L_2E estimation of mixture complexity

by

Thayasivam Umashanger

(Under the direction of T.N. Sriram)

Abstract

Finite mixture models provide a flexible way to model data coming from population consisting of finite number of homogeneous subpopulations. These models are particularly useful in determining clusters or subgroups within a data. Model selection is a crucial step in every statistical data analysis and especially so for data coming from unknown number of subpopulations. In this thesis, we focus squarely on determining parsimonious finite mixture models using a model selection criterion based on L_2 distance.

In many applications, the scientific information available may not be sufficient to determine the number of components in finite mixture models; hence, it is important to find mixtures with fewest number of components, known as the *mixture complexity*, that provide satisfactory fit to the data. Estimation of mixture complexity is a fundamental yet challenging problem that has received an enormous attention in the past few decades. In this thesis, we treat the estimation of mixture complexity as a model selection problem and construct an estimator of mixture complexity as a by-product of minimizing a Information Criterion based on L_2 distance for both count and continuous data. The estimator of mixture complexity, called $\hat{m}_n^{L_2E}$, is shown to be consistent when the form of component densities are unknown but are postulated to be members of some parametric family. The estimator is also

shown to be robust against model misspecification via simulations. When the model is correctly specified, Monte Carlo simulations for a wide variety of normal and Poisson mixtures show that our estimator is very competitive with several others in the literature in correctly identifying the true mixture complexity. The performance of this method is illustrated for several simulated data and well-known real datasets. We begin the thesis with a survey of methods available in the literature. Detailed description of the methods and associated results can be found in the respective chapters.

INDEX WORDS: Finite mixtures; L_2E estimation; information criterion; algorithm; threshold; consistency; efficiency; robustness.

L_2E estimation of mixture complexity

by

Thayasivam Umashanger

B.S., The University of Colombo, 2001

M.S., University of Georgia, 2004

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

2009

© 2009

Thayasivam Umashanger

All Rights Reserved

L_2E estimation of mixture complexity

by

Thayasivam Umashanger

Approved:

Major Professor: T.N. Sriram

Committee: William McCormick
Jaxk Reeves
Stephen Rathbun
Lynne Seymour

Electronic Version Approved:

Dr. Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2009

Dedication

To my parents Dr. and Mrs. Thayasivam; my wife, Rooby, and my lovely daughter, Nethiri.

Acknowledgments

Any concise acknowledgment will unfortunately exclude people who certainly provided invaluable support throughout the time taken to compose this thesis. Nonetheless, several specific people deserve recognition for their assistance, and I would like to take this opportunity to acknowledge their contribution.

I would first like to express my sincere appreciation to my major professor, Dr. T.N. Sriram, for all the support and guidance he has provided throughout this entire research process with his deep insight of the field, enlightened suggestions, kindness, and patience. Besides, Dr. Sriram, taught me how to develop an academic career, how to collaborate with other researchers and most of all how to influence a student positively. With deep gratefulness, I thank him for his immense support both professionally and personally to channel me in right directions.

I am truly grateful to my committee members, Dr. Bill Mccormick, Dr. Jaxk Reeves, Dr. Stephen Rathbun, and Dr. Lynne Seymour, for their timely assistance and valuable suggestions.

The faculty in the Department of Statistics is outstanding; their enthusiasm and approachability create a very special environment for learning, which made attending UGA such a great experience. Special thanks to Dr. John Stufken (Head), Dr. Gauri Datta and Dr. Lynne Seymour for their help, encouragement and advice. I would like to acknowledge and thank Mr. Jimmy Cretney and Mr. Jesse Bowling for all their technological support and Mr. Tim Cheek, Ms. Daphney Smith and Ms. Germaine Cahoon for their administrative assistance.

I would like to express my special gratitude to Dr. Bala Rajaratnam at Stanford University for his support and encouragement to pursue my career in academics. Interesting conversations I had with him and the learning I gained through those are invaluable.

I would like to express my gratitude to Dr. W.N. Wickremasinghe and all other faculty in the Department of Statistics, University of Colombo, Sri Lanka for their support and encouragement to pursue a doctoral degree.

I have many friends who have been very helpful during my graduate study. I would also like to thank them for their support.

I am also indebted to my family, my parents, my brothers, in-laws, nieces and my grand mother for their support and encouragement throughout my whole life. Certainly, I would say that my father implanted mathematics inside me. Moreover, he also channeled me to observe and understand the world via mathematics. My mother exemplified the determination necessary to successfully achieve my ambitions. My brothers offered immeasurable support throughout my education. Last, but not least, I am deeply grateful to my wife, Rooby, and my daughter, Nethiri; without their understanding, love and patience, I would not have been able to finish this thesis.

Table of Contents

Page

Acknowledgments v

List of Figures ix

List of Tables x

Chapter

1 INTRODUCTION AND LITERATURE REVIEW 1

1.1 INTRODUCTION 1

1.2 LITERATURE REVIEWS 4

1.3 OUTLINE OF THE THESIS 15

1.4 References 17

2 L_2E ESTIMATION OF MIXTURE COMPLEXITY : COUNT DATA 23

2.1 INTRODUCTION 24

2.2 L_2E ESTIMATOR 26

2.3 CONSISTENCY THEOREM 29

2.4 COMPUTATIONAL DETAILS 30

2.5 SIMULATION STUDIES 32

2.6 DATA ANALYSIS 37

2.7 CONCLUSION 42

2.8 SUPPLEMENTAL MATERIALS 44

2.9 APPENDIX 52

2.10 References 55

2.11 Tables and Figures 58

3 L_2E ESTIMATION OF MIXTURE COMPLEXITY : THE CONTINUOUS CASE 65

3.1 INTRODUCTION 66

3.2 L_2E ESTIMATOR 70

3.3 CONSISTENCY THEOREM 73

3.4 COMPUTATIONAL DETAILS 73

3.5 SIMULATION STUDIES 77

3.6 DATA ANALYSIS 83

3.7 SUMMARY AND CONCLUSIONS 87

3.8 SUPPLEMENTAL MATERIALS 90

3.9 APPENDIX 100

3.10 References 105

3.11 Tables and Figures 111

4 Conclusions 122

4.1 SUMMARY 122

4.2 Future Research 123

Bibliography 124

List of Figures

3.1 Fitted normal mixture for SLC data. 119

3.2 Fitted Normal mixture for Acidity Data 120

3.3 Fitted Normal mixture for Enzyme Data 121

List of Tables

2.1 Relative frequencies of estimated number of components based on 500 replications. 58

2.2 Samples drawn from 2-component Negative Binomial mixture in (2.5.12) with $\boldsymbol{\theta}_2 = (\pi_1, \lambda_1, \lambda_2)$ and $r = 10$. 60

2.3 Samples drawn from 2-component Negative Binomial mixture in (2.5.12) with $\boldsymbol{\theta}_2 = (\pi_1, \lambda_1, \lambda_2)$. 61

2.4 Parameter estimates for Poisson mixture models: Spanish bank data. 62

2.5 Comparison of observed frequencies and expected frequencies: Spanish bank data. 62

2.6 Parameter estimates for Poisson mixture models: Death notice data. 63

2.7 Comparison of observed frequencies and expected frequencies: Death notice data. . 63

2.8 Parameter estimates for Poisson mixture models: Accident data. 64

2.9 Comparison of observed frequencies and expected frequencies: Accident data. . . . 64

3.1 Mixture Complexity Estimation results for three component normal mixture in (3.5.11) 111

3.2 Mixture Complexity Estimation results for the Marron and Wand densities 2 & 4-9 112

3.3 Mixture Complexity Estimation results for $t(4)$ components 114

3.4 Mixture Complexity Estimation results for $t(2)$ components 115

3.5 Mixture Complexity Estimation results for Rescaled $t(3)$ components 116

3.6 Mixture Complexity Estimation results for Rescaled $t(4)$ components 117

3.7 SLC Data Parameter Estimates 118

3.8 Acidity Data Parameter Estimates 118

3.9 Enzyme Data Parameter Estimates 118

Chapter 1

INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

Finite mixture models have been used in applications for several decades. The area has seen renewed popularity during the last decade due to increase in computing power and applications in Bio-informatics. Applications of mixture model extend beyond Bio-informatics to include a wide range of areas such as biology, medicine, physics, economics and marketing. These models are particularly suitable for datasets, where observations originate from different groups but the group affiliations are not known. Furthermore, the exact number of groups present in a dataset may not be available, making the selection of appropriate finite mixture model a challenging task.

There is an enormous body of literature concerning the theory, computation and application aspects of finite mixture models when the number of components (groups) is known in advance. Over the last three decades, a variety of estimation approaches have been adopted for mixture models. These include the method of moments, the maximum likelihood (ML) method, minimum distance methods and Bayesian methods. If the number of mixture components is known and the component densities are assumed to belong to a specified parametric family, the EM algorithm of Dempster et al. (1977) is a useful way to compute ML estimates. However, when there is a small perturbation in one of the component densities, ML estimates become highly unstable (Aitkin and Wilson, 1980).

Robust methods such as M-estimation are not easily adapted for mixtures, and these generally achieve robustness at the cost of efficiency at the parametric model. For continuous data modeled by finite mixtures, Cutler and Cordero-Braňa (1996) developed a minimum

Hellinger distance (MHD) estimator (Beran, 1977) of unknown parameters and showed that their estimator is efficient at the parametric model and robust under gross-error contaminations. For count data, Karlis and Xekalaki (1998) developed MHD estimation of parameters in Poisson mixtures. While the MHD estimation method does lead to efficiency and robustness, as noted in Scott (1999 and 2001) and Markatou (2000 and 2001), the method (in the continuous case) involves some practical challenges such as selection of an appropriate nonparametric kernel density estimator and associated bandwidth.

Scott (1998, 1999, 2001 and 2004) introduced an alternative minimum distance estimation method based on integrated squared error criterion, termed L_2E, which avoids the use of nonparametric kernel density estimators; see section 1.2.3. The L_2E approach is a special case of a general method introduced by Basu et al. (1998), who devised a whole continuum of divergence estimators that begin with the MLE and interpolate to the L_2E estimator and beyond; see section 1.2.3. Markatou (2000 and 2001), on the other hand, used the weighted likelihood estimation approach of Markatou, Basu and Lindsay (1998) to address the effects of misspecification of mixture model on parameter estimates and provided a heuristic way to identify the number of components in mixture models.

A complication in many applications is that our scientific knowledge may not be sufficient to determine the number of mixture components, termed *mixture complexity*. Estimation of mixture complexity is a fundamental problem because correct identification of mixture complexity followed by efficient estimation of all parameters would lead to finding a mixture with the fewest possible components. Developing methods to determine mixture complexity has been an area of intense research in the recent years; see, for example, Schlattmann and Böhning (1993), Roeder (1994), Pauler et al. (1996), Dellaportas et al. (1997), Karlis and Xekalaki (1999), James et al. (2001), Ishwaran et al. (2001) and references therein.

Recently, Woo and Sriram (2006, 2007) introduced a Hellinger Information Criterion (HIC), which formed the basis for constructing their MHD-based estimator of mixture complexity. More specifically, by treating the estimation of mixture complexity as a model

selection problem, they constructed an estimator of mixture complexity as a by-product of minimizing the HIC. Because the basic construction is rooted in an approach based on minimum Hellinger distance, their estimator is shown to inherit the property of robustness against model misspecification while consistently estimating the true mixture complexity for parametric family of mixtures. Their approach is such that it not only provides a consistent estimate of the mixture complexity for a given dataset but also provides consistent MHD estimates of the mixture parameters; see section 1.2.3 for more details.

While the MHD-based estimator of mixture complexity has attractive large sample and robustness features, there are difficult computational issues associated with the implementation of the MHD algorithm described in Woo and Sriram (2006, 2007), the first of which concerns the precise nature of the nonparametric density estimator. When all the mixture parameters are unknown, Cutler and Cordero-Braňa (1996) point out that it is necessary to use some form of adaptive density estimator in order to avoid severe bias problems with the scale estimates. Secondly, one needs to carefully choose the bandwidth for the (adaptive) nonparametric density estimators. Undoubtedly, these selections put an extra burden on the computation of MHD estimates.

In this thesis, we focus squarely on the estimation of parameters in finite mixture models when the number of components is not known. More specifically, we propose a comprehensive estimation procedure for all the parameters involved in a finite mixture model (including the unknown number of components) based on a familiar L_2 distance. Our proposed L_2 estimation method, called L_2E henceforth, avoids the use of nonparametric density estimator altogether, but is shown to possess robustness property which is comparable to that of a procedure based on minimum Hellinger distance. In addition, it has distinct computational advantage over the MHD. The thesis illustrates the scope and use of L_2 estimation method in applications, thereby providing a competitive alternative to other procedures in the literature.

1.1.1 BASIC DEFINITIONS

Consider a parametric family of probability density or mass functions (p.d.f. or p.m.f.) $\mathcal{F}_m = \{f_{\boldsymbol{\theta}_m} : \boldsymbol{\theta}_m \in \Theta_m \subseteq R^p\}$ such that $f_{\boldsymbol{\theta}_m}$ can be represented as a finite mixture of the form

$$f_{\boldsymbol{\theta}_m}(x) = \sum_{i=1}^{m} \pi_i f(x|\boldsymbol{\phi}_i), \quad x \in \mathcal{X} \subseteq \mathcal{R}, \tag{1.1.1}$$

where $m > 0$ is a finite integer, $f(x|\phi_i)$ is the component p.d.f. (or p.m.f.), $\boldsymbol{\phi}_i \in \Phi \subseteq R^s$, the mixing proportions $\pi_i \geq 0$, $\sum_{i=1}^{m} \pi_i = 1$ for $i = 1, \ldots, m$ and $\boldsymbol{\theta}_m = (\pi_1, \ldots, \pi_{m-1}, \boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_m^T)^T$. For each m, the functional form of component p.d.f. (or p.m.f.) is known, but $\boldsymbol{\theta}_m$ is unknown. In the discrete case, $\mathcal{X} = \{0, 1, 2, \ldots\}$. The class $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$ for all m and we denote $\mathcal{F} = \bigcup_{m=1}^{\infty} \mathcal{F}_m$.

Suppose $\mathbf{X}_n = (X_1, \ldots, X_n)$ is a random sample from an unknown p.d.f. or p.m.f. f_0. Define the *index of the economical representation* of f_0, relative to the family of mixtures $\mathcal{F}_m$, as

$$m_0 = m(f_0) = \min\{m : f_0 \in \mathcal{F}_m\}. \tag{1.1.2}$$

If indeed f_0 is a finite mixture defined in (1.1.1), 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. Before describing our research in detail, we give a brief survey of available literature on estimation approaches for finite mixture models.

1.2 LITERATURE REVIEWS

1.2.1 Estimation in Mixture Models

Over the past years, a variety of methods have been developed for estimating the parameters in finite mixture models. The following four estimation methods are widely used for mixture models: Method of moments, Maximum likelihood method, Minimum-distance method, and Bayesian method.

THE METHOD OF MOMENTS

The first published investigation relating to estimation of finite mixture models appears to be that of Pearson (1894), who considered the method of moments (MOM) estimation of the parameters in a mixture of two univariate normal densities. Years later, Pollard (1934) obtained MOM estimates for parameters in a mixture of three univariate normal densities, and Cooper (1967), Day (1969) and John (1970) obtained MOM estimates for mixture of multivariate normals. Also, see Pearson (1915), Muench (1936), Schilling (1947), Gumbel (1939), Arley and Buch (1950), Rider (1962), Blischke (1962), Cohen (1963) and Kabir (1968) for the development of MOM estimates for parameters in finite mixtures of Binomial and Poisson.

THE METHOD OF MAXIMUM LIKELIHOOD AND THE EM ALGORITHM

With the arrival of increasingly powerful computers and increasingly sophisticated numerical methods during the 1960's, the method of maximum likelihood (MLE) became the widely preferred approach to estimation of parameters in finite mixture models. Despite some initial success, the problem of obtaining MLE was generally considered to be completely intractable for computational reasons.

For mixture models, computational difficulties with respect to MLE arise because of the complex dependence of the likelihood function on the parameters to be estimated. More specifically, the likelihood equations are almost always nonlinear and beyond hope of solution by analytic means. Consequently, one must resort to an approximate solution via some iterative procedure. There are, of course, many general iterative procedures which are suitable for finding an approximate solution of the likelihood equations such as Newton's method and its variants, and conjugate gradient methods.

Incomplete data often result in complicated likelihood functions, where MLE usually has to be computed iteratively. In such situations, algorithms such as the Newton-type methods may turn out to be more complicated. The Expectation-Maximization algorithm proposed

by Dempster et al. (1977) in a seminal paper, popularly known as the EM algorithm, is a broadly applicable approach to the iterative computation of MLE. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found in the E step. The parameters found in the M step are then used to begin another E step, and the process is repeated.

In most instances, EM has the advantage of reliable global convergence, low cost per iteration, economy of storage and ease of programming, as well as certain heuristic appeal. Unfortunately, its convergence can be very slow in simple problems which are often encountered in practice. Also, as mentioned in the introduction, when there is a small perturbation in one of the component densities, the ML estimates become highly unstable.

MINIMUM DISTANCE ESTIMATION

If model assumptions are violated, minimum distance estimators are usually more robust than the MLE. Minimum distance estimators are obtained by minimizing some specified distances between the parametric and empirical densities or distributions. A variety of minimum distance estimation methods have been considered for mixture models. Choi (1968) proposed the minimum Wolfowitz distance estimator for mixing proportions with known component distributions. MacDonald (1971) and Woodward et al. (1984) examined a similar method of minimizing the Cramer-von Mises distance to estimate the mixing proportions in mixture of normal distributions. Clarke and Heathcote (1994) developed explicit estimators for mixing proportions in mixture normal distributions by minimizing the L_2 distance between parametric and empirical distribution functions. Woodward et al. (1995) proposed the MHD for mixing proportion in the mixture of two normals and Karlis and Xekalaki (1998) examined the case of finite Poisson mixtures. Most of the methods mentioned above

give estimators for mixing proportions that are not in explicit form. Some of them need complicated numerical techniques to calculate the estimators.

The MHD is one of the minimum distance estimation approaches, which has received considerable attention in mixture models. For the mixture of normal distributions and the Poisson mixtures, Woodward et al. (1995) and Karlis and Xekalaki (1998), respectively, showed that the MHD is asymptotically normally distributed with full efficiency under model assumptions, and is more robust to departures from the underlying assumptions than the MLE. Cutler and Cordero-Braňa (1996) developed a minimum Hellinger distance (MHD) estimator of unknown parameters and showed that their estimator is efficient at the parametric model and robust under gross-error contaminations.

The integrated squared distance has been used as the goodness-of-fit criterion in nonparametric density estimation for a long time. Scott (1999) introduced an alternative minimum distance estimation method based on integrated squared error criterion, termed L_2E, which avoids the use of nonparametric kernel density estimators. In his paper, Scott showed that the L_2E is especially suited for parameter-rich models such as mixture models. Scott (1999) showed that the L_2E approach, whose genesis may be traced to the pioneering work of Rudemo (1982) and Bowman (1984), is computationally feasible and it leads to robust estimators like all other minimum distance methods. Scott (1999) points out that L_2E is a special class of robust estimators like the median-based estimators, which sacrifice some asymptotic efficiency for substantial computational benefits in difficult estimation problems. In fact, Scott (2001) showed that L_2E estimator performs much better than the MHD estimator, under data contamination.

The L_2E estimator belongs to the family of minimum density power divergence ($MDPD$) estimators introduced by Basu et al. (1998) with the tuning parameter $\alpha = 1$. The tuning parameter α in an $MDPD$ estimator controls the trade-off between robustness and efficiency. Basu et al. (1998) also show that that the robustness of the L_2E estimator is achieved at a fairly stiff price in asymptotic efficiency. They showed that for the normal, exponen-

tial and Poisson distributions with small values of α ($\leq$ 0.10), the $MDPD$ has strong robustness properties and retains high asymptotic relatively efficiency (ARE) with respect to MLE. Nonetheless, within the family of density-based power divergence measures, the L_2E approach has the distinct advantage that a key integral can be computed in a closed form, especially for Normal mixtures. Moreover, Scott (1999) also showed that the L_2E is more robust than the MHD against gross-error contamination.

BAYESIAN ESTIMATION

In the framework of the Bayesian approach, one needs to assume that a prior distribution is available. Using Bayes' theorem, we can obtain the posterior density . As summarized in Fruhwirth-Schnatter (2006), there are two main reasons why people may be interested in using the Bayesian method in finite mixture models. Firstly, including a suitable prior distribution for the parameter in the framework of the Bayesian approach may avoid spurious modes when maximizing the log-likelihood function. The idea for the penalized MLE in Chen et al. (2007) can be seen as putting a proper prior distribution on the variance parameters. Secondly, when the posterior distribution for the unknown parameters is available, the Bayesian method can yield valid inference without relying on the asymptotic normality. As warned by McLachlan and Peel (2000, p.68), the asymptotic theory of the MLE can apply only when the sample size n is very large. Hence the second advantage of the Bayesian method become obvious when the sample size n is small. Unfortunately, for the likelihood function, it is impossible to find the conjugate prior for, which means whatever prior we choose, the posterior distribution may not belong to any tractable distribution family. This problem no longer poses a serious obstacle to the application of Bayesian method after the widespread use of Markov Chain Monte Carlo ($MCMC$) methods. The main idea of Bayesian estimation using the $MCMC$ methods followed Dempster et al. (1977) by realizing a mixture model is a special case of incomplete data problem with the missing component indicator variables. The idea of Bayesian estimation was to estimate the augmented parameter by sampling from

the complete-data posterior distribution In many situations, we can simulate the parameter by using Gibbs sampling.

1.2.2 Estimation of Mixture Complexity in Finite Mixture Models

Frequentist Approach

A survey of literature shows that in the continuous and discrete cases, developing methods to determine mixture complexity has been an area of intense research for many years. In the continuous case, a variety of approaches for determining the mixture complexity have been discussed in the literature; see, for example, James et al. (2001) and Ishwaran et al. (2001), and references therein. James et al. (2001), for instance, used the Kullback-Leibler (KL) distance to construct a consistent estimator of mixture complexity, when the component densities are normal.

For the count data case, Schlattmann and Böhning (1993) used the resampling approach of McLachlan (1987) to determine the mixture complexity in their application of Poisson mixtures to disease mapping. Also, Pauler et al. (1996) used this method to determine the mixture complexity in their modeling of anticipatory saccade counts from schizophrenic patients and controls. Karlis and Xekalaki (1999) determined the mixture complexity using a sequential testing procedure based on likelihood ratio test (LRT) that utilizes a resampling approach.

Model Selection Approaches

Henna (1985) considered a model selection approach for estimation of mixture complexity in finite mixtures. Common model selection methods based on a penalized likelihood, including AIC and BIC have been considered in fitting mixture models by Leroux (1992). He considered a sequence of nested mixture models with possible number of components $k = 1, ..., n$, and proposed an estimator $\hat{k}$ for the true value of k. The penalized maximum likelihood methods

usually help reduce overestimation of the model. However, Leroux pointed out that the estimated number of components is at least as large as the true number.

Chen and Kalbfleisch (1996) proposed a method to estimate the number of mixture components based on the penalized minimum distance, and showed that their estimator is consistent. They calculated the distance between the CDF of the fitting distribution, $F(x|G)$, and the CDF of the empirical distribution, $F_n(x)$, with the penalty term of the summation of the log weights, then chose the model with the minimum distance.

Recently, Chen and Khalili(2006,2008) proposed a new method, for estimating the number of mixture components in finite mixture models by combining the strength of two existing methods. The first is the Modified likely hood proposed by Chen and Kalbfleisch(1996). The second is the variable selection method called the smoothly clipped absolute deviation or SCAD, by Fan and Li(2001). For this reason , they called the new method as $MSCAD$.

BAYESIAN METHODS

Richardson and Green (1997) described a fully Bayesian treatment for mixture modeling. They jointly modeled the number of components k, the identity (or label) of the group from which each observation is drawn (the unobserved indicator) z, and the component parameters θ, π. The inference of k is made based on the simulated posterior probabilities. The difficulty with a full model is that distributions with different k will have different parameter dimensions, which violates a necessary condition for the convergence of the usual $MCMC$ method. Richardson and Green used the reversible jump $MCMC$ methods developed by Green (1995) to sample from mixtures with varying number of components. Their program moves between models with different k by splitting one mixture component into two or combining two into one.

Stephens (2000) proposed a new $MCMC$ algorithm for the mixture problem when k is unknown. He considered the parameters of the model as a point process with each point rep-

resenting a component density, and constructed a continuous time Markov birth-death process. Stephens (2000)'s method is competitive to the reverse jump MCMC both in results and computation complexity. McGrory and Titterington(2007) shows the variational approach to model selection in the case of mixtures of Gaussian distributions leads to an automatic choice of model complexity. They use the Deviance Information Criterion(DIC)of Spiegelhalter et al. (2002). They have shown that it is a practical and useful alternative to $MCMC$ for the analysis of mixtures of Gaussian.

Recently, Ishwaran et al. (2001) considered estimation of mixture complexity using a Bayesian model selection. They assumed that there is a fixed upper bound K for the true mixture complexity, and decomposed the marginal density for the data into K densities, each corresponding to the contribution from the prior with $k = 1, \ldots, K$ components. The weighted Bayes factor was then used for selecting the dimension k. They used an i.i.d. generalized weighted Chinese restaurant ($GWCR$) Monte Carlo algorithm, and proved that the posterior distribution is consistent.

1.2.3 ROBUST ESTIMATION OF MIXTURE COMPLEXITY

This thesis concentrates on estimation methods that are less sensitive to model misspecification and extreme values, and seeks a semi-parametric density estimator of the form

$$\hat{f}_n^*(x) = \sum_{i=1}^{\hat{m}_n} \hat{\pi}_i f(x|\hat{\boldsymbol{\phi}}_i), \tag{1.2.3}$$

with the property that $\hat{m}_n \to m_0$ almost surely (a.s.) as $n \to \infty$. Consequently, if $f_0 \in \mathcal{F}_m$, 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 $\hat{f}_n^* \to f_0$.

As mentioned earlier, in many applications, there is not much *a priori* information about the mixture complexity and, hence, has to be inferred from the data. Estimation of mixture complexity is a fundamental, yet challenging, problem. Correct identification of mixture complexity followed by efficient estimation of all the mixture parameters would lead to finding a mixture with the fewest possible components which provides a satisfactory fit.

MHD APPROACH

Recently, Woo and Sriram (2006, 2007) used the Hellinger distance between p.d.f. (or p.m.f.) f and g defined by

$$H^2(f,g) = ||f^{1/2} - g^{1/2}||_2^2, \tag{1.2.4}$$

where $||\cdot||_2$ is the L_2 norm, as the basis for constructing an estimator of mixture complexity m_0; see equation (1.1.2). We now briefly review their construction; see Woo and Sriram (2006) for more details. Suppose $\hat{f}_n$ is a kernel density estimator (or an empirical mass function in the discrete case) of f_0. For each integer $m > 0$, define $\hat{f}^m = \arg\min_{f\in\mathcal{F}_m} H(f, \hat{f}_n)$ and $f_0^m = \arg\min_{f\in\mathcal{F}_m} H(f, f_0)$. When $m > 0$ is known, the MHD estimator of $\boldsymbol{\theta}_m$ is denoted by $\hat{\boldsymbol{\theta}}_{n,m}^{MHD} = \arg\min_{\boldsymbol{t}_m\in\Theta_m} H(f_{\boldsymbol{t}_m}, \hat{f}_n)$. Note that $\hat{f}^m = f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}$. By treating estimation of m_0 as a model selection problem, Woo and Sriram (2006) introduced a Hellinger Information Criterion

$$HIC = H^2(\hat{f}^m, \hat{f}_n) + n^{-1}b(n)\nu(m), \tag{1.2.5}$$

where $b(n)$ depends only on n and $\nu(m)$ is the number of parameters in the mixture model, and motivated the following estimator of m_0 defined by

$$\hat{m}_n^{MHD} = \min\{m : H^2(\hat{f}^m, \hat{f}_n) \le H^2(\hat{f}^{m+1}, \hat{f}_n) + \alpha_{n,m}\}. \tag{1.2.6}$$

Here, $\{\alpha_{n,j}; j \ge 1\}$ are positive sequences of threshold values chosen in such a way that they converge to zero as $n \to \infty$.

Treating the continuous case and the discrete case separately, Woo and Sriram (2006, 2007) established the following result under certain regularity conditions: If f_0 is a finite mixture with mixture complexity $m_0 < \infty$, then for any sequence $\alpha_{n,m} \to 0$ the estimator $\hat{m}_n^{MHD}$ is strongly consistent, i.e., $\hat{m}_n^{MHD} \to m_0$ a.s. as $n \to \infty$. If f_0 is *not* a finite mixture, then $\hat{m}_n^{MHD} \to \infty$ a.s. Furthermore, Woo and Sriram (2006) showed via Monte Carlo simulations for a wide variety of normal mixtures that, when the model is correctly specified, the performance of their estimator is competitive with several others in the literature in

correctly identifying the true mixture complexity. Similar studies for a variety of Poisson mixtures were carried out in Woo and Sriram (2007). Also, Woo and Sriram (2006) showed that their basic construction, being firmly rooted in the minimum Hellinger distance approach, enables their estimator to naturally inherit the property of robustness and correctly determine the mixture complexity, even when the postulated model is a mixture of normals but the data are generated from mixtures with moderate to more extreme symmetric departure from component normality. In the discrete case, Woo and Sriram (2007) showed that similar robustness results hold, when the postulated model is a Poisson mixture but the data comes from negative binomial mixtures with moderate to more extreme overdispersion in one of its components.

For the Monte Carlo simulations and data analysis in the continuous case, Woo and Sriram (2006) used the threshold value $\alpha_{n,m} = 3/n$, which they motivated as a choice based on the Akaike Information Criterion (AIC). For the discrete case, Woo and Sriram (2007) used two threshold values $\alpha_{n,m} = 2/n$ and $ln(n)/n$, motivated as choices based on the AIC and Schwarz Bayesian Criterion (SBC), respectively. Woo and Sriram (2006, 2007) also illustrated the performance of $\hat{m}_n^{MHD}$ for a hypertension data analyzed in Roeder (1994) and for three count data sets (two of which with zero-inflation) analyzed in Karlis and Xekalaki (1998, 1999 and 2001).

While the MHD estimation method does lead to efficiency and robustness, as noted in Scott (1999 and 2001) and Markatou (2000 and 2001), the method (in continuous case) requires a nonparametric kernel density estimator with proper choice of bandwidth and involves numerical integration, which makes the method computationally intensive, especially in the context of finite mixture models. Scott (1999) showed that the L_2E approach is relatively simple to setup even with some very complex model specifications, computationally feasible and leads to robust estimators like all other minimum distance methods. In fact, Scott (2001) showed that L_2E estimator performs much better than the MHD estimator, under data contamination. Motivated by simplicity and computational benefits associated with the

L_2E approach, next we proceed to propose an alternative estimator of mixture complexity m_0 based on L_2E, which is an important special case of the family of $MDPD$ measures (Basu et al. 1998).

$MDPD$ AND L_2E APPROACH

One way to avoid the challenges associated with MHD estimation is to use the robust estimation approach introduced by Basu et al. (1998), known as minimum density power divergence ($MDPD$) estimation. Basu et al. (1998) defined the following new class of distances known as *density power divergences*

$$d_\alpha(g,f) = \int \left[f^{1+\alpha}(x) - (1+1/\alpha)g(x)f^\alpha(x) + (1/\alpha)g^{1+\alpha}(x)\right] dx. \tag{1.2.7}$$

Let $\mathbf{F}$ denote the class of all distributions F with corresponding density f (of X_1) belonging to a class of density functions, say, $\mathcal{F}_0$. Define a density power divergence functional $T^{DPD}_{\alpha,m}$ on $\mathbf{F}$ by the requirement that for every $F \in \mathbf{F}$

$$T^{DPD}_{\alpha,m}(F) = \arg\min_{\boldsymbol{\theta}_m} \left[\int f^{1+\alpha}_{\boldsymbol{\theta}_m}(x)dx \; - (1+1/\alpha)\int f^{\alpha}_{\boldsymbol{\theta}_m}(x)dF(x)\right]. \tag{1.2.8}$$

Let $\hat{F}_n$ denote the empirical distribution of $\{X_i, i = 1,\ldots,n\}$. Then, we define a *Minimum Density Power Divergence* ($MDPD$) estimator $\hat{\boldsymbol{\theta}}^{MDPD}_{n,m} = T^{DPD}_{\alpha,m}(\hat{F}_n)$, where

$$\hat{\boldsymbol{\theta}}^{MDPD}_{\alpha,n,m} = \arg\min_{\boldsymbol{\theta}_m} \left[\int f^{1+\alpha}_{\boldsymbol{\theta}_m}(x)dx - (1+1/\alpha)n^{-1}\sum_{i=1}^{n} f^{\alpha}_{\boldsymbol{\theta}_m}(X_i)\right]. \tag{1.2.9}$$

In the discrete case, $\int f^{1+\alpha}_{\boldsymbol{\theta}_m}(x)dx$ in the definition of $\hat{\boldsymbol{\theta}}^{MDPD}_n$ will be replaced by $\sum_{k=0}^{\infty} f^{1+\alpha}_{\boldsymbol{\theta}_m}(k)$. Note that, when $\alpha = 1$, $d_1(f,g) = L_2(f,g) = \int[f(x) - g(x)]^2dx$. Scott (2001) considered this case and defined an L_2E estimator of $\boldsymbol{\theta}_m$ defined by

$$\hat{\boldsymbol{\theta}}^{L_2E}_{n,m} = \arg\min_{\boldsymbol{\theta}_m} \left[\int f^2_{\boldsymbol{\theta}_m}(x)dx - 2n^{-1}\sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(X_i)\right]. \tag{1.2.10}$$

Now, for each m, let $L(\boldsymbol{\theta}_m) = \left[\int f^2_{\boldsymbol{\theta}_m}(x)dx - 2n^{-1}\sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(X_i)\right]$. Then, we propose the following L_2E estimator of mixture complexity m_0 defined by

$$\hat{m}^{L_2E}_n = \min\{m : L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m}) \leq L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m+1}) + \alpha_{n,m}\}. \tag{1.2.11}$$

where, the sequence $\{\alpha_{n,m}\}$ is chosen so that it goes to 0 as $n \to \infty$.

Unlike the definition of $\hat{m}_n^{MHD}$ in (1.2.6), our definition of $\hat{m}_n^{L_2E}$ avoids specification of kernel density estimator and associated choice of bandwidths in the continuous case. This would make $\hat{m}_n^{L_2E}$ computationally more feasible, resulting in substantial reduction in computation time, than its MHD counterpart.

As mentioned earlier, the L_2E method of Scott (1999) is a special case of Basu et al. (1998), who suggest that for values of $\alpha \in [0, 1/4]$ the $MDPD$ estimators are more robust and have modest loss of efficiency. Scott (1999) makes an interesting observation that, while the $MDPD$ estimators may be more efficient than L_2E estimators for $\alpha \in [0, 1/4]$, for mixture models, the key integral $\int f_{\boldsymbol{\theta}_m}^{1+\alpha}(x)dx$ in (1.2.8) or (1.2.9) cannot be computed in a closed form except at α values 0 and 1, and the numerical integration involved for other values of α are computationally intensive.

In this thesis, we focus only on $\hat{m}_n^{L_2E}$. We treat the continuous case and the discrete case separately and show that $\hat{m}_n^{L_2E}$ is strongly consistent. Furthermore, we carry out several Monte Carlo studies and data analysis, and compare the performance of $\hat{m}_n^{L_2E}$ with other procedures in the literature. Moreover, we extensively study the robustness properties of $\hat{m}_n^{L_2E}$ under model misspecification, as done in Woo and Sriram (2006, 2007), and compare the results with those of $\hat{m}_n^{MHD}$.

1.3 OUTLINE OF THE THESIS

In Chapter 2, we consider mixture complexity estimation for the count data based on the L_2E. Here, we establish the strong consistency of the mixture complexity estimator under certain regularity conditions and assess their robustness against model misspecification via extensive Monte Carlo simulations. Also, we illustrate the performance of our estimator for three real count datasets with overdispersion and/or zero-inflation.

In Chapter 3, we consider mixture complexity estimation for the continuous case and propose based on the L_2E. Here, once again, we establish the strong consistency of the

mixture complexity estimator under certain regularity conditions and assess their robustness against model misspecification via extensive Monte Carlo simulations. Once again, analyze three datasets arising in wide variety of fields.

Each chapter is self-contained in terms of describing and highlighting the performance of the above mentioned methods, but we give a concluding summary of both methods and discuss future work in Chapter four.

1.4 References

[1] Aitkin, M., and Wilson, G. T. (1980), "Mixture Models, Outliers, and the EM Algorithm," *Technometrics*, 22, 325-331.

[2] Arley and Buch, (1950), *Introduction to the Theory ofProbability andStatistics*, New York: Wiley.

[3] Basu, A., Harris, I.R., Hjort, H.L., and Jones, M.C. (1998), "Robust and efficient estimation by minimizing a density power divergence," *Biometrika*, 85, 549 - 560.

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

[5] Blischke, W.R. (1962), "Moment estimatorsfor the parameters of a mixture of two binomial distributions," *The Annals of Math Statistics*, 33, 444-454.

[6] Bowman ,A.W., (1984), "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, 71, 353 - 360.

[7] Chen,J. and Kalbfleisch. J. D. (1996), "Penalized minimum distance estimates in finite mixture models". *Canad. J Statist* , 24 , 167-175.

[8] Chen, J.and Khalili, A. (2006), "Order selection in finite mixture models," *Technical Report*, Department of Statistics and Actuarial Science University of Waterloo, Canada.

[9] Chen, J.and Khalili, A. (2008), "Order selection in finite mixture models with a nonsmooth penalty," *Journal of the American Statistical Association*, 103, 1674-1683.

[10] Chen, J., Li, P., Tan and Xianming. (2007), "Inference for von Mises mixtures in mean direction and concentration parameters"'. *J. Syst. Sci. Math. Sci.* , 27, No. 1, 59-67

[11] Choi and BulGren . (1968), "An estimation procedurefor mixtures of distributions," *Royal Statist. Soc. Ser*, 30, 444-460.

[12] Clarke, B. R., and Heathcote, C. R. (1994), "Robust Estimation of k- Component Univariate Normal Mixtures," *Annals of the Institute of Statistical Mathematics*, 46, 83-93.

[13] Cohen, A.C. (1963), *Estimation in mixtures of discrete distributions*, in Proc. of the InternationalSymposium on Classical and Contagious Discrete Distributions, Pergamon Press, New York, 351-372.

[14] Cooper P. W.,(1967), Some topics on nonsupervised adaptive detection for multivariate normal distributions, in Computer and Information Sciences, 11, J. T. Tou, ed., Academic Press, New York,143-146.

[15] Cutler, A., and Cordero-Braňa, O. I. (1996), "Minimum Hellinger distance estimation for finite mixture models," *Journal of the American Statistical Association* 91, 1716-1723.

[16] Day , N.E. (1969), Estimating the components of a mixture of normal distributions, "'*Biometrika*, 56, 463-474.

[17] Dellaportas, P., Karlis, D and Xekalaki, E. (1997), " Bayesian analysis of finite Poisson mixtures," Technical Report, Department of Statistics, Athens University of Economics and Business.

[18] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum-Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

[19] Fan,J. and Li,R. (2001),"Variable selection via non-concave penalized likelihood and its oracle properties, " *Journal of the American Statistical Association*, 96, 1348-1360.

[20] Fruhwirth-Schnatter, S. (2006). "Finite Mixture and Markov Switching Models," *Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.*

[21] Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82,711-732.

[22] Gumbel (1939), "La dissection d'une repartition," *Annales de I'Universite de Lyon*, 3 39-51.

[23] Henna,J. (1985). " On estimating of the number of constitueilts of a finite mixture of continuous distributions". *Ann. Inst. Statist. Math.*, 37 , 235-240.

[24] Ishwaran, H., James, L. F., and Sun, J. (2001),"Bayesian model selection in finite mixtures by marginal density decompositions," *Journal of the American Statistical Association*, 96, 1316-1332.

[25] James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent estimation of mixture complexity," *The Annals of Statistics*, 29, 1281-1236.

[26] John,S.(1970), On identifying the population of origin of each observation in a mixture of observationsfrom two normalpopulations," *Technometrics*, 12 , 553-563.

[27] Kabir, A.B.M, (1968), "On Estimation of parameter of a finite mixture of distributions," *Royal Statist. Soc*, 30, 472-482.

[28] Kaestner, R. (1999), "Health insurance, the quantity and quality of prenatal care, and infant health,", *Inquiry*, 36, 162-175.

[29] Karlis D., Xekalaki E. (1998) "Minimum Hellinger distance estimation for finite Poisson mixtures,". *Computational Statistics and Data Analysis* , 29, 81-103.

[30] Karlis, D. and Xekalaki, E. (1999), "On testing for the number of components in a mixed Poisson model," *Annals of Institute of Statistical Mathematics*, 51, 149-162.

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

[32] Leroux, B. G. (1992)., "Consistent estimation of a mixing distribution". *Ann. Statist*, 20 , 1350-1360.

[33] Macdonald ,P.D.M. (1971), "An estimation procedure for mixtures of distribution", *J. Royal Statist. Soc. Ser.*, 33, 326-329.

[34] Markatou, M. (2000), "Mixture models, robustness and the weighted likelihood methodology", *Biometrics*, 56, 483-486.

[35] Markatou, M. (2001), "A closer look at the weighted likelihood in the context of mixtures", *Probability and Statistical Models with Applications*, Charalambides, C.A., Koutras, M.V. and Balakrishnan, N. (eds), Chapman and Hall/CRC, 4447-467.

[36] Markatou, M., Basu, A., and Lindsay, B. G. (1998), "Weighted likelihood estimating equations with a bootstrap root search", *Journal of the American Statistical Association*, 93, 740-750.

[37] McGrory, C.A. and Titterington, D. M. (2007),"Variational approximations in Bayesian model selection for finite mixture distributions", *Computational Statistics and Data Analysis*, 51, 5352-5367.

[38] McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Journal of the Royal Statistical Society Series C (Applied Statistics) 36, 318-324.

[39] McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

[40] Munech, H. (1936), 'Probability distribution of protection test results", *Journal of the American Statistical Association*, 31, 677-689.

[41] Pauler, D. K., Escobar, M. D., Sweeney, J. A. and Greenhouse, J. (1996), "Mixture models for eye-tracking data: A case study," *Statistics in Medicine*, 15, 1365-1376.

[42] Pearson (1894), Contributions to the mathematical theory of evolution, Phil. Trans. Royal Soc., 185A , 71-110.

[43] Pearson (1915), On certain types of compound frequency distributions in which the components can be individually described by binomial series, "'*Biometrika*, 11 , 139-144.

[44] Polllard(1934), On the relative stability of the median and the arithmetic mean, with particular reference to certain frequency distributions which can be dissected into normal distributions,, "'*Ann.Math.Statist*, 5 , 227-262.

[45] Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with unknown number of components (with discussion)". *Journal of the Royal Statistical Society*, Series B, 59, 731–92

[46] Rider, P.R. (1961), " The method of moments applied to a mixture of two exponential distributions," *Ann. Math. Statist*, 32, 143-147

[47] Roeder, K. (1994), " A graphical technique for determining the number of components in a mixture of normals," *Journal of the American Statistical Association*, 89, 487-495.

[48] Rudemo, M. (1982), "Empirical choice of histograms and kernel density estimators,"*Scand. J.Statist.*, 9,65-78.

[49] Schilling, W. (1947), " A frequency distribution represented as the sum of two Poisson distributions," *Journal of the American Statistical Association*, 42, 407-424.

[50] Schlattmann, P. and Böhning, D. (1993), "Mixture models and disease mapping," *Statistics in Medicine*, 12, 943-950.

[51] Scott, D. W. (1998), "On fitting and adapting of density estimates," *Computing Science and Statistics*, S. Weisberg, Ed., 30, 124 - 133.

[52] Scott, D.W. (1999), "Remarks on fitting and interpreting mixture models," *Computing Science and Statistics*, K. Berk and M. Pourahmadi, Eds., 31, 104-109.

[53] Scott, D. W. (2001), "Parametric statistical modeling by minimum integrated square error," *Technometrics*, 43, 274-285.

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

[55] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002). "Bayesian measures of model complexity and fit (with discussion)"',*Journal of the Royal Statistical Society Series B*, 64 583-639.

[56] Stephens, M. (2000), "Bayesian Analysis of Mixture Models with an Unknown Number of Components: An Alternative to Reversible Jump Methods"'. *Annals of Statistics* 28, 40-74

[57] Woo, Mi-Ja, and Sriram, T. N. (2006), "Robust estimation of mixture complexity," *Journal of American Statistical Association*, 101, 1475-1486.

[58] Woo, Mi-Ja and Sriram, T. N. (2007), "Robust estimation of mixture complexity for count data," *Computational Statistics and Data Analysis*, 51, 4379-4392.

[59] Woodward, W. A., Parr, W. C., Schucany. W. R. and Lindsay, H. (1984, "A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion." *Journal o f the American Statistical Association*, 79, 590-598.

[60] Woodward, W, A,, Whitney, P,, and Eslinger, P, (1995,, Hellinger Distance Estimation of Mixture Proportions unpublished manuscript,

Chapter 2

L_2E ESTIMATION OF MIXTURE COMPLEXITY FOR COUNT DATA[1]

[1]Umashanger, T. and Sriram, T. N. Accepted by: *Computational Statistics and Data Analysis.* Reprinted here with permission of publisher, 05/22/2009.

ABSTRACT

For count data, robust estimation of the number of mixture components in finite mixtures is revisited using L_2 distance. An information criterion based on L_2 distance is shown to yield an estimator, which is also shown to be strongly consistent. Monte Carlo simulations show that our estimator is competitive with other procedures in correctly determining the number of components when the data comes from Poisson mixtures. When the data comes from a negative binomial mixture but the postulated model is a Poisson mixture, simulations show that our estimator is highly competitive with the minimum Hellinger distance (MHD) estimator in terms of robustness against model misspecification. Furthermore, we illustrate the performance of our estimator for real datasets with overdispersion and/or zero-inflation. Computational simplicity combined with robustness property makes the L_2E approach an attractive alternative to other procedures in the literature.

Key words and Phrases: Finite mixtures; Mixture complexity; Information criterion; Threshold; Consistency; Robustness.

2.1 INTRODUCTION

Applied statisticians face an array of practical issues when analyzing data, the most vexing of which is identification of data points that are outliers. Such data points if not appropriately down-weighted can dramatically affect parameter estimates, leading to poorly fitted models and incorrect interpretations. In such instances, robust variations of estimation are the only feasible alternatives.

It has been known for some time that likelihood methodology can be replaced by minimum distance criteria, which yield estimators that are inherently robust. Minimum distance methods for finite mixtures with fixed number of components are well studied. Cutler and Cordero-Braňa (1996) developed a minimum Hellinger distance (MHD) estimator (Beran, 1977) of unknown component parameters, and Karlis and Xekalaki (1998) developed a MHD

estimator of parameters in Poisson mixtures. Both showed that MHD estimators achieve efficiency at the true model density and simultaneously possess desirable robustness properties under gross-error contaminations, thereby reconciling the conflicting concepts of robustness and efficiency.

A complication in many applications is that there is not much *a priori* information about the number of mixture components, termed *mixture complexity*. Estimation of mixture complexity is a fundamental problem because correct identification of mixture complexity followed by efficient estimation of all parameters would lead to finding a mixture with the fewest possible components. Developing methods to determine mixture complexity has been an active area of research in recent years; see, for example, Schlattmann and Böhning (1993), Roeder (1994), Pauler et al. (1996), Dellaportas et al. (1997), Basu et al. (1998), Karlis and Xekalaki (1999), James et al. (2001), Ishwaran et al. (2001), Shen (2004), Chen and Khalili (2008) and references therein. Recently, Woo and Sriram (2006, 2007) treated the estimation of mixture complexity as a model selection problem and constructed an estimator of mixture complexity as a by-product of minimizing a Hellinger Information Criterion (HIC). They showed that their estimator of mixture complexity is consistent and also illustrated through simulations the ability of their estimator to correctly determine the number of components when the postulated mixture model is correct. In addition, they showed that their estimator continues to perform well even when the data comes from a model that is somewhat different from the postulated mixture model; see Woo and Sriram (2006, 2007) for more details.

While the MHD-based estimator of mixture complexity has attractive large sample and robustness features, the implementation of the MHD algorithm for continuous and count data require specifications which place severe burden on the computation of MHD estimates. Scott (1998, 1999, 2001 and 2004) introduced an alternative minimum distance estimation method based on an integrated squared error criterion, termed L_2E, which has many computational advantages over MHD. The L_2E approach is a special case of a general method introduced by Basu et al. (1998), who devised a whole continuum of density-based power

divergence estimators that begin with the MLE and interpolate to the L_2E estimator and beyond. While the L_2E approach has computational advantages, it suffers from moderate loss of efficiency at the parametric model relative to MHD and maximum likelihood estimators. Nonetheless, within the family of density-based power divergence measures, the L_2E approach has the distinct advantage that a key integral can be computed in closed form, especially for finite mixtures; see Scott (2001). These findings and others discussed in section 2.4 below motivate us to investigate the L_2E approach for the estimation of mixture complexity, when all the component parameters are unknown.

In section 2.2, we introduce the L_2E criterion due to Scott and propose an estimator of mixture complexity based on an L_2 model selection criterion. A consistency theorem for the estimator is stated in section 2.3 but proved in the Appendix. Computational details and advantages concerning our estimator are given in section 2.4. In section 2.5.1, we carry out extensive Monte Carlo studies for correctly specified 2-, 3- and 4- component Poisson mixtures and, in each case, compare the ability of our estimator in correctly determining the mixture complexity with other procedures in the literature. In section 2.5.2, we examine the robustness of our estimator, when the postulated mixture model is incorrect. In section 2.6.1, 2.6.2 and 2.6.3 we analyze three different count data sets with overdispersion and zero-inflation. Overall conclusions are given in section 2.7. We begin with some basic notations and definitions.

2.2 L_2E ESTIMATOR

The integrated squared distance has been used as the goodness-of-fit criterion in nonparametric density estimation for a long time. Scott (1999) introduced an alternative minimum distance estimation method based on integrated squared error criterion, termed L_2E, which avoids the use of nonparametric kernel density estimators. Scott showed that the L_2E is especially well suited for parameter-rich models such as mixture models.

The following discussion introduces the basic notations, L_2E criterion and an estimator of mixture complexity. Consider a parametric family of probability mass functions (p.m.f.) $\mathcal{F}_m = \{f_{\boldsymbol{\theta}_m} : \boldsymbol{\theta}_m \in \Theta_m \subseteq R^p\}$ concentrated on $\mathcal{X} = \{0, 1, 2, \ldots\}$ such that $f_{\boldsymbol{\theta}_m}$ can be represented as a finite mixture of the form

$$f_{\boldsymbol{\theta}_m}(x) = \sum_{i=1}^{m} \pi_i f(x|\boldsymbol{\phi}_i), \quad x \in \mathcal{X}, \tag{2.2.1}$$

where $m \geq 1$ is a finite integer, $f(x|\boldsymbol{\phi}_i)$ is the component p.m.f., $\boldsymbol{\phi}_i \in \Phi \subseteq R^s$, mixing proportions $\pi_i \geq 0$, $\sum_{i=1}^{m} \pi_i = 1$ for $i = 1, \ldots, m$, $\boldsymbol{\theta}_m = (\pi_1, \ldots, \pi_{m-1}, \boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_m^T)^T$ and R^p is the p-dimensional Euclidean space. For each m, the functional form of component p.m.f. is known, but $\boldsymbol{\theta}_m$ is unknown. Note that $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$ for all m.

Let $X_1, \ldots, X_n$ be independent random variables taking values in $\mathcal{X}$ with an unknown p.m.f. $f_0 \in \Gamma$, where Γ denotes the set of all p.m.f.'s defined on $\mathcal{X}$. Define the *index of the economical representation* of f_0 relative to $\mathcal{F}_m$ as

$$m_0 = m(f_0) = \min\{m : f_0 \in \mathcal{F}_m\}. \tag{2.2.2}$$

If indeed f_0 is a finite mixture defined in (2.2.1), 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. Our goal is to find a semi-parametric estimator of the form

$$\hat{f}_n^*(x) = \sum_{i=1}^{\hat{m}_n} \hat{\pi}_i f(x|\hat{\boldsymbol{\phi}}_i), \tag{2.2.3}$$

with the property that $\hat{m}_n \to m_0$ almost surely (a.s.) as $n \to \infty$. Consequently, 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 $\hat{f}_n^* \to f_0$.

To this end, define the squared L_2 distance between two p.m.f.'s $g, f \in \Gamma$ as

$$\begin{aligned} L_2(g, f) &= \sum_{x=0}^{\infty} (g(x) - f(x))^2 \\ &= \sum_{x=0}^{\infty} g^2(x) - 2\sum_{x=0}^{\infty} g(x)f(x) + \sum_{x=0}^{\infty} f^2(x). \end{aligned} \tag{2.2.4}$$

For each fixed m, define a L_2E functional $T_m^{L_2E}$ on Γ by the requirement that for every $f \in \Gamma$

$$T_m^{L_2E}(f) = \{\boldsymbol{\theta}_m \in \Theta_m : L_2(f_{\boldsymbol{\theta}_m}, f) = \min_{\boldsymbol{t}_m \in \Theta_m} L_2(f_{\boldsymbol{t}_m}, f)\}. \tag{2.2.5}$$

Since $\sum_{x=0}^{\infty} f^2(x)$ in (2.2.4) does not involve $\boldsymbol{\theta}_m$, the functional

$$T_m^{L_2E}(f) = \arg\min_{\boldsymbol{\theta}_m} \left[\sum_{x=0}^{\infty} f_{\boldsymbol{\theta}_m}^2(x) - 2 \sum_{x=0}^{\infty} f_{\boldsymbol{\theta}_m}(x) f(x) \right]. \tag{2.2.6}$$

Let $\hat{f}_n$ be the empirical mass function given by

$$\hat{f}_n(x) = n^{-1} \sum_{i=1}^{n} I_{\{X_i = x\}}, \quad x = 0, 1, \ldots, \tag{2.2.7}$$

where I_A is the indicator of set A. Then, we can define the L_2E estimator of $\boldsymbol{\theta}_m$ as

$$\hat{\boldsymbol{\theta}}_{n,m}^{L_2E} = T_m^{L_2E}(\hat{f}_n) = \arg\min_{\boldsymbol{\theta}_m} \left[\sum_{x=0}^{\infty} f_{\boldsymbol{\theta}_m}^2(x) - 2n^{-1} \sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(X_i) \right]. \tag{2.2.8}$$

In order to propose an estimator of m_0 in (2.2.2), as in Woo and Sriram (2006 or 2007, section 2), we introduce a model selection criterion based on $L_2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}, \hat{f}_n)$ defined by

$$LIC = L_2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}, \hat{f}_n) + n^{-1} \ln g(m),$$

where $g(m)$ is a penalty function depending on m. The definition of LIC is motivated by the work of Poland and Shachter (1994; see section 4 and 5). Here, the value of m yielding the minimum LIC specifies the best model. Since $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$, we have $L_2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}, \hat{f}_n) \geq L_2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{L_2E}}, \hat{f}_n)$. 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 less than that for model $(m+1)$. That is, this heuristic stops when

$$L_2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}, \hat{f}_n) + n^{-1} \ln g(m) \leq L_2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{L_2E}}, \hat{f}_n) + n^{-1} \ln g(m+1)$$

or, equivalently,

$$L_2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}, \hat{f}_n) - L_2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{L_2E}}, \hat{f}_n) \leq n^{-1} \ln[g(m+1)/g(m)].$$

Now, if we let $g(m) = am^k$ (Poland and Shachter, 1994; see section 4 and 5) then $n^{-1} \ln[g(m+1)/g(m)] = n^{-1}k \ln((m+1)/m)$. This heuristic naturally leads to the following estimator of m_0 defined by

$$\hat{m}_n^{L_2E} = \min\{m : L_2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}, \hat{f}_n) \leq L_2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{L_2E}}, \hat{f}_n) + \alpha_{n,m}\}, \tag{2.2.9}$$

where $\{\alpha_{n,m}\}$ is a sequence such that it goes to 0 as $n \to \infty$. For simulations and data analysis, we set $k = 0.6$ and define $\alpha_{n,m} = n^{-1}0.6 \ln((m+1)/m)$, which is referred in the rest of the article as the $L_2E(LIC)$ threshold. In section 2.5.2, we also use a SBC-type threshold, $\alpha_{n,m} = n^{-1}0.6 \ln(n) \ln((m+1)/m)$, which is referred in the rest of the article as the $L_2E(SBC)$ threshold. Our empirical studies with different values of k in $\alpha_{n,m}$ showed that $k = 0.6$ performs the best in all our simulations and data analysis given in section 2.5. This is why we set $k = 0.6$ in our threshold.

2.3 CONSISTENCY THEOREM

The main theoretical result of the article is the consistency of $\hat{m}_n^{L_2E}$, which is stated as a theorem below. First, we state a Proposition giving regularity conditions for the existence and uniqueness of $T_m^{L_2E}(f)$ in (2.2.5). The proof of the Proposition and the theorem are given in the Appendix.

Theorem. Suppose the assumptions of the Proposition (see Appendix) hold. If f_0 is a finite mixture with mixture complexity $m_0 < \infty$, then for any sequence $\alpha_{n,m} \to 0$

$$\hat{m}_n^{L_2E} \to m_0 \quad \text{a.s.}$$

as $n \to \infty$, where $\hat{m}_n^{L_2E}$ and m_0 are as defined in (2.2.9) and (2.2.2), respectively. If f_0 is *not* a finite mixture, then $\hat{m}_n^{L_2E} \to \infty$ a.s.

2.4 COMPUTATIONAL DETAILS

Computation of an estimate of mixture complexity using (2.2.9) is clearly an iterative procedure. To this end, note that we can re-write $\hat{m}_n^{L_2E}$ in (2.2.9) as

$$\hat{m}_n^{L_2E} = \min\{m : L(\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}, \hat{f}_n) \leq L(\hat{\boldsymbol{\theta}}_{n,m+1}^{L_2E}, \hat{f}_n) + \alpha_{n,m}\}, \tag{2.4.10}$$

where

$$L(\boldsymbol{\theta}_m, \hat{f}_n) = \left[\sum_{x=0}^{\infty} f_{\boldsymbol{\theta}_m}^2(x) - 2n^{-1}\sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(X_i)\right]. \tag{2.4.11}$$

Note that our L_2E objective function, $L(\boldsymbol{\theta}_m, \hat{f}_n)$, depends only on the postulated parametric mixture model and the data $X_1, \cdots, X_n$. This simple structure enables us to use the built-in nonlinear minimization (nlm) routine in R to minimize the objective function with respect $\boldsymbol{\theta}_m$ for each $m \geq 1$.

The procedure starts by assuming that the data comes from a mixture with single component $f_{\boldsymbol{\theta}_1}$. Then, an estimate $\hat{\boldsymbol{\theta}}_{n,1}^{L_2E}$ which minimizes $L(\boldsymbol{\theta}_1, \hat{f}_n)$ (see (2.4.11)) is computed. This yields $L(\hat{\boldsymbol{\theta}}_{n,1}^{L_2E}, \hat{f}_n)$. Next, another component is added yielding a mixture with two components ($m = 2$) and an estimate $\hat{\boldsymbol{\theta}}_{n,2}^{L_2E}$ which minimizes $L(\boldsymbol{\theta}_2, \hat{f}_n)$ is computed, yielding $L(\hat{\boldsymbol{\theta}}_{n,2}^{L_2E}, \hat{f}_n)$. The difference $L(\hat{\boldsymbol{\theta}}_{n,1}^{L_2E}, \hat{f}_n) - L(\hat{\boldsymbol{\theta}}_{n,2}^{L_2E}, \hat{f}_n)$ is then compared with the threshold value $\alpha_{n,1}$. The above procedure of adding one more component to the previous mixture is repeated until the first value $m = m^*$ for which the difference $L(\hat{\boldsymbol{\theta}}_{n,m^*}^{L_2E}, \hat{f}_n) - L(\hat{\boldsymbol{\theta}}_{n,m^*+1}^{L_2E}, \hat{f}_n)$ falls below the threshold value α_{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 $\hat{\boldsymbol{\theta}}_{n,m^*}^{L_2E}$.

Common problems faced during minimization of objective functions involving finite mixtures are possible existence of equal components and empty components. Here, we seldom observed equal components. Although empty components do exist during minimizations, the estimate $\hat{m}_n^{L_2E}$ does not result in empty components whether or not our procedure correctly detects the true mixture complexity. Note that this is consistent with our theoretical definition of mixture complexity in (2.2.2), which does not allow empty components.

An important step in our iterative method is the choice of initial values. For our L_2E methodology, extensive preliminary simulations indicated that the final estimate of mixture complexity is not affected by the choice of initial values. More specifically for given m, we chose initial values for the remaining parameters using three different methods, namely K-Means, H-cluster and sample(x,n) routines in R, for our preliminary simulations and found that the estimates of mixture complexity were not sensitive to different initial value choices. Therefore, we used a K-means method for all our simulations and data analysis given in this article.

With respect to computing time, on a typical desktop it took on the average about 6 seconds to obtain one value of $\hat{m}_n^{L_2E}$ based on a simulated dataset of size $n = 500$ from a Poisson mixture model with 4 components, which is the largest number components considered in Table 2.1 in section 2.11. Since our algorithm automatically provides L_2E estimates of the component parameters, the time reported above also includes the estimation of parameters and the overhead of generating a dataset. Furthermore, the number of iterations required for *nlm* in R to converge was usually no more than 10. The time reported here is based on using K-Means to choose initial values; this is slightly different for other initial value choices.

The L_2E method has distinct advantages over the other methods compared in this article. First, the objective function of L_2E is simpler than those for MHD and $MSCAD$, two methods to which L_2E is compared in the article. More specifically, the objective function for MHD, $-2\sum_{x=0}^{\infty} f_{\boldsymbol{\theta}_m}^{1/2}(x)\hat{f}_n^{1/2}(x)$, is a relatively more complicated function to minimize. In fact, Karlis and Xekalaki (1998) developed an EM type algorithm known as HELMIX to compute MHD estimates. Chen and Khalili (2008) developed a new penalized likelihood approach called $MSCAD$, which deviates from information-based methods such as AIC, SBC, HIC and Robust Information Criterion (RIC) due to Basu et al. (1998) [also see Shen, 2004]. The objective function for $MSCAD$ is also relatively more complicated because it involves a SCAD-type penalty, hence the name $MSCAD$. The $MSCAD$ method is also based on a revised EM algorithm, which uses the penalized likelihood instead of the log-likelihood.

Secondly, as for the choice of initial values, observations made by Karlis and Xekalaki (1998) show that the MHD parameter estimates are sensitive to the choice of initial values, which in turn affects the estimate of mixture complexity. Furthermore, HELMIX algorithm also shares some of the weaknesses of the EM algorithm in terms of slow convergence. Karlis and Xekalaki (1999), on the other hand, use a sequential testing procedure based on Likelihood Ratio Test (LRT) along with bootstrapping to determine the number of components. Since LRT also involves an EM algorithm, it shares the same drawbacks as above. While not mentioned explicitly in Chen and Khalili (2008), the fact that $MSCAD$ is also an EM type algorithm, it is also likely to share some of the drawbacks of EM in terms of slow convergence and choice of initial values. Furthermore, the $MSCAD$ procedure also requires a careful choice of tuning parameters for their SCAD penalty (Fan and Li, 2001). Finally, as for computing time, the time reported above for L_2E is substantially lower than those for MHD (Woo and Sriram, 2006, Section 7) and $MSCAD$ (Chen and Khalili, 2008, Section 4). These computational advantages make our L_2E approach a more attractive alternative to MHD, $MSCAD$ and LRT.

2.5 SIMULATION STUDIES

In this section, we carry out two different simulation studies in order to assess the ability of our estimator $\hat{m}_n^{L_2E}$ to correctly determine the number of components. In both the studies, the postulated model is a Poisson mixture. The first study assumes that the model is correctly specified, that is, data are generated from a Poisson mixture model. The second study examines the robustness of our estimator under model misspecification, that is, data are generated from a negative binomial mixture model, where one of the components is subject to low to severe overdispersion. These are described in the following two subsections, respectively.

2.5.1 POISSON MIXTURES

In order to assess the performance of $\hat{m}_n^{L_2E}$, we generated data from 2-, 3- and 4-component Poisson mixtures and compared the performance of our estimator with MHD, $MSCAD$ and LRT procedures mentioned in section 2.4. More specifically, we consider a total of 6 cases consisting of 2-, 3- and 4-component Poisson mixtures with respective parameter vectors:

$$\boldsymbol{\theta}_2 = (0.5, 1, 9); (0.8, 1, 9); (0.95, 1, 10)$$

$$\boldsymbol{\theta}_3 = (0.33, 0.33, 1, 5, 10); (0.45, 0.45, 1, 5, 10)$$

$$\boldsymbol{\theta}_4 = (0.25, 0.25, 0.25, 1, 5, 10, 15).$$

For each of the target mixtures, we implemented our computational algorithm described in section 2.4 for sample sizes $n = 100, 500$ using only the $L_2E(LIC)$ threshold (but not the $L_2E(SBC)$ threshold) defined in section 2.2. The reason for using only the $L_2E(LIC)$ threshold is that when the model is correctly specified, it performs better than the $L_2E(SBC)$ threshold. However, in section 2.5.2 we use both the thresholds; see section 2.7 for more discussion on the use of the two thresholds. For each sample size, we performed 500 Monte Carlo replications of our algorithm, each yielding an estimate of mixture complexity. Table 2.1 of section 2.11 gives the relative frequencies (out of 500 replications) of the number of components determined by our method for each parameter vector and sample size. For comparative purposes, Table 2.1 also lists the relative frequencies based on $MHD(AIC)$ (see Tables 1-2 in section 5.1 of Woo and Sriram, 2007), $MSCAD$ (see Tables 2 and 6-8 of Chen and Khalili, 2008) and LRT from Karlis and Xekalaki (1999). In Table 2.1 of section 2.11, 50% or above correct identifications are given in bold with an asterisk beside it.

Note from Table 2.1 that for the 2-component cases, our $L_2E(LIC)$ and the other procedures perform well, except in the case when $n = 100$ and one of the mixing proportions is small, the $MSCAD$ and LRT perform better than $L_2E(LIC)$ and $MHD(AIC)$. As for the 3-component cases, the performance of $L_2E(LIC)$ is slightly better than that of $MHD(AIC)$ but similar to $MSCAD$ and LRT. However, once again, when $n = 100$ and one of the

mixing proportions is small, the $MSCAD$ and LRT perform better than $L_2E(LIC)$ and $MHD(AIC)$. Finally, for the 4-component case considered here, only the $L_2E(AIC)$ and $MSCAD$ perform well for $n = 500$, but all the procedures fail to perform well when $n = 100$.

As noted in Woo and Sriram (2007), the failure of $L_2E(LIC)$ to detect the correct number of components when n and/or one of the mixing proportions is small may be attributable to the inherent robustness property of L_2E to ignore the presence of a component with small mixing proportion, especially for small samples. Overall, our L_2E procedure provides a competitive yet computationally simpler alternative to the MHD, $MSCAD$ and LRT methods.

2.5.2 ROBUSTNESS

Here, we assess the robustness of $\hat{m}_n^{L_2E}$ in terms of its ability to correctly identify the true mixture complexity when the postulated Poisson mixture model is incorrect. As in Woo and Sriram (2007), we assess the robustness of $\hat{m}_n^{L_2E}$ when the postulated model is a 2-component Poisson mixture $f_{\boldsymbol{\theta}_2}(x)$ with λ_1 and λ_2 as its component means, but the data are generated from a 2-component negative binomial mixture given by

$$f(x) = \pi f_1(x) + (1-\pi) f_2(x), \tag{2.5.12}$$

where, for $i = 1, 2$, $f_i(x) = \begin{pmatrix} r+x-1 \\ x \end{pmatrix} p_i^r (1-p_i)^x, x = 0, 1, \ldots$. Let f_1 and f_2 be the p.m.f.s associated with random variables, say, X_1 and X_2, respectively. Then, $E(X_i) = r(1-p_i)/p_i$, $Var(X_i) = r(1-p_i)/p_i^2$, for $i = 1, 2$. Furthermore, if for each $i = 1, 2$, $r \to \infty$ and $p_i \to 1$ such that $r(1-p_i) \to \lambda_i$, then $E(X_i) \to \lambda_i$ and $Var(X_i) \to \lambda_i$. This shows that the negative binomial family of distributions includes the Poisson distribution as a limiting case.

In our simulation studies, we consider two scenarios. In both the scenarios, we set the component mean of the sampling model to be the same as that of the postulated model, that is, $r(1-p_i)/p_i = \lambda_i$, for $i = 1, 2$. In the first scenario, we set $r = 10$ and $\lambda_1 = 1$ (this sets $E(X_1) = 1$ and $Var(X_1) = 1.1$), but vary the values of $E(X_2) = \lambda_2 = 2, 5, 7$

with corresponding values of $Var(X_2) = 2.4, 7.5, 11.9$. Notice that the values of $Var(X_2)$ are progressively much larger compared to the corresponding values of λ_2, creating a low to severe overdispersion in the second negative binomial component.

In the second scenario, we set $\lambda_1 = 1$ and $\lambda_2 = 10$ (this sets $E(X_1) = 1$ and $E(X_2) = 10$), but vary the values of $r = 10$, 20 and 40, which yield corresponding values of $(Var(X_1), Var(X_2)) = (1.1, 20), (1.05, 15), (1.025, 12.5)$. Note that, as the value of r decreases, the values of $Var(X_1)$ stay close to $E(X_1) = 1$ but the values of $Var(X_2)$ become much larger compared to $E(X_2) = 10$, once again creating a low to severe overdispersion in the second (negative binomial) component. Finally, in each of these two scenarios, we set the mixing proportion $\pi = 0.25$ and 0.5.

For each of the above set of parameter values in each scenario, count data are generated from the negative binomial mixture in (2.5.12), but the computational algorithm described in section 2.4 is implemented under the assumption that the class $\mathcal{F}_m$ defined in section 2.2 is a family of Poisson mixtures. Here, we perform simulation studies for three sample sizes $n = 100, 500, 1000$ using both the $L_2E(LIC)$ and $L_2E(SBC)$ thresholds defined in section 2.2. As before, we performed 500 Monte Carlo replications of our algorithm, each yielding an estimate of mixture complexity. Table 2.2 of section 2.11, gives the relative frequencies (out of 500 replications) of the number of components determined by our method for the first scenario and Table 2.3 of section 2.11 gives similar results for the second scenario. Once again, the percentage (50% or above) of correct identification is given in bold with an asterisk beside it.

The two scenarios considered in Tables 2.2 and 2.3, respectively, can be broadly classified into three types of overdispersion: Low ($\lambda_2 = 2$ or $r = 40$), Moderate ($\lambda_2 = 5$ or $r = 20$) and Severe ($\lambda_2 = 7$ or $r = 10$). The low overdispersion cases from Table 2.2 ($\lambda_2 = 2$) show that when $n = 100$, all the procedures fail, but when $n = 500$, both $L_2E(LIC)$ and $MHD(AIC)$ perform better than their respective SBC versions. However, when $n = 1000$, all the procedures perform well. The low overdispersion cases from Table 2.3 ($r = 40$) show

that all the procedures perform well at all sample sizes, except in the case when $n = 1000$ and $\pi_1 = 0.25$, both $L_2E(SBC)$ and $MHD(SBC)$ perform better than $L_2E(LIC)$ and $MHD(AIC)$.

The moderate overdispersion cases from Table 2.2 ($\lambda_2 = 5$) show that when $n \leq 500$, all the procedures perform well. However, when $n = 1000$ and $\pi_1 = 0.25$, both $L_2E(SBC)$ and $MHD(SBC)$ perform better than $L_2E(LIC)$ and $MHD(AIC)$. When $n = 1000$ and $\pi_1 = 0.5$, all the procedures considered here perform well, except the $MHD(AIC)$. The moderate overdispersion cases from Table 2.3 ($r = 20$) show that when $n = 100$, all the procedures perform well. However, when $n = 500$ and $\pi_1 = 0.25$, both $L_2E(SBC)$ and $MHD(SBC)$ perform better than $L_2E(LIC)$ and $MHD(AIC)$. When $n = 500$ and $\pi_1 = 0.5$, all the procedures perform well, except the $MHD(AIC)$. When $n = 1000$ and $\pi_1 = 0.5$, both $L_2E(SBC)$ and $MHD(SBC)$ perform better than $L_2E(LIC)$ and $MHD(AIC)$. However, when $n = 1000$ and $\pi_1 = 0.25$, only the $L_2E(SBC)$ procedure performs well.

The severe overdispersion cases from Table 2.2 ($\lambda_2 = 7$) show that when $n = 100$, all the procedures perform well. However, when $n = 500$, both $L_2E(SBC)$ and $MHD(SBC)$ perform better than $L_2E(LIC)$ and $MHD(AIC)$. When $n = 1000$ only the $L_2E(SBC)$ procedure performs well. The severe overdispersion cases from Table 2.3 ($r = 10$) show that when $n = 100$, all the procedures perform well. However, when $n = 500$, only the $L_2E(SBC)$ procedure performs well. When $n = 1000$ and $\pi_1 = 0.25$, all the procedures fail, but when $\pi_1 = 0.5$, only the $L_2E(SBC)$ procedure performs well.

These findings show that when there is low overdispersion, the overall performance of $L_2E(LIC)$ is comparable to $MHD(AIC)$ but better than the SBC versions of L_2E and MHD. This is not surprising because the low overdispersion case is almost similar to the"correctly specified model" case in section 2.5.1, and as we noted there, in such cases the $L_2E(LIC)$ performs better than the SBC versions. However, when there is moderate or severe overdispersion, the performance of $L_2E(SBC)$ is significantly better than the rest. Therefore, in cases where the experimenter suspects the presence of severe overdispersion

in one (or more) of the components and the sample size is large, we recommend use of $L_2E(SBC)$ to obtain a parsimonious fit. These results serve as a testament that our L_2E-based estimate of mixture complexity is highly robust under model misspecification and its performance is better than that of MHD, especially when there is moderate to severe overdispersion.

A referee pointed out that when there is large overdispersion in one of the components, usually one may need perhaps more than one Poisson component to describe it. However, in this case, our L_2E procedure is able to detect the true number of components despite model misspecification, as shown in Tables 2.2 and 2.3 of section 2.11. Note that the model misspecification is used only to illustrate the robustness of $\hat{m}_n^{L_2E}$. For a real life count data that is highly overdispersed, we recommend determining an L_2E estimate of mixture complexity using, say, mixture of negative binomials, and then fitting a negative binomial mixture to the data. In fact, for a cross-section count data on the demand for medical care with high overdispersion, Deb and Trivedi (1997) fit a two-component negative binomial mixture using maximum likelihood estimation.

2.6 DATA ANALYSIS

Here, we consider overdispersed count datasets which have been modeled using Poisson mixtures.

2.6.1 Spanish Bank Data

Here, we consider a count data that gives the number of defaulted installments in a financial institution in Spain (see Table 2.5 of section 2.11). The sample size for this data is $n = 4691$. This data was originally considered by Dionne, Artis and Guillen (1996). Karlis and Xekalaki (2001) concluded that a Poisson mixture would be plausible for modeling this data. Based on plots of Hellinger gradient function for different values of mixture complexity, Karlis and Xekalaki (2001) concluded that a semiparametric MHD estimate of the mixing

distribution supports a 6-component Poisson mixture model for the data. Woo and Sriram (2007) used their MHD estimate of mixture complexity and determined that a 3- or 4-component Poisson mixture would fit the data well. Noting that there is significant zero-inflation, they also determined a MHD estimate of mixture complexity based on zero-inflated Poisson (ZIP) mixtures. They showed that a 4-component ZIP mixture would provide a good fit of the data. Recently, Chen and Khalili (2008) analyzed this data and determined a 4-component Poisson mixture and a 5-component ZIP mixture using their $MSCAD$ method. Once again, we reanalyzed this data by determining an L_2E estimate of mixture complexity, providing a fitted Poisson mixture and a fitted ZIP mixture. The details are given below.

As for the data on the number of defaulted installments in a financial institution in Spain, our L_2E estimate of mixture complexity is $\hat{m}_n^{L_2E} = 4$ for both the $L_2E(SBC)$ and $L_2E(LIC)$ thresholds. As for ZIP mixtures, our L_2E estimate of mixture complexity is $\hat{m}_n^{L_2E} = 5$ for both the thresholds. As done in Chen and Khalili (2008), we also computed the chi-square (χ^2) goodness-of-fit for our fitted Poisson mixture (with degrees of freedom 10) and ZIP mixture (with degrees of freedom 9) using L_2E estimates and compared it with those for the $MSCAD$ and MHD methods. All these are given in Table 2.4 of section 2.11, along with the parameter estimates corresponding to each of these methods. Based on the estimates and the goodness-of-fit values, we conclude that the L_2E method is as good as the $MSCAD$ (perhaps slightly better) under both model assumptions, but superior to the MHD method. Table 2.5 of section 2.11, gives the observed and the expected frequencies based on the L_2E and $MSCAD$ methods. While the expected frequencies of the two methods are similar for $x \leq 15$, the expected frequency of L_2E for $x \geq 16$ is much closer to the observed frequency than that of the $MSCAD$.

2.6.2 Death Notice Data

The Death notice data, taken from Schilling(1947), consists of the number of death notices for women aged 80 years and over, which appeared in the London "Times" newspaper on

each day for the 3-year period from 1910 to 1912 (Titterington et al., 1985, p.89). Hasselblad (1969) indicated that this data could possibly be thought of as an example where the death rate during winter months is higher than in summer months, thereby fitting a mixture of two Poisson distributions. Shen (2004) also fitted a mixture of two Poisson for this data by estimating the component parameters using L_2E and compared her estimates with those obtained using MLE and MHD methods. We reanalyze this data by first estimating the number of Poisson mixture components and then determining a fitted Poisson mixture using L_2E. We compare our results with the $MSCAD$ method of Chen and Khalili (2006) mentioned earlier, who also determine the mixture complexity and obtain a fitted Poisson mixture. The details are given below.

For the Death notice data, our estimate $\hat{m}_n^{L_2E}$ (see 2.2.9) of mixture complexity based on the $L_2E(SBC)$ and the $L_2E(LIC)$ thresholds are 1 and 2, respectively; the $MSCAD$ method detects a two-component Poisson mixture. Table 2.6 of section 2.11 gives the L_2E estimates of parameters in the two-component Poisson mixture along with estimates obtained using the $MSCAD$ method of Chen and Khalili (2006), and the MLE and MHD estimates from Shen (2004). Table 2.6 of section 2.11 also gives the chi-square goodness-of-fit statistic value (with degrees of freedom 7) for each method. It is evident from the parameter estimates and the associated chi-square values that all the methods listed in the table provide a good fit of the data. Table 2.7 of section 2.11 gives the observed and the expected frequencies based on the L_2E, MLE, MHD and $MSCAD$ methods. The table shows that our method not only detects two components (as expected), but also provides estimates that are competitive with other methods.

Shen (2004) also illustrated the robustness of L_2E for the Death notice data (n=1096) by adding one large value ranging from 10 to 20 (the maximum number of observed death notices is 9). They noted that addition of single value to the data changes their MLE estimate dramatically when the value is far away from the original data. They showed that their MHD estimates also change considerably if the contaminating value is moderately far away from

the original data, but are not affected if the value is very far away. Compared to MLE and MHD estimates, Shen observed that the L_2E is more robust to the existence of a contaminating value. Since our focus is on the estimation of the number of components, we added 50 values located at 10 (about 4.36% contamination at 10) and numerically studied the effect of inclusion of a spurious component on $\hat{m}_n^{L_2E}$. Once again, we obtained $\hat{m}_n^{L_2E} = 2$ using the LIC criterion, and this continued to be the case even when we increased the contamination percentage to about 8%. However, when the contamination percentage was larger than 6%, L_2E estimates of proportion and component parameters were somewhat affected. This shows that our $\hat{m}_n^{L_2E}$ is not influenced by a small spurious component.

Also, as per a referee's suggestion, we examined robustness against outliers by inflating the data by adding new observations generated from a Poisson component. Recall that a 2-component Poisson based on L_2E, MHD, $MSCAD$ or MLE fits the data well with component means $\lambda_1 \approx 1.3$ and $\lambda_2 \approx 2.6$. We carried out a robustness study adding 50 new observations (to the original sample size $n = 1096$) from a Poisson component with mean $\lambda_3 = 5$, 6, 7, 8 or 9. When $5 \leq \lambda_3 \leq 8$, our $\hat{m}_n^{L_2E} = 2$ with the new threshold, but when $\lambda_3 = 9$, our $\hat{m}_n^{L_2E} = 3$. This shows that the procedure $\hat{m}_n^{L_2E}$ is not influenced by small spurious Poisson component located at $\lambda_3 = 5$, 6, 7 or 8.

2.6.3 Accident Data

This example concerns the number of accidents incurred by 414 machinists over a period of three months. This count data (see Table 2.9 below) is taken from the classical paper of Greenwood and Yule (1920) and has been analyzed by several authors including Karlis and Xekalaki (1999). Greenwood and Yule noted that the fit provided by single Poisson distribution to this data is very poor. Using a sequential testing procedure based on likelihood ratio test (LRT) that utilizes a resampling approach, Karlis and Xekalaki (1999) determined that a 3-component Poisson mixture provides a better fit to the data. Observe from Table 2.9 of section 2.11 that this data contains excessive number of zeros, indicating that a (Poisson)

mixture model that simultaneously addresses the excess zeros and overdispersion, referred here as a zero-inflated Poisson (ZIP) mixture model (see definition below), may also be appropriate for this data. Woo and Sriram (2007) used their MHD estimate of mixture complexity and determined that a 2-component Poisson mixture would fit the data well. Noting that there is significant zero-inflation, they also determined a MHD estimate of mixture complexity based on zero-inflated Poisson (ZIP) mixtures. They showed that a 3-component ZIP mixture would provide a good fit of the data. Once again, we reanalyzed this data by determining an L_2E estimate of mixture complexity, providing a fitted Poisson mixture and a fitted ZIP mixture. The details are given below.

As for the accident data, our L_2E estimate of mixture complexity is $\hat{m}_n^{L_2E} = 3$ for the $L_2E(LIC)$ and $L_2E(SBC)$ thresholds. As for ZIP mixtures, our L_2E estimate of mixture complexity is $\hat{m}_n^{L_2E} = 3$ for both the thresholds. We also computed the chi-square (χ^2) goodness-of-fit for our fitted Poisson mixture (with degrees of freedom 1) and ZIP mixture(with degrees of freedom 2) using L_2E estimates and compared it with those for the MHD methods. All these are given in Table 2.8 of section 2.11 along with the parameter estimates corresponding to each of these methods. We conclude that our 3-component ZIP mixture fit and the 3-component Poisson mixture fit based on L_2E estimates provide the best fit to the data. However, from the point of view of slight parsimony (because λ_1 is set to 0 in the 3-component ZIP mixture), we would prefer the 3-component ZIP mixture fit (based on L_2E estimates) for the data. We also computed expected frequencies based on a 3-component ZIP mixture using these estimates. Table 2.9 of section 2.11 gives the observed and the expected frequencies based on these methods. Based on the estimates and the goodness-of-fit values (with degrees of freedom 1), undoubtedly both L_2E and $L_2E(ZIP)$ methods provide the best fit.

2.7 CONCLUSION

For the count data, we have introduced an estimator of the unknown number of components in finite mixtures. This estimator is derived as a by-product of minimizing an information criterion based on L_2 distance, where the penalty is a logarithmic function of number of components. The estimator, called L_2E, is shown to be strongly consistent under certain regularity conditions. Two distinctive features of the L_2E estimator are that it is easy to compute and its performance is on par with two recently proposed estimators known as MHD and $MSCAD$. Furthermore, the performance of L_2E is on par or better than that of MHD in terms of robustness against model misspecification.

Computation of L_2E is iterative and its eventual value is determined using a threshold, which is a slowly decreasing function of m. For computations and data analysis, we have suggested two different thresholds referred to as $L_2E(LIC)$ and $L_2E(SBC)$. These thresholds are different from the ones suggested for MHD, but more appropriate for the L_2 distance under consideration. In most applications, we recommend using the $L_2E(LIC)$ threshold for all sample sizes. However, in situations where an experimenter suspects the presence of severe overdispersion in count data and the sample size is large, we recommend the use of $L_2E(SBC)$ to obtain a parsimonious fit.

With respect to computation, the L_2E procedure has many distinct advantages over MHD and $MSCAD$. For example, the L_2E objective function has a simple structure which enables us to use the built-in nlm routine in R for minimization. Furthermore, the L_2E estimates are not affected by the choice of initial values and it requires less computing time. Thus, transparency, ease of use and efficiency in achieving computational speed combined with competitive performance makes the L_2E estimator an attractive alternative to other existing methods in the literature.

A similar L_2E approach can be developed for the estimation of mixture complexity in the continuous case. We do not present the details of the continuous case here because the definition of the L_2E functional, proof of consistency of mixture complexity estimator, and

assessment of robustness are different from the ones given here. These details will be reported in a subsequent article.

ACKNOWLEDGMENTS

We would like to thank the two referees whose suggestions led to a greatly improved paper. Sriram was supported in part by a NSA grant MSPF-08G-072. This study was supported in part by resources provided by the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer.

2.8 SUPPLEMENTAL MATERIALS

Here we give a list of computer codes used in the simulation and data analysis. We used *R*2.8 for all simulations and data analysis in this article.

```
#########################################################################
# This function is to randomly generate from a Poisson Mixture
#########################################################################

rpoismix<- function(n,probs,lambda) {

  out<- rep(0,n)
for (i in 1:n) {
u<- runif(1)

    k<-length(probs)
    indg<- 1:k
   cp = cumsum(probs)

   j = min(indg[u <= cp])

   out[i] <-rpois(1,lambda[j])
}
return(out)
}

#########################################################################
# This function is to calculate
# mixture density function for the Poisson Mixture
#########################################################################

dpoismix <- function(x2,probs,lambda)
{
density <- rep(0,length(x2))
for (i in 1: length(probs))
density <- density + probs[i]* dpois(x2,lambda[i])
return(density)
}

#########################################################################
# This function is to calculate
# mixture density function for the ZIP-Poisson Mixture
```

```
#############################################################################

dzpoismix <- function(x2,probs,lambda)
{
density <- rep(0,length(x2))
for (i in 2: length(probs))
{density <- density + probs[i]* dpois(x2,lambda[i-1])}
            density=  density + probs[1]*(x2==0)
return(density)

}


#############################################################################
#
#This the main function: For an input data x
#and a given number of components k, this function computes the L2E
#function value for a Poisson mixture and outputs the
#estimates of the parameters and corresponding minimum
#value of the function.
#
#############################################################################
#############################################################################
#
# Inputs:
#          x -input data
#          K - number of components desired (K=1 default)
#          prms1- initial input for parameters
#
# Output:
# list containing estimated parameters and the minimum
# (lambda=lambda w=w value=lmin)
#############################################################################

mixpois.l2e<- function(x1,prms1,k)
{

   # L2E function to minimize
   p.crit<-function(k,prms1)
   {
    mu<-prms1[1:k]
    w<-prms1[(k+1):(2*k)]

   xmax <-max(200,5*max(x1))
```

```
    f1x<-dpoismix(0:xmax,w,mu)
    f2x<-dpoismix(x1,w,mu)

    p.crit= (sum(f1x^2) - 2*mean(f2x))

    }


#nonlinear minimization routine using -nlm
#ans<-nlm(p.crit,prms1,fscale=length(x),print.level=0)

  #pr<-ans$est
   #lamda<-pr[1:k]
   # w<-pr[(k+1):(2*k)]
   #w<- w/sum(w)


   #ans<-nlm(p.crit,pr,fscale=length(x),print.level=0)
   #pr<-ans$est
   #lmin<-ans$min

#Can use the following non-linear minimization routine as well

  lower=0; upper= max(x)+5
  ans<-nlminb(prms1,p.crit,lower=0,upper=upper)
  pr<-ans$par

  ans<-nlminb(pr,p.crit,lower=0,upper=upper)
   pr<-ans$par
   lmin<- ans$obj

   lamda<-pr[1:k]
   w<-pr[(k+1):(2*k)]
   w<- w/sum(w)
   list(l=lamda,w=w,value=lmin)
}

#############################################################################
# This function is used to
# calculate the L2E value for the Poisson mixtures
#############################################################################
#Input x-data, m=mean, w=mixing proportions
#output theL2E function value
#############################################################################
```

```
l2ecal<-function (x,m,w) {
l2e.mixval<- function(x2,prms2,k)
{

    mu<-prms2[1:k]
    w<-prms2[(k+1):(2*k)]

    xmax <-max(200,5*max(x2))

    f1x<-dpoismix(0:xmax,w,mu)
    f2x<-dpoismix(x2,w,mu)

     l2eval= (sum(f1x^2) - 2*mean(f2x))
     return(l2eval)

    }


#########################################################################
#This function can be used to input initial guesses for parameters. Initial
#guesses can be either "‘K-means"’ or "‘Random sample"’or "Hierarchical-Clustering"
#########################################################################

init<-function(type="km",x,k){


      if(type=="km"){
      #K-Means
      mm<-kmeans(x,k)
      g<-mm$cluster
      mu<-mm$center
      s1<-mm$size
      w<- s1/ss
      init=c(mu,w)
           }

    if(type=="rs"){
      #Random sample
      g <-sample(0:k,ss,T)
      memb=g
      n=length(x)
      nk <- rep(0,k)
      mu<-rep(0,k)

      for(i in 1:k) {
```

```
      ii <- seq(n)[memb==i]; nk[i] <- length(ii)
      mu[i] <- mean(x[ii,drop=F])
      w <- nk/n
    }

   init=c(mu,w)
  }

  if(type=="hc"){
     library(amap)

    #Hierarchical-Clustering
    hc<-hcluster(x, method = "euclidean", diag = FALSE, upper = FALSE,
             link = "complete", members = NULL, nbproc = 2,
             doubleprecision = TRUE)

    memb <- cutree(hc, k = k)
    g=memb
    n=length(x)
    nk <- rep(0,k)
    mu<-rep(0,k)

    for(i in 1:k) {
      ii <- seq(n)[memb==i]; nk[i] <- length(ii)
      sig[i]<- var(x[ii,drop=F])
      mu[i] <- mean(x[ii,drop=F])
      w <- nk/n
    }
    init=c(mu,w)
  }


  list(init)
}


#############################################################################
#This function is used to test the mixture complexity value.
# We need to have other functions such as mix.pois.l2e.r,
# init.r, dpoismix.r,.....
#############################################################################

start=proc.time()# to see the computation time

cc=500                  #No of MC tests
```

```
ss=1000                  #Sample size
count <- rep(0,cc)    #results(K) holder

for (ct in  1:cc) {

  k=1
  k=1
 n=ss

  #Input data-
  #simulate from Poisson mixture or negative binomial mixture or real data


 x=rpoismix(n,probs,lam)

 m=mean(x)
 w=1

#Here we can use  the initial value function
#for the initial input for the parameters

prms<-c(mu,w)

xx<-mixpois.l2e(x,prms,k)

l<-xx$l
w<-xx$w

l2e <-rep(0,10)

#l2e[k]<-  xx$value # directly use the function value
l2e[k]<-l2e.mixval(x,pr,k) # we calculate the minimum value
   cat("l2e=",l2e[k])
    cat("   ")
repeat {
k=k+1
n=ss

#Initial input using function init

 iVal = init("km",xd,k) # using the kmeans to get the initial value
 prms<-iVal$init

 xx<-mixpois.l2e(x,prms,k)
 l<-xx$l
```

```
 w<-xx$w



 #l2e[k]<-xx$value # directly use the function value
 l2e[k]<-l2e.mixval(x,pr,k)  # calculate the minimum value

   cat("l2e=",l2e[k]) # printing the l2e values

   diff<- l2e[k-1]-l2e[k] # calculating the difference in l2e

   cat("diff",diff) # printing the difference
   cat("   ")
   if ( diff <= th) break  # compare it with the threshold value here
   # used the LIC or BIC
  }

 k= k-1

 cat("k=",k)

 count[ct]=k


}
 last=proc.time() - start  # time taken to run 500 Montecarlo

last

count
table(count)/cc


#############################################################################
#The real data used in this chapter are:
######################
#The Spanish Bank Data
#############################################################################

x=c(rep(0,3002),rep(1,502),rep(2,187),rep(3,138),rep(4,233),
    rep(5,160),rep(6,107),rep(7,80),rep(8,59),rep(9,53),rep(10,41),
    rep(11,28),rep(12,34),rep(13,10),rep(14,13),rep(15,11),rep(16,4),
    rep(17,5),rep(18,8),rep(19,6),rep(20,3),rep(21,0),rep(22,1),rep(23,0),
    rep(24,1),rep(25,0),rep(26,0),rep(27,0),rep(28,1),rep(29,1),rep(30,1),
    rep(31,1),rep(32,0),rep(33,0),rep(34,1))
```

```
#########################################################################
#The Accident Data
#########################################################################
x=c(rep(0,296),rep(1,74),rep(2,26),rep(3,8),rep(4,4),
          rep(5.4),rep(6,1),rep(7,0),rep(8,1))

#########################################################################
#The Death Notice Data
#########################################################################

x=c(rep(0,162),rep(1,267),rep(2,271),rep(3,185),rep(4,111),rep(5,61),
     rep(6,27),rep(7,8),rep(8,3),rep(9,1))
#########################################################################
```

2.9 APPENDIX

PROOFS

Proposition. Let $\tilde{\Gamma} \subset \Gamma$ denote the sub-class of p.m.f.'s defined on $\mathcal{X}$ for which the following condition holds: For each m, there is a compact set $C_m \subseteq \Theta_m$ such that for every $f \in \tilde{\Gamma}$,

$$\inf_{\boldsymbol{t}_m \in \Theta_m - C_m} L_2(f_{\boldsymbol{t}_m}, f) > L_2(f_{\boldsymbol{\theta}_m^*}, f),$$

for some $\boldsymbol{\theta}_m^* \in C_m$. If, for each m, Θ_m is compact then $C_m = \Theta_m$. For each m, we will assume that $f_{\boldsymbol{\theta}_m}(x)$ is continuous in $\boldsymbol{\theta}_m$ for each $x \in \mathcal{X}$, and the class $\mathcal{F}_m$ is identifiable. Then the following hold for the functional $T_m^{L_2E}$ defined in (2.2.5) and $f \in \tilde{\Gamma}$:

(i) $T_m^{L_2E}(f)$ exists satisfying (2.2.5).

(ii) If $T_m^{L_2E}(f)$ is unique, then the functional $T_m^{L_2E}$ is continuous at f in L_2 topology.

(iii) $T_m^{L_2E}(f_{\boldsymbol{\theta}_m}) = \boldsymbol{\theta}_m$ uniquely for every $\boldsymbol{\theta}_m \in \Theta_m$.

Proof. For $f \in \tilde{\Gamma}$, let $h(\boldsymbol{t}_m) = ||f_{\boldsymbol{t}_m} - f||_2$ for $\boldsymbol{t}_m \in \Theta_m$. Suppose $\{\boldsymbol{t}_{n,m}\} \in \Theta_m$ is any sequence such that $\boldsymbol{t}_{n,m} \to \boldsymbol{t}_m$ as $n \to \infty$. Then, by the Minkowski's inequality

$$\begin{aligned} |h(\boldsymbol{t}_{n,m}) - h(\boldsymbol{t}_m)|^2 &= \left| ||f_{\boldsymbol{t}_{n,m}} - f||_2 - ||f_{\boldsymbol{t}_m} - f||_2 \right|^2 \\ &\leq ||(f_{\boldsymbol{t}_{n,m}} - f_{\boldsymbol{t}_m})||_2^2. \end{aligned}$$

Since $f_{\boldsymbol{t}_m}(x)$ and $f_{\boldsymbol{t}_{n,m}}(x)$ are p.m.f.s and $f_{\boldsymbol{t}_m}(x)$ is assumed to be continuous in $\boldsymbol{t}_m$ for each x, we have that

$$\begin{aligned} ||(f_{\boldsymbol{t}_{n,m}} - f_{\boldsymbol{t}_m})||_2^2 &\leq 2 \sum_{x=0}^{\infty} |(f_{\boldsymbol{t}_{n,m}}(x) - f_{\boldsymbol{t}_m}(x))| \\ &\to 0 \end{aligned}$$

by the Glick's theorem (Devroye and Györfi, 1985, p.10). Hence $h(\boldsymbol{t}_m)$ is continuous in $\boldsymbol{t}_m$. This, together with the assumptions made above, implies that $T_m^{L_2E}(f)$ exists.

Proof of parts (ii) and (iii) are similar to those in Theorem 1 of Beran (1977). Hence, we omit the details. □

Proof of the Theorem. Recall from (2.2.9) that

$$\hat{m}_n^{L_2E} = \min\{m : L_2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}, \hat{f}_n) \leq L_2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{L_2E}}, \hat{f}_n) + \alpha_{n,m}\}.$$

Now, since $\hat{f}_n(x) \to f_0(x)$ a.s. for each x, another application of the Glick's theorem (Devroye and Györfi, 1985, p.10) yields

$$\begin{aligned} ||\hat{f}_n - f_0||_2^2 &\leq 2\sum_{x=0}^{\infty} |\hat{f}_n(x) - f_0(x)| \\ &\to 0 \qquad a.s. \end{aligned} \tag{2.9.13}$$

Henceforth, we will suppress "a.s.", as it will be clear from the context. Define $h_n(\boldsymbol{t}) = ||f_{\boldsymbol{t}} - \hat{f}_n||_2$ and $h(\boldsymbol{t}) = ||f_{\boldsymbol{t}} - f_0||_2$. Once again, applying the Minkowski's inequality we get $|h_n(\boldsymbol{t}) - h(\boldsymbol{t})| \leq ||\hat{f}_n - f_0||_2$. Hence,

$$\sup_{\boldsymbol{t}} |h_n(\boldsymbol{t}) - h(\boldsymbol{t})| \leq ||\hat{f}_n - f_0||_2 \quad \to \quad 0,$$

by (2.9.13). Let $\boldsymbol{\theta}_{0,m} = T_m^{L_2E}(f_0)$ and $\hat{\boldsymbol{\theta}}_{n,m}^{L_2E} = T_m^{L_2E}(\hat{f}_n)$. Then, it is possible to show that

$$|\min_{\boldsymbol{t}\in\Theta_m} h_n(\boldsymbol{t}) - \min_{\boldsymbol{t}\in\Theta_m} h(\boldsymbol{t})| \to 0. \tag{2.9.14}$$

That is, $||\hat{f}_n - f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}||_2 \to ||f_0 - f_{\boldsymbol{\theta}_{0,m}}||_2$. Therefore, from the definitions above

$$L_2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}}, \hat{f}_n) - L_2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{L_2E}}, \hat{f}_n) \to L_2(f_{\boldsymbol{\theta}_{0,m}}, f_0) - L_2(f_{\boldsymbol{\theta}_{0,m+1}}, f_0) = d_m. \tag{2.9.15}$$

Note from (2.2.2) and (2.9.15) that

$$m_0 \quad = \quad min\{m : L_2(f_{\boldsymbol{\theta}_{0,m}}, f_0) - L_2(f_{\boldsymbol{\theta}_{0,m+1}}, f_0) = d_m \leq 0\}.$$

If f_0 is not a finite mixture, then $m_0 = \infty$. This implies that $d_m > 0$ for all $m > 0$. Therefore, by (2.2.9) and (2.9.15) it follows that $\hat{m}_n \to \infty$ almost surely. If f_0 is a finite mixture, that is $f_0 = f_{\boldsymbol{\theta}_{m_0}}$, then we will show that $d_m > 0$ for $m < m_0$ and $d_m = 0$ for $m \geq m_0$.

Let $m \geq m_0$, since $f_0 \in \mathcal{F}_{m_0} \subseteq \mathcal{F}_j, j \geq m_0$

$$L_2(f_{\hat{\boldsymbol{\theta}}_{n,j}^{L_2E}}, \hat{f}_n) \leq L_2(f_{\boldsymbol{\theta}_{m_0}}, \hat{f}_n) \to 0, \tag{2.9.16}$$

by (2.9.13). Therefore, by (2.9.15) we have that $d_m = 0$ for $m \geq m_0$.

Let $m < m_0$ then by the definition of m_0, we have $f_0 \in \mathcal{F}_{m_0}$ but $f_0 \notin \mathcal{F}_m$ for $m < m_0$. Suppose, on the contrary, $d_m = 0$ for some $m < m_0$, that is, $L_2(f_{\boldsymbol{\theta}_{0,m}}, f_0) = L_2(f_{\boldsymbol{\theta}_{0,m+1}}, f_0)$. Then, for all $\mathbf{t}_{m+1} \in \Theta_{m+1}$

$$L_2(f_{\boldsymbol{\theta}_{0,m}}, f_0) \leq L_2(f_{\boldsymbol{t}_{m+1}}, f_0). \tag{2.9.17}$$

For an arbitrary $\epsilon \in (0,1)$ and $\boldsymbol{\phi} \in \Phi$, let $f_{\mathbf{t}_{m+1}}(x) = (1-\epsilon)f_{\boldsymbol{\theta}_{0,m}}(x) + \epsilon f(x|\boldsymbol{\phi})$. Then, $f_{\boldsymbol{t}_{m+1}} \in \mathcal{F}_{m+1}$ and from (2.9.17)

$$\sum_{x=0}^{\infty} |f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)|^2 - \sum_{x=0}^{\infty} |(1-\epsilon)f_{\boldsymbol{\theta}_{0,m}}(x) + \epsilon f(x|\boldsymbol{\phi}) - f_0(x)|^2 \leq 0.$$

Now, using the identity $x^2 - y^2 = (x-y)(x+y)$ and algebraic calculations we get

$$\epsilon \sum_{x=0}^{\infty} [f_{\boldsymbol{\theta}_{0,m}}(x) - f(x|\boldsymbol{\phi})] \left\{ 2[f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)] + \epsilon[f(x|\boldsymbol{\phi}) - f_{\boldsymbol{\theta}_{0,m}}(x)] \right\} \leq 0,$$

which implies

$$2\epsilon \sum_{x=0}^{\infty} [f_{\boldsymbol{\theta}_{0,m}}(x) - f(x|\boldsymbol{\phi})][f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)] \;\; \leq \;\; \epsilon^2 \sum_{x=0}^{\infty} [f_{\boldsymbol{\theta}_{0,m}}(x) - f(x|\boldsymbol{\phi})]^2. \tag{2.9.18}$$

Dividing both sides of (2.9.18) by ϵ and letting $\epsilon \to 0$ we get

$$\sum_{x=0}^{\infty} [f_{\boldsymbol{\theta}_{0,m}}(x) - f(x|\boldsymbol{\phi})][f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)] \leq 0,$$

which implies

$$\sum_{x=0}^{\infty} f_{\boldsymbol{\theta}_{0,m}}(x)[f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)] \leq \sum_{x=0}^{\infty} f(x|\boldsymbol{\phi})[f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)]. \tag{2.9.19}$$

Since $f_0 \in \mathcal{F}_{m_0}$, we can write $f_0(x) = \sum_{i=1}^{m_0} \pi_i^0 f(x|\boldsymbol{\phi}_i^0)$ and (2.9.19) holds for each $\boldsymbol{\phi} = \boldsymbol{\phi}_i^0$, $i = 1, \ldots, m_0$. Since $\sum_{i=1}^{m_0} \pi_i^0 = 1$, from (2.9.19)

$$\sum_{x=0}^{\infty} f_{\boldsymbol{\theta}_{0,m}}(x)[f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)] \leq \sum_{x=0}^{\infty} f_0(x)[f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)],$$

which implies that $\sum_{x=0}^{\infty} [f_{\boldsymbol{\theta}_{0,m}}(x) - f_0(x)]^2 = 0$. This contradicts the fact that $f_0 \notin \mathcal{F}_m$ for $m < m_0$. Therefore, $d_m = 0$ for $m < m_0$. Hence the Theorem. □

2.10 References

[1] Basu, A., Harris, I.R., Hjort, H.L., and Jones, M.C. (1998), "Robust and efficient estimation by minimizing a density power divergence," *Biometrika*, 85, 549 - 560.

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

[3] Chen, J.and Khalili, A. (2006), "Order selection in finite mixture models," *Technical Report*, Department of Statistics and Actuarial Science University of Waterloo, Canada.

[4] Chen, J.and Khalili, A. (2008), "Order selection in finite mixture models with a nonsmooth penalty," *Journal of the American Statistical Association*, 103, 1674-1683.

[5] Cutler, A., and Cordero-Braňa, O. I. (1996), "Minimum Hellinger distance estimation for finite mixture models," *Journal of the American Statistical Association*, 91, 1716-1723.

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

[7] Dellaportas, P.,Karlis, D and Xekalaki, E. (1997), "Bayesian analysis of finite Poisson mixtures," *Technical Report*, Department of Statistics, Athens University of Economics and Business.

[8] Devroye, L. P., and Györfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*, New York: Wiley.

[9] Dionne, G. Artis, M.,and Guillen, M., (1996), "Count data models for a credit scoring system," *Journal of Empirical Finance*, 3, 303-325.

[10] Fan,J. and Li,R. (2001),"Variable selection via non-concave penalized likelihood and its oracle properties, " *Journal of the American Statistical Association*, 96, 1348-1360.

[11] Greenwood, M., and Yule, G., (1920), "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents," *Journal of Royal Statistical Society, Ser. A* , 83, 255-279.

[12] Hasselblad,V. (1969),"Estimation of finite mixtures of distributions from the exponential family, " *Journal of the American Statistical Association*, 64, 1459-1471.

[13] Ishwaran, H., James, L. F., and Sun, J. (2001),"Bayesian model selection in finite mixtures by marginal density decompositions," *Journal of the American Statistical Association*, 96, 1316-1332.

[14] James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent estimation of mixture complexity," *The Annals of Statistics*, 29, 1281-1236.

[15] Karlis, D. and Xekalaki, E. (1998), "Minimum Hellinger distance estimation for finite Poisson mixtures," *Computational Statistics and Data Analysis*, 29, 81-103.

[16] Karlis, D. and Xekalaki, E. (1999), "On testing for the number of components in a mixed Poisson model," *Annals of Institute of Statistical Mathematics*, 51, 149-162.

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

[18] Pauler, D. K., Escobar, M. D., Sweeney, J. A. and Greenhouse, J. (1996), "Mixture models for eye-tracking data: A case study," *Statistics in Medicine*, 15, 1365-1376.

[19] Poland, W. B., and Shachter, R. D. (1994), "Three approaches to probability model selection", *In Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference*, San Mateo, CA: Morgan Kaufmann, 478-483.

[20] Roeder, K. (1994), "A graphical technique for determining the number of components in a mixture of normals," *Journal of the American Statistical Association*, 89, 487-495.

[21] Schilling, W. (1947), " A frequency distribution represented as the sum of two Poisson distributions, " *Journal of the American Statistical Association*, 42, 407-424.

[22] Schlattmann, P.and Böhning, D. (1993), "Mixture models and disease mapping," *Statistics in Medicine*, 12, 943-950.

[23] Scott, D. W. (1998),"On fitting and adapting of density estimates," *Computing Science and Statistics*, S. Weisberg, Ed., 30, 124 - 133.

[24] Scott, D.W. (1999),"Remarks on fitting and interpreting mixture models," *Computing Science and Statistics*, K. Berk and M. Pourahmadi, Eds., 31, 104-109.

[25] Scott, D. W. (2001),"Parametric statistical modeling by minimum integrated square error, " *Technometrics*, 43, 274-285.

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

[27] Shen, S. (2004), "The minimum L_2E distance estimator in Poisson mixtures," *Ph.D. Thesis*, Dedmon College, Southern Methodist University.

[28] Titterington, D.M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

[29] Woo, Mi-Ja and Sriram, T. N. (2006), "Robust estimation of mixture complexity," *Journal of American Statistical Association*, 101, 1475-1486.

[30] Woo, Mi-Ja and Sriram, T. N. (2007),"Robust estimation of mixture complexity for count data," *Computational Statistics and Data Analysis*, 51, 4379-4392.

2.11 Tables and Figures

Table 2.1: Relative frequencies of estimated number of components based on 500 replications.

	2-components					
	$\boldsymbol{\theta}_2 = (0.5, 1, 9)$					
Sample Size	$n = 100$			$n = 500$		
α_n	1	2	3	1	2	3
$L_2E(LIC)$	0.004	***.958**		0	***.984**	.016
$MHD(AIC)$	0	***.998**	.002	0	***1.00**	
$MSCAD$	0	***.988**	.012	0	***1.00**	0
LRT	0	***.95**	.05	0	***.96**	.04
	$\boldsymbol{\theta}_2 = (0.8, 1, 9)$					
Sample Size	$n = 100$			$n = 500$		
α_n	1	2	3	1	2	3
$L_2E(LIC)$	.004	***.928**	.068	0	***.944**	.056
$MHD(AIC)$	0	***.998**	.002	0	***1.00**	
$MSCAD$	.002	***.986**	.012		***.990**	.008
LRT	0	***.95**	.05	0	***.96**	.04
	$\boldsymbol{\theta}_2 = (0.95, 1, 10)$					
Sample Size	$n = 100$			$n = 500$		
α_n	1	2	3	1	2	3
$L_2E(LIC)$	.568	.402	.03	.096	***.832**	.072
$MHD(AIC)$	.616	.384		0	***1.00**	
$MSCAD$	.052	***.868**	.080		***.994**	.004
LRT	0	***.93**	.07	0	***.95**	.05

Table 2.1 (continued)

3-components						
$\boldsymbol{\theta}_3 = (0.33, 0.33, 1, 5, 10)$						
Sample Size	$n = 100$			$n = 500$		
α_n	1	2	3	1	2	3
$L_2E(LIC)$	0	.48	***.52**	0.1	.038	***.952**
$MHD(AIC)$	0	.84	.16	0	.018	***.982**
$MSCAD$	0	.2	***.780**	0	.016	***.964**
LRT	0	.30	***.66**	0	0	***.94**
$\boldsymbol{\theta}_3 = (0.45, 0.45, 1, 5, 10)$						
Sample Size	$n = 100$			$n = 500$		
α_n	1	2	3	1	2	3
$L_2E(LIC)$	.004	.83	.166	0	.374	***.626**
$MHD(AIC)$	0	.966	.034	0	.162	***.838**
$MSCAD$	0	.280	***.692**	0	.082	***.896**
LRT	0	.39	***.58**	0	0	***.94**
4-components						
$\boldsymbol{\theta}_4 = (0.25, 0.25, 0.25, 1, 5, 10, 15)$						
Sample Size	$n = 100$			$n = 500$		
α_n	≤ 3	4	5	≤ 3	4	5
$L_2E(LIC)$	.956	.044		.46	***.54**	
$MHD(AIC)$	.994	.004		.924	.076	
$MSCAD$	.512	.460	.028	.110	***.812**	.078
LRT	.88	.12		.59	.4	.01

Table 2.2: Samples drawn from 2-component Negative Binomial mixture in (2.5.12) with $\boldsymbol{\theta}_2 = (\pi_1, \lambda_1, \lambda_2)$ and $r = 10$.

$\boldsymbol{\theta}_2 = (.25, 1, 2)$									
Sample Size	$n = 100$			$n = 500$			$n = 1000$		
α_n	1	2	3	1	2	3	1	2	3
$L_2E(LIC)$	.554	.446		.07	***.924**	.006	.02	***.954**	.044
$L_2E(SBC)$	.968	.028		.812	.188		.482	***.518**	
$MHD(AIC)$	.908	.092		.218	***.782**		.008	***.992**	
$MHD(SBC)$	.998	.002		.888	.112		.474	***.526**	
$\boldsymbol{\theta}_2 = (.25, 1, 5)$									
Sample Size	$n = 100$			$n = 500$			$n = 1000$		
α_n	1	2	3	1	2	3	1	2	3
$L_2E(LIC)$	.002	***.84**	.158	0	***.584**	.416	.002	.187	.811
$L_2E(SBC)$	.218	***.782**		0	***.998**	.002	0	***.976**	.024
$MHD(AIC)$	.006	***.978**	.016	0	***.724**	.276	0	.252	.748
$MHD(SBC)$	.078	***.922**		0	***.996**	.004	0	***.948**	.052
$\boldsymbol{\theta}_2 = (.25, 1, 7)$									
Sample Size	$n = 100$			$n = 500$			$n = 1000$		
α_n	1	2	3	1	2	≥ 3	1	2	≥ 3
$L_2E(LIC)$	0	***.69**	.31	0	0.11	.89	0	0.03	.97
$L_2E(SBC)$	.066	***.932**	.002	0	***.928**		0	***.646**	.354
$MHD(AIC)$	0	***.916**	.084	0	.068	.932	0	0	1.00
$MHD(SBC)$	.002	***.992**	.006	0	***.718**	.284	0	.112	.888
$\boldsymbol{\theta}_2 = (.5, 1, 2)$									
Sample Size	$n = 100$			$n = 500$			$n = 1000$		
α_n	1	2	3	1	2	3	1	2	3
$L_2E(LIC)$	.528	.472		.068	***.924**	.008	0	***.964**	.036
$L_2E(SBC)$	.95	.05		.766	.234		.356	***.644**	
$MHD(AIC)$	.882	.118		.14	***.86**		.002	***.998**	
$MHD(SBC)$	.994	.006		.802	.198		.276	***.724**	
$\boldsymbol{\theta}_2 = (.5, 1, 5)$									
Sample Size	$n = 100$			$n = 500$			$n = 1000$		
α_n	1	2	3	1	2	3	1	2	3
$L_2E(LIC)$	.04	***.96**		0	***.87**	.13	0.004	***.736**	.26
$L_2E(SBC)$	.044	***.956**		0	***.984**	.016	0	***.976**	.024
$MHD(AIC)$	0	***.992**	.008	0	***.808**	.192	0	.40	.60
$MHD(SBC)$	.006	***.992**	.002	0	***1.00**		0	***.984**	.016
$\boldsymbol{\theta}_2 = (.5, 1, 7)$									
Sample Size	$n = 100$			$n = 500$			$n = 1000$		
α_n	1	2	3	1	2	3	1	2	3
$L_2E(LIC)$	0.028	***.794**	.178	0	.274	.726	0	.156	.844
$L_2E(SBC)$	0	***.992**	.008	0	***.962**	.038	0	***.87**	.13
$MHD(AIC)$	0	***.946**	.054	0	.234	.766	0	.006	.994
$MHD(SBC)$	0	***.998**	.002	0	***.914**	.086	0	.474	.526

Table 2.3: Samples drawn from 2-component Negative Binomial mixture in (2.5.12) with $\boldsymbol{\theta}_2 = (\pi_1, \lambda_1, \lambda_2)$.

$\boldsymbol{\theta}_2 = (.25, 1, 10);\quad n = 100$									
r		$r = 10$			$r = 20$			$r = 40$	
α_n	1	2	3	1	2	3	1	2	3
$L_2E(LIC)$	0	***.56**	.44	0.008	***.728**	.264	.004	***.894**	.102
$L_2E(SBC)$	.084	***.9**	.016	.142	***.852**	.006	.186	***.814**	
$MHD(AIC)$	.002	***.594**	.404	0	***.92**	.08	0	***.99**	.01
$MHD(SBC)$	.002	***.942**	.056	0	***.996**	.004	0	***1.00**	
$\boldsymbol{\theta}_2 = (.25, 1, 10);\quad n = 500$									
r		$r = 10$			$r = 20$			$r = 40$	
α_n	1	2	$\geq$3	1	2	$\geq$3	1	2	$\geq$3
$L_2E(LIC)$	0	0	1.00	0	.13	.87	0	***.6**	.4
$L_2E(SBC)$	0	***.54**	.46	0	**.984**	.016	0	***1.00**	
$MHD(AIC)$	0	0	1.00	0	.124	.876	0	***.766**	.234
$MHD(SBC)$	0	.022	.978	0	***.782**	.218	0	***.998**	.002
$\boldsymbol{\theta}_2 = (.25, 1, 10);\quad n = 1000$									
r		$r = 10$			$r = 20$			$r = 40$	
α_n	1	2	$\geq$3	1	2	$\geq$3	1	2	$\geq$3
$L_2E(LIC)$	0	.0	1.00	0	.02	.98	.01	.286	.714
$L_2E(SBC)$	0	.12	.88	0	***.84**	.16	0	***.998**	.002
$MHD(AIC)$	0	0	1.00	0	.004	.996	0	.282	.718
$MHD(SBC)$	0	0	1.00	0	.192	.808	0	***.966**	.034
$\boldsymbol{\theta}_2 = (.5, 1, 10);\quad n = 100$									
r		$r = 10$			$r = 20$			$r = 40$	
α_n	1	2	$\geq$3	1	2	$\geq$3	1	2	$\geq$3
$L_2E(LIC)$	0.022	***.678**	.3	0	***.82**	.18	0	***.89**	.11
$L_2E(SBC)$	0	***.996**	.004	0	***.998**	.002	0	***1.00**	
$MHD(AIC)$	0	***.76**	.240	0	***.94**	.06	0	***.99**	.01
$MHD(SBC)$	0	***.984**	.016	0	***.998**	.002	0	***1.00**	
$\boldsymbol{\theta}_2 = (.5, 1, 10);\quad n = 500$									
r		$r = 10$			$r = 20$			$r = 40$	
α_n	1	2	≥ 3	1	2	≥ 3	1	2	≥ 3
$L_2E(LIC)$	0	.2	.8	0	**.68**	.32	0	***.74**	.26
$L_2E(SBC)$	0	***.87**	.13	0	***.994**	.006	0	***1.00**	0
$MHD(AIC)$	0	.006	.994	0	.386	.614	0	***.896**	.104
$MHD(SBC)$	0	.252	.748	0	***.942**	.058	0	***1.00**	
$\boldsymbol{\theta}_2 = (.5, 1, 10);\quad n = 1000$									
r		$r = 10$			$r = 20$			$r = 40$	
α_n	1	2	≥ 3	1	2	≥ 3	1	2	≥ 3
$L_2E(LIC)$	0	0	1.00	0	0.058	.942	0	***.638**	.332
$L_2E(SBC)$	0	***.568**	.432	0	***.982**	.018	0	***.996**	.004
$MHD(AIC)$	0	0	1.00	0	.04	.96	0	***.608**	.392
$MHD(SBC)$	0	0	1.00	0	***.622**	.378	0	***.994**	.006

Table 2.4: Parameter estimates for Poisson mixture models: Spanish bank data.

$Method(\hat{m})$	$(\hat{\pi_1}, \hat{\lambda_1})$	$(\hat{\pi_2}, \hat{\lambda_2})$	$(\hat{\pi_3}, \hat{\lambda_3})$	$(\hat{\pi_4}, \hat{\lambda_4})$	$(\hat{\pi_5}, \hat{\lambda_5})$	χ^2
$L_2E(4)$	(.736,.147)	(.204,4.05)	(.055,10.05)	(.005,24.09)	-	34.61
$MSCAD(4)$	(.733,.147)	(.200,3.98)	(.060,9.52)	(.007,19.72)	-	34.81
$MHD(4)$	(.742,.15)	(.204,4.15)	(.053,10.43)	(.001,23.18)	-	43.57
$L_2E-ZIP(5)$	(.342,0)	(.401,3.16)	(.198,4.23)	(.054,10.08)	(.005,20.38)	33.99
$MSCAD-ZIP(5)$	(.328,0)	(.417,.302)	(.193,4.19)	(.055,9.78)	(.007,20.01)	34.68
$MHD-ZIP(4)$	(.373,0)	(.385,.36)	(.199,4.52)	(.043,11.26)	-	45.44

Table 2.5: Comparison of observed frequencies and expected frequencies: Spanish bank data.

x	Observed	Expected Frequencies			
		$m=4$ L_2E	$m=4$ $MSCAD$	$m=5$ L_2E-ZIP	$m=5$ $MSCAD$-ZIP
0	3002	2996.2	2986.0	2990.4	2998.6
1	502	506.3	506.3	491.3	494.4
2	187	169.4	171.9	190.1	187.0
3	138	187.9	188.7	185.4	177.0
4	233	191.6	190.4	179.6	182.2
5	160	160.9	159.4	161.7	158.5
6	107	118.3	118.1	122.9	120.8
7	80	82.3	84.1	87.1	86.5
8	59	59.1	62.0	62.3	62.6
9	53	46.1	48.8	47.4	48.1
10	41	38.2	39.8	38.4	38.7
11	28	32.0	32.3	31.7	31.4
12	34	25.8	25.1	25.5	24.7
13	10	19.7	18.7	19.6	18.7
14	13	14.2	13.3	14.4	13.6
15	11	9.7	9.4	10.2	9.7
$\geq$16	33	32.9	36.7	32.9	38.3

Table 2.6: Parameter estimates for Poisson mixture models: Death notice data.

$Method$	π_1	π_2	λ_1	λ_2	χ^2
L_2E	.4213	.5787	1.36119	2.7418	1.2204
$MSCAD$	.34	.66	1.23	2.64	1.29
MLE	.3599	.6401	1.2561	2.6634	1.180
MHD	.3375	.6625	1.2196	2.6302	1.234

Table 2.7: Comparison of observed frequencies and expected frequencies: Death notice data.

X	0	1	2	3	4	5	6	7	8	9
$Frequency$	162	267	271	185	111	61	27	8	3	1
$MHD(m=2)$	161.575	270.867	262.243	191.714	114.413	57.345	24.560	9.128	2.986	0.870
$MLE(m=2)$	161.230	271.346	262.073	191.199	114.191	57.548	24.859	9.335	3.089	0.911
$L_2E(m=2)$	159.247	273.207	263.319	190.193	113.195	57.397	25.168	9.652	3.273	0.992

Table 2.8: Parameter estimates for Poisson mixture models: Accident data.

$Method$	π_1	π_2	π_3	λ_1	λ_2	λ_2	χ^2
$L_2E(3)$	.4377	.5203	.0420	.000001	.6355	3.8415	0.04074985
$L_2E(ZIP(3))$	.4376	.5204	.0420	0	.6355	3.8415	0.04068744
$MHD(2)$	.8796	.1204		.22749	2.1859		1.040910
$MHD(ZIP(3))$	.42335	.52580	.05084	0	.5896	3.0449	1.049298

Table 2.9: Comparison of observed frequencies and expected frequencies: Accident data.

X	0	1	2	3	4	≥ 5
$Frequency$	296	74	26	8	4	6
$MHD(m=2)$	295.66	78.23	20.89	10.32	5.36	3.54
$MHD(ZIP(m=3))$	297.02	74.20	25.62	8.84	4.19	4.13
$L_2E(m=3)$	295.6599	73.9419	25.7963	8.409	4.1635	5.7334
$L_2E(ZIP(m=3))$	295.6595	73.9423	25.7962	8.4087	4.1634	5.7338

Chapter 3

L_2E ESTIMATION OF MIXTURE COMPLEXITY: CONTINUOUS CASE[2]

[2]Umashanger, T. and Sriram, T. N. To be submitted to: *Computational Statistics and Data Analysis.*

ABSTRACT

In many applications, there may not be sufficient information about the number of mixture components, termed *mixture complexity*, to determine a satisfactory finite mixture model fit for a dataset. This article focuses on continuous data and develops an estimator of mixture complexity which is consistent when the form of the component densities are unknown but are postulated to belong to a parametric family, and which is simultaneously robust against model misspecification. We construct an estimator of mixture complexity as a by-product of minimizing an information criterion based on L_2 distance. When the model is correctly specified, Monte Carlo simulations for a wide variety of normal mixtures show that our estimator correctly identifies the true mixture complexity. Robustness of the estimator is examined via simulations under symmetric departures from postulated component normality. The performance of our estimator is assessed through simulations and comparisons are made with other procedures in the literature. It is shown that our estimator performs better than all other procedures including the minimum Hellinger distance estimator of Woo and Sriram (2006). Three well-known real datasets are examined to illustrate the performance of this method.

Key words and Phrases: Finite mixtures; Mixture complexity; Information criterion; Threshold; Consistency; Robustness.

3.1 INTRODUCTION

Ever since the work of Pearson (1894), finite mixture models have been widely used in many disciplines such as astronomy, biology, engineering, genetics, medicine, and social sciences, among others. Finite mixture models are applicable in situations where datasets consist of two or more subpopulations. Due to this flexibility in modeling, researchers continue to study finite mixture models theoretically and identify new applications areas such as Bioinformatics. A comprehensive account of statistical inference for mixture models with

applications can be found in Everitt and Hand (1981), Titterington, Smith and Makov (1985), and McLachlan and Basford (1988), Lindsay (1995), Böhning (1999), McLachlan and Peel (2000), and Böhning and Seidel (2003).

When the number of components, termed mixture complexity, is assumed to be known and there is no contamination in the data, there is an enormous body of literature giving a variety of approaches to estimate the unknown component parameters in finite mixture models. EM algorithm of Dempster, Laird and Rubin (1977) is undoubtedly a useful way to compute maximum likelihood estimates (MLE) of all the component parameters. However, the MLE becomes highly unstable (Aitkin and Wilson 1980) when there is a small perturbation in one of the component densities, thereby affecting the quality and interpretability of fitted mixture models. To address the issue of robust estimation of component parameters, a variety of methods such as M-estimation (McLachlan and Basford 1988), robust version of the EM algorithm (De Veaux and Krieger 1990; Windham and Cutler 1994) and several minimum distance estimation methods (Woodward et al. (1984); McCann and Sarkar (2000)) possessing some degree of automatic robustness (Donoho and Liu 1988) have been studied in the literature as alternative approaches. For each estimation approach, the literature offers associated theory along with computational methodologies.

Robust methods such as M-estimation are not easily adapted for mixtures and these generally achieve robustness at the cost of efficiency at the true parametric model density. One possible way to partially reconcile the conflicting concepts of robustness and efficiency is to use a density-based minimum Hellinger distance (MHD) estimator introduced by Beran (1977). In the context of mixtures, Cutler and Cordero-Braňa (1996) developed a MHD estimator for all parameters when the exact form of the component densities are unknown but are thought to be close to members of some parametric family. They proposed a new computational algorithm, somewhat similar to the EM algorithm, and an adaptive density estimate to compute the MHD estimates. In addition to studying basic asymptotic properties, they showed via simulations that their MHD estimates are also robust under gross-error

contaminations. Woodward, Whitney and Eslinger (1995) also studied MHD estimators in the case of estimation of mixing proportion in a mixture of two normals.

In many situations, the mixture complexity is not known in advance. In such scenarios, fitting a parsimonious finite mixture model becomes considerably more challenging. Examples of scenarios where mixture complexity is not known are plentiful; see, for instance, Bogardus et al. (1989), McLaren et al.(1991), Roeder (1994), McLachlan, McLaren and Matthews (1995), McLaren (1996) Richardson and Green (1997), McLachlan and Peel (1997, 2000). Developing methods of estimation for mixture complexity has been an area of intense research over the last two decades; see Henna (1985); McLachlan (1987); Roeder (1994); Escobar and West (1995); Chen and Kalbfleisch (1996); Dacunha-Castelle and Gassiat (1997, 1999); Roeder and Wasserman (1997); Keribin (2000); Priebe and Marchette (2000); and Ishwaran, James and Sun (2001); Roeder (1994), Dellaportas et al. (1997), Basu et al. (1998), Karlis and Xekalaki (1999), James et al. (2001), Ishwaran et al. (2001), McGrory and Titterington(2007); Chen and Khalili (2008) and references therein. Once again, the estimation methodologies proposed in the above articles are sensitive to model misspecification and presence of outliers in datasets.

Recently, Woo and Sriram (2006) treated the estimation of mixture complexity as a model selection problem and constructed an estimator of mixture complexity as a by-product of minimizing a Hellinger Information Criterion (HIC). They showed that their estimator of mixture complexity($MHDE$) is consistent and also illustrated through simulations the ability of their estimator to correctly determine the number of components when the postulated mixture model is correct. Furthermore, they showed that their estimator continues to perform well even when the data comes from a model that is somewhat different from the postulated mixture model; see Woo and Sriram (2006) for more details.

Undoubtedly, the $MHDE$ estimator of mixture complexity considered in Woo and Sriram (2006) has attractive large sample and robustness features. However, the implementation of the $MHDE$ algorithm requires specification of an adaptive nonparametric density estimator

and careful choice of bandwidth; see Woo and Sriram, 2006 for details. Clearly, these specifications severely impact the computations of $MHDE$ estimates, especially if the true mixture complexity is more than two.

To overcome computational difficulties associated with $MHDE$, Scott (1998, 1999, 2001 and 2004) introduced an alternative minimum distance estimation method based on integrated squared error criterion, termed L_2E, which avoids the use of nonparametric kernel density estimators. The L_2E approach is a special case of a general method introduced by Basu et al. (1998), who devised a whole continuum of density-based power divergence estimators that begin with the MLE and interpolate to the L_2E estimator and beyond. While the L_2E approach has the advantage of not requiring any nonparametric density estimator, L_2E estimators suffer from moderate loss of efficiency at the parametric model relative to $MHDE$ and maximum likelihood estimators. Nonetheless, within the family of density-based power divergence measures, the L_2E approach has the distinct advantage that a key integral can be computed in a closed form, especially for finite mixtures; see equation (3.2.7) below and Scott (2001). These findings motivate us to investigate the L_2E approach for the estimation of mixture complexity, when all the component parameters are unknown.

This paper describes a new algorithm for estimating mixture complexity based on L_2E distance. As a member of the family of minimum distance estimators, the L_2E criterion is by nature robust and hence less influenced by outliers. Our primary aim is to develop an estimator of mixture complexity based on L_2E distance which is not only consistent and robust, but also computationally simpler than $MHDE$. By treating the estimation of mixture complexity as a model selection problem, we construct an estimator of mixture complexity as a by-product of minimizing a Information Criterion (LIC) based on L_2E distance introduced in section 3.2; see display (3.2.8) and details below it.

In section 3.2, we introduce the L_2E criterion due to Scott and propose an estimator of mixture complexity using this criterion. The main theorem concerning the consistency of the estimator is stated in section 3.3 but proved in the Appendix. Computational details

concerning our estimator are given in section 3.4. In the first two subsections of section 3.5, we carry out extensive Monte Carlo studies for a variety of mixtures with normal components, in order to support the consistency result and compare the performance of our estimator with those available in the literature. In section 3.5.3, we examine the robustness of our estimator through extensive simulations. Here, we examine the robustness of our estimator against model misspecification and compare them with the estimator of James et al. (2001) and $MHDE$ of Woo and Sriram(2006). In section 3.6, we estimate the mixture complexity for three well known real data sets and compare our performance with those in the literature. Overall summary and conclusions are given in section 3.7. We begin with some basic notations and definitions.

3.2 L_2E ESTIMATOR

The L_2 distance estimators, termed L_2E, were introduced by Scott(1998, 1999), who convincingly argued that the estimation method is particularly appropriate for analyzing large data sets in which an estimator is expected to be robust to the existence of gross errors and still retain acceptable level of efficiency. In this section, we introduce some basic notations, the L_2 estimation approach and then propose an estimator of the mixture complexity.

Consider a parametric family of distribution functions $\mathcal{F}_m = \{F_{\boldsymbol{\theta}_m} : \boldsymbol{\theta}_m \in \Theta_m \subseteq R^p\}$ for each fixed $m < \infty$ such that $F_{\boldsymbol{\theta}_m}$ can be represented as a finite mixture of the form

$$F_{\boldsymbol{\theta}_m}(x) = \sum_{i=1}^{m} \pi_i F(x|\boldsymbol{\phi}_i), \quad x \in \mathcal{X} \subseteq \mathcal{R}, \tag{3.2.1}$$

and $\boldsymbol{\theta}_m = (\pi_1, \ldots, \pi_{m-1}, \boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_m^T)^T$. 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 $f_{\boldsymbol{\theta}_m}(x)$ denote the mixture density function corresponding to $F_{\boldsymbol{\theta}_m}(x)$ with component densities denoted by $f(x|\boldsymbol{\phi}_i)$, for $i = 1 \cdots m$. That is, $f_{\boldsymbol{\theta}_m}(x) = \sum_{i=1}^{m} \pi_i f(x|\boldsymbol{\phi}_i)$.

Let $X_1, \ldots, X_n$ be independent and identically distributed 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

$$m(G) = \min\{m : G \in \mathcal{F}_m\}. \tag{3.2.2}$$

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 find a semi-parametric estimator of the form

$$\hat{f}_n^*(x) = \sum_{i=1}^{\hat{m}_n} \hat{\pi}_i f(x|\hat{\boldsymbol{\phi}}_i), \tag{3.2.3}$$

with the property that $\hat{m}_n \to m_0$ almost surely (a.s.) as $n \to \infty$. Consequently, 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 $\hat{f}_n^* \to f_0$.

To this end, define the squared L_2 distance between two density functions g, f as

$$\begin{aligned} L_2(g,f) &= \int_{-\infty}^{\infty} (g(x) - f(x))^2 dx \\ &= \int_{-\infty}^{\infty} g^2(x)dx - 2\int_{-\infty}^{\infty} g(x)f(x)dx + \int_{-\infty}^{\infty} f^2(x)dx. \end{aligned} \tag{3.2.4}$$

Let

$$L(\boldsymbol{\theta}_m, F) = \left[\int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_m}^2(x) - 2\int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_m}(x)dF(x)\right] \tag{3.2.5}$$

for each fixed integer $m > 0$, and define a L_2E functional $T_m^{L_2E}$ on $\mathcal{F}$ by the requirement that for every $F \in \mathcal{F}$

$$T_m^{L_2E}(F) = \{\boldsymbol{\theta}_m \in \Theta_m : L(\boldsymbol{\theta}_m, F) = \min_{\boldsymbol{t}_m \in \Theta_m} L(\boldsymbol{t}_m, F)\}. \tag{3.2.6}$$

Let $\hat{F}_n$ denote the empirical distribution of $\{X_i, i = 1, \ldots, n\}$. Then, an L_2E estimator of $\boldsymbol{\theta}_m$ is one that minimizes $L(\boldsymbol{\theta}_m, \hat{F}_n) = \left[\int_{-\infty}^{\infty} f^2_{\boldsymbol{\theta}_m}(x)dx - 2n^{-1}\sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(X_i)\right]$ with respect to $\boldsymbol{\theta}_m$. That is, we define

$$\hat{\boldsymbol{\theta}}^{L_2E}_{n,m} = T^{L_2E}_m(\hat{F}_n) = \arg\min_{\boldsymbol{\theta}_m}\left[\int_{-\infty}^{\infty} f^2_{\boldsymbol{\theta}_m}(x)dx - 2n^{-1}\sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(X_i)\right], \tag{3.2.7}$$

with $L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m}, \hat{F}_n) = \min_{\boldsymbol{\theta}_m} L(\boldsymbol{\theta}_m, \hat{F}_n)$. In order to propose an estimator of m_0, as in Woo and Sriram (2006 or 2007, section 2), we introduce a model selection criterion based on $L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m}, \hat{F}_n)$ defined by

$$LIC = L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m}, \hat{F}_n) + n^{-1}b(n)\nu(m), \tag{3.2.8}$$

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 $L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m}, \hat{F}_n) \geq L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m+1}, \hat{F}_n)$. 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 lesser than that for model $(m + 1)$. That is, this heuristic stops

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

or, equivalently,

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

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

$$\hat{m}^{L_2E}_n = \min\{m : L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m}, \hat{F}_n) \leq L(\hat{\boldsymbol{\theta}}^{L_2E}_{n,m+1}, \hat{F}_n) + \alpha_{n,m}\}. \tag{3.2.9}$$

Note that in equation (3.2.9) the threshold value $\alpha_{n,m}$ 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 $\alpha_{n,m}$

values decrease. Since an $\hat{m}_n^{L_2E}$ value determines the mixture complexity of the final mixture model, choice of $\alpha_{n,m}$ 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 $\alpha_{n,m} = 3/n$ to numerically study the performance of $\hat{m}_n^{L_2E}$ throughout the article.

3.3 CONSISTENCY THEOREM

The main theoretical result of the article is the consistency of $\hat{m}_n^{L_2E}$, which is stated as a theorem below. First, we state a Proposition giving regularity conditions for the existence and uniqueness of $T_m^{L_2E}(F)$ in (3.2.6). The proof of the Proposition and the theorem are given in the Appendix.

Theorem: Suppose the assumptions of the Proposition (see Appendix) hold. If f_0 is a finite mixture with mixture complexity $m_0 < \infty$, then for any sequence $\alpha_{n,m} \to 0$

$$\hat{m}_n^{L_2E} \to m_0 \quad \text{a.s.}$$

as $n \to \infty$, where $\hat{m}_n^{L_2E}$ and m_0 are as defined in (3.2.9) and (3.2.2), respectively. If f_0 is *not* a finite mixture, then $\hat{m}_n^{L_2E} \to \infty$ a.s.

3.4 COMPUTATIONAL DETAILS

Computation of an estimate of mixture complexity using (3.2.9) is clearly an iterative procedure which can be used for any mixture density. The computation of the integral term in $L(\boldsymbol{\theta}_m, \hat{F}_n) = \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_m}^2(x)dx - 2n^{-1}\sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(X_i)$ during minimization can be difficult for some mixture densities. However, as noted in Scott (2001), computation of the integral term in the L_2E criterion is particularly easy for normal mixtures with the use of the following identity

$$\int_{-\infty}^{\infty} \phi(x|\ \mu_1, {\sigma_1}^2)\phi(x|\ \mu_2, {\sigma_2}^2)dx = \phi(\mu_1 - \mu_2|\ 0, {\sigma_1}^2 + {\sigma_2}^2),$$

where $\phi(x|\ \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2. The identity above is one of many useful formulas given in Wand and Jones (1995). The following shows

that for normal mixtures, $f_{\boldsymbol{\theta}_m}(x) = \sum_{i=1}^{m} \pi_i \phi(x|\ \mu_i, {\sigma_i}^2)$, the use of the above identity reduces the key integral to

$$\int_{-\infty}^{\infty} f^2_{\boldsymbol{\theta}_m}(x)dx = \sum_{i=1}^{m} \frac{1}{2\sqrt{\pi}\sigma_i}\pi_i^2 + 2\sum_{j<k}^{m} \pi_j\pi_k \ \ \phi(\mu_j - \mu_k|\ 0, {\sigma_j}^2 + {\sigma_k}^2),$$

making the integral tractable and thereby significantly reducing the computations involved in minimizing $L(\boldsymbol{\theta}_m, \hat{F}_n)$. Thus, the L_2E criterion for normal mixture has the following analytical form:

$$\begin{aligned} L(\boldsymbol{\theta}_m, \hat{F}_n) &= \sum_{i=1}^{m} \frac{1}{2\sqrt{\pi}\sigma_i}\pi_i^2 + 2\sum_{j<k}^{m} \pi_j\pi_k \ \ \phi(\mu_j - \mu_k|\ 0, {\sigma_j}^2 + {\sigma_k}^2) \\ &- 2/n \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_j \phi(X_i|\ \mu_j, {\sigma_j}^2). \end{aligned} \tag{3.4.10}$$

Now, using the logarithmic transformations for the variances and a logistic-like transformation for the mixing proportions, π_i, in (3.4.10) leads to an unconstrained optimization problem that is solved with standard built-in quasi-Newton method algorithms such as *nlm* in R.

We will now describe our algorithm for normal mixtures; however, one could adopt this algorithm for any family of finite mixture models.

3.4.1 $L_2MIXAlgorithm$

- **Step 1** : Start with $m = 1$, i.e data comes from a single normal density, and formulate $L(\boldsymbol{\theta}_1, \hat{F}_n)$ in equation (3.4.10). Using the *nlm* or *nlminb* routine in R with a choice of initial value (to be discussed below) for $\boldsymbol{\theta}_1$, compute $\hat{\boldsymbol{\theta}}_{n,1}^{L_2E}$ which minimizes $L(\boldsymbol{\theta}_1, \hat{F}_n)$. Now use $\hat{\boldsymbol{\theta}}_{n,1}^{L_2E}$ as an initial value and recompute an L_2E estimate of $\boldsymbol{\theta}_1$. This yields the minimum value $L(\hat{\boldsymbol{\theta}}_{n,1}^{L_2E}, \hat{F}_n)$.

- **Step 2** : Now set $m = 2$ and compute $L(\boldsymbol{\theta}_2, \hat{F}_n)$ in equation (3.4.10). Using the *nlm* or *nlminb* routine in R once again with a choice of initial value for $\boldsymbol{\theta}_2$, compute $\hat{\boldsymbol{\theta}}_{n,2}^{L_2E}$.

Once again, use $\hat{\boldsymbol{\theta}}_{n,2}^{L_2E}$ as an initial value and recompute an L_2E estimate of $\boldsymbol{\theta}_2$. This yields the minimum value of the function, $L(\hat{\boldsymbol{\theta}}_{n,2}^{L_2E}, \hat{F}_n)$.

- **Step 3** : Calculate the difference $L(\hat{\boldsymbol{\theta}}_{n,1}^{L_2E}, \hat{F}_n) - L(\hat{\boldsymbol{\theta}}_{n,2}^{L_2E}, \hat{F}_n)$ and compare it with the threshold value $\alpha_{n,m}$ in (3.2.9). If this difference is less than $\alpha_{n,m}$ $(= 3/n$ in all simulations and data analysis) then stop and report $\hat{m}_n^{L_2E} = 1$. Otherwise go to **Step 4**.

- **Step 4** : Repeat **Steps 2 and 3** by adding one more component to the previous mixture and comparing the difference until the first value $m = m^*$ for which the difference $L(\hat{\boldsymbol{\theta}}_{n,m^*}^{L_2E}, \hat{F}_n) - L(\hat{\boldsymbol{\theta}}_{n,m^*+1}^{L_2E}, \hat{F}_n)$ falls below the threshold value α_{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 $\hat{\boldsymbol{\theta}}_{n,m^*}^{L_2E}$.

As mentioned above, an important step in our iterative method is the choice of initial values. For our L_2E methodology, extensive preliminary simulations indicated that the final estimate of mixture complexity is not severely affected by the choice of initial values. This is also because our estimation algorithm recomputes to arrive at the final estimate; see **Step 1**, for instance. In our numerical studies given m, we chose initial values for the remaining parameters using three different methods, namely K-Means, H-cluster and sample(x,n) routines in R. In our studies, we found that the estimates of mixture complexity were not sensitive to different initial value choices. We used K-means method for most of our simulations and data analysis given in this article. However, in few cases with one of the components having a small mixing proportion, we received warning/error messages concerning insufficient group size or matrix singularity, which were overcome by using sample(x,n) routines to generate initial values.

With respect to computing time, on a typical desktop it took on the average about 5 seconds to obtain one value of $\hat{m}_n^{L_2E}$ based on a simulated dataset of size $n = 1000$

from a normal mixture model with 3 components, which is the largest number components considered in Tables 3.1 and 3.2 in section 3.11. Since our algorithm automatically provides L_2E estimates of the component parameters, the time reported above also includes the estimation of parameters and the overhead of generating a dataset. Furthermore, the number of iterations required for *nlm* or *nlminb* in R to converge was usually no more than 10. The time reported here is based on using K-Means to choose initial values; this is slightly different for other initial value choices.

The L_2E method has distinct advantages over the other methods compared in this article. Firstly, the L_2E criterion could be reduced to a closed form expression given in (3.4.10) for normal mixtures. Whereas, the numerical integration required to compute MHD in HMIX algorithm (see Cutler and Cordero-Braňa, 1996; Woo and Sriram, 2006) places a practical limitation not only in the computation time but also in obtaining sufficient accuracy to perform quasi-Newton optimization. Similar comments as for MHD also apply to NKE and MKE estimators of James et al. (2001). Chen and Khalili (2006,2008) developed a new penalized likelihood approach called $MSCAD$, which deviates from information-based methods such as AIC and SBC. The objective function for $MSCAD$ is also relatively more complicated because it involves a SCAD-type penalty, hence the name $MSCAD$. The $MSCAD$ method is also based on revised EM algorithm, which uses the penalized likelihood instead of the log-likelihood.

Secondly, as for the choice of initial values, observations made by Cutler and Cordero-Braňa (1996) and Woo and Sriram (2006) show that the $MHDE$ parameter estimates are sensitive to the choice of initial values, which in turn affects the estimate of mixture complexity. Furthermore, the $MHDE$ algorithm also shares some of the weaknesses of the EM algorithm in terms of slow convergence. Similar comments as for MHDapply to NKE and MKE estimators of James et al. (2001), the Bayesian algorithm of Roeder and Wasserman (1997) denoted by $R\&W$; the Bootstrap algorithm of McLachlan (1987) denoted by *Bootstrap*; and the CDF method of Henna (1985) denoted by *Henna*. While not mentioned explic-

itly in Chen and Khalili (2008), the fact that $MSCAD$ is also an EM type algorithm, it is also likely to share some of the drawbacks of EM in terms of slow convergence and choice of initial values. Furthermore, $MSCAD$ procedure also requires careful choice of tuning parameters for their SCAD penalty (Fan and Li, 2001). Finally, as for computing time, the time reported above for L_2E is substantially lower than those for $MHDE$ (Woo and Sriram, 2006, Section 7) and all other competing procedures considered here. These computational advantages make our L_2E approach a more attractive alternative to all other available procedures in the literature.

3.5 SIMULATION STUDIES

In this section, we conduct a variety of Monte Carlo simulations to validate the consistency Theorem by assessing the performance of $\hat{m}_n^{L_2E}$ in (3.2.9) for moderate to large sample sizes. We carry out simulation studies for two different scenarios, but in both the postulated model is a member of family of normal mixtures. In the first scenario, the data are generated from a normal mixture model, whereas in the second the data are generated from a mixture model with symmetric departures from component normality. Note that the first scenario would examine the efficiency of our estimator when the model is correctly specified, while the second would assess the robustness of our estimator against model misspecification.

For the first scenario, we perform the two simulation studies discussed in Woo and Sriram (2006) and in each study compare the performance of our estimator with six other estimators for mixture complexity available in the literature. The first simulation demonstrates the performance on a target density, which is a three-component mixture of normal densities, for a variety of sample sizes. The second is a simulation study on a variety of normal mixtures from Marron and Wand (1992) for a fixed sample size.

For the second scenario, we perform four different simulation studies to assess the robustness of our estimator under symmetric departures from postulated component normality. In these simulations, the samples are drawn from mixtures with two components, where the

component densities are those of scale and location transformations, respectively, of a Student's t random variable with two or four degrees of freedom, or a rescaled t random variable with three or four degrees of freedom. In addition, we consider varying degrees of separation (or equivalently, *overlap*) between the two component densities. The setup for our robustness analysis is similar to those described in Woo and Sriram (2006), Woodward et al. (1984) and Markatou (2001); also see Woodward et al. (1995) and McCann and Sarkar (2000). In each of these simulations, robustness of our estimator of mixture complexity under model misspecification is also compared with the estimator of mixture complexity defined in James et al. (2001)and the $MHDE$ estimator defined in Woo and Sriram (2006).

3.5.1 **Three-component mixture**

The first simulation demonstrates the performance of (3.2.9) for the target density given by

$$f(x) = (1/2)\phi(x|(0, 10)) + (1/4)\phi(x|(-0.3, 0.05)) + (1/4)\phi(x|(0.3, 0.05)), \qquad (3.5.11)$$

where ϕ denotes the normal density with mean and variance identified inside the parentheses. Here, one of the components has a large variance and the other two have small variances. We implement our computational algorithm for sample sizes n = 50, 250, 500 and 1000 drawn from (3.5.11). For each sample size, we perform 100 Monte Carlo replications of our algorithm, each yielding an estimate of mixture complexity. We then tally the estimated number of components (out of 100 replications).

These counts are reported for each sample size in Table 3.1 of section 3.11, where L_2E corresponds to the estimate given by (3.2.9). In addition, for comparison purposes, we also provide in Table 3.1 the counts obtained via the $MHDE$ algorithm of Woo and Sriram (2006); the NKE and MKE algorithm of James et al. (2001); the Bayesian algorithm of Roeder and Wasserman (1997) denoted by $R\&W$; the Bootstrap algorithm of McLachlan (1987) denoted by *Bootstrap*; and the CDF method of Henna (1985) denoted by *Henna*. In

this case, the true mixture complexity is 3. In Table 3.1 of section 3.11, we denote the highest percentage (50% or above) of correct identifications in bold with an asterisk beside it.

The simulation results in Table 3.1 of section 3.11 show the following: For sample size $n = 50$, only the $R\&W$ algorithm correctly identifies a large percentage of times, while all the other algorithms largely underestimate the true mixture complexity. While the $R\&W$ procedure correctly identifies the true mixture complexity, it should be noted that in this case, it also overestimates the true mixture complexity about 12% of the times. For sample size $n = 250$, only our L_2E and the $R\&W$ algorithm correctly identifies a large percentage of times (50% or above). While the L_2E and $R\&W$ procedure correctly identify the true mixture complexity, it should be noted that in this case, $R\&W$ procedure also overestimates the true mixture complexity about 40% of the times, where as our L_2E over estimates only about 5 % of the times.

For sample size $n = 500$, only our L_2E and the $MHDE$ algorithm correctly identify a large percentage of times. However, $MHDE$ underestimates the mixture complexity about 35% of times. Out of the two, our L_2E was indisputably the best, as it correctly identifies the mixture complexity substantially higher percentage of times. For sample size $n = 1000$, only our L_2E, $MHDE$ and MKE algorithms perform well. However, both $MHDE$ and MKE underestimate the mixture complexity substantially more than that of L_2E. In addition, it should be noted that in this case, MKE algorithm also overestimates the true mixture complexity about 19% of the times. In comparison, our L_2E neither overestimates nor underestimates severely. Overall, when the model is correctly specified, when the sample size is larger than 250, our L_2E is the best as it correctly identifies the mixture complexity substantially higher percentage of times than all the others considered here.

3.5.2 Marron and Wand mixtures

Here we carry out similar studies as in section 3.5.1, where target densities are normal mixture models #2 and # 4 - #9 considered in Marron and Wand (1992, see pages 717

and 718). These mixtures exhibit a range of unimodal, skewed and multimodal densities appropriate for testing the performance of the above algorithms. As in Woo and Sriram(2006) and James et al. (2001), we compare the performance of all the algorithms mentioned in Table 3.1 of section 3.11 based on percentage correct identification of the true mixture complexity. The sample size for this study is $n = 1000$. Once again, in Table 3.2 of section 3.11, we denote the highest percentage (50% or above) of correct identifications in bold with an asterisk beside it.

The simulation results in Table 3.2 of section 3.11 show the following: When the true number of components $m = 2$, as in mixtures model # 4 - 8, all the algorithms perform well by correctly identifying a large percentage of times, except the R & W algorithm which overestimates the true mixture complexity for model # 4. For model # 5, while the $R\&W$ procedure correctly identifies the true mixture complexity a large percentage of times, it should be noted that in this case, it also overestimates the true mixture complexity about 45% of the times. Once again, for model # 8, despite correct identification, the $R\&W$ procedure also overestimates the true mixture complexity about 18% of the times.

In the case of mixture model # 2 ($m = 3$), our L_2E was indisputably the best as it is the only one that correctly identifies the true mixture complexity a large percentage of times, while all the other algorithms largely underestimate the true mixture complexity. However, it should be noted that in this case, our L_2E also underestimates the true mixture complexity about 46% of the times.

In the case of mixture model #9 ($m = 3$), our L_2E, the MKE, the $MHDE$ and the *Bootstrap* algorithms perform well. And our L_2E out performed all other methods by a healthy margin. While the L_2E and *Bootstrap* procedure correctly identifies the true mixture complexity , it should be noted that in this case, *Bootstrap* procedure overestimates the true mixture complexity about 12% of the times and L_2E procedure also overestimates the true mixture complexity about 19% of the times. On the other hand, while the MKE and $MHDE$ procedure correctly identifies the true mixture complexity , it should be noted that in this

case, MKE procedure underestimates the true mixture complexity about 38% of the times and $MHDE$ procedure also underestimates the true mixture complexity about 49% of the times.

Overall, our L_2E is the only procedure which correctly identifies large percentage of times in all the seven models considered here. These show that, when the model is correctly specified, the L_2E algorithm provides a useful way to estimate the mixture complexity for a variety of mixtures. Undoubtedly, L_2E algorithm is the best among those considered here and associated computations are considerably less intensive compared to those considered in the article.

3.5.3 **Robustness**

In this section, we describe an approach to assess the robustness of $\hat{m}_n$ in terms of its ability to correctly identify the true mixture complexity when the postulated mixture model is misspecified. We assess the robustness of $\hat{m}_n^{L_2E}$ when the postulated model is a mixture of normals but the data are generated from a mixture with symmetric departure from component normality. As in Woo and Sriram (2006), we consider two slightly different setups for our simulation study. The first setup is as described in Woodward et al. (1984) for the estimation of mixing proportions (also see Woodward et al. (1995) and McCann and Sarkar (2000)). The second setup is as described in Section 29.3.3 of Markatou (2001); also see section 4 of Markatou (2000). More specifically, for our simulation study, we consider a mixture with two components given by

$$f_{\boldsymbol{\theta}_2}(x) = \pi f_1(x) + (1-\pi) f_2(x), \tag{3.5.12}$$

where f_1 is the density associated with a random variable $X_1 = aY$ and f_2 is the density associated with a random variable $X_2 = Y + b$ for some $a > 0$ and $b > 0$. Here, the postulated distribution for Y is standard normal but, in the first setup, the samples are generated from the mixture in (3.5.12) when Y is a Student's $t(df)$-random variable with degrees of freedom $df = 2$ or 4. For our first setup, we set $\pi = 0.25, 0.50$ and 0.75, $a = 1$ and $\sqrt{2}$, and for each

pair of (π, a) values, we choose the values of b so that the *overlap* (see Woodward et al. 1984 for definition) between the two t-component densities in (3.5.12) is either 0.10 or 0.03. We will not explicitly give these b values, except in three cases (see below), but refer to these b values as t-overlap in Tables 3.3 and 3.4 of section 3.11. Note that the general shapes of such a two-component postulated (normal mixture) model and a two-component t-mixture model from which the data are generated are markedly different for some values of π, a and b (see, e.g., Figure 1 in McCann and Sarkar (2000) for $\pi = 0.75, a = \sqrt{2}$, overlap $= 0.10$ and $df = 4$). In addition, the component densities in the sampling model have much heavier tail than those in the postulated (normal) mixture model.

Our second simulation setup differs slightly from the one above in that the samples are generated from the mixture in (3.5.12) when Y is a rescaled Student's $t(df)$-random variable with degrees of freedom $df = 3$ or 4. As in Markatou (2001), by a rescaled Student's $t(df)$ we mean a $t(df)$-random variable that is rescaled to have variance 1. Also, for each pair of (π, a) values given above, we choose the values of b so that the overlap between the two normal-component densities in (3.5.12) is either 0.10 or 0.03. That is, we use the b values that are given in Table 2 of Cutler and Cordero-Braňa (1996). We will refer to these b values as N-overlap in Tables 3.5 and 3.6 of section 3.11.

The sample size for this study is $n = 1000$ and we performed 100 Monte Carlo replications of our L_2E, $MHDE$ algorithm of Woo and Sriram (2006) and the MKE algorithm of James et al. (2001), with $\alpha_{n,m} = 3/n$. Tables 3.3 to 3.6 give a tally of estimated number of components for the L_2E, $MHDE$ and MKE algorithms, for each choice of a, π and b given above. In all these cases the true mixture complexity is 2 and we denote the highest percentage (50% or above) of correct identifications in bold with an asterisk beside it, in Tables 3.3 to 3.6.

The simulations presented here span over a variety of moderate to more extreme symmetric departures from component normality along with two different types and amounts of separation between the component densities. In all, there are 40 different cases of model

misspecifications considered here. Our L_2E algorithm was indisputably the best as it performed well in all 40 cases. Although the $MHDE$ algorithm was not the best but nearly so, it performed well in about 36 cases out of 40. The MKE algorithm, as expected, performed well only in 9 cases in terms of correctly identifying the true mixture complexity $m_0 = 2$. However, when the $t(2)$ components are poorly separated (t-overlap = 0.10) and in the following three cases, $(\pi, a) = (0.5, 1), (0.5, \sqrt{2})$ and $(0.75, \sqrt{2})$, Table 3.4 shows that (our L_2E) and the MKE perform better than the $MHDE$ algorithm.

In Tables 3.1 and 3.2 of section 3.11, we noticed that our L_2E algorithm overestimates the true mixture complexity in some instances. However, Tables 3.3 to 3.6 of section 3.11, show that our L_2E algorithm overestimates slightly in some instances but rather severely in some other cases. We do not observe much underestimation with our L_2E algorithm here at all. However, Tables 3.3 to 3.6 of section 3.11, show that in many instances the MKE algorithm rather severely underestimates the true mixture complexity and in some instances $MHDE$ algorithm also rather severely underestimates. Given the extreme nature of symmetric departures from component normality considered in our simulations, the results in Table 3.3 to Table 3.6 of section 3.11, serve as a testament that our L_2E algorithm is highly robust against model misspecification, and simply the best when the computational aspect is also taken into account.

3.6 DATA ANALYSIS

The goals in analyzing finite mixture models are two-fold: (1) to determine what model best fits the data at hand (eg, a mixture of 1, 2, or 3 normal distributions) and (2) to estimate the parameters of that best-fitting model. In practice, these steps are performed in reverse order: parameters are first estimated, and the solutions for different models are then compared. However, we perform the data analysis in such a way we simultaneously determine an estimate of mixture complexity and estimates of the component parameters.

In this section, we analyze three well-known real datasets to further demonstrate the use of our L_2E method.

3.6.1 **SLC data**

Red blood cell sodium-lithium contertransport (SLC) activity data collected from 190 individuals was analyzed originally in Dudley et al. (1991). The SLC is measured as the difference in lithium efflux rate from lithium-loaded cells into sodium chloride and sodium-free media. Roeder (1994) discussed 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 p and $1-p$. Furthermore, Roeder (1994) argued that red blood cell SLC is believed to follow one of the following two competing genetic models.

Model I : (Simple dominance model) Genotypes A_1A_1 and A_1A_2 have pheno-type $\boldsymbol{\theta}_1$, where as A_2A_2 have phenotype $\boldsymbol{\theta}_2$. Hence $P(\Theta = \boldsymbol{\theta}_1) = p^2 + 2p(1-p)$ and $P(\Theta = \boldsymbol{\theta}_2) = (1-p)^2$.

Model II : (Additive model) Each of the three genotypes yields a distinct phenotype with $P(\Theta = \boldsymbol{\theta}_1) = p^2$, $P(\Theta = \boldsymbol{\theta}_2) = 2p(1-p)$ and $P(\Theta = \boldsymbol{\theta}_3) = (1-p)^2$. Furthermore, $\boldsymbol{\theta}_1 < \boldsymbol{\theta}_2 < \boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_3 - \boldsymbol{\theta}_2 = \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1$.

Geneticists are interested in SLC because it is correlated with blood pressure and hence may be an important cause of hypertension. Roeder(1994) fitted a mixture of normal with three components to this data. Her fit corresponds to the additive model *Model II* above. Ishwaran, James and Sun (2001) adopted a Bayesian approach to estimating the mixture complexity and proposed two algorithms called the *generalized weighted Chinese restaurant* (GWCR) and *blocked Gibbs sampler*. Their analysis of SLC data showed that GWCR supported a three component mixture while the blocked Gibbs sampler based on Bayes Information Criterion penalty supported a two-component mixture. Recently, Woo and Sriram (2006) analyzed this data using $MHDE$ and suggested a two-component mixture. Also Chen and

Khalili (2006) used their $MSCAD$ procedure and fitted a three-component mixture which is similar to one of the models reported in Ishwaran, James and Sun (2001). Recently, Fujisawa and Eguchi (2006) proposed robust parameter estimates, called β-estimates, for normal mixtures using a modified likelihood approach suggested in Basu et al. (1998); they also analyzed the SLC data using their method. Here, we revisit the SLC data using our L_2E algorithm.

We used our L_2E algorithm to first determine an estimate of mixture complexity and simultaneously obtain estimates of the component parameters for the SLC data. Our postulated model is a normal mixture with unknown means, (unequal) variances and mixing proportions, and we used our L_2E algorithm with threshold value $\alpha_{n,m} = 3/n$. Our analysis yielded an estimate $\hat{m}_n^{L_2E} = 3$ of the mixture complexity for the SLC data. When our L_2E algorithm stops and reports $\hat{m}_n^{L_2E} = 3$, it also automatically provides L_2E estimates of all the parameters in the three-component mixture. These L_2E estimates corresponding to the best fitting three-component normal mixture density are given in Table 3.7 of section 3.11. For the SLC data, Fujisawa and Eguchi (2006) computed robust estimates for the parameters (optimal β-estimate) assuming that the underlying distribution has a normal mixture model with $m = 3$. Interestingly, their fitted mixture model is almost identical to our L_2E fit. The fitted mixture density using our L_2E method along with the $MHDE$ method of Woo and Sriram (2006), the MKE method given in James et al. (2001), the $MSCAD$ method of Chen and Khalili (2006), the optimal β-estimate of Fujisawa and Eguchi (2006) and $Kernel$ density estimate of the data are superimposed over the histogram of the data in Figure 3.1 of section 3.11. Mixture fit given by $L_2E(m = 3)$, $MHDE(m = 2)$, $MSCAD(m = 3)$, $MKE(m = 2)$ and optimal β given in the Table 3.7. From these graphs and tables, we conclude that the three-component L_2E provides a better fit of the data.

Note that all procedure considered in the Table 3.7 with the excepetion of $MSCAD$ assume unequal component varainces. With equal variance assumption, the $MSCAD$ procedure seems to satisfy the additive model assumtion approximately. It should be noted that all the procedures considered in Table 3.7 (with the exception of $MSCAD$) do not satisfy

the additive model. Nevertheless, Figure 3.1 show the L_2E and optimal β estimate fit the data better than the rest.

3.6.2 **Acidity Data**

The acidity data involves an acid-neutralizing capacity (ANC) index measured in a sample of 155 lakes in North-central Wisconsin, United States. Acidification is an environment problem and identifying different subpopulations of lakes(e.g. at-risk lakes, not-at-risk lakes) can be useful in determining which lake characteristics, if any can be used to predict higher acidification. The data have been previously analyzed using a mixture of normal distributions on the log scale; see Crawford et al. (1992), Richardson and Green (1997), McLachlan an Peel(1997b) and Ishwaran et al. (2001). Recently Mcgrory and Titterington (2007) analyzed this data using Deviance Information criterion (DIC) based on Bayesian measures of the complexity. Chen and Khalili (2006) also analyzed the acidity data using their $MSCAD$ procedure.

Richardson and Green(1997), McLachlan an Peel(1997b), Ishwaran et al. (2001) and Chen and Khalili (2006) suggested a three-component normal mixture for the acidity data. Mcgrory and Titterington (2007) based on their DIC suggested a two-component normal mixture. We used our L_2E algorithm to determine an estimate of mixture complexity and simultaneously determine an estimate of the component parameters. For this, we assumed normal mixture models with unknown means, (unequal) variances and mixing proportions, and used our L_2E algorithm with threshold value $\alpha_{n,m} = 3/n$. Our analysis yielded an estimate $\hat{m}_n^{L_2E} = 3$. The L_2E estimates corresponding to the best fitting three-component normal mixture density are given in Table 3.8 of section 3.11 along with the $DIC(m = 2)$ estimates from Mcgrory and Titterington (2007) and $MSCAD(m = 3)$ estimates of Chen and Khalili (2006). We also graph all these fitted densities along with the kernel density estimate superimposed over the histogram of the data in Figure 3.2 of section 3.11. Once

again, from these graphs and tables, we conclude that the three-component L_2E provides a better fit of the data than the others.

3.6.3 **Enzyme Data**

The Enzyme data concerns the distribution of enzymatic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances, among group of 245 unrelated individuals. The study was undertaken to validate the caffeine as a probe drug to establish the genetic status of rapid metabolisers and slow metabolisers, and to use such subgroups as a marker of genetic polymorphism in the general population.

Richardson and Green (1997) also analyzed the enzyme data and suggested mixture models with 3 to 5 components. Ishwaran et al. (2001) suggested a mixture model with 8 components. Recently, Mcgrory and Titterington (2007) used their DIC method and fitted a four-component normal mixture. We used our L_2E algorithm to determine an estimate of mixture complexity and simultaneously determine an estimate of the component parameters for the Enzyme data. We postulated a normal mixture model with unknown means, (unequal) variances and mixing proportions, and used our L_2E algorithm with threshold value $\alpha_{n,m} = 3/n$. Our analysis yielded an estimate $\hat{m}_n^{L_2E} = 3$. The L_2E estimates corresponding to the best fitting three-component normal mixture density are given in Table 3.9 of section 3.11 along with the $DIC(m = 4)$ estimates from Mcgrory and Titterington (2007). Also the fitted densities of these estimates along with the kernel density are superimposed over the histogram of the data in Figure 3.3 of section 3.11. All these make a compelling case that our three-component mixture density based on the L_2E estimates provides a good and parsimonious fit of the Enzyme data.

3.7 SUMMARY AND CONCLUSIONS

An information criterion approach based on minimum L_2 distances is used to construct an estimator of unknown number of components in finite mixtures, when the form of component

densities are unknown but are postulated to be members of some parametric family. This estimator, termed as L_2E, is consistent for any parametric family of finite mixture models. When the postulated normal mixture model is same as the model from which samples are drawn, simulations show that our estimator competes well with other procedures available in the literature, and particularly well against an estimator based on Kullback-Leibler distance introduced by James et al. (2001) and $MHDE$. The most distinguishing feature of our estimator is that it continues to identify the mixture complexity correctly even when the sampling model is a (moderate to more extreme) symmetric departure from postulated component normality, while the estimator of James et al. (2001) becomes highly unstable in these situations. Furthermore, the L_2E turns out to have better overall robustness property than the $MHDE$ of Woo and Sriram (2006). In addition to better performance than the $MHDE$, the L_2E is computationally much simpler, avoiding delicate choice of adaptive density estimator and associated bandwidths as needed in Woo and Sriram (2006). It should be noted that the conclusions based on our numerical study of robustness (see section 3.5.3) are by no means definitive. However, we do believe that our findings on robustness are of sufficient substance to raise the interesting theoretical questions such as behavior of influence functions and breakdown points.

Choice of threshold values $\alpha_{n,m}$ undoubtedly has an impact on the final estimate of the unknown mixture complexity. In our numerical studies we motivate our choice of $\alpha_{n,m} = 3/n$ based on the AIC criterion. More work remains to be done on the choice of $\alpha_{n,m}$ for our estimator. However, it is shown that this choice of threshold yields a parsimonious mixture model fit for three real datasets, which are superior to fits provided by other competing methods in the literature.

Finally, with respect to computation, the L_2E procedure has many distinct advantages over MHD and other procedures in the literature. For example, the L_2E criterion has a simple structure which enables us to use the built-in *nlm* and *nlmbin* routines in R for minimization. Furthermore, the L_2E estimates are not affected by the choice of initial values

and it requires less computing time. Thus, transparency, ease of use and efficiency in achieving computational speed combined with competitive performance and robustness feature makes the L_2E estimator stand out as an attractive alternative to other existing methods in the literature.

3.8 SUPPLEMENTAL MATERIALS

Here we give a list of computer codes used in the simulation and data analysis. We used *R*2.8 for all simulations and data analysis in this article.

```
#############################################################################
#This function is to randomly generate the normal mixture
#############################################################################

rmixnorm<- function(n,probs,means,sigma) {

  out<- rep(0,n)
    for (i in 1:n) {
    u<- runif(1)
    k<-length(probs)
    indg<- 1:k
   cp = cumsum(probs)
   j = min(indg[u <= cp])
   out[i] <-rnorm(1,means[j],sigma[j])
}
return(out)
}

#############################################################################
#This function for simulating values from t-mixture for robustness study
#############################################################################

rtmix <- function(n, prob=0.5,df1=1, df2=1,a=1,b=1){
    u <- runif(n)
    out <- numeric(n)
    for(i in 1:n) out[i] <- if(u[i] < prob) a*rt(1,df1) else (b+ rt(1,df2))
    return(out)
}

#############################################################################
#This function for simulating values from re-scaled t-mixture for robustness study
#############################################################################

rtmixrscale <- function(n, prob=0.5,df1=1, df2=1,a=1,b=1){

     u <- runif(n)
   out <- numeric(n)
   for(i in 1:n)
```

```
    out[i] <- if(u[i] < prob) (a*rt(1,df1)/sqrt(3))
                else (b+ (rt(1,df2)/sqrt(3)))
     return(out)
}

#########################################################################
#
#This is the main function: For an input data x
#and given number of components k, this function computes the L2E
#function value for normal mixture and outputs the
#estimates of parameters and the corresponding minimum
#value of the function.
#
#########################################################################
# Thanks to Professor David W. Scott for providing this function.
# Inputs:
# x - n x d input data matrix (d=1 vector OK)
# K - number of components desired (K=1 default)
# grps -optional way of inputting initial guesses  (data labels 1,2,...,K)
#       vector of length n    labels = 0 are ignored (useful if w.sum=F)
# w.sum - constrain weights to sum to 1  (T/F)
# mu - input guess for means  d x K (matrix)
# sig - d x d x K input guess for covariance matrices (array)
# w - input guess for K weights (always length K, even with w.sum=T)
# nit - max number of iterations for nlm
# nev - max number of function evaluations for nlm
# pl - print level for nlm (0=none 1=some 2=lots)
# Output:
#list containing estimated parameters and the minimum
# (m=mean s=sig w=w,lmin=lmin)
#########################################################################

mix.pdc <- function(X,K=1,grps,w.sum=T,mu,sig,w,nit=100,nev=200,pl=1) {
  if(!is.matrix(X)) { X <- cbind(X) };  n <- nrow(X); d <- ncol(X)

  #Evaluate the initial values for mu & sigma if grps is given
  if(!missing(grps))
  { if(length(grps)!=n)
    stop("Invalid grps length.(grps length should be same as the sample.)")
    if(max(grps)!=K) print("Warning -- max of grps does not match K")
    nk <- rep(0,K); mu <- matrix(0,d,K); sig <- array(0,c(d,d,K))
    for(k in 1:K) { ii <- seq(n)[grps==k]; nk[k] <- length(ii)
      if( length(ii)<2 )
      stop(paste("grps insufficient for class ",as.character(k)))
      mu[,k] <- apply(X[ii,,drop=F],2,"mean"); sig[,,k] <- var(X[ii,]) }
```

```
    w <- nk/n }

  # Input Validation
  if(d==1) { mu <- matrix(c(mu),1,K); sig <- array(c(sig),c(1,1,K)) }
  if(any(dim(mu)!=c(d,K))) stop("input mu matrix wrong dimension")
  if(any(dim(sig)!=c(d,d,K))) stop("input sig array wrong dimension")
  if(length(w)!=K) stop("input w vector wrong length")

# L2E function to minimize
# x is  a list of parameters of mixtures
# x = [mus:sigmas:weights]

pdc <- function(x) {          # first extract parameters from  x
    onen <- rep(1,n); oned <- rep(1,d)
    mu <- matrix(x[1:(d*K)],d,K)  # assumes fortran order
    ns <- d*(d+1)/2; # no of sigmas in a MVG-d (sigx sigy sigxy d=2, ns=3)
    sigi.u <- matrix(x[(d*K+1):((d*K)+ (ns*K))],ns,K)
    # will make an ns size array
    tw <- x[-(1:(d*K+ns*K))];
    # total of weigts =1
    if(length(tw)==K) {
w <- exp(tw) # if length=K then no sum constraint
    } else {
tw <- c(exp(tw),1);
w <- tw/sum(tw) # if length!=K then sum constraint
    } #end if tw.length==K

    tot1 <- 0;
    tot2 <- 0;
    cc <- 2^(d/2);
    sig <- array(0,c(d,d,K))
    deti <- rep(0,K)   # determinant inverse sq root

    for(k in 1:K) {
    muk <- mu[,k];
      U <- matrix(0,d,d)
      U[row(U)<=col(U)] <- sigi.u[,k]; #upper triangle of U
      deti[k] <- exp(sum(diag(U)))
      dU <- exp(diag(U)); #variance should be positive
      diag(U) <- dU;
      sig[,,k] <- solve( t(U)%*%U )
      #solve(x) returns the inverse of matrix x

      tot1 <- tot1 + w[k]^2*deti[k]/cc
      tot2 <- tot2 + w[k] * deti[k] *
```

```
            sum( exp( (-.5* ( (X-outer(onen,muk))%*%t(U) )**2) %*%oned ) )

      if(k>1) {
     for(m in 1:(k-1)) {
     mum <- mu[,m];
                     sigi.km <- solve( sig[,,k]+sig[,,m] )
        U0 <- chol( (sigi.km+t(sigi.km))/2 );
                     deti.km <- exp(sum(log(diag(U0))))
           dd <- U0%*%(muk-mum);
                     tot1 <- tot1 + 2*w[k]*w[m]*deti.km*exp(-.5*sum(dd^2))
       }#end for
    }#end if k>1
    }#end for k
    tot <- tot1 - 2*tot2/n
}

  x0 <- c(mu)    # assumes fortran column order
  for(k in 1:K) {
sig0 <- sig[,,k]
     if(d==1) {
xU0 <- log(1/sqrt(sig0))
} else {
tmp <- solve(sig0);
sigi <- (tmp+t(tmp))/2
     U0 <- chol(sigi); diag(U0) <- log(diag(U0));
xU0 <- U0[row(U0)<=col(U0)]
}
      x0 <- c(x0,xU0)
  }
  if(w.sum) { if(K==1)
  {w0 <- NULL} else { w0 <- log(w[-K]/w[K]) } } else { w0 <- log(w) }
  x0 <- c(x0,w0)    # w0 of length K-1 if sum constraint on (NULL if K=1)

  #nonlinear minimization routine-nlm
  ans <- nlm(pdc,x0,print.level=pl,iterlim=nit)
  pr<-ans$est
  ans <- nlm(pdc,x0,print.level=pl,iterlim=nit)


  #Can use the following non-linear minimization routine as well
  #lower=-Inf; upper= Inf
  #lower= min(x)-5; upper=max(x)+5
  #ans<-nlminb(x0,pdc,lower=0,upper=upper)
  #pr<-ans$par
   #ans<-nlminb(pr,pdc,lower=0,upper=upper)
```

```
    pr<- ans$est
    lmin<-ans$min/sqrt(pi)
    xx<-ans$est

  #If using the nlminb use the following
   # pr<-ans$par
   #lmin<- ans$obj/sqrt(pi)
   # xx <- ans$par

  ns <- d*(d+1)/2
  mu <- matrix(xx[1:(d*K)],d,K);
  xx <- xx[-(1:(d*K))]
  for(k in 1:K) {
    U.ans <- matrix(0,d,d);
    U.ans[row(U.ans)<=col(U.ans)] <- xx[1:ns]
    diag(U.ans) <- exp(diag(U.ans));
    sigi <- t(U.ans) %*% U.ans;
    sig[,,k] <- solve(sigi);
    xx <- xx[-(1:ns)]
  }
    s<-as.vector(sqrt(sig))
  w <- xx;
  if(length(w)==K) { w <- exp(w) } else { w <- exp(c(w,0)); w <- w/sum(w) }
  list(m=mu,s=s,w=w,lmin=lmin)
}

#############################################################################
#  Initial guess for the parameters
#############################################################################
init<-function(type="km",x,k){

      if(type=="km"){
      #K-Means
      mm<-kmeans(x,k)
      g<-mm$cluster
      mu<-mm$center
      s1<-mm$size
      w<- s1/ss
      sig<-mm$withinss /s1
      #sig<- sqrt(sig)
    }

    if(type=="rs"){
      #Random sample
```

```
      g <-sample(0:k,ss,T)
      #list(g=g)
      mu=0
      sig=0
      w=0
    }

    if(type=="hc"){
       library(amap)

      #Hierarchical-Clustering
      hc<-hcluster(x, method = "euclidean", diag = FALSE, upper = FALSE,
                link = "complete", members = NULL, nbproc = 2,
                doubleprecision = TRUE)

      memb <- cutree(hc, k = k)
      g=memb
      n=length(x)
      nk <- rep(0,k)
      mu<-rep(0,k)

      for(i in 1:k) {
        ii <- seq(n)[memb==i]; nk[i] <- length(ii)
        sig[i]<- var(x[ii,drop=F])
        mu[i] <- mean(x[ii,drop=F])
        w <- nk/n
      }
    }

    list(g=g,mu=mu,sig=sig,w=w)
}

#############################################################################
#This is the function used to test for the mixture complexity.
# We need to have other functions such as mix.pdc.r,
# init.r, rmixnorm.r, rtmix.r,....
#############################################################################

cc=100                  #No of MC tests
ss=1000               #Sample size
count <- rep(0,cc)    #results(K) holder

for (ct in  1:cc) {

  k=1
```

```
  #input the data-using rmixnorm for the mixture of normal or
  # for robust study us rmix or input  for the real data
  xd<- rtmix(ss,0.5,4,4,1,3.066)

  #initial guess
  imu = mean(xd)
  isig=var(xd)
  iw=1

  # can also use the init function
  #g<-init("hc",xd,k)

  #call the main function to compute the estimates
  # note her sigma here is the variance
  l2emix<- mix.pdc (xd,K=k,w.sum=T,mu=imu,sig=isig,w=iw,nit=100,nev=200,pl=1);

  # Can input the initial guess as g instead of (mu, sigma, w)

  #l2emix<- mix.pdc (xd,K=k,grps=g,w.sum=T,nit=500,nev=200,pl=1);

  mu<-l2emix$m
  w<-l2emix$w
  s<-l2emix$s
  s<-sqrt(s)
  l2e<-rep(0,50)
  #the minimum value fo the l2e
  l2e[k]<-l2emix$lmin

  repeat {
      k=k+1

       iVal = init("km",xd,k)

    #call the main function to compute the estimates
  # note her sigma here is the variance

#l2emix<- mix.pdc (xd,K=k,w.sum=T,mu=iVal$mu,
                       sig=ival$sig,w=iVal$w,nit=100,nev=200,pl=1);

# Can input the initial guess as g instead of (mu, sigma, w)

 l2emix<- mix.pdc (xd,K=k,grps=iVal$g,w.sum=T,nit=500,nev=200,pl=1);

   mu<-l2emix$m
   w<-l2emix$w
```

```
   s<-l2emix$s

     #the minimum value of the l2e
     l2e[k]<-l2emix$lmin
    # compute the difference between the minimum l2e for (k-1) and k
     diff<- l2e[k-1]-l2e[k]
    #th is the threshold value we use "3/n" here
    # Check for the m
   if( diff <= (th)) break
    } #end of repeat
    k= k-1
   #k
   cat("k=",k)
   count[ct]=k
   } # end for MC

  count
  table(count)

#############################################################################
This function is for calculating the L2E value-mixture of normal
#############################################################################
#Input x-data, m=mean, w=mixing proportions, s=standard deviation
#output theL2E function value
#############################################################################
l2ecal<-function (x,m,w,s) {
    p=outer(w,w)
    c=p[upper.tri(p)]

    c1=0
    c2=0
    e=outer(m,m,FUN="-")
    e1=e[upper.tri(e)]
    sig=outer(s,s,FUN="+")
    esig=sig[upper.tri(sig)]

    c1=sum(2*c*dnorm(0,e1,esig))
    c2= sum((w^2)/(2*sqrt(pi)*s))

    ad=0
    fd=0
    for( j in 1: length(w)){
      ad[j]=sum(w[j]*dnorm(x,m[j],s[j]))
    }
    n=length(x)
```

```
    fd= (2/n)*sum(ad)
    l2e=c1+c2- fd
}

#########################################################################
The real data used in this chapter are:
######################
The Enzyme Data
#########################################################################

xd=c(0.130, 0.080, 1.261, 0.224, 0.132, 1.052, 0.085, 0.124, 0.718,
     0.280, 0.687,0.106, 0.088, 0.137, 0.096, 0.124, 0.126, 1.279,
     1.007, 0.195, 0.167, 0.213,0.108, 1.371, 0.190, 0.184, 1.298,
     1.036, 0.205, 1.950, 1.018, 0.172, 0.148,0.292, 0.113, 0.185,
     0.129, 1.329, 0.149, 0.236, 2.545, 1.073, 0.162, 2.518,0.142,
     2.880, 0.178, 1.075, 0.128, 0.083, 0.409, 0.340, 0.246, 1.195,
     1.452,1.123, 1.361, 0.222, 0.962, 0.875, 0.078, 0.520, 0.194,
     1.195, 0.709, 0.021,0.166, 0.081, 0.265, 0.159, 0.308, 1.604,
     0.179, 0.172, 0.131, 0.305, 0.215,0.214, 0.853, 0.137, 0.466,
     1.419, 2.016, 1.944, 1.040, 1.200, 0.255, 0.232,0.200, 0.240,
     0.216, 0.277, 2.427, 0.320, 0.142, 0.134, 0.198, 0.126, 1.173,
     0.342, 1.672, 0.193, 1.633, 0.860, 1.293, 0.207, 1.811, 1.741,
     1.488 ,0.124,1.326, 0.148, 0.109, 1.848, 1.310, 0.118, 1.004,
     0.204, 0.192, 0.299, 1.885,0.264, 0.230, 0.250, 0.061, 0.953,
     0.138, 0.313, 0.174, 1.768, 1.369, 0.130,1.113, 0.320, 0.190,
     0.818, 1.461, 0.149, 0.291, 0.225, 1.622, 0.185, 0.198,0.360,
     0.387, 2.338, 1.713, 0.368, 1.573, 0.309, 0.232, 0.347, 0.325,
     1.861,0.258, 0.258, 1.625, 0.291, 1.169, 0.210, 0.241, 0.112,
     0.183, 0.258, 0.357,1.176, 0.111, 0.978, 0.279, 1.742, 0.184,
     0.230, 0.275, 2.183, 2.264, 1.405,0.408, 0.126, 0.263, 0.162,
     0.902, 1.516, 0.293, 0.198, 0.118, 0.305, 0.031,0.192, 0.151,
     0.182, 0.909, 0.379, 1.010, 0.167, 0.929, 0.083, 0.179, 1.567,
     1.241, 0.077, 0.166, 1.271, 0.100, 1.229, 0.152, 1.374, 0.157,
     1.003, 0.084,0.171, 0.953, 0.192, 0.967, 1.300, 0.122, 1.036,
     0.200, 0.070, 0.998, 0.176,0.673, 0.839, 0.867, 0.985, 0.096,
     0.238, 0.933, 1.231, 0.162, 0.044, 0.175,0.132, 1.166, 0.144,
     0.180, 0.945, 0.180, 0.152, 0.108, 0.923 ,0.192, 0.895,0.176,
     0.191, 1.161)
#########################################################################
The Acidity Data
#########################################################################
y1=c(2.928524,3.910021,3.732896,3.688879,3.822098 ,3.735286,4.143135,
 4.276666, 3.931826 ,4.077537 ,4.779123 ,4.234107, 4.276666, 4.543295,
 6.467388, 4.127134 ,3.977811, 4.264087, 4.007333, 3.921973, 5.384495,
 4.912655 ,4.046554 ,4.043051, 4.406719 ,4.505350, 3.931826 ,6.752270,
```

```
 6.928538, 5.994460 ,4.248495 ,4.060443 ,4.727388 , 6.047372 ,4.082609,
 4.244200, 4.890349 ,4.416428, 5.743003, 4.127134 , 5.489764 ,4.778283,
 5.249652 , 4.855929 ,4.128746,4.442651 ,4.025352 ,4.290459, 4.593098 ,
 4.652054 , 4.178992 , 4.382027 ,5.569489, 5.049856 ,4.188138, 6.629363,
 4.647271, 4.784989, 4.348987, 5.361292 ,4.574711, 4.442651, 6.120297,
 4.060443 ,4.143135, 4.510860, 6.049733, 4.510860, 4.406719 ,6.343880,
 4.430817 ,5.929589 ,5.973301, 4.481872, 4.301359, 6.452680, 4.204693,
 4.143135 ,6.603944, 4.644391, 5.863631, 4.025352, 5.717028, 5.308268,
 6.267201 ,4.060443, 5.017280, 4.510860, 5.834811, 4.330733, 4.007333,
 6.806829 ,5.257495 ,4.624973 ,4.781641, 4.099332 , 7.044382,3.914021,
 4.330733 ,4.016383 , 5.572154 ,4.043051, 4.843399 ,4.110874,4.454347,
 4.356709 ,6.154858, 6.284321, 6.978214 ,4.301359, 5.929855 ,4.465908,
 6.035481 ,6.726473 ,7.105130, 6.014937 ,4.882802 ,7.032095 ,4.518522,
 6.476665 ,6.125558 ,4.189655, 5.323498, 4.938065 ,6.313548 ,5.853925,
 6.278146 ,7.020191 ,5.023881, 4.262680, 6.725634 ,6.489205 ,5.743003,
 6.739337 ,6.466145, 6.855409, 5.120983, 5.913773 ,6.516932 ,4.058717,
 6.213608 ,6.554218 ,6.155707 ,4.314818 ,6.662494 ,6.749931, 6.100319,
 4.112512 ,6.946014 ,4.131961 ,6.234411 ,6.595781 ,6.683861 ,6.957973,
 4.497585 )
##########################################################################
The SLC Data
##########################################################################
x=c(.467,.430,.192,.192,.293,.160,.164,.126,.328,.202,.282,.328,.247,.132,
  .138,.224,.512,.221,.252,.193,.263,.186,.346,.219,.177,.349,.272,.245,
  .213,.197,.229,.245,.210,.281,.175,.273,.439,.471,.451,.237,.313,.136,
  .245,.391,.349,.158,.252,.416,.232,.183,.254,.195,.141,.151,.073,.300,
  .231,.075,.208,.267,.187,.244,.245,.231,.167,.337,.251,.209,.181,.411,
  .191,.288,.280,.119,.394,.443,.423,.534,.393,.273,.149,.225,.159,.170,
  .329,.183,.262,.250,.179,.329,.253,.270,.310,.321,.333,.284,.380,.222,
  .178,.265,.289,.199,.309,.279,.194,.203,.139,.162,.251,.619,.343,.155,
  .340,.332,.412,.218,.304,.261,.206,.231,.182,.267,.198,.191,.258,.179,
  .197,.188,.202,.150,.201,.255,.293,.255,.189,.414,.292,.253,.168,.295,
  .215,.213,.267,.216,.264,.138,.239,.288,.311,.414,.462,.361,.623,.199,
  .215,.321,.273,.259,.206,.376,.228,.155,.186,.097,.179,.174,.386,.393,
  .198,.243,.326,.250,.590,.461,.361,.321,.236,.139,.316,.313,.263,.180,
  .184,.354,.264,.269,.171,.359,.338,.163)
##########################################################################
```

3.9 APPENDIX

Here, we state and prove a Proposition under certain regularity conditions followed by a proof of the consistency Theorem stated in section 3.3.

Proposition: For each m, assume that the parameter space Θ_m can be embedded in a compact subset of R^p, the class $\mathcal{F}_m$ is identifiable for $\boldsymbol{\theta}_m \in \Theta_m$, and for almost every x, the component densities $f(x|\boldsymbol{\phi}_i)$ are continuous in $\boldsymbol{\phi}_i$, for each $i = 1, \cdots, m$. Assume also that there exists a function $g(x)$ (independent of m) such that $|\frac{\partial}{\partial x} f_{\boldsymbol{t}_m}(x)| \leq g(x)$ and $\int_{-\infty}^{\infty} g(x)dx < \infty$.Furthermore, we assume that the function $L(\mathbf{t}_m, F)$ (see equation (3.2.5)) is continuous in $\mathbf{t}_m \in \Theta_m$. Then the following hold for the functional $T_m^{L_2E}$ defined in (3.2.6).

(i) For every $F \in \mathcal{F}$, there exists $T_m^{L_2E}(F) \in \Theta_m$ satisfying (3.2.6).

(ii) If $T_m^{L_2E}(F)$ is unique, then the functional $T_m^{L_2E}$ is continuous at F under the supremum norm defined by $\sup_x |F(x) - G(x)|$, for distributions F and $G \in \mathcal{F}$.

(iii) $T_m^{L_2E}(F_{\boldsymbol{\theta}_m}) = \boldsymbol{\theta}_m$ uniquely for every $\boldsymbol{\theta}_m \in \Theta_m$.

Proof: Part (i) directly follows from our assumption that the function $L(\mathbf{t}_m, F)$ is continuous in $\mathbf{t}_m \in \Theta_m$ and that Θ_m can be embedded in a compact subset of R^p. Part (iii) follows from our identifiability assumption. Therefore, it only remains to prove the assertion in (ii).

For this, let us suppose that a sequence $\{F_n\}$ and F belong to $\mathcal{F}$ such that $\sup_x |F_n(x) - F(x)| \to 0$ as $n \to \infty$. We wish to show that $T_m(F_n) \to T_m(F)$ as $n \to \infty$. Before we show this, let us examine the large sample behavior of the difference $L(\boldsymbol{t}_m, F_n) - L(\boldsymbol{t}_m, F)$, as $n \to \infty$. By (3.2.5) and integration by parts we have

$$\begin{aligned} L(\boldsymbol{t}_m, F_n) - L(\boldsymbol{t}_m, F) &= -2\int_{-\infty}^{\infty} f_{\boldsymbol{t}_m}(x) d[F_n(x) - F(x)] \\ &= 2\int_{-\infty}^{\infty} [F_n(x) - F(x)] \, \frac{\partial}{\partial x} f_{\boldsymbol{t}_m}(x) dx. \end{aligned}$$

Therefore,

$$|L(\boldsymbol{t}_m; F_n) - L(\boldsymbol{t}_m; F)| \quad \leq \quad 2 \sup_x |F_n(x) - F(x)| \int_{-\infty}^{\infty} |\frac{\partial}{\partial x} f_{\boldsymbol{t}_m}(x)| dx.$$

By the assumption that $|\frac{\partial}{\partial x} f_{\boldsymbol{t}_m}(x)| \leq g(x)$ and $\int_{-\infty}^{\infty} g(x)dx < \infty$, we have

$$\sup_{\boldsymbol{t}_m} |L(\boldsymbol{t}_m, F_n) - L(\boldsymbol{t}_m, F)| \to 0 \ \text{ as } \ n \to \infty. \tag{3.9.1}$$

Since $L(\boldsymbol{t}_m, F)$ is assumed to be continuous in $\boldsymbol{t}_m$ and also Θ_m is assumed to be embedded in a compact subset of R^p, we have that

$$|\min_{\boldsymbol{t}_m} L(\boldsymbol{t}_m, F_n) - \min_{\boldsymbol{t}_m} L(\boldsymbol{t}_m, F)| \to 0 \ \text{ as } \ n \to \infty. \tag{3.9.2}$$

Or, equivalently

$$|L(T_m^{L_2E}(F_n), F_n) - L(T_m^{L_2E}(F), F)| \to 0. \tag{3.9.3}$$

Also, by (3.9.1)

$$|L(T_m^{L_2E}(F_n), F_n) - L(T_m^{L_2E}(F_n), F)| \to 0. \tag{3.9.4}$$

Therefore, by (3.9.3), (3.9.4) and the triangle inequality

$$|L(T_m^{L_2E}(F_n), F) - L(T_m^{L_2E}(F), F)| \to 0. \tag{3.9.5}$$

Now, using standard subsequence arguments, compactness and the continuity of $L(\boldsymbol{t}_m, F)$ in $\boldsymbol{t}_m$ (see Theorem 1 of Beran(1977), after display (2.4)) it is possible to show that

$$T_m^{L_2E}(F_n) \to T_m^{L_2E}(F) \ \text{ as } \ n \to \infty.$$

□

Proof of the Theorem. Recall from (3.2.9) that

$$\hat{m}_n^{L_2E} = \min\{m : L(\hat{\boldsymbol{\theta}}_{n,m}^{L_2E}, \hat{F}_n) \leq L(\hat{\boldsymbol{\theta}}_{n,m+1}^{L_2E}, \hat{F}_n) + \alpha_{n,m}\}. \tag{3.9.6}$$

Clearly, by (3.9.3)

$$L(T_m^{L_2E}(\hat{F}_n), \hat{F}_n) - L(T_{m+1}^{L_2E}(\hat{F}_n, \hat{F}_n) \to L(T_m^{L_2E}(F_0), F_0) - L(T_{m+1}^{L_2E}(F_0), F_0) = d_m. \tag{3.9.7}$$

Note from (3.2.5) that

$$\begin{aligned} d_m &= L(T_m^{L_2E}(F_0), F_0) - L(T_{m+1}^{L_2E}(F_0), F_0) \\ &= \int_{-\infty}^{\infty} f^2_{T_m^{L_2E}(F_0)}(x) - 2\int_{-\infty}^{\infty} f_{T_m^{L_2E}(F_0)}(x) dF_0(x) \\ &- \int_{-\infty}^{\infty} f^2_{T_{m+1}^{L_2E}(F_0)}(x) - 2\int_{-\infty}^{\infty} f_{T_{m+1}^{L_2E}(F_0)}(x) dF_0(x). \end{aligned}$$

Since F_0 has an associated density f_0, we have

$$\begin{aligned} d_m &= [\int_{-\infty}^{\infty} f^2_{T_m^{L_2E}(F_0)}(x) - 2\int_{-\infty}^{\infty} f_{T_m^{L_2E}(F_0)}(x) f_0(x) d(x) + \int f_0^2(x)dx] \\ &- [\int_{-\infty}^{\infty} f^2_{T_{m+1}^{L_2E}(F_0)}(x) - 2\int_{-\infty}^{\infty} f_{T_{m+1}^{L_2E}(F_0)}(x) f_0(x) d(x) + \int f_0^2(x)dx] \\ &= \int_{-\infty}^{\infty} |f_{T_m^{L_2E}(F_0)}(x) - f_0(x)|^2 dx - \int_{-\infty}^{\infty} |f_{T_{m+1}^{L_2E}(F_0)}(x) - f_0(x)|^2 dx. \end{aligned}$$

Note that the true mixture complexity

$$\begin{aligned} m_0 &= min\{m : L(T_m^{L_2E}(F_0), F_0) - L(T_{m+1}^{L_2E}(F_0), F_0) \le 0\} \\ &= min\{m : \int_{-\infty}^{\infty} |f_{T_m^{L_2E}(F_0)}(x) - f_0(x)|^2 dx - \int_{-\infty}^{\infty} |f_{T_{m+1}^{L_2E}(F_0)}(x) - f_0(x)|^2 dx \le 0\} \\ &= min\{m : d_m \le 0\}, \end{aligned}$$

where d_m is defined as in (3.9.7). Let $m \ge m_0$. Since $F_0 \in \mathcal{F}_{m_0} \subseteq \mathcal{F}_j, j \ge m_0$ we have that for each $j \ge m_0$

$$L(T_j^{L_2E}(\hat{F}_n), \hat{F}_n) \le L(T_{m_0}^{L_2E}(F_0), \hat{F}_n). \tag{3.9.8}$$

Therefore, for $m \ge m_0$, the expression on the right side of (3.9.7)

$$\begin{aligned} 0 &\le L(T_m^{L_2E}(\hat{F}_n), \hat{F}_n) - L(T_{m+1}^{L_2E}(\hat{F}_n), \hat{F}_n) \\ &= L(T_m^{L_2E}(\hat{F}_n), \hat{F}_n) + \int f_0^2(x)dx - \{L(T_{m+1}^{L_2E}(\hat{F}_n), \hat{F}_n) + \int f_0^2(x)dx\}. \end{aligned}$$

Note from (3.9.8) and that $T_{m_0}^{L_2E}(F_0) = \boldsymbol{\theta}_{m_0}$ (see Proposition (iii)), which implies $f_{T_{m_0}^{L_2E}(F_0)} = f_{\boldsymbol{\theta}_{m_0}} = f_0$, we have

$$L(T_m^{L_2E}(\hat{F}_n), \hat{F}_n) + \int f_0^2(x)dx \le L(T_{m_0}^{L_2E}(F_0), \hat{F}_n) + \int f_0^2(x)dx$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} f^2_{T_{m_0}^{L_2E}(F_0)}(x)dx - 2/n \sum_{i=0}^{n} f_{T_{m_0}^{L_2E}(F_0)}(X_i) + \int f_0^2(x)dx \\
&\to \int_{-\infty}^{\infty} f^2_{\boldsymbol{\theta}_{m_0}}(x)dx - 2\int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_{m_0}}(x) f_{\boldsymbol{\theta}_{m_0}}(x)dx + \int f^2_{\boldsymbol{\theta}_{m_0}}(x)dx \\
&= 0,
\end{aligned}
$$

where the convergence is as $n \to \infty$. Similarly, $L(T_{m_0}^{L_2E}(F_0), \hat{F}_n) + \int f_0^2(x)dx \to 0$. Therefore, we have that $d_m = 0$ for $m \geq m_0$.

Let $m < m_0$. Then by the definition of m_0, we have $F_0 \in \mathcal{F}_{m_0}$ but $F_0 \notin \mathcal{F}_m$ for $m < m_0$. Now, we want to show that $d_m > 0$ for all $m < m_0$. Suppose on the contrary that $d_m = 0$ for some $m < m_0$. Then, by (3.9.7), $L(T_m^{L_2E}(F_0), F_0) = L(T_{m+1}^{L_2E}(F_0), F_0)$ for $m < m_0$. Then,

$$
L(T_m^{L_2E}(F_0), F_0) \leq L(\boldsymbol{t}_{m+1}, F_0), \quad \text{for all} \quad \boldsymbol{t}_m \in \Theta_{m+1}
$$

which implies

$$
\int_{-\infty}^{\infty} |f_{T_m^{L_2E}(F_0)}(x) - f_0(x)|^2 dx \leq \int_{-\infty}^{\infty} |f_{\boldsymbol{t}_{m+1}}(x) - f_0(x)|^2 dx. \tag{3.9.9}
$$

For an arbitrary $\epsilon \in (0,1)$ and $\boldsymbol{\phi}$, let $f_{\mathbf{t}_{m+1}}(x) = (1-\epsilon) f_{T_m^{L_2E}(F_0)}(x) + \epsilon f(x|\boldsymbol{\phi})$. Then, the associated distribution $F_{\boldsymbol{t}_{m+1}} \in \mathcal{F}_{m+1}$ and from (3.9.9)

$$
\int_{-\infty}^{\infty} |f_{T_m^{L_2E}(F_0)}(x) - f_0(x)|^2 - \int_{-\infty}^{\infty} |(1-\epsilon) f_{T_m^{L_2E}(F_0)}(x) + \epsilon f(x|\boldsymbol{\phi}) - f_0(x)|^2 \leq 0.
$$

Now, using the identity $x^2 - y^2 = (x-y)(x+y)$ and algebraic calculations we get

$$
\epsilon \int_{-\infty}^{\infty} [f_{T_m^{L_2E}(F_0)}(x) - f(x|\boldsymbol{\phi})] \left\{ 2[f_{T_m^{L_2E}(F_0)}(x) - f_0(x)] + \epsilon [f(x|\boldsymbol{\phi}) - f_{T_m^{L_2E}(F_0)}(x)] \right\} dx \leq 0,
$$

which implies

$$
2\epsilon \int_{-\infty}^{\infty} [f_{T_m^{L_2E}(F_0)}(x) - f(x|\boldsymbol{\phi})][f_{T_m^{L_2E}(F_0)}(x) - f_0(x)]dx \leq \epsilon^2 \int_{-\infty}^{\infty} [f_{T_m^{L_2E}(F_0)}(x) - f(x|\boldsymbol{\phi})]^2 dx. \tag{3.9.10}
$$

Dividing both sides of (3.9.10) by ϵ, letting $\epsilon \to 0$, and applying Fatou's lemma, we get

$$
\int_{-\infty}^{\infty} [f_{T_m^{L_2E}(F_0)}(x) - f(x|\boldsymbol{\phi})][f_{T_m(F_0)}(x) - f_0(x)] \leq 0,
$$

which implies

$$\int_{-\infty}^{\infty} f_{T_m^{L_2E}(F_0)}(x)[f_{T_m^{L_2E}(F_0)}(x) - f_0(x)]dx \leq \int_{-\infty}^{\infty} f(x|\boldsymbol{\phi})[f_{T_m^{L_2E}(F_0)}(x) - f_0(x)]dx. \quad (3.9.11)$$

Since $f_0 \in \mathcal{F}_{m_0}$, we can write $f_0(x) = \sum_{i=1}^{m_0} \pi_i^0 f(x|\boldsymbol{\phi}_i^0)$ and (3.9.11) holds for each $\boldsymbol{\phi} = \boldsymbol{\phi}_i^0$, $i = 1, \ldots, m_0$. Since $\sum_{i=1}^{m_0} \pi_i^0 = 1$, from (3.9.11)

$$\int_{-\infty}^{\infty} f_{T_m^{L_2E}(F_0)}(x)[f_{T_m^{L_2E}(F_0)}(x) - f_0(x)]dx \leq \int_{-\infty}^{\infty} f_0(x)[f_{T_m^{L_2E}(F_0)}(x) - f_0(x)]dx,$$

which implies that $\int_{-\infty}^{\infty} [f_{T_m^{L_2E}(F_0)}(x) - f_0(x)]^2 = 0$. This contradicts the assumption that $F_0 \notin \mathcal{F}_m$ for $m < m_0$. Therefore, $d_m > 0$ for $m < m_0$. Hence the Theorem follows from the above arguments. □

3.10 References

[1] Aitkin, M., and Wilson, G. T. (1980), "Mixture Models, Outliers, and the EM Algorithm," *Technometrics*, 22, 325-331.

[2] Basu, A., Harris, I.R., Hjort, H.L., and Jones, M.C. (1998), "Robust and efficient estimation by minimizing a density power divergence," *Biometrika*, 85, 549 - 560.

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

[4] Bogardus, C., Lillioja, S., Nyomba, B. L., Zurlo, F., Swinburn, B., Puente, A. E. -D., Knowler, W. C., Ravussin, E., Mott, D. M., and Bennett, P. H. (1989), "Distribution of in vivo insulin action in Pima-Indians as mixture of 3 normal-distributions", *Diabetes* 38, 1423-1432.

[5] Böhning, D. (1999), *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*, New York: Chapman & Hall/CRC.

[6] Böhning, D., and Seidel, W. (2003), "Editorial: Recent Developments in Mixture Models," *Computational Statistics and Data Analysis*, 41, 349-257.

[7] Chen, J. and Kalbfleisch, J. D. (1996), "Penalized minimum distance estimates in finite mixture models," *Canadian Journal of Statistics*, 24, 167-175.

[8] Chen, J.and Khalili, A. (2006), "Order selection in finite mixture models," *Technical Report*, Department of Statistics and Actuarial Science University of Waterloo, Canada.

[9] Chen, J.and Khalili, A. (2008), "Order selection in finite mixture models with a non-smooth penalty", *Journal of the American Statistical Association*, 103, 1674-1683.

[10] Crawford, S.L., M.H. Degroot, J.B. Kadane, AND M.J. Small. (1992), "Modeling lake-chemistry distributions: Approximate Bayesian methods for estimating a finite-mixture model," *Technometrics* 34:441453.

[11] Cutler, A., and Cordero-Braňa, O. I. (1996), "Minimum Hellinger distance estimation for finite mixture models," *Journal of the American Statistical Association* 91, 1716-1723.

[12] Dacunha-Castelle, D. and Gassiat, E. (1997), "The estimation of the order of a mixture model," *Bernoulli*, 3, 279-299.

[13] Dacunha-Castelle, D. and Gassiat, E. (1999), "Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes," *The Annals of Statistics*, 27, 1178-1209.

[14] Dellaportas, P., Karlis, D and Xekalaki, E. (1997), " Bayesian analysis of finite Poisson mixtures," Technical Report, Department of Statistics, Athens University of Economics and Business.

[15] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum-Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

[16] DeVeaux, R. D., and Krieger, A. M. (1990), "Robust Estimation of a Normal Mixture," *Statistics and Probability Letters*, 10, 1-7.

[17] Donoho, D. L., and Liu, R. C. (1988), "The 'Automatic' Robustness of Minimum Distance Functionals," *The Annals of Statistics*, 16, 552-586.

[18] Dudley, C.R.K., Giuffra, L.A., Raine, A.E.G., Reeders, S.T., (1991), " Assessing the role of APNH, a gene encoding for a human amiloride-sensitive Na+/H+ antiporter, on the interindividual variation in red cell Na+/Li+ countertransport,"' J. Am. Soc. Nephrol., 2, 937943.

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

[20] Everitt, B. S. and Hand, D. J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

[21] Fan,J. and Li,R. (2001),"Variable selection via non-concave penalized likelihood and its oracle properties, " *Journal of the American Statistical Association*, 96, 1348-1360.

[22] Fujisawa, H., Eguchi, S. (2006), "Robust estimation in the normal mixture model," *Journal of Statistical Planning and Inference*, 136, 3989 4011

[23] Henna, J. (1985), "On estimating of the number of constituents of a finite mixture of continuous distributions," *Annals of the Institute of Statistical Mathematics*, 37, 235-240.

[24] Ishwaran, H., James, L. F., and Sun, J. (2001),"Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," *Journal of the American Statistical Association*, 96, 1316-1332.

[25] James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent Estimation of Mixture Complexity," *The Annals of Statistics*, 29, 1281-1236.

[26] Karlis, D. and Xekalaki, E. (1999), "On testing for the number of components in a mixed Poisson model," *Annals of Institute of Statistical Mathematics*, 51, 149-162.

[27] Keribin, C. (2000), "Consistent estimation of the order of mixture models," *Sankhyā, Ser. A* 62, 49-62.

[28] Leroux, B. G. (1992), "Consistent estimation of a mixing distribution," *The Annals of Statistics*, 20, 1350-1360.

[29] Lindsay, B. G., (1995),"Mixture Models: Theory, Geometry, and Applications," NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics, Hayward.

[30] Markatou, M. (2000), "Mixture models, robustness and the weighted likelihood methodology", *Biometrics*, 56, 483-486.

[31] Markatou, M. (2001), "A closer look at the weighted likelihood in the context of mixtures", *Probability and Statistical Models with Applications*, Charalambides, C.A., Koutras, M.V. and Balakrishnan, N. (eds), Chapman and Hall/CRC, 4447-467.

[32] Marron, J. S. and Wand, M. P. (1992), "Exact mean integrated squared error," *The Annals of Statistics*, 20, 712-736.

[33] McCann, M. and Sarkar, S. (2000), "Minimum Negative Exponential Disparity Estimation of Mixture Proportions," *Journal of Statistical Planning and Inference* 87, 187-197.

[34] McGrory, C.A. and Titterington, D. M. (2007),"Variational approximations in Bayesian model selection for finite mixture distributions", *Computational Statistics and Data Analysis*, 51, 5352-5367.

[35] McLachlan, G. J. (1987), "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *Journal of the Royal Statistical Society, Ser. C (Applied Statistics)* 36, 318-324.

[36] McLachlan, G. J. and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

[37] McLachlan, G. J., McLaren, C. E., and Matthews, D. (1995), "An algorithm for the likelihood ratio test of one versus two components in a mixture model fitted to grouped and truncated data," *Communications in Statistics – Simulation and Computation*, 24, 965-985.

[38] McLachlan, G. J. and Peel, D. (1997), "On a resampling approach to choosing the number of components in normal mixture models," in *Computing Science and Statistics*, Vol. 28, eds. L. Billard and N. I. Fisher (Eds.). Fairfax Station, Virginia: Interface Foundation of North America, pp. 260-266.

[39] McLachlan, G. J. and Peel, D. (1997b), Contribution to the discussion of paper by S. Richardson and P. J. Green, *Journal of the Royal Statistical Society, Ser. B*, 59, 779-780.

[40] McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

[41] McLaren, C. E. (1996), "Mixture models in haematology: a series of case studies," *Statistical Methods in Medical Research*, 5, 129-153.

[42] McLaren, C. E., Wagstaff, M., Brittenham, G. M., and Jacobs, A. (1991),"Detection of Two Component Mixtures of Lognormal Distributions in Grouped Doubly-truncated Data: Analysis of Red Blood Cell Volume Distributions," *Biometrics*, 47, 607-622.

[43] Pearson (1894), Contributions to the mathematical theory of evolution, Phil. Trans. Royal Soc., 185A , 71-110.

[44] Priebe, C. E. and Marchette, D. J. (2000), "Alternating kernel and mixture density estimates," *Computational Statistics and Data Analysis*, 35, 43-65.

[45] Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)," *Journal of the Royal Statistical Society Ser. B*, 59, 731-792. Correction (1998). *Journal of the Royal Statistical Society Ser. B*, 60, 661.

[46] Roeder, K. (1994), " A graphical technique for determining the number of components in a mixture of normals," *Journal of the American Statistical Association*, 89, 487-495.

[47] Roeder, K. and Wasserman, L. (1997), "Practical Bayesian density estimation using mixtures of normals," *Journal of the American Statistical Association*, 92, 894-902.

[48] Scott, D. W. (1998),"On fitting and adapting of density estimates," *Computing Science and Statistics*, S. Weisberg, Ed., 30, 124 - 133.

[49] Scott, D.W. (1999),"Remarks on fitting and interpreting mixture models," *Computing Science and Statistics*, K. Berk and M. Pourahmadi, Eds., 31, 104-109.

[50] Scott, D. W. (2001),"Parametric statistical modeling by minimum integrated square error, " *Technometrics*, 43, 274-285.

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

[52] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

[53] Wand, M. P. and Jones, M. C. (1995), "Kernel Smoothing " , London Chapman and Hall.

[54] Windham, M. P. and Cutler, A. (1994), "Mixture Analysis with Noisy Data," in *New Approaches in Classification and Data Analysis*, eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Berlin: Springer-Verlag.

[55] Woo, Mi-Ja and Sriram, T. N. (2006), "Robust estimation of mixture complexity", *Journal of American Statistical Association*, 101, 1475-1486.

[56] Woo, Mi-Ja and Sriram, T. N. (2007),"Robust estimation of mixture complexity for count data", *Computational Statistics and Data Analysis*, 51, 4379-4392.

[57] Woodward, W. A., Parr, W. C., Schucany, W. R., and Lindsay, H. (1984), "A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion," *Journal of the American Statistical Association*, 79, 590-598.

[58] Woodward, W. A., Whitney, P., and Eslinger, P. (1995), "Minimum Hellinger Distance Estimation of Mixture Proportions," *Journal of Statistical Planning and Inference*, 48, 303-319.

3.11 Tables and Figures

Table 3.1: Mixture Complexity Estimation results for three component normal mixture in (3.5.11)

	Estimated number of components							
	1	2	3	4	5	6	7	8
$n = 50$								
L_2E	89	11						
$MHDE$	80	20						
NKE	44	56						
MKE	44	53	3					
$R\&W$	22	7	***59**	10	1	1		
$Bootstrap$	0	96	4					
$Henna$	25	68	6	1				
$n = 250$								
L_2E	0	22	***73**	5				
$MHDE$	16	39	45					
NKE	0	99	1					
MKE	0	87	11	1	1			
$R\&W$	0	0	***60**	22	18			
$Bootstrap$	0	83	16	1				
$Henna$	0	90	10					
$n = 500$								
L_2E		8	***89**	3				
$MHDE$	0	35	***65**					
NKE	0	97	3					
MKE	0	58	34	6	2			
$R\&W$	0	0	22	12	61	5		
$Bootstrap$	0	74	20	6				
$Henna$	0	85	15					
$n = 1000$								
L_2E		2	***97**	1				
$MHDE$	0	26	***74**					
NKE	0	86	14					
MKE	0	18	***63**	10	2	3	1	3
$R\&W$	0	0	0	1	89	10		
$Bootstrap$	0	79	15	4	2			
$Henna$	0	78	15	5	1	0	1	

Table 3.2: Mixture Complexity Estimation results for the Marron and Wand densities 2 & 4-9

	Estimated number of components									
	1	2	3	4	5	6	7	8	9	10
Mixture2										
L_2E	0	46	***52**	2						
$MHDE$	0	78	22							
NKE	0	99	1							
MKE	0	99	1							
$R\&W$	3	96	1							
$Bootstrap$	0	89	11							
$Henna$	0	100								
Mixture4										
L_2E	0	***100**								
$MHDE$	0	***100**								
NKE	0	***99**	1							
MKE	0	***91**	6	3						
$R\&W$	0	0	0	0	75	18	5	2		
$Bootstrap$	0	***95**	5							
$Henna$	0	***88**	12							
Mixture5										
$MHDE$	0	***100**								
L_2E	0	***98**	2							
NKE	0	***96**	4							
MKE	0	***91**	8	1						
$R\&W$	0	***55**	45							
$Bootstrap$	0	***95**	5							
$Henna$	1	***97**	1	0	0	0	0	0	1	

Table 3.2 (continued)

	Estimated number of components									
	1	2	3	4	5	6	7	8	9	10
Mixture6										
L_2E	0	***100**								
$MHDE$	0	***100**								
NKE	0	***100**								
MKE	0	***98**	2							
$R\&W$	0	***100**								
$Bootstrap$	0	***95**	5							
$Henna$	0	***97**	3							
Mixture7										
L_2E	0	***99**	1							
$MHDE$	0	***100**								
NKE	0	***100**								
MKE	0	***96**	4							
$R\&W$	0	***100**								
$Bootstrap$	0	***93**	6	1						
$Henna$	0	***96**	4							
Mixture8										
L_2E	0	***97**	2	1						
$MHDE$	0	***97**	2	1						
NKE	0	***100**								
MKE	0	***97**	3							
$R\&W$	0	***80**	20							
$Bootstrap$	0	***93**	7							
$Henna$	0	***99**	1							
Mixture9										
L_2E	0	1	***80**	18	1					
$MHDE$	0	49	***51**							
NKE	0	94	6							
MKE	0	38	***59**	2						
$R\&W$	0	91	9							
$Bootstrap$	0	13	***75**	12						
$Henna$	0	82	18							

Table 3.3: Mixture Complexity Estimation results for $t(4)$ components

			t-overlap=.10					t-overlap=.03				
			Estimated number of components					Estimated number of components				
π	a		1	2	3	4	5	1	2	3	4	5
.25	1	L_2E	0	***99**	1			0	***88**	12		
		$MHDE$	0	***100**				0	***100**			
		MKE	33	***60**	7			2	23	75		
.25	$\sqrt{2}$	L_2E	1	***99**				0	***91**	9		
		$MHDE$	0	***92**	8			0	***100**			
		MKE	0	***74**	26			0	35	64	1	
.50	1	L_2E	0	***99**	1			0	***97**	3		
		$MHDE$	0	***95**	5			0	***100**			
		MKE	97	3				100				
.50	$\sqrt{2}$	L_2E	0	***100**				0	***99**	1		
		$MHDE$	0	***100**				0	***100**			
		MKE	94	4	2			99	1			
.75	$\sqrt{2}$	L_2E	1	***99**				0	***98**	1	1	
		$MHDE$	0	***100**				0	***100**			
		MKE	80	19	1			61	8	31		

Table 3.4: Mixture Complexity Estimation results for $t(2)$ components

			t-overlap=.10					t-overlap=.03				
			Estimated number of components					Estimated number of components				
π	a		1	2	3	4	5	1	2	3	4	5
.25	1	L_2E	0	***80**	20			0	***59**	37	4	
		$MHDE$	3	***97**				0	***98**	2		
		MKE	6	***91**	2	1		72	24	4		
.25	$\sqrt{2}$	L_2E	0	***96**	4			0	***71**	28	1	
		$MHDE$	0	***100**				0	***99**	1		
		MKE	8	***89**	1	1	1	79	21			
.50	1	L_2E	2	***87**	11			1	***70**	27	2	
		$MHDE$	89	11				0	***100**			
		MKE	9	***77**	14			59	40	1		
.50	$\sqrt{2}$	L_2E	2	***95**	3			0	***84**	16		
		$MHDE$	77	23				0	***100**			
		MKE	15	***76**	9			88	12			
.75	$\sqrt{2}$	L_2E	1	***91**	8			0	***81**	19		
		$MHDE$	63	35	2			0	***100**			
		MKE	9	***86**	2	3		75	24	1		

Table 3.5: Mixture Complexity Estimation results for Rescaled $t(3)$ components

			N-overlap=.10					N-overlap=.03				
			Estimated number of components					Estimated number of components				
π	a		1	2	3	4	5	1	2	3	4	5
.25	1	L_2E	0	***70**	26	4		0	***53**	46	1	
		$MHDE$	0	***97**	3			0	***100**			
		MKE	45	41	14			14	41	45		
.25	$\sqrt{2}$	L_2E	0	***61**	36	3		0	***53**	40	7	
		$MHDE$	0	***60**	40			0	***100**			
		MKE	10	***63**	20	2		14	44	38	4	
.50	1	L_2E	0	***86**	12	2		0	***60**	34	6	
		$MHDE$	0	***69**	31			0	***97**	3		
		MKE	99	1				97	3			
.50	$\sqrt{2}$	L_2E	0	***90**	9	1		0	***81**	16	3	
		$MHDE$	0	***91**	9			0	***96**	4		
		MKE	98	2				98	2			
.75	$\sqrt{2}$	L_2E	0	***87**	13			0	***65**	35		
		$MHDE$	1	***91**	8			0	***100**			
		MKE	80	18	1	1		66	17	17		

Table 3.6: Mixture Complexity Estimation results for Rescaled $t(4)$ components

			N-overlap=.10					N-overlap=.03				
			Estimated number of components					Estimated number of components				
π	a		1	2	3	4	5	1	2	3	4	5
.25	1	L_2E	0	***79**	19	2		0	***68**	28	4	
		$MHDE$	1	***99**				0	*100			
		MKE	35	34	31			0	26	74		
.25	$\sqrt{2}$	L_2E	0	***73**	27			0	***68**	30	2	
		$MHDE$	0	***88**	12			0	***100**			
		MKE	55	44	1			0	34	64	2	
.50	1	L_2E	0	***86**	13	1		0	***80**	14	6	
		$MHDE$	2	***98**				0	***99**	1		
		MKE	100					100				
.50	$\sqrt{2}$	L_2E	0	***96**	4			0	***90**	10		
		$MHDE$	1	***99**				0	***100**			
		MKE	100					99	1			
.75	$\sqrt{2}$	L_2E	0	***86**	13	1		0	***82**	17	1	
		$MHDE$	23	***77**				0	***100**			
		MKE	91	9				56	10	34		

Table 3.7: SLC Data Parameter Estimates

	π_1	π_2	π_3	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
$L_2E(m=3)$	0.082	0.794	0.124	0.187	0.238	0.418	0.01	0.07	0.053
$Beta(m=3)$	0.076	0.584	0.340	0.187	0.227	0.336	0.010	.062	0.108
$MHDE(m=2)$	0.695	0.305		0.222	0.352		0.060	0.106	
$MKE(m=2)$	0.754	0.246		0.225	0.378		0.060	0.102	
$MSCAD(m=3)$	0.75	0.22	0.03	0.221	0.372	0.564	0.057	0.057	0.057

Table 3.8: Acidity Data Parameter Estimates

	π_1	π_2	π_3	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
$L_2E(m=3)$	.085	.487	0.428	4.07	4.34	6.27	.053	.332	0.607
$DIC(m=2)$	.59	.41		4.32	6.23		.38	.55	
$MSCAD(m=3)$	.59	.14	.27	4.32	5.69	6.51	.37	.37	.37

Table 3.9: Enzyme Data Parameter Estimates

	π_1	π_2	π_3	π_4	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4
$L_2E(m=3)$	0.562	0.097	0.341		0.172	1.036	1.216		0.069	0.156	0.603	
DIC$(m=4)$	0.48	0.13	0.17	0.22	0.16	0.31	1.05	1.49	0.055	0.055	0.184	0.531

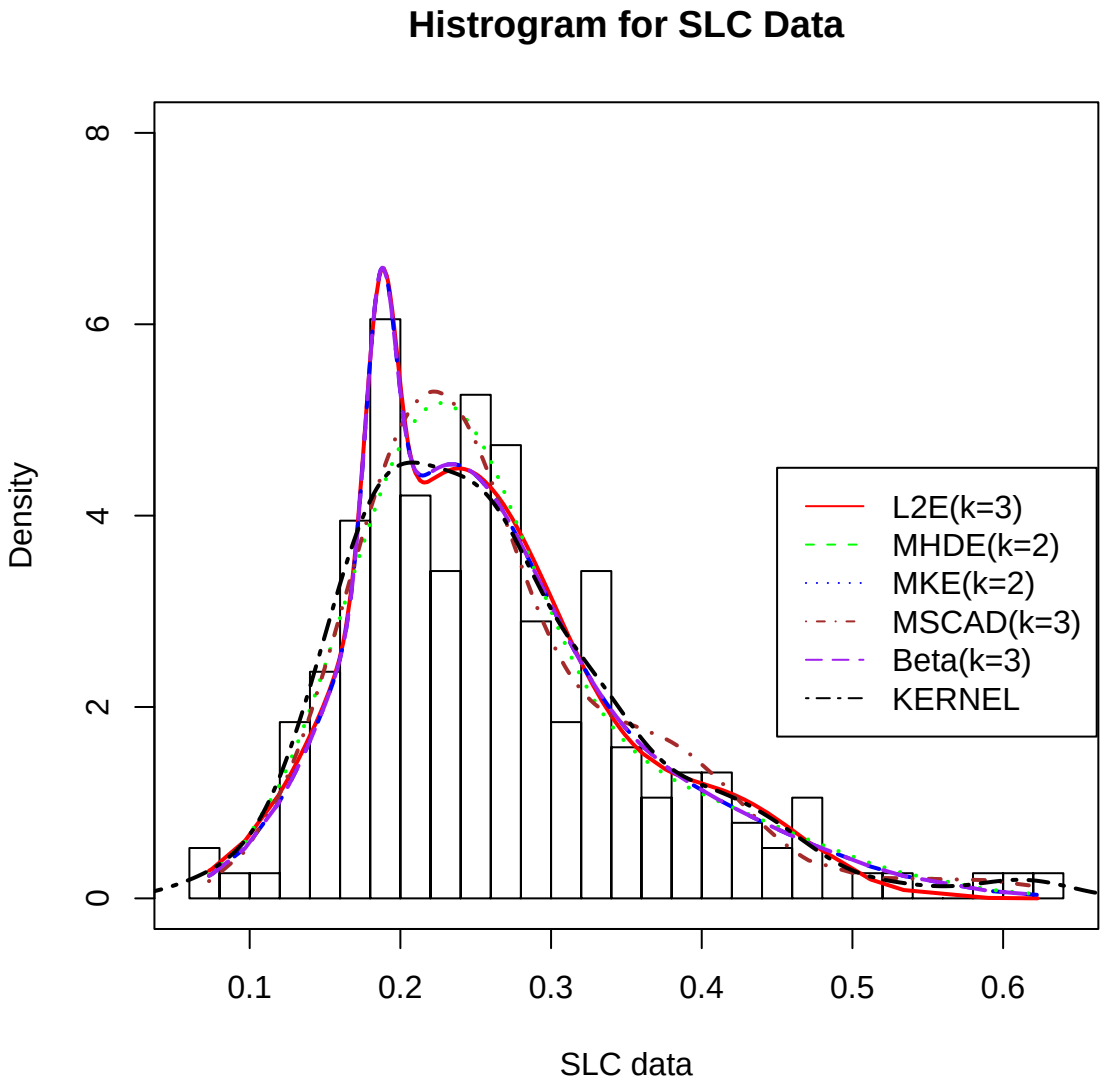

Figure 3.1: Fitted normal mixture for SLC data.

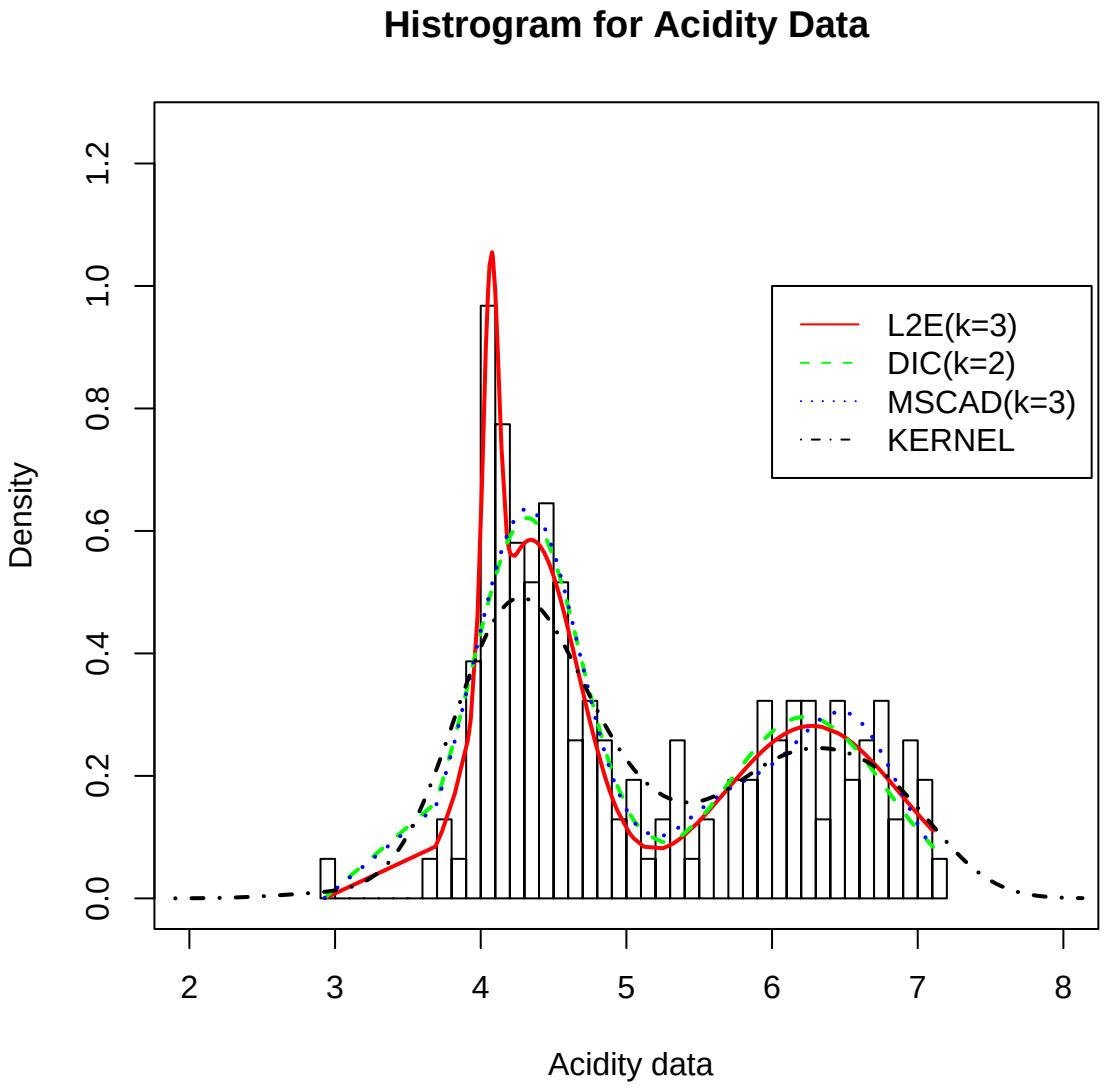

Figure 3.2: Fitted Normal mixture for Acidity Data

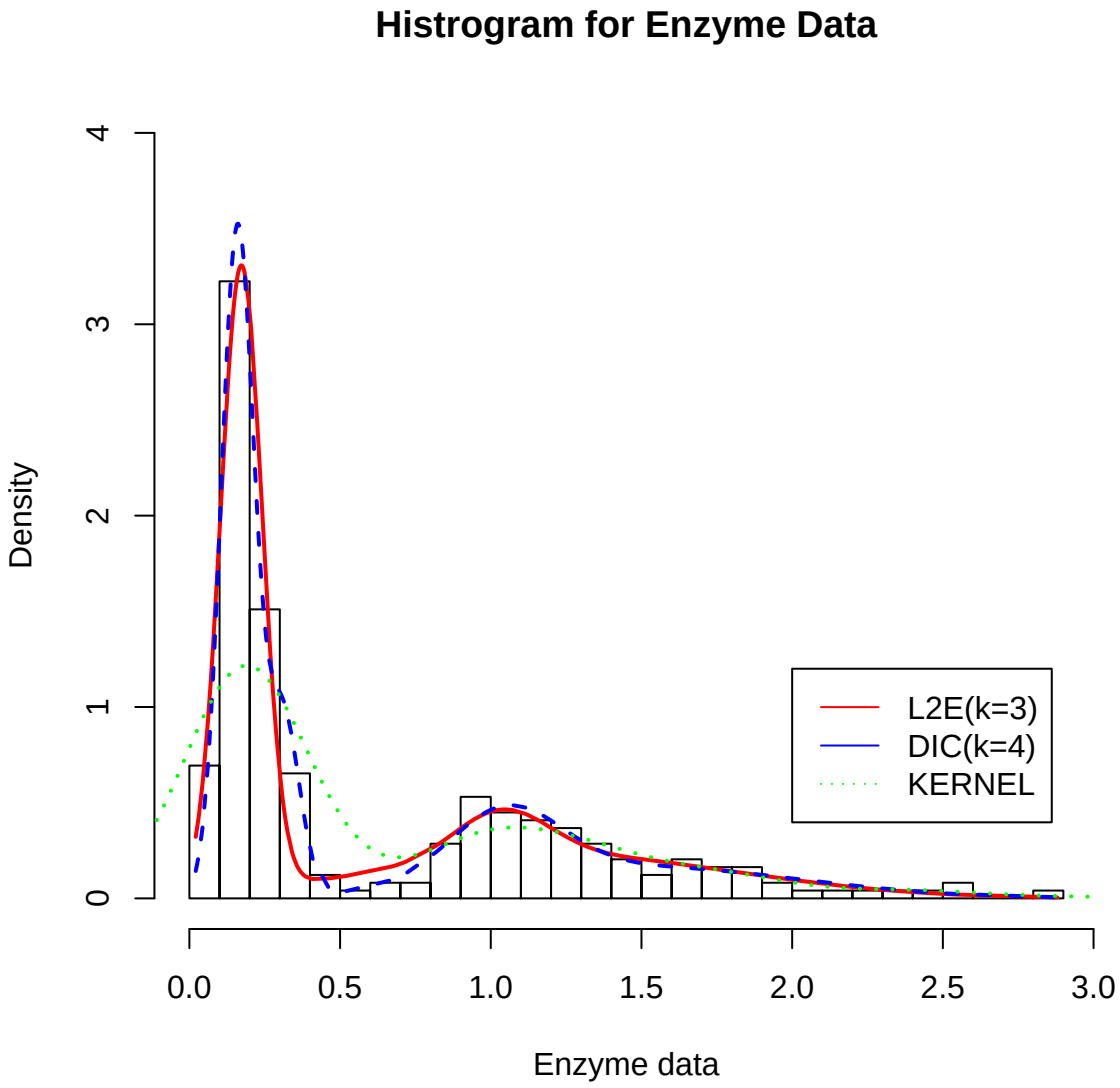

Figure 3.3: Fitted Normal mixture for Enzyme Data

Chapter 4

Conclusions

In this chapter we give brief concluding remarks regarding the robust estimation of mixture complexity based on the L_2E method developed in this thesis.

4.1 SUMMARY

In Chapter 2, we have introduced an estimator of the unknown number of components in finite mixtures for count data. This estimator is derived as a by-product of minimizing an information criterion constructed using the L_2 distance plus a penalty, which is a logarithmic function of the number of components. Our L_2E estimator is shown to be strongly consistent under certain regularity conditions. In comparison with other estimation methods available in the literature, our L_2E estimator has many distinctive features such as transparency, ease of use, efficiency in achieving computational speed and robustness against model misspecification. These features combined with competitive performance makes the L_2E estimator an attractive alternative to other existing methods in the literature.

In Chapter 3, we have developed a similar L_2E approach for the robust estimation of mixture complexity in the continuous case. Once again, we construct an estimator of mixture complexity by minimizing an information criterion based on L_2 distance plus a penalty function, which is similar to the well-known AIC criterion. Our L_2E estimator is once again shown to be strongly consistent under certain regularity conditions. When the data is continuous, our L_2E estimator has many distinct advantages over the other methods compared in this article. Firstly, the L_2E estimating function has a closed form expression given in (3.4.10) for normal mixtures. Secondly, parameter estimates are not sensitive to the choice of

initial values. Furthermore, our procedure avoids use of kernel density estimators and choice associated bandwidth, and continues to be robust against model misspecification.

Overall, we have illustrated via a variety of statistical applications that our procedure offers an excellent addition to the practitioners toolbox.

4.2 Future Research

We propose to extend our L_2E to the multivariate case and also consider different penalty functions to improve performance. We also propose to investigate the robustness properties theoretically and study influence functions and breakdown points of the estimators involved.

Bibliography

[1] Aitkin, M., and Wilson, G. T. (1980), "Mixture Models, Outliers, and the EM Algorithm," *Technometrics*, 22, 325-331.

[2] Arley and Buch, (1950), *Introduction to the Theory ofProbability andStatistics*, New York: Wiley.

[3] Basu, A., Harris, I.R., Hjort, H.L., and Jones, M.C. (1998), "Robust and efficient estimation by minimizing a density power divergence," *Biometrika*, 85, 549 - 560.

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

[5] Blischke, W.R. (1962), "Moment estimatorsfor the parameters of a mixture of two binomial distributions," *The Annals of Math Statistics*, 33, 444-454.

[6] Bogardus, C.,Lillioja, S., Nyomba, B. L., Zurlo, F., Swinburn, B., Puente, A. E. -D., Knowler, W. C., Ravussin, E., Mott, D. M., and Bennett, P. H. (1989), "Distribution of in vivo insulin action in Pima-Indians as mixture of 3 normal-distributions", *Diabetes* 38, 1423-1432.

[7] Böhning, D.(1999), *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*, New York: Chapman & Hall/CRC.

[8] Böhning, D., and Seidel, W. (2003), "Editorial: Recent Developments in Mixture Models," *Computational Statistics and Data Analysis*, 41, 349-257.

[9] Bowman ,A.W., (1984), "An alternative method of cross-validation for the smoothing of density estimates, " *Biometrika*, 71, 353 - 360.

[10] Chen, J. and Kalbfleisch, J. D. (1996), "Penalized minimum distance estimates in finite mixture models," *Canadian Journal of Statistics*, 24, 167-175.

[11] Chen, J.and Khalili, A. (2006), "Order selection in finite mixture models," *Technical Report*, Department of Statistics and Actuarial Science University of Waterloo, Canada.

[12] Chen, J.and Khalili, A. (2008), "Order selection in finite mixture models with a non-smooth penalty", *Journal of the American Statistical Association*, 103, 1674-1683.

[13] Chen, J., Li, P., Tan and Xianming. (2007), "Inference for von Mises mixtures in mean direction and concentration parameters"'. *J. Syst. Sci. Math. Sci.* , 27, No. 1, 59-67

[14] Choi and BulGren . (1968), "An estimation procedurefor mixtures of distributions," *Royal Statist. Soc. Ser*, 30, 444-460.

[15] Clarke, B. R., and Heathcote, C. R. (1994), "Robust Estimation of k- Component Univariate Normal Mixtures," *Annals of the Institute of Statistical Mathematics*, 46, 83-93.

[16] Cohen, A.C. (1963), *Estimation in mixtures of discrete distributions*, in Proc. of the InternationalSymposium on Classical and Contagious Discrete Distributions, Pergamon Press, New York, 351-372.

[17] Cooper P. W.,(1967), Some topics on nonsupervised adaptive detection for multivariate normal distributions, in Computer and Information Sciences, 11, J. T. Tou, ed., Academic Press, New York,143-146.

[18] Crawford, S.L., M.H. Degroot, J.B. Kadane, AND M.J. Small. (1992), "Modeling lake-chemistry distributions: Approximate Bayesian methods for estimating a finite-mixture model," *Technometrics* 34:441453.

[19] Cutler, A., and Cordero-Braňa, O. I. (1996), "Minimum Hellinger distance estimation for finite mixture models," *Journal of the American Statistical Association* 91, 1716-1723.

[20] Dacunha-Castelle, D. and Gassiat, E. (1997), "The estimation of the order of a mixture model," *Bernoulli*, 3, 279-299.

[21] Dacunha-Castelle, D. and Gassiat, E. (1999), "Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes," *The Annals of Statistics*, 27, 1178-1209.

[22] Day , N.E. (1969), Estimating the components of a mixture of normal distributions, "'*Biometrika*, 56, 463-474.

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

[24] Dellaportas, P., Karlis, D and Xekalaki, E. (1997), " Bayesian analysis of finite Poisson mixtures," Technical Report, Department of Statistics, Athens University of Economics and Business.

[25] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum-Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

[26] DeVeaux, R. D., and Krieger, A. M. (1990), "Robust Estimation of a Normal Mixture," *Statistics and Probability Letters*, 10, 1-7.

[27] Devroye, L. P., and Györfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*, New York: Wiley.

[28] Dionne, G. Artis, M.,and Guillen, M., (1996), "Count data models for a credit scoring system," *Journal of Empirical Finance*, 3, 303-325.

[29] Donoho, D. L., and Liu, R. C. (1988), "The 'Automatic' Robustness of Minimum Distance Functionals," *The Annals of Statistics*, 16, 552-586.

[30] Dudley, C.R.K., Giuffra, L.A., Raine, A.E.G., Reeders, S.T., (1991), " Assessing the role of APNH, a gene encoding for a human amiloride-sensitive Na+/H+ antiporter, on the interindividual variation in red cell Na+/Li+ countertransport,"' J. Am. Soc. Nephrol., 2, 937943.

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

[32] Everitt, B. S. and Hand, D. J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

[33] Fan,J. and Li,R. (2001),"Variable selection via non-concave penalized likelihood and its oracle properties, " *Journal of the American Statistical Association*, 96, 1348-1360.

[34] Fruhwirth-Schnatter, S. (2006). "Finite Mixture and Markov Switching Models," *Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.*

[35] Fujisawa, H., Eguchi, S. (2006), "Robust estimation in the normal mixture model," *Journal of Statistical Planning and Inference*, 136, 3989 4011

[36] Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82,711-732.

[37] Greenwood, M., and Yule, G., (1920), "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents," *Journal of Royal Statistical Society, Ser. A* , 83, 255-279.

[38] Gumbel (1939), "La dissection d'une repartition,"*Annales de I'Universite de Lyon*, 3 39-51.

[39] Hasselblad,V. (1969),"Estimation of finite mixtures of distributions from the exponential family, " *Journal of the American Statistical Association*, 64, 1459-1471.

[40] Henna, J. (1985), "On estimating of the number of constituents of a finite mixture of continuous distributions," *Annals of the Institute of Statistical Mathematics*, 37, 235-240.

[41] Ishwaran, H., James, L. F., and Sun, J. (2001),"Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," *Journal of the American Statistical Association*, 96, 1316-1332.

[42] James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent Estimation of Mixture Complexity," *The Annals of Statistics*, 29, 1281-1236.

[43] John,S.(1970), On identifying the population of origin of each observation in a mixture of observationsfrom two normalpopulations," *Technometrics*, 12 , 553-563.

[44] Kabir, A.B.M, (1968), "On Estimation of parameter of a finite mixture of distributions," *Royal Statist. Soc*, 30, 472-482.

[45] Kaestner, R. (1999), "Health insurance, the quantity and quality of prenatal care, and infant health,", *Inquiry*, 36, 162-175.

[46] Karlis D., Xekalaki E. (1998) "Minimum Hellinger distance estimation for finite Poisson mixtures,". *Computational Statistics and Data Analysis* , 29, 81-103.

[47] Karlis, D. and Xekalaki, E. (1999), "On testing for the number of components in a mixed Poisson model," *Annals of Institute of Statistical Mathematics*, 51, 149-162.

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

[49] Keribin, C. (2000), "Consistent estimation of the order of mixture models," *Sankhyā, Ser. A* 62, 49-62.

[50] Leroux, B. G. (1992), "Consistent estimation of a mixing distribution," *The Annals of Statistics*, 20, 1350-1360.

[51] Lindsay, B. G., (1995),"Mixture Models: Theory, Geometry, and Applications," NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics, Hayward.

[52] Macdonald ,P.D.M. (1971), "An estimation procedure for mixtures of distribution", *J. Royal Statist. Soc. Ser.*, 33, 326-329.

[53] Markatou, M. (2000), "Mixture models, robustness and the weighted likelihood methodology", *Biometrics*, 56, 483-486.

[54] Markatou, M. (2001), "A closer look at the weighted likelihood in the context of mixtures", *Probability and Statistical Models with Applications*, Charalambides, C.A., Koutras, M.V. and Balakrishnan, N. (eds), Chapman and Hall/CRC, 4447-467.

[55] Markatou, M., Basu, A., and Lindsay, B. G. (1998), "Weighted likelihood estimating equations with a bootstrap root search", *Journal of the American Statistical Association*, 93, 740-750.

[56] Marron, J. S. and Wand, M. P. (1992), "Exact mean integrated squared error," *The Annals of Statistics*, 20, 712-736.

[57] McCann, M. and Sarkar, S. (2000), "Minimum Negative Exponential Disparity Estimation of Mixture Proportions," *Journal of Statistical Planning and Inference* 87, 187-197.

[58] McGrory, C.A. and Titterington, D. M. (2007),"Variational approximations in Bayesian model selection for finite mixture distributions", *Computational Statistics and Data Analysis*, 51, 5352-5367.

[59] McLachlan, G. J. (1987), "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *Journal of the Royal Statistical Society, Ser. C (Applied Statistics)* 36, 318-324.

[60] McLachlan, G. J. and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

[61] McLachlan, G. J., McLaren, C. E., and Matthews, D. (1995), "An algorithm for the likelihood ratio test of one versus two components in a mixture model fitted to grouped and truncated data," *Communications in Statistics – Simulation and Computation*, 24, 965-985.

[62] McLachlan, G. J. and Peel, D. (1997), "On a resampling approach to choosing the number of components in normal mixture models," in *Computing Science and Statistics*, Vol. 28, eds. L. Billard and N. I. Fisher (Eds.). Fairfax Station, Virginia: Interface Foundation of North America, pp. 260-266.

[63] McLachlan, G. J. and Peel, D. (1997b), Contribution to the discussion of paper by S. Richardson and P. J. Green, *Journal of the Royal Statistical Society, Ser. B*, 59, 779-780.

[64] McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

[65] McLaren, C. E.(1996), "Mixture models in haematology: a series of case studies," *Statistical Methods in Medical Research*, 5, 129-153.

[66] McLaren, C. E., Wagstaff, M., Brittenham, G. M., and Jacobs, A. (1991),"Detection of Two Component Mixtures of Lognormal Distributions in Grouped Doubly-truncated Data: Analysis of Red Blood Cell Volume Distributions," *Biometrics*, 47, 607-622.

[67] Munech, H. (1936), 'Probability distribution of protection test results", *Journal of the American Statistical Association*, 31, 677-689.

[68] Pauler, D. K., Escobar, M. D., Sweeney, J. A. and Greenhouse, J. (1996), "Mixture models for eye-tracking data: A case study," *Statistics in Medicine*, 15, 1365-1376.

[69] Pearson (1894), Contributions to the mathematical theory of evolution, Phil. Trans. Royal Soc., 185A , 71-110.

[70] Pearson (1915), On certain types of compound frequency distributions in which the components can be individually described by binomial series, "'*Biometrika*, 11 , 139-144.

[71] Poland, W. B., and Shachter, R. D. (1994), "Three approaches to probability model selection", *In Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference*, San Mateo, CA: Morgan Kaufmann, 478-483.

[72] Polllard(1934), On the relative stability of the median and the arithmetic mean, with particular reference to certain frequency distributions which can be dissected into normal distributions,, "'*Ann.Math.Statist*, 5 , 227-262.

[73] Priebe, C. E. and Marchette, D. J. (2000), "Alternating kernel and mixture density estimates," *Computational Statistics and Data Analysis*, 35, 43-65.

[74] Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)," *Journal of the Royal Statistical Society Ser. B*, 59, 731-792. Correction (1998). *Journal of the Royal Statistical Society Ser. B*, 60, 661.

[75] Roeder, K. (1994), " A graphical technique for determining the number of components in a mixture of normals," *Journal of the American Statistical Association*, 89, 487-495.

[76] Roeder, K. and Wasserman, L. (1997), "Practical Bayesian density estimation using mixtures of normals," *Journal of the American Statistical Association*, 92, 894-902.

[77] Scott, D. W. (1998),"On fitting and adapting of density estimates," *Computing Science and Statistics*, S. Weisberg, Ed., 30, 124 - 133.

[78] Scott, D.W. (1999),"Remarks on fitting and interpreting mixture models," *Computing Science and Statistics*, K. Berk and M. Pourahmadi, Eds., 31, 104-109.

[79] Scott, D. W. (2001),"Parametric statistical modeling by minimum integrated square error, " *Technometrics*, 43, 274-285.

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

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

[82] Stather, G. R. (1981), "Robust statistical inference using Hellinger distance methods," unpublished Ph.D. dissertation, LaTrobe University, Australia, Department of Mathematical Statistics.

[83] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

[84] Wand, M. P. and Jones, M. C. (1995), "Kernel Smoothing " , London Chapman and Hall.

[85] Windham, M. P. and Cutler, A. (1994), "Mixture Analysis with Noisy Data," in *New Approaches in Classification and Data Analysis*, eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Berlin: Springer-Verlag.

[86] Woo, Mi-Ja and Sriram, T. N. (2006), "Robust estimation of mixture complexity", *Journal of American Statistical Association*, 101, 1475-1486.

[87] Woo, Mi-Ja and Sriram, T. N. (2007),"Robust estimation of mixture complexity for count data", *Computational Statistics and Data Analysis*, 51, 4379-4392.

[88] Woodward, W. A., Parr, W. C., Schucany, W. R., and Lindsay, H. (1984), "A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion," *Journal of the American Statistical Association*, 79, 590-598.

[89] Woodward, W. A., Whitney, P., and Eslinger, P. (1995), "Minimum Hellinger Distance Estimation of Mixture Proportions," *Journal of Statistical Planning and Inference*, 48, 303-319.