

There are also situations where the intrinsic range of the variable is the nonnegative real line. In addition, the data being modeled may exhibit skewness and multimodality, and there may be presence of heavy tail in the observed data (Dey et al., 1995; Wiper et al., 2001; Sahu and Cheng, 2003; Wijaya et al., 2008). In these cases, a finite mixture of gamma densities, for example, may be more appropriate for the data.

We carried out simulation studies for two three-component gamma mixtures, the first of which was studied in Wiper et al. (2001) given by

$$f_{\theta}(x) = 0.2 \mathcal{G}(x|40,1/20) + 0.6 \mathcal{G}(x|6,1) + 0.2 \mathcal{G}(x|200,1/20),$$

where $\mathcal{G}(x|\mu,\sigma)$ denotes the gamma distribution with shape parameter $\mu$ and rate parameter $\sigma$. The second mixture model is given by

$$f_{\theta}(x) = 0.2 \mathcal{G}(x|40,1/20) + 0.4 \mathcal{G}(x|10,1) + 0.4 \mathcal{G}(x|200,1/20),$$

where $\mathcal{G}(x|\mu,\sigma)$ denotes the gamma distribution with shape parameter $\mu$ and rate parameter $\sigma$. These models are used to illustrate the performance of the $L_2E_p$ criterion in detecting small components and estimating the mixture parameters.
where $G(x|\alpha, \beta)$ is a gamma distribution defined in (3.11). As mentioned in Wiper et al. (2001), this is a mixture of a rather diffused gamma distribution (second component) with mean 6 and two lower weighted distributions with smaller variances and means of 2 and 10, respectively (See upper panel of Figure 3.2). For sample of size $n = 400$, we find that $\hat{m}_{L2E} = 2$ about 23% of the time, but 69% of the time $\hat{m}_{L2E} = 3$, correctly detecting the true mixture complexity. However, when the second gamma component is not diffused as in (See lower panel of Figure 3.2)

$$f_{\theta_3}(x) = 0.2 \ G(x|10, 1/10) + 0.6 \ G(x|50, 1/10) + 0.2 \ G(x|80, 1/10),$$

then for $n = 400$ we find that the correct detection, $\hat{m}_{L2E} = 3$, increases to 89%.

Figure 3.2: Generating density functions for two different gamma mixtures
In summary, these simulation studies indicate that our $L_2E$ estimator of mixture complexity performs well even for non-normal mixture models. Also, the $L_2E_p$ helps improve the efficiency of $\hat{m}_n^{L_2E_p}$, when the true mixture model has a small fraction of extreme values.

3.5 Data Analysis

In this section, we use the $L_2E$ approach to analyze two data sets, the first of which has been analyzed using normal mixtures by many authors in the literature. Each data set has a small fraction of extreme values, making them ideal candidates to test the $L_2E$ and $L_2E_p$ methodology.

3.5.1 Red Blood Cell Sodium-Lithium Countertransport Activity

Red blood cell sodium-lithium countertransport (SLC) activity data collected from 190 individuals was originally analyzed in Dudley, Giuffra, Raine, and Reeders (1991). The SLC is measured as the difference in lithium efflux rate from lithium-loaded cells into sodium chloride and sodium-free media. Geneticists are interested in SLC because it is correlated with blood pressure; hence may be an important cause of hypertension. Assume that a trait such as blood pressure is determined by simple mode of inheritance compatible with the action of a single action gene with two alleles, $A_1$ and $A_2$, which occur with probabilities $\pi$ and $1 - \pi$. It is believed that SLC measurements are derived from one of two competing genetic models (Roeder, 1994): Simple dominance or Additive. In the simple dominance model, genotypes ($A_1A_1$ and $A_1A_2$) have phenotype $\mu_1$, whereas $A_2A_2$ has phenotype $\mu_2$. In the additive model, the three genotypes ($A_1A_1$), ($A_1A_2$) and ($A_2A_2$) have phenotypes $\mu_1$, $\mu_2$ and $\mu_3$, respectively. Thus, the Simple dominance model corresponds to a two-component mixture and the Additive model to a three-component mixture, which is given below:

Additive Model: Each of the three genotypes yields a distinct phenotype with $\pi_1 = \pi^2$, $\pi_2 = 2\pi(1 - \pi)$ and $\pi_3 = (1 - \pi)^2$, where $\pi_i$’s are mixing proportions. Furthermore, $\mu_1 < \mu_2 < \mu_3$ and $\mu_3 - \mu_2 = \mu_2 - \mu_1$. 
Roeder (1994) fitted a three-component normal mixture to the SLC data assuming equal component variances and showed that her fit approximately satisfies the conditions of the *additive model*. Later, Ishwaran, James, and Sun (2001) and Chen and Khalili (2006) adopted a blocked Gibbs sampler approach and the MSCAD method, respectively, which supported a three-component normal mixture with equal variances. They also showed that their three-component normal mixtures approximately satisfy the *additive model*. In Table 3.2, we provide the parameter estimates corresponding to the three fitted mixture models, identified as *Roeder*, *Gibbs*, and *MSCAD*. Note from Table 3.2 that each of the fitted mixture has a small third component located around extreme data values.

Table 3.2: SLC data parameter estimates

<table>
<thead>
<tr>
<th></th>
<th>$\pi_1$</th>
<th>$\pi_2$</th>
<th>$\pi_3$</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\mu_3$</th>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
<th>$\sigma_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_2E_p$</td>
<td>0.771</td>
<td>0.210</td>
<td>0.019</td>
<td>0.219</td>
<td>0.366</td>
<td><strong>0.514</strong></td>
<td>0.061</td>
<td>0.061</td>
<td>0.061</td>
</tr>
<tr>
<td><em>Roeder</em></td>
<td>0.774</td>
<td>0.202</td>
<td>0.024</td>
<td>0.223</td>
<td>0.379</td>
<td>0.577</td>
<td>0.058</td>
<td>0.058</td>
<td>0.058</td>
</tr>
<tr>
<td><em>MSCAD</em></td>
<td>0.750</td>
<td>0.220</td>
<td>0.030</td>
<td>0.221</td>
<td>0.372</td>
<td>0.564</td>
<td>0.057</td>
<td>0.057</td>
<td>0.057</td>
</tr>
<tr>
<td><em>Gibbs</em></td>
<td>0.760</td>
<td>0.210</td>
<td>0.030</td>
<td>0.222</td>
<td>0.373</td>
<td>0.567</td>
<td>0.057</td>
<td>0.057</td>
<td>0.057</td>
</tr>
</tbody>
</table>

For the SLC data, several other authors also fitted a three-component normal mixture model with unequal variances using different estimation criteria. For example, Cutler and Cordero-Braña (1996, see Section 2, Figure 1) fitted a three-component normal mixture using MHDE and MLE, and concluded that the MLE has a small third component located around extreme data values, whereas their fit based on MHDE largely ignores these values. Recently, Fujisawa and Eguchi (2006) and Woo and Sriram (2006) used a modified likelihood approach (Basu et al., 1998) and MHDE approach, respectively, which supported a three-component normal mixture model for the SLC data. However, none of these fits come close to satisfying the conditions of the *additive model*. Hence, we do not provide these fits in Table 3.2.

Here, we revisit the analysis of SLC data using our $L_2E$ approach. If we assume unequal variances, then our $L_2E$ algorithm with threshold value $\zeta_{n,m} = 3/n$ yields $\hat{m}_n^{L_2E} = 3$, but the resulting three-component normal mixture does not satisfy the conditions of the *additive*
Interestingly, our three-component normal mixture model obtained using $L_2E$ (with unequal variances assumption) is almost identical to that of Fujisawa and Eguchi (2006)'s fit. However, if we assume equality of variances, then $\zeta_{n,m} = 3/n$ yields $\hat{m}_n^{L_2E} = 2$.

We observed earlier that each of the three-component normal mixture fits (Roeder, Gibbs, and MSCAD) given in Table 3.2 has a small third component located around extreme values. This observation combined with the performance of $\hat{m}_n^{L_2E_p}$ [see (3.10)] for the four-component normal mixture simulation in Section 3.4 prompts us to estimate the mixture complexity for the SLC data using $\hat{m}_n^{L_2E_p}$. Assuming normal mixtures with equal variances and using $\zeta_{n,m} = 3/n$ we obtained $\hat{m}_n^{L_2E_p} = 3$. The parameter estimates corresponding to the three-component normal mixture fit using $L_2E_p$ are also given in Table 3.2. Note from Table 3.2 that the $L_2E_p$ also identifies a small third component. While all four fits in Table 3.2 are similar, it is important to note that the $L_2E_p$ estimate of $\mu_3$ is much less affected by the small fraction of extreme values, indicating that $L_2E_p$ is more robust than the other procedures in Table 3.2. Furthermore, for the fits in Table 3.2, we computed all the values needed to check how well each model conforms to the additive model; these are given in Table 3.3. While each fit in Table 3.3 satisfies the conditions of the additive model reasonably well, our $L_2E_p$ three-component normal mixture fit almost exactly corresponds to the additive model. Thus, in terms of sensitivity to extreme values and conformity to the additive model, we conclude that our $L_2E_p$ fit is better than all the others considered here.

<table>
<thead>
<tr>
<th>Table 3.3: Conformity to the additive model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_1 - \pi^2$</td>
</tr>
<tr>
<td>$L_2E_p$</td>
</tr>
<tr>
<td>Roeder</td>
</tr>
<tr>
<td>MSCAD</td>
</tr>
<tr>
<td>Gibbs</td>
</tr>
</tbody>
</table>
3.5.2 Italian Surgical Procedure Times

Hospitals are under increased economic pressure to schedule elective surgeries efficiently to contain costs associated with surgical services. Surgical scheduling is often complicated by variability inherent in the caliber of surgeons and/or the duration of surgical procedures. A good statistical model for total surgical procedure times, the time from entry into the operating suite until exit following emergence from anesthesia, is essential for several purposes including generating accurate time estimates. For instance, a good model could be used retrospectively to identify those surgeons (or procedures) whose total surgical procedure times are unusually long or short. Identifying these may allow surgical managers to eliminate sources of variability or, alternatively, to identify the extreme times and schedule them separately. Resulting estimates are crucial operationally to improve surgery room utilization and to strategically identify surgeons, procedures, or patients whose duration of surgeries differ from what might be expected (Strum, May, and Vargas, 2000; Strum, May, Sampson, Vargas, and Spangler, 2003). Due to inherent variability in surgeons’ competence, diagnoses, total surgical procedure times and presence of extreme values, one commonly observes heterogeneous mixed populations of surgical procedure times.

Here, we consider data on total surgical procedure times (in minutes) for \( n = 433 \) patients collected in 2005 at a hospital in Genoa, Italy. This data set also includes confidential information such as each patient’s date of birth, gender, diagnosis, operating surgeon’s name along with his/her specialty, and time of entry into operating room until exit including various stages of anesthesia. The total surgical procedure times (in minutes) range from 17 to 475 with median and mean times 110 and 140, respectively. In addition, the 90th, 95th and 99th percentiles of times are 275, 315 and 410, respectively, with four observed times over 425, indicating possible presence of subpopulations and extreme values. For convenience, we let \( X = \text{total surgical procedure times}/1440 \) (total minutes in a day) and build a best fitting finite mixture model for \( X \). Based on the histogram of \( X \) values, we postulated a gamma mixture model and a lognormal mixture model. Using \( \zeta_{n,m} = 3/n \) we obtained
\( \hat{m}_n^{L_2E} = 2 \) for both the postulated mixtures. The parameter estimates corresponding to the fitted two-component gamma mixture and lognormal mixture using \( L_2E \) are given in Table 3.4. In Figure 3.3, we also provide plots of the two fitted mixtures overlaid on the histogram of \( X \) values.

Table 3.4: Estimates of gamma and lognormal mixtures: Italian surgical procedure times

<table>
<thead>
<tr>
<th></th>
<th>( \pi_1 )</th>
<th>( \pi_2 )</th>
<th>( \alpha_1 )</th>
<th>( \alpha_2 )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( L_2E )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>0.79242</td>
<td>0.20758</td>
<td>4.27306</td>
<td>21.51017</td>
<td>0.01699</td>
<td>0.00884</td>
<td>-5.97445</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.91575</td>
<td>0.08425</td>
<td>-2.59252</td>
<td>-1.64859</td>
<td>0.35020</td>
<td>0.00950</td>
<td>-5.93894</td>
</tr>
</tbody>
</table>

Figure 3.3: Plots of fitted gamma and lognormal mixtures: Italian surgical procedure times

From Table 3.4, it can be seen that the estimated component means for the gamma mixture are 0.0726 (\( \approx 105 \) minutes) and 0.190 (\( \approx 274 \) minutes), which are similar to the
estimated component means of the lognormal mixture: 0.0796 (≈ 115 minutes) and 0.192 (≈ 276 minutes), respectively. However, the mixing proportion for the two fits are different and the $L_2E$ value of gamma mixture fit is slightly smaller than that of the lognormal mixture fit; see Table 3.4. Note that the second component means of the fitted gamma and lognormal mixtures, respectively, are not influenced by the presence of extreme values. From these and the fitted density plots in Figure 3.3, we conclude that both the mixture models provide a good fit; however, the gamma mixture may be slightly better.

3.6 DISCUSSION

We have revisited Umashanger (2009)’s work on mixture complexity estimation using $L_2E$, but focused on detection of small yet plausible mixture component and robust estimation in non-normal mixtures. We have developed a modified estimator of mixture complexity by adding penalty to the $L_2E$ estimation function considered in Umashanger (2009). The use of this modified estimator is illustrated via simulations and analysis of Red blood cell sodium-lithium countertransport (SLC) activity data.

We have also extended Umashanger’s simulations and data analyses to include finite mixtures of gamma or lognormal distributions, which are of considerable interest when modeling phenomena with positive responses, possibly with long asymmetric tails. We have assessed the performance of the $L_2E$ estimator of mixture complexity via simulations for gamma mixtures and through analysis of Italian surgical procedure times data.

The $L_2E$ or $L_2E_p$ estimator of mixture complexity has been shown to have many distinct advantages over other procedures in the literature. We conclude that, transparency, ease of use and efficiency in achieving computational speed combined with competitive performance and robustness makes the $L_2E$ or $L_2E_p$ estimator stand out as an attractive alternative to other existing methods in the literature. Finally, data sets and computational codes used in this chapter may be obtained from the author.
3.7 References


Chapter 4

Conclusions

In this dissertation, we have developed a robust estimation method for finite mixture of regression models based on a minimum integrated $L_2$ distance for count response. The estimators, called the $L_2E$, of a parameter vector and mixture complexity have desirable theoretical properties such as consistency and asymptotic normality. Through extensive simulation studies, we have shown that the $L_2E$ is highly competitive to the maximum likelihood (ML) estimator and a better, yet viable, alternative to a minimum Hellinger distance (MHD) estimator, when the model is correctly specified. In addition, the $L_2E$ has been shown to be more robust compared to the ML and MHD estimators. For the poorly-separated model with a small fraction of extreme values, we have introduced a modified $L_2E$ criterion, called the $L_2E_p$ with a penalty function. Once again, the $L_2E_p$ estimator has been shown to be consistent and asymptotically normal for the case of two-component mixture regression models. In addition, we have theoretically shown strong consistency of the $L_2E_p$ estimator of mixture complexity. Furthermore, we have illustrated extensively via real data analysis that our criteria, $L_2E$ and $L_2E_p$, outperform the ML and MHD methods.

We have also considered the problem of $L_2E$ estimation of mixture complexity for continuous data. In other words, we have studied detection of small yet plausible mixture component and robust estimation in non-normal mixtures. To this end, we have proposed a modified $L_2$ criterion with a penalty function when there is a small fraction of extreme values in the data. The simulation studies and data analysis have shown that the $L_2E_p$ and $L_2E$ estimators correctly detect true mixture complexity when there is a small component in finite mixtures and when the true model is non-normal mixture, respectively. We believe
that our $L_2E$ and $L_2E_{p}$ methods will be useful for analyzing count and continuous data sets arising in contexts as well.