2018-01-18

Inferences for Two-Component Mixture Models with Stochastic Dominance

Abedin, Tasnima

UNIVERSITY OF CALGARY

Inferences for Two-Component Mixture Models with Stochastic Dominance

by

Tasnima Abedin

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

January, 2018

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

Abstract

In this thesis, we studied a two-component nonparametric mixture model with a stochastic dominance constraint, which is a model that arises naturally from genetic studies. For this model, we proposed and studied nonparametric estimation based on cumulative distribution functions (c.d.f.s) and maximum likelihood estimation (MLE) through multinomial approximation. In order to incorporate the stochastic dominance constraint, we introduced a semiparametric model structure for which we proposed and investigated both MLE and minimum Hellinger distance estimation (MHDE). We also proposed a hypothesis testing to test the validity of the semiparametric model. For the proposed methods, we investigated their asymptotic properties such as consistency and asymptotic normality theoretically and through simulation studies. Our numerical studies demonstrated that (1) all the proposed estimation methods work well; (2) the semiparametric model structure incorporates nicely the stochastic dominance constraint and thus the MLE and MHDE based on it are superior in terms of efficiency than the two estimation techniques that do not use this model structure; (3) the MHDE is much more robust than the MLE. To demonstrate the use of these methods, we applied them to several real data including publicly available grain data (Smith et al., 1986) and malaria data (Vonatsou et al., 1998).

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Jingjing Wu, without whose support this journey would not be possible. Dr. Wu guided me throughout my graduate program with her valuable advice and feedback. Whenever I have any question or doubts about my research, she always gave me time and encouragement and steer me into the right direction. She not only inspired me to excel in academic arena but also gave me unbelievable support when I dealt with any personal problem. I can never thank her enough for her outstanding support as a mentor.

I would also like to thank my committee members Dr. Xuewen Lu, Dr. Alexander de Leon and Dr. Wenyuan Liao for their valuable comments and advice which helped me a lot to improve my thesis. I also want to thank Dr. Dan Nettleton for giving the motivation of the model we consider for our research and for his constructive and thoughtful advice. I am also grateful to the Department of Mathematics and Statistics for giving me the opportunity to pursue my degree here, where I have learned a great deal through its excellent graduate program. All the professors and the staff members were extremely helpful in every way.

Lastly my heartfelt gratitude to my family and friends, who supported me in every step of my graduate life. Specially my parents who always encouraged me for higher studies and my husband who took care of me and the family during this whole time. I would like to thank my friends Nana Xing, Arfan Afzal and Mili Roy for their support and friendship.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
MSE	mean squared error
MLE	maximum likelihood estimation/estimator/estimate
MHDE	minimum Hellinger distance estimation/estimator/estimate
WLLN	Weak Law of Large Number
CLT	Central Limit Theorem
DCT	Dominated Convergence Theorem
IF	influence function
MR	misclassification rate
OMR	optimal misclassification rate
p.d.f.	probability distribution function
c.d.f.	cumulative distribution function
i.i.d.	independent and identically distributed
$N(\mu,\sigma)$	normal distribution with mean μ and standard deviation σ
$Po(\lambda)$	Poisson distribution with mean λ
$U(a,b)$	uniform distribution over interval $[a,b]$
$LN(\mu,\sigma)$	log-normal distribution with mean μ and standard deviation σ
$I_A(x)$	the indicator function over set A
$\|\cdot\|$	L^2-norm
$\xrightarrow{\mathscr{P}}$	converge in probability
$\xrightarrow{a.e.}$	converge almost everywhere
$\xrightarrow{\mathscr{L}}$	converge in law

Chapter 1

Introduction

In this chapter, we give an introduction to the model under our consideration and the research presented in this thesis. In Section 1.1, we present a literature review of varying mixture models. In Section 1.2, we introduce the two-component mixture model under our consideration and present some motivating examples. Finally in Section 1.3 we lay out the structure of the thesis.

1.1 Review of mixture models

Statistical modeling is a very important and critical tool to understand many biological problems and human diseases. Particularly, mixture models arise frequently in this area due to the nature of the problems. A mixture model is a probabilistic model which represents the presence of subpopulations and which subpopulation each individual observation belongs to is not identified. In this thesis, I will focus on a two-component mixture model with stochastic dominance constraint on the two mixing components.

The m-component mixture model has the probability distribution function (p.d.f.)

$$\sum_{j=1}^{m} \pi_j f_j(x), \quad x \in \mathbb{R}^k, \tag{1.1}$$

where f_j is the j-th component p.d.f., π_j is the mixing proportion associated with f_j, and the mixing proportion vector $\pi = (\pi_1, \ldots, \pi_m)^\top$ satisfies $\sum_{j=1}^{m} \pi_j = 1$. When m is unknown, there are various articles that discuss the selection of m; see, for example, Roeder (1994), McLachlan and Peel (2000), Chen et al. (2001 & 2004) and Chen and Li (2009). We assume throughout this proposal that m is fixed and known to be two. Note that model (1.1) is generally unidentifiable if no restrictions are placed on f_j, simply due to the fact that f_j alone could be another mixture of several distributions.

When all components f_j's belong to a parametric family, which means that the space of unknown parameters is reduced to a Euclidean set, then (1.1) becomes parametric mixture model for which an extensive literature is available. The monographs of parametric mixture models include Everitt and Hand (1981), Lindsay (1995), Titterington et al. (1985), McLachlan and Basford (1988), Böhning (1999), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006) among others. When a parametric model is assumed for the components distributions, likelihood theory applies and the EM algorithm typically supplies the computations. Some of the well-known methods that have been proposed include maximum likelihood (Cohen, 1967; Lindsay, 1983a&b; Redner and Walker, 1984), minimum chi-square (Day, 1969), method of moments (Lindsay and Basak, 1993), Bayesian approaches (Diebolt and Robert, 1994; Escobar and West, 1995) and techniques based on moment generating function (Quandt and Ramsey, 1978).

In practice, however, the choice of parametric family is difficult when little is known about sub-populations. Many researchers have been trying to relax parametric assumptions on mixing components, and as a result, several important semiparametric mixture models were proposed and investigated in the last decade. Cruz-Medina and Hettmansperger (2004), Bordes et al. (2006b), Bordes et al. (2007) and Hunter et al. (2007) considered a semiparametric location-shifted mixture model for univariate case. Song et al. (2010) studied a two-component mixture model with one component specified as normal up to a scale parameter. Robin et al. (2007) considered a two-component mixture model of which the first component is known and the second is completely unknown, while Bordes et al. (2006a) and Bordes and Vandekerkhove (2010) considered a similar model where the second component is known to be symmetric. Leung and Qin (2006) adopted the 'exponential tilt' model for two-component bivariate case with the assumption of within-individual independent and identically distributed (i.i.d.) structure (i.e. conditionally independent repeated measurements), while Hammel (2010) extended their results to m-component mixture and conditionally independent but different marginal distributions for coordinates. Zou et al. (2002) employed the exponential tilt model for multi independent samples. Qin (1999), Zhang (2002, 2006) and Zhang (2005) considered the exponential tilt in two-component mixture model in a different direction when training samples are available. Deng et al. (2009) proposed an improved goodness of fit test for the model discussed in Zhang (1999). Qin and Liang (2011) considered testing the mixing proportion of a two-sample mixture model where a sample is available from the first component and another sample is available from the mixture, and Di et al. (2017) discussed testing the homogeneity of a similar model. Chen and Wu (2013) considered a two-sample semiparametric model with exponential tilt for classifying leukemia patients based on gene expression levels. Li, Liu and Qin (2017) discussed a different semiparametric mixture model with the same exponential tilt but known mixing proportion.

Comparatively, nonparametric mixture models have been given less attention due to the fact that they are reputed nonparametrically nonidentifiable. For nonparametric mixture without training data, methods for estimating mixing proportions have been developed specifically for each particular form of model. Assuming within-individual i.i.d. structure, Hettmansperger and Thomas (2000), Cruz-Medina (2001), Thomas and Hettmansperger (2001), Cruz-Medina et al. (2004), and Elmore et al. (2004) reduced the nonparametric multivariate mixture model to binomial or multinomial mixture model by discretization. However, some information is lost in the discretization step and for this reason it becomes difficult to obtain density distribution estimates of components. By assuming the vectors of observations are conditionally independent but, unlike repeated measurements, may have different marginal distributions, Hall and Zhou (2003) proposed a minimum distance estimator based on weighted-bootstrap estimation, Hall et al. (2005) investigated the inversion of mixture models in order to recover component distributions, while Benaglia et al. (2009) proposed an EM-like algorithm extended from Bordes et al. (2007). Recently, Hohmann and Holzmann (2013) considered the framework of conditional mixtures for the model discussed by Hall and Zhou (2003). Jochman et al. (2016) considered a two-sample mixture model with tail restrictions. Chauveau and Hoang (2016) discussed a nonparametric mixture model with conditionally independent multivariate densities. Other approaches to nonparametric estimation for mixture models rely on training samples. Training samples are supplementary data observed directly from the components. Estimation of nonparametric finite mixture models with training data had been studied mainly back in the seventies and eighties. Hosmer (1973) was the first to consider the use of training samples; however, he restricted attention to normal mixtures. Murray and Titterington (1978) considered nonparametric estimation using density estimates for Hosmer's model M2. Hall (1981,1983) described minimum distance estimators based on empirical distribution function, while Titterington (1983) considered minimum quadratic distance estimators based on density estimations. Hall and Titterington (1984) constructed a sequence of multinomial approximations and related maximum likelihood estimators (MLEs) for Hosmer's models M2. Cruz-Medina (2001) applied the discretization approach to Hosmer's models M1 and M2. Karunamuni and Wu (2009) proposed minimum Hellinger distance estimator (MHDE) for Hosmer's model M1. All these literatures assume that data from each of the components are available. However, training samples may be available for some but not all components.

1.2 Two-component mixture model with stochastic dominance

The research problem under my consideration is described as follows. Suppose there is a random sample from a two-component mixture population $h = (1-\lambda)f + \lambda g$ and independently another sample from the first component f, i.e.

$$\begin{aligned} X_1, \ldots, X_m &\overset{\text{i.i.d.}}{\sim} f(x), \\ Y_1, \ldots, Y_n &\overset{\text{i.i.d.}}{\sim} h(x) = (1-\lambda)f(x) + \lambda g(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.2}$$

where the unknown mixing proportion $\lambda \in (0,1)$, and f and g are two unknown p.d.f. satisfying the stochastic dominance constraint $F \geq G$. Here we denote F, G and H the corresponding cumulative distribution functions (c.d.f.) of f, g and h, and thus $H = (1-\lambda)F + \lambda G$. In many situations, λ is a value close to zero and a sample from the abundant population F is readily available. The problem of our interest is to make inferences for the mixing proportion λ and estimate the likelihood of an observation being from the second component.

The introduction of this model is motivated by the problem of identifying differentially expressed genes under two or more conditions (e.g. healthy tissue vs. diseased tissue) in microarray data. For this purpose the same test is used for each gene. Under the null hypothesis, corresponding to a lack of difference in expression level, the test statistics usually has a specified distribution F (e.g. normal or Student's t). However under the alternative hypothesis, corresponding to the presence of difference in expression level, the distribution G of the test statistic is unknown. For each of thousands of genes, either differentially expressed or not, the test statistic value is calculated. Treating each test statistic value as a response from the corresponding gene, the thousands of responses of all genes come from a mixture of two distributions, the known distribution F (under the null for the not differentially expressed genes) and another unknown G (under the alternative for the differentially expressed genes), with some unknown mixing proportion λ. Once mixing proportion λ and G have been estimated, one can estimate the probability that a gene is not differentially expressed, i.e. belongs to F. Thus, using a classification criterion we can classify each gene as either differentially expressed or not differentially expressed with estimated misclassification rate. Based on all the identified differentially expressed genes together, i.e. marker genes, one could build a classification rule, say based on weighted average, to classify each subject (e.g. healthy vs. diseased). More generally, F might be also unknown in practice but a training sample from F is immediately available. In the above setup of microarray test of genes, this means that particular genes have been confidently identified by pathologists or experts as not differentially expressed, i.e. from F, the distribution that is generally unknown or not exactly the same for small sample size as the postulated distribution for large sample size but otherwise for which the information is contained in the identified non differentially expressed genes. This generalization makes the model more robust than when F is assumed known. The stochastic dominance between F and G arise naturally in many situations where one believes that the test statistics for marker genes tend to be larger or smaller than those for non-marker genes. For example, the most often used Student's t (strictly $|t|$) and ANOVA F statistics satisfy the stochastic ordering $F \geq G$.

Besides the motivating example, model (1.2) could be used to model many other real data structure such as the following. Clinical malaria can be diagnosed by the presence of parasites and fever. However in endemic areas children can tolerate malaria parasite without the development of any sign of disease, and they may have fever due to some other reason. We can consider a mixture model where the mixture consists of parasite densities in children with fever due to malaria or due to other causes. One component of the mixture corresponds to children without malaria and the other corresponds to children with malaria. Parasite levels in children from a community could be available and used as a training sample, i.e. a sample that comes from the component of the mixture corresponding to children without clinical malaria but have parasites in their body and hence fever. Here the mixing proportion is the proportion of children whose fever is attributable to malaria. Irion et al. (2002) and Smith et al. (1994) discussed such example of disease conditions.

Many biomedical assays involve classifying individuals into two groups according to whether some output (e.g. optical density, titer, parasite density, amount of radioactive label) exceeds a given cutoff. Many such assays do not classify all samples correctly because there is an overlap between the distributions of the output from the two groups. Optimally cutoffs are evaluated by determining the misclassification probabilities of samples with known diagnoses (the 'Gold Standard'). Often, however, a sample from the distribution of true negatives (F) is available but there is no Gold Standard for the true positives (G), which can only be identified by using the assay itself, and then with uncertainty.

Let $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda)$ with $Z_i = 1$ if Y_i comes from G and $Z_i = 0$ if Y_i comes from F. Then model (1.2) gives $Y_i|Z_i \sim (1-Z_i)F + Z_i G$, $i = 1,\ldots,n$. Note that Z_i's are unobservable, i.e. missing data. By Bayes' rule, given an observation $Y = y$ the probability of it being from G is

$$p(y) \;:=\; P(Z=1|Y=y) \;=\; \frac{\lambda g(y)}{(1-\lambda)f(y)+\lambda g(y)}. \tag{1.3}$$

In the motivating example, λ and $p(y)$ correspond to the proportion of marker genes and the chance of being a marker gene given a test statistic value y respectively. In biomedical assay case, they correspond to the proportion of patients with particular disease and the chance of having this disease given an assay index value y.

To our best knowledge, there is not even a single work on model (1.2) in the literature. The closest work that has been done related to this model is given in Smith and Vounatsou (1997). However their model did not take the stochastic dominance constraint but instead assumed that the probability function in (1.3) is monotonically increasing. For this model, Smith and Vounatsou (1997) proposed several estimators of the mixing proportion λ in diagnostic assays based on a logistic power model, a nonparametric monotone regression model and latent class model, while Vounatsou et al. (1998) proposed a Bayesian method for discretized samples. Other literatures for the same model are in medicine, e.g. Smith et al. (1994), Irion et al. (2002) and Nagelkerke et al. (2003), which used above methods for analyzing different diseases such as malaria and legionella pneumonia. The monotone assumption on function p in these works is stronger than the stochastic dominance constraint in our model (1.2). To see this, note that function p being monotonic increasing generally implies that $F \geq G$. But the implication of the other direction is not true. A counter-example is $f(x) = 0.5I_{[0,2]}(x)$ and $g(x) = (1-0.25x)I_{[1,3]}(x)$ with I the indicator function, for which simple calculation shows that $F \geq G$ but p is not a monotonic function. Therefore, our model (1.2) is a generalization of their model.

1.3 Organization of thesis

Since model (1.2) is the focus of this thesis, each chapter of this thesis is devoted to an inference method for this model. Specifically, Chapters 2 and 3 propose two nonparametric estimation methods and Chapters 4 and 5 propose two semiparametric estimation methods with an exponential tilt imposed on the two components. Chapter 6 tests the exponential tilt structure assumed in Chapters 4 and 5, while Chapter 7 illustrate applications of all the methods in previous chapters to real data sets and presents concluding remarks.

In Chapter 2, we propose a nonparametric estimator of λ in model (1.2) based on c.d.f.s and an estimator of the lower bound of $p(y)$ defined in (1.3). We provide sufficient conditions for model (1.2) being identifiable, under which we demonstrate that the proposed estimator of λ is consistent. To assess its finite-sample performance, we carry out Monte Carlo simulation studies for different mixtures and compare our estimator with two existing estimators proposed by Smith and Vonastsou (1997).

In Chapter 3, we propose and investigate another nonparametric MLE through multinomial approximation. We prove that the proposed estimator is consistent when both the number of partitions and sample sizes go to infinity. Simulation studies are conducted to assess its finite-sample performance and to compare it with the estimator proposed in Chapter 2.

In order to accommodate the stochastic dominance constraint on the two components in model (1.2), in Chapter 4 we introduce the exponential tilt link between the two components. More specifically, the log ratio of the two component density functions is assumed of regression form which results in a two-sample semiparametric mixture model. With well defined parameter space, the stochastic dominance condition is automatically satisfied. For this model, we construct the MLE and prove that it is asymptotically normally distributed. Through simulation studies, we demonstrate that the proposed estimator works efficiently.

In Chapter 5, for the same model as in Chapter 4, we propose and investigate a robust MHDE of the unknown parameters. Asymptotic properties such as consistency and asymptotic normality of the proposed MHDE are presented. Through simulation studies, we compare the performance of the MHDE with that of MLE in Chapter 4 and observe that both are very competitive in efficiency but the MHDE is much more robust than MLE.

Since in Chapters 4 and 5 we assume the exponential tilt link, in Chapter 6 we propose a Kolmogorov-Smirnov type hypothesis testing to test the validity of the link. We develop two test statistic, based on the MLE and MHDE in Chapters 4 and 5 respectively, and discuss through simulation studies the estimated level of significance and power of the proposed test statistics.

In Chapter 7, we analyze two real data sets, a grain data set and a malaria data set, using the methods and procedures proposed in previous chapters to demonstrate their implementations. The remarks and discussion of future work are presented at the end.

Chapter 2

Nonparametric Estimation I based on C.D.F.

In this chapter, we propose a nonparametric estimator of λ in model (1.2). The estimator utilizes the dominance condition applied on nonparametric estimations of the c.d.f.s of the two components. In Section 2.1, we derive a sufficient condition for model (1.2) being identifiable. In Section 2.2, we present the construction of the nonparametric estimator. Its asymptotic properties are discussed in Section 2.3, while its finite-sample performance is presented in Section 2.4 through a simulation study.

2.1 Model identifiability

A nonparametric mixture model is generally unidentifiable. Robin et al. (2007) considered a two-component mixture model where the first component is known and the second is completely unknown. However, they didn't discuss the identifiability of the model. In fact, that model is not identifiable if no assumptions are put on the components. Bordes et al. (2006a) and Bordes and Vandekerkhove (2010) considered a similar model where the second component is known to be symmetric. Bordes et al. (2006a) showed that under moment and symmetry conditions the model is identifiable. In univariate cases, all discussions on identifiability are for mixture models with symmetric components; see, for example, Bordes et al. (2006b) and Hunter et al. (2007). In our model (1.2), both of the two components f and g could be arbitrarily of any form (not necessarily symmetric); however we have a training sample from the first component f. Thus f could be estimated consistently and at a certain convergence rate (e.g. $m^{-2/5}$ for kernel density estimator), i.e. f can be 'identified' by the training sample. As a result, when investigating the identifiability of model (1.2), we can simply, and equivalently, assume that f is known and we have a single sample from the mixture.

The thus reduced mixture model,

$$h(x) = (1-\lambda)f(x) + \lambda g(x), \quad x \in \mathbb{R}$$

with f known, λ and g unknown and $F \geq G$ (i.e. the model in Robin et al., 2007 with dominance), nevertheless is still generally unidentifiable. To see this, note that for any $\lambda^{'} \in (\lambda, 1)$,

$$(1-\lambda)f + \lambda g \;=\; (1-\lambda^{'})f + \lambda^{'}\left[\left(1-\frac{\lambda}{\lambda^{'}}\right)f + \frac{\lambda}{\lambda^{'}}g\right]$$

and $F \geq (1-\frac{\lambda}{\lambda^{'}})F + \frac{\lambda}{\lambda^{'}}G$ if $F \geq G$. To make any estimation of the mixture model in (1.2) meaningful, we need it to be identifiable. The following theorem gives a sufficient condition for model (1.2) being identifiable. Let D_f and D_g denote the lower limit of the support of function f and g respectively. Note that D_f and D_g could possibly be $-\infty$. Since $F \geq G$, we have $D_f \leq D_g$ and thus $g(x)/f(x)$ is well defined and $g(x)/f(x) \leq 1$ as $x \to D_f^+$.

Theorem 2.1. *Assume that $p(x) \to 0$ or equivalently $g(x)/f(x) \to 0$ as $x \to D_f^+$ and m is sufficiently large. Then the mixture model (1.2) is identifiable.*

Proof. As explained above, we can equivalently assume f is known. Suppose $h = (1-\lambda)f + \lambda g$ could also be represented as $h = (1-\lambda_1)f + \lambda_1 g_1$ with $0 < \lambda_1 < 1$ and g_1 a p.d.f. such that $g_1(x)/f(x) \to 0$ as $x \to D_f^+$. Then we have

$$\frac{g_1(x)}{f(x)} \;=\; \frac{\lambda_1 - \lambda}{\lambda_1} + \frac{\lambda}{\lambda_1}\frac{g(x)}{f(x)} \;\to\; \frac{\lambda_1 - \lambda}{\lambda_1} = 0,$$

and thus $\lambda_1 = \lambda$ and $g_1 = g$. $\square$

Remark 2.1. *The sufficient condition in Theorem 2.1 is quite weak and also easy to check. For example, if $D_f \neq D_g$, then the condition holds. When $D_f = D_g$ (either $-\infty$ or finite), the condition needs to be checked case by case.*

2.2 Construction of nonparametric estimation I

Note that for any $\alpha \in (0,1)$,

$$1 - H(F^{-1}(\alpha))/\alpha \;=\; 1 - \frac{(1-\lambda)\alpha + \lambda G(F^{-1}(\alpha))}{\alpha} \;=\; \lambda\left[1 - \frac{G(F^{-1}(\alpha))}{\alpha}\right].$$

Since $F \geq G$, we have $\frac{G(F^{-1}(\alpha))}{\alpha} \leq 1$ and then

$$1 - H(F^{-1}(\alpha))/\alpha \leq \lambda.$$

Thus a lower bound estimate of λ is given by

$$\hat{\lambda}_\alpha \; = \; 1 - \frac{H_n(F_m^{-1}(\alpha))}{\alpha},$$

where F_m and H_n are some appropriate nonparametric estimators of F and H respectively based on the samples X_i's and Y_i's. If the discrepancy between F and G is large enough, then $\frac{G(F^{-1}(\alpha))}{\alpha}$ may be near zero at some α value and as a result $1 - H(F^{-1}(\alpha))/\alpha$ will be close to λ. Especially when the sufficient condition for identifiability in Theorem 2.1, i.e. $\lim_{x \to D_f^+} g(x)/f(x) = 0$, holds, $\frac{G(F^{-1}(\alpha))}{\alpha}$ will be very close to zero for small α values. Thus, intuitively, we propose an estimator of λ given by

$$\hat{\lambda} \; = \; \sup_{\alpha \in (0,1)} \hat{\lambda}_\alpha \; = \; 1 - \inf_{\alpha \in (0,1)} \frac{H_n(F_m^{-1}(\alpha))}{\alpha}. \tag{2.1}$$

The function $p(y)$ in (1.3) now can be estimated by

$$\hat{p}(y) = 1 - (1 - \hat{\lambda}) \frac{f_m(y)}{h_n(y)}, \tag{2.2}$$

where f_m and h_n are the corresponding p.d.f.s of F_m and H_n respectively. In this thesis, we use kernel density estimators

$$f_m(x) \; = \; \frac{1}{mb_m} \sum_{i=1}^{m} K_0 \left(\frac{x - X_i}{b_m} \right), \tag{2.3}$$

$$h_n(x) \; = \; \frac{1}{nb_n} \sum_{j=1}^{n} K_1 \left(\frac{x - Y_j}{b_n} \right), \tag{2.4}$$

where K_0 and K_1 are kernel p.d.f.s and bandwidths b_n and b_m are positive sequences such that $b_m \to 0$ as $m \to \infty$ and $b_n \to 0$ as $n \to \infty$.

2.3 Asymptotic properties

The following theorem gives the asymptotic properties of the proposed estimator $\hat{\lambda}$ in (2.1).

Theorem 2.2. *Suppose that both f and g are uniformly continuous and the bandwidths b_m and b_n make $\sum_{m=1}^{\infty}\exp(-rmb_m^2)$ and $\sum_{n=1}^{\infty}\exp(-rnb_n^2)$ converge for every $r>0$. Then as $m,n\to\infty$,*

$$\hat{\lambda}_\alpha \xrightarrow{a.e.} \lambda - \lambda\frac{G(F^{-1}(\alpha))}{\alpha} \tag{2.5}$$

for any $\alpha\in(0,1)$, and as a result

$$\hat{\lambda} \xrightarrow{a.e.} \lambda - \lambda \inf_{\alpha\in(0,1)}\frac{G(F^{-1}(\alpha))}{\alpha}. \tag{2.6}$$

Proof. Since both f and g are uniformly continuous, h is also uniformly continuous. By Rao (1983), we have $\sup_x|f_m(x)-f(x)| \xrightarrow{a.e.} 0$ as $m\to\infty$ and $\sup_x|h_n(x)-h(x)| \xrightarrow{a.e.} 0$ as $n\to\infty$. Since $\int f(t)dt - \int f_m(t)dt = 1-1=0$, we have $\int[f(t)-f_m(t)]^+dt = \int[f(t)-f_m(t)]^-dt$ and thus $\int|f(t)-f_m(t)|dt = \int[f(t)-f_m(t)]^+dt + \int[f(t)-f_m(t)]^-dt = \int 2[f(t)-f_m(t)]^+dt$. By the Dominated Convergence Theorem (DCT), $\int|f(t)-f_m(t)|dt \xrightarrow{a.e.} 0$ as $m\to\infty$, and thus

$$\begin{aligned}
\sup_x|F_m(x)-F(x)| &= \sup_x\left|\int_{-\infty}^{x}[f_m(t)-f(t)]dt\right| \\
&\leq \sup_x\int_{-\infty}^{x}|f_m(t)-f(t)|dt \\
&\leq \int_{-\infty}^{\infty}|f_m(t)-f(t)|dt \\
&\xrightarrow{a.e.} 0.
\end{aligned}$$

Similarly $\sup_x|H_n(x)-H(x)| \xrightarrow{a.e.} 0$ as $n\to\infty$. As a result, $\sup_x|F(F_m^{-1}(x))-F(F^{-1}(x))| = \sup_x|x-F(F_m^{-1}(x))| = \sup_x|F_m(F_m^{-1}(x))-F(F_m^{-1}(x))| \xrightarrow{a.e.} 0$ as $m\to\infty$. Since f is uniformly continuous, we have $\sup_x|F_m^{-1}(x)-F^{-1}(x)| \xrightarrow{a.e.} 0$ as $m\to\infty$. Therefore, as $m,n\to\infty$,

$$\begin{aligned}
\hat{\lambda}_\alpha &= 1-\frac{H(F_m^{-1}(\alpha))}{\alpha} - \frac{H_n(F_m^{-1}(\alpha))-H(F_m^{-1}(\alpha))}{\alpha} \\
&\xrightarrow{a.e.} 1-\frac{H(F^{-1}(\alpha))}{\alpha} \\
&= 1-\frac{(1-\lambda)F(F^{-1}(\alpha))+\lambda G(F^{-1}(\alpha))}{\alpha} \\
&= \lambda-\lambda\frac{G(F^{-1}(\alpha))}{\alpha},
\end{aligned}$$

i.e. (2.5) holds.

Since $\sup_x |F_m(x) - F(x)| \xrightarrow{a.e.} 0$ as $m \to \infty$ and $\sup_x |H_n(x) - H(x)| \xrightarrow{a.e.} 0$ as $n \to \infty$, $F_m(x) - H_n(x) \xrightarrow{a.e.} F(x) - H(x) \geq 0$ as $F \geq G$, and thus $F_m(x) \overset{a.e.}{\geq} H_n(x)$ as $m,n \to \infty$ for any x. As a result, $0 \overset{a.e.}{\leq} \hat{\lambda} \leq 1$ as $m,n \to \infty$ and thus $0 \overset{a.e.}{\leq} \underline{\lim}_{m,n\to\infty} \hat{\lambda} \leq \overline{\lim}_{m,n\to\infty} \hat{\lambda} \leq 1$. By (2.5) we have $\hat{\lambda} \geq \hat{\lambda}_\alpha \xrightarrow{a.e.} \lambda - \lambda \dfrac{G(F^{-1}(\alpha))}{\alpha}$ for any $\alpha \in (0,1)$, and then

$$\underline{\lim}_{m,n\to\infty} \hat{\lambda} \overset{a.e.}{\geq} \lambda - \lambda \inf_{\alpha\in(0,1)} \frac{G(F^{-1}(\alpha))}{\alpha}.$$

On the other hand, we will show in the following that $\overline{\lim}_{m,n\to\infty} \hat{\lambda} \overset{a.e.}{\leq} \lambda - \lambda \inf_{\alpha\in(0,1)} \frac{G(F^{-1}(\alpha))}{\alpha}$ and therefore $\hat{\lambda} \xrightarrow{a.e.} \lambda - \lambda \inf_{\alpha\in(0,1)} \dfrac{G(F^{-1}(\alpha))}{\alpha}$ as $m,n \to \infty$, i.e. (2.6) holds. Let $\alpha_{m,n}$ denote a subsequence such that $\lim_{m,n\to\infty} \hat{\lambda}_{\alpha_{m,n}} \overset{a.e.}{=} \overline{\lim}_{m,n\to\infty}\hat{\lambda}$. Since the interval $(0,1)$ is finite, $\{\alpha_{m,n}\}$ must has a convergent subsequence, and we will still use $\alpha_{m,n}$ to denote this convergent subsequence without confusion. If $\alpha_{m,n} \to a_0 \neq 0$, then by the fact that $\sup_x |F_m^{-1}(x) - F^{-1}(x)| \xrightarrow{a.e.} 0$ we have

$$\frac{H(F^{-1}(\alpha_{m,n}))}{\alpha_{m,n}} \xrightarrow{a.e.} \frac{H(F^{-1}(a_0))}{a_0} = 1 - \lambda + \lambda \frac{G(F^{-1}(a_0))}{a_0},$$

$$\frac{H_n(F_m^{-1}(\alpha_{m,n})) - H(F_m^{-1}(\alpha_{m,n}))}{\alpha_{m,n}} \xrightarrow{a.e.} 0,$$

and thus

$$\begin{aligned} \hat{\lambda}_{\alpha_{m,n}} &= 1 - \frac{H(F_m^{-1}(\alpha_{m,n}))}{\alpha_{m,n}} - \frac{H_n(F_m^{-1}(\alpha_{m,n})) - H(F_m^{-1}(\alpha_{m,n}))}{\alpha_{m,n}} \\ &\xrightarrow{a.e.} \lambda - \lambda \frac{G(F^{-1}(a_0))}{a_0}. \end{aligned}$$

If $a_0 = 0$, then since $H_n(F_m^{-1}(\alpha_{m,n})) \overset{a.e.}{\leq} \alpha_{m,n}$, L'Hospital's rule gives

$$\lim_{m,n\to\infty} \frac{H_n(F_m^{-1}(\alpha_{m,n}))}{\alpha_{m,n}} \overset{a.e.}{=} \lim_{m,n\to\infty} \frac{h_n(F_m^{-1}(\alpha_{m,n}))}{f_m(F_m^{-1}(\alpha_{m,n}))} \overset{a.e.}{=} \lim_{x\to D_f^+} \frac{h(x)}{f(x)} = 1 - \lambda + \lambda \lim_{x\to D_f^+} \frac{g(x)}{f(x)}.$$

Since $G(F^{-1}(x)) \leq x$ by $F \geq G$, L'Hospital's rule gives

$$\lim_{x\to 0} \frac{G(F^{-1}(x))}{x} = \lim_{x\to 0} \frac{g(F^{-1}(x))}{f(F^{-1}(x))} = \lim_{x\to D_f^+} \frac{g(x)}{f(x)} \tag{2.7}$$

and thus $\hat{\lambda}_{\alpha_{m,n}} \xrightarrow{a.e.} \lambda - \lambda \lim_{x \to D_f^+} \frac{g(x)}{f(x)} = \lambda - \lambda \lim_{x \to 0} \frac{G(F^{-1}(x))}{x}$. For whichever the case of a_0, we always have $\overline{\lim}_{m,n\to\infty} \hat{\lambda} \overset{a.e.}{\leq} \lambda - \lambda \inf_{\alpha \in (0,1)} \frac{G(F^{-1}(\alpha))}{\alpha}$. □

Corollary 2.1. *Suppose the conditions in Theorem 2.2 are satisfied and in addition* $g(y)/f(y) \to 0$ *as* $y \to D_f^+$. *Then* $\hat{\lambda} \xrightarrow{a.e.} \lambda$ *as* $m,n \to \infty$.

Proof. The equations (2.6) and (2.7) give the result. □

Remark 2.2. *By Theorem 2.2, the estimator $\hat{\lambda}$ defined in (2.1) is generally biased. However when the sufficient condition for identifiability given in Theorem 2.1 is satisfied, then model (1.2) is identifiable and at the same time the estimator $\hat{\lambda}$ is consistent by Corollary 2.1.*

2.4 Simulation studies

We assess the efficiency of the proposed estimator $\hat{\lambda}$ using the following Monte Carlo simulation study and compare its finite-sample performance with the only few other methods available in literature. In our simulation study, we consider the five two-component mixture models given in Table 2.1. We can easily check that all the five models satisfy the stochastic dominance condition. Even though the focus of this thesis is on continuous mixture models, we also want to check the performance of the proposed methods for discrete mixture models such as M3 and M4.

Table 2.1: Mixture models considered in simulation study.

M1	$(1-\lambda)N(0,1)+\lambda N(1,1)$	mixture of normals that are close
M2	$(1-\lambda)N(0,1)+\lambda N(5,1)$	mixture of normals that are apart
M3	$(1-\lambda)Po(2)+\lambda Po(4)$	mixture of Poissons that are close
M4	$(1-\lambda)Po(2)+\lambda Po(6)$	mixture of Poissons that are apart
M5	$(1-\lambda)U(0,4)+\lambda U(2,6)$	mixture of uniforms

For each of the five mixture models, we consider varying values of $\lambda = 0.05, 0.20, 0.50, 0.80, 0.95$. We use two different sets of sample sizes $(m,n) = (30,30)$ and $(100,100)$. We take replication number $N = 1000$ as the number of random samplings. In the kernel density estimators f_m and h_n given in (2.3) and (2.4) respectively, we use Gaussian kernel function $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ for both K_0 and K_1. The bandwidths b_m and b_n in (2.3) and (2.4) respectively are chosen the same as in Silverman's (1986), i.e.

$$\begin{aligned} b_m &= 0.9m^{-1/5}\min\left[SD_X, \frac{IQR_X}{1.34}\right], \\ b_n &= 0.9n^{-1/5}\min\left[SD_Y, \frac{IQR_Y}{1.34}\right], \end{aligned} \tag{2.8}$$

where SD_X, SD_Y, IQR_X and IQR_Y are the sample standard deviation and interquartile range of samples X_i's and Y_i's respectively. To evaluate the finite-sample performance of the estimator $\hat{\lambda}$, we estimate the bias and mean squared error (MSE) by

$$Bias(\hat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\lambda}_i - \lambda\right) \tag{2.9}$$

and

$$MSE(\hat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\lambda}_i - \lambda\right)^2 \tag{2.10}$$

respectively, where N is the number of replications and $N = 1000$ in our simulation.

To examine the performance of the estimator $\hat{p}(y)$ given in (2.2) of the function $p(y)$, the probability of an observation y being from G, we check the classification results of a simple classification rule based on $\hat{p}$. Here the hard threshhold of 0.5 is used as the classification rule, i.e. an individual with observation y is classified as from G if $\hat{p}(y) > 0.5$ and F if otherwise. Then we use the misclassification rate (MR), the fraction of misclassified observations, as a measure of the performance of $\hat{p}$. However, we can expect that the MR will be high for some models, such as M1, when the two components have a large part that is in common and thus a high chance of misclassification exists. Considering this fact, we use the optimal misclassification rate (OMR) as the baseline to compare with. The OMR is the misclassification rate calculated when the function $p(y)$ is assumed completely specified and the same threshhold of 0.5 is used as the classification rule, which is the best scenario for this classification rule.

More specifically,

$$\text{OMR} = (1-\lambda)\int_{\{y:\lambda g(y)>(1-\lambda)f(y)\}} f(y)dy + \lambda\int_{\{y:\lambda g(y)<(1-\lambda)f(y)\}} g(y)dy. \tag{2.11}$$

Note that for uniform mixture model M5, $\lambda g(y) = (1-\lambda)f(y)$ for any $y \in [2,4]$. In this case we classify y as from f if $y \in [2,3]$ and g if $y \in (3,4]$, i.e. classify it to whichever the population of which it is closer to the centre. These OMRs and simulation results are presented in Table 2.2.

Table 2.2: Bias and MSE of $\hat{\lambda}$ and MR (%) of a classification rule based on $\hat{p}$.

		$m=n=30$		$m=n=100$		
Model	λ	$Bias(\hat{\lambda})(MSE(\hat{\lambda}))$	MR	$Bias(\hat{\lambda})(MSE(\hat{\lambda}))$	MR	OMR
M1	0.05	0.052 (0.031)	8.03	0.049 (0.019)	7.37	4.99
	0.20	0.093 (0.074)	26.30	0.085 (0.052)	26.09	18.61
	0.50	0.067 (0.078)	36.63	0.067 (0.052)	37.05	30.85
	0.80	-0.009 (0.047)	22.83	0.008 (0.026)	21.29	18.61
	0.95	-0.052 (0.029)	9.07	-0.027 (0.012)	6.80	4.99
M2	0.05	0.053 (0.022)	2.20	0.048 (0.015)	2.28	0.24
	0.20	0.095 (0.032)	3.97	0.065 (0.018)	3.02	0.48
	0.50	0.082 (0.015)	2.73	0.049 (0.005)	2.12	0.62
	0.80	0.061 (0.014)	6.93	0.057 (0.009)	6.21	0.48
	0.95	0.034 (0.002)	7.40	0.036 (0.002)	4.16	0.24
M3	0.05	0.203 (0.089)	28.40	0.035 (0.005)	5.04	4.76
	0.20	0.133 (0.064)	32.27	0.034 (0.013)	16.69	13.90
	0.50	0.019 (0.039)	34.13	-0.015 (0.001)	15.54	19.05
	0.80	-0.096 (0.034)	24.17	-0.084 (0.190)	25.20	13.35
	0.95	-0.152 (0.042)	14.40	-0.119 (0.023)	14.24	6.06
M4	0.05	0.211 (0.092)	26.30	0.044 (0.006)	3.03	3.14
	0.20	0.178 (0.071)	25.27	0.071 (0.012)	9.61	7.03
	0.50	0.117 (0.034)	21.60	0.048 (0.009)	15.54	10.19
	0.80	0.019 (0.010)	13.77	0.009 (0.006)	13.34	7.82
	0.95	-0.026 (0.004)	5.73	-0.006 (0.002)	5.81	3.21
M5	0.05	0.249 (0.136)	30.83	0.018 (0.002)	3.09	2.50
	0.20	0.151 (0.078)	29.90	0.017 (0.005)	10.99	10.00
	0.50	0.067 (0.048)	30.43	0.016 (0.008)	24.24	25.00
	0.80	0.031 (0.022)	16.20	0.012 (0.006)	12.76	10.00
	0.95	0.009 (0.006)	4.33	0.012 (0.002)	3.87	2.50

From Table 2.2 we can see that for all the five models, when sample sizes increase, the performance of both $\hat{\lambda}$ in terms of bias and MSE and $\hat{p}$ in terms of MR improves. Especially for models M3-M5, the performance of $\hat{\lambda}$ and $\hat{p}$ is not very good for $m = n = 30$ but improves dramatically when $m = n = 100$. When $m = n = 100$, $\hat{\lambda}$ has good efficiency in terms of small Bias and small MSE and the MRs are quite close to the OMRs. Considering the fact that the calculation of OMR assumes that the function p is completely specified, we can conclude that for relatively large sample sizes we are able to classify individuals equivalently well when $\hat{p}$ is used as when p is used.

The relatively worse performance for models M3-M5 when $m = n = 30$ could be possibly explained by the fact that these models don't satisfy the conditions listed in Corollary 2.1 for $\hat{\lambda}$ being consistent. For M3 and M4, they are not continuous mixture models and the condition $\lim_{x \to D_f} \frac{g(x)}{f(x)} = 0$ doesn't hold. Actually M3 and M4 are not identifiable. Even though model M5 satisfies the sufficient condition to be identifiable, the two components f and g are piecewise continuous but not globally continuous. Fortunately when sample sizes become large, the proposed estimator works well for models M3-M5. Comparatively, normal mixtures M1 and M2 satisfy all the conditions in Corollary 2.1 for being identifiable and for the estimator being consistent.

The proposed estimator $\hat{\lambda}$ in (2.1) is compared with two other estimators proposed by Smith and Vounatsou (1997). The first one assumes that both f and g are categorical populations with a baseline category $x = 0$ (i.e. the smallest possible value that the random variable can take). The estimator is based on the observed odds ratio and is defined as

$$\lambda_+ = [1 - \tilde{h}(0)](\tilde{\Psi}_+ - 1)/\tilde{\Psi}_+, \tag{2.12}$$

where $\tilde{h}(0)$ is the proportion of observations in the mixture sample that belong to the baseline category $x = 0$ and $\tilde{\Psi}_+$ is the odds ratio defined as

$$\tilde{\Psi}_+ = \frac{[1 - \tilde{h}(0)]\tilde{f}(0)}{[1 - \tilde{f}(0)]\tilde{h}(0)}.$$

The λ_+ is equivalent to population attributable fraction from case-control data. The second method proposed by Smith and Vounatsou (1997) is based on logistic power model and is defined as

$$\lambda_{el} = (1/n)\sum_i [\exp(\beta_1 X_i^\tau) - 1] / \exp(\beta_1 X_i^\tau), \tag{2.13}$$

where β_1 and τ are the MLEs of a non-linear logistic regression. Smith and Vonatsou (1997) considered two-component mixture of Poisson populations, mixture of log-normal populations and mixture of uniform populations. A total of repetition $N = 100$ data sets were simulated with $m = 100$ observations from f and $n = 100$ observations from h for varying λ values. Since their methods were developed particularly for categorical data, for comparison purpose, we will use their data generating scheme. Thus as what Smith and Vonatsou (1997) did, we discretize mixture of log-normal and mixture of uniform distributions. Specifically, for each repetition we pool X_i's and Y_j's together and the observed range of the pooled data is divided into 10 intervals with equal number of observations (20 in our case) falling into each interval. Then k is used as the observed data if the original value falls in the k^{th} interval, along with the indication whether an observed data is from population f or h. The category $k = 0$ corresponds to the baseline category used in the estimator λ_+. The simulation results of our estimator (2.1) in comparison with these two methods are given in Table 2.3.

From Table 2.3 we can see that for each model under consideration, our proposed estimator $\hat{\lambda}$ is very competitive with λ_+ and λ_{el}. Both λ_+ and λ_{el} tend to perform worse for larger λ values than smaller values, while $\hat{\lambda}$ tends to perform worse for smaller λ values. The $\tilde{\Psi}_+$ in the definition of λ_+ is sometimes less than one and as a result λ_+ is frequently negative. Smith and Vonatsou (1997) used 0 as the estimate whenever λ_+ is negative. Our estimator $\hat{\lambda}$ seldom give negative estimate of λ and thus no need of this trimming. This may explains the smaller bias and MSE of λ_+ relative to $\hat{\lambda}$ in some cases. Even though λ_+ is easy to calculate, it is often hard to define the baseline category and it does not exist if $h(0) = 0$. The λ_{el} has the advantage of model fitting using logistic regression, but the simulation studies in Smith and Vonatsou (1997) showed that it tends to underestimate λ, which is also demonstrated in Table 2.3 by those negative biases.

Table 2.3: Bias and MSE of $\hat{\lambda}$ in (2.1), λ_+ in (2.12) and λ_{el} in (2.13).

Model	λ	$Bias(\hat{\lambda})$ $(MSE(\hat{\lambda}))$	$Bias(\lambda_+)$ $(MSE(\lambda_+))$	$Bias(\lambda_{el})$ $(MSE(\lambda_{el}))$
	0.05	0.132 (0.047)	0.067 (0.029)	0.086 (0.019)
	0.20	0.061 (0.016)	0.003 (0.039)	0.136 (0.027)
$(1-\lambda)Po(2)+\lambda Po(6)$	0.50	0.004 (0.008)	-0.035 (0.049)	0.116 (0.019)
	0.80	-0.054 (0.007)	-0.048 (0.025)	0.040 (0.004)
	0.95	-0.042 (0.005)	-0.136 (0.043)	-0.003 (0.001)
	0.05	0.368 (0.263)	0.171 (0.109)	0.071 (0.019)
	0.20	0.181 (0.106)	0.039 (0.153)	0.052 (0.020)
$(1-\lambda)Po(4)+\lambda Po(6)$	0.50	0.001 (0.051)	-0.259 (0.163)	-0.029 (0.017)
	0.80	-0.106 (0.049)	-0.535 (0.388)	-0.145 (0.029)
	0.95	-0.145 (0.048)	-0.676 (0.564)	-0.196 (0.046)
	0.05	0.164 (0.070)	0.101 (0.052)	0.074 (0.009)
	0.20	0.086 (0.029)	0.036 (0.055)	0.063 (0.005)
$(1-\lambda)LN(2,1.4)+\lambda LN(6,2.5)$	0.50	0.010 (0.011)	-0.043 (0.059)	-0.043 (0.006)
	0.80	-0.063 (0.010)	-0.088 (0.025)	-0.088 (0.012)
	0.95	-0.049 (0.007)	-0.079 (0.013)	-0.108 (0.014)
	0.05	0.141 (0.049)	0.099 (0.989)	-0.002 (0.003)
	0.20	0.063 (0.019)	0.030 (0.091)	-0.055 (0.009)
$(1-\lambda)U(0,4)+\lambda U(2,6)$	0.50	0.032 (0.017)	-0.034 (0.115)	-0.074 (0.013)
	0.80	0.011 (0.008)	-0.020 (0.040)	-0.051 (0.005)
	0.95	0.015 (0.002)	-0.021 (0.046)	-0.090 (0.010)

Our proposed estimator $\hat{\lambda}$ defined in (2.1) provides an easy way to estimate λ without imposing the monotonicity of p or the stochastic dominance constraint $F \geq G$ (already incorporated in the definition of our estimator). Our simulation studies show that it performs well for both mixtures of continuous populations and mixtures of discrete populations. On the other hand, $\hat{\lambda}$ is a crude estimator of λ and we only have the consistency under some conditions but no results on asymptotic distributions. For small λ and sample sizes and mixture of discrete populations, our simulation shows that $\hat{\lambda}$ is biased. To improve the performance, we introduce in next chapter another nonparametric estimator of λ.

Chapter 3

Nonparametric Estimation II based on Multinomial Approximation

In this chapter, we introduce a second nonparametric estimator of the mixing proportion. This estimator is a MLE of a multinomial approximation to the mixture model. In Section 3.1 we present the idea of using a multinomial distribution to approximate the mixture model (1.2), and then construct the MLE of the approximated multinomial distribution. In Section 3.2 we study its consistency while the simulation study is given in Section 3.3.

3.1 Construction of MLE

Hall and Titterington (1984) constructed a sequence of multinomial approximations and studied related MLE for a model similar to (1.2). The model under their consideration is Hosmer's (1973) model M2 where the sample consists of both mixed (a sample from mixture) and known data (a sample from each of the two components), and the known data contains information about the mixing proportion. They derived a Cram*é*r-Rao lower bound for the nonparametric estimator of the mixing proportion and thereby characterized asymptotically optimal estimators. Karunamuni and Wu (2009) exploited the same method for Hosmer's (1973) model M1 where the sample consists of both mixed and known data, but no information about mixing proportion is contained in the known data. Elmore et al. (2004) utilized the same idea for multivariate mixture models. We propose to apply the same idea to our model (1.2). Model (1.2) is more complex in the sense that no sample is available from the second component, and due to the unidentifiability, it is not possible to find the MLE of the mixing proportion λ based on multinomial approximation. However, we will give an MLE of the lower bound of

λ.

Following Hall and Titterington (1984), we first partition the support of h in model (1.2) into L regions $R_1, \ldots, R_L$ so that each observation will fall uniquely into one region. Define, for $l = 1, \ldots, L$,

$$\begin{aligned} p_{1l} &= \int_{R_l} f(x)dx, \\ p_{2l} &= \int_{R_l} g(x)dx, \\ p_{3l} &= \int_{R_l} h(x)dx = (1-\lambda)p_{1l} + \lambda p_{2l}. \end{aligned} \tag{3.1}$$

Obviously $\sum_{l=1}^{L} p_{il} = 1$, $i = 1, 2, 3$. Let m_l and n_l denote the number of observations out of m and n respectively that fall into region R_l. Then when L is large, model (1.2) could be approximated closely by the multinomial populations given in (3.1). Based on this multinomial approximation, the likelihood is given by

$$\prod_{l=1}^{L} (p_{1l})^{m_l} \left[(1-\lambda)p_{1l} + \lambda p_{2l}\right]^{n_l}. \tag{3.2}$$

To obtain the MLE of the parameter $\theta = (\lambda, p_{1l}, \ldots, p_{1L}, p_{2l}, \ldots, p_{2L})^\top$, we take the partial derivatives of the log-likelihood function and make them equal to zero. This yields the following system of estimating equations:

$$\frac{m_l}{p_{1l}} - \frac{m_L}{p_{1L}} + \frac{n_l(1-\hat{\lambda})}{(1-\hat{\lambda})p_{1l} + \hat{\lambda}p_{2l}} - \frac{n_L(1-\hat{\lambda})}{(1-\hat{\lambda})p_{1L} + \hat{\lambda}p_{2L}} = 0, \quad l = 1, \ldots, L-1 \text{ and } p_{1l} \neq 0, \tag{3.3}$$

$$\frac{n_l\hat{\lambda}}{(1-\hat{\lambda})p_{1l} + \hat{\lambda}p_{2l}} - \frac{n_L\hat{\lambda}}{(1-\hat{\lambda})p_{1L} + \hat{\lambda}p_{2L}} = 0, \quad l = 1, \ldots, L-1, \tag{3.4}$$

$$\sum_{l=1}^{L} \frac{n_l(p_{2l} - p_{1l})}{(1-\hat{\lambda})p_{1l} + \hat{\lambda}p_{2l}} = 0, \tag{3.5}$$

$$\sum_{l=1}^{L} p_{1l} = 1, \tag{3.6}$$

$$\sum_{l=1}^{L} p_{2l} = 1, \tag{3.7}$$

subject to constraints $p_{il} \geq 0$, $i = 1,2$ and $l = 1,\ldots,L$. Let $\hat{\theta} = (\hat{\lambda}_L, \hat{p}_{11}, \ldots, \hat{p}_{1L}, \hat{p}_{21}, \ldots, \hat{p}_{2L})^\top$ denote the solution to the system (3.3)-(3.7). From (3.3) and (3.4) we get $\frac{m_l}{p_{1l}} = \frac{m_L}{p_{1L}}$, i.e.

$$\hat{p}_{1l} = \frac{m_l}{m_L} p_{1L}, \;\; l = 1,\ldots,L-1.$$

Since $\sum_{l=1}^{L} p_{1l} = 1$, summing up the above equation over l gives $\hat{p}_{1L} = \frac{m_L}{m}$, and thus we have

$$\hat{p}_{1l} \;=\; \frac{m_l}{m}, \;\; l = 1,\ldots,L. \tag{3.8}$$

Now plug (3.8) into (3.4) and use the constraint $\sum_{l=1}^{L} p_{2l} = 1$, we obtain

$$\hat{p}_{2L} \;=\; \frac{1}{\lambda} \cdot \frac{n_L}{n} - \frac{1-\lambda}{\lambda} \cdot \frac{m_L}{m}$$

and further

$$\hat{p}_{2l} \;=\; \frac{1}{\lambda} \cdot \frac{n_l}{n} - \frac{1-\lambda}{\lambda} \cdot \frac{m_l}{m}, \;\; l = 1,\ldots,L. \tag{3.9}$$

If we plug the MLEs $\hat{p}_{1l}$ and $\hat{p}_{2l}$, in (3.8) and (3.9) respectively, into (3.3)-(3.7), all give identities that do not involve λ. Thus the MLE of the estimating system does not exist. However, since $\hat{p}_{2l} \geq 0$, we have from (3.9) that $\lambda \geq 1 - \frac{n_l}{n} \cdot \frac{m}{m_l}$ for each l such that $m_l \neq 0$. Then the lower bound of MLE of λ is given by

$$\hat{\lambda}_L \;=\; 1 - \min_{\substack{l=1,\ldots,L \\ m_l \neq 0}} \left\{ \frac{n_l/n}{m_l/m} \right\}. \tag{3.10}$$

The MLE of $p(y)$ in (1.3) is given by

$$\hat{p}_L(y) = \frac{\hat{\lambda}_L \hat{p}_{2i}}{(1-\hat{\lambda}_L)\hat{p}_{1i} + \hat{\lambda}_L \hat{p}_{2i}}$$

if y falls in the i^{th} region R_l.

3.2 Asymptotic properties

In this section we discuss the asymptotic properties of the proposed MLE $\hat{\lambda}_L$ in (3.10) of the lower bound of λ. We will prove that under certain conditions, $\hat{\lambda}_L$ is the MLE of λ and thus is consistent. We need the following two lemmas for the consistency of $\hat{\lambda}_L$.

Lemma 3.1. *If $P(A_i) > 1 - p_i$ for all $i = 1, \ldots, L$, then $P\left(\bigcap_{i=1}^{L} A_i\right) > 1 - \sum_{i=1}^{L} p_i$.*

Proof. Since $P(A_i^c) < p_i$, $P\left((\bigcap_{i=1}^{L} A_i)^c\right) = P\left(\bigcup_{i=1}^{L} A_i^c\right) \leq \sum_{i=1}^{L} P(A_i^c) < \sum_{i=1}^{L} p_i$ and thus the result. □

Lemma 3.2. *If $X_{ni} \xrightarrow{\mathscr{P}} c_i$ as $n \to \infty$, $i = 1, \ldots, L$, then $\min_{i=1,\ldots,L}\{X_{ni}\} \xrightarrow{\mathscr{P}} \min_{i=1,\ldots,L}\{c_i\}$ as $n \to \infty$.*

Proof. For any $\varepsilon > 0$, we will show $P\left(\min_i\{X_{ni}\} - \min_i\{c_i\} < -\varepsilon\right) \to 0$ and $P\left(\min_i\{X_{ni}\} - \min_i\{c_i\} > \varepsilon\right) \to 0$. As a result $P\left(\left|\min_i\{X_{ni}\} - \min_i\{c_i\}\right| \leq \varepsilon\right) \to 1$, i.e. $\min_{i=1,\ldots,L}\{X_{ni}\} \xrightarrow{\mathscr{P}} \min_{i=1,\ldots,L}\{c_i\}$.

Since $\min_i\{X_{ni} - c_i\} + \min_i\{c_i\} \leq \min_i\{X_{ni}\}$, we have

$$\min_i\{X_{ni} - c_i\} + \varepsilon \;\leq\; \min_i\{X_{ni}\} - \min_i\{c_i\} + \varepsilon. \tag{3.11}$$

Since $X_{ni} \xrightarrow{\mathscr{P}} c_i$, we have $P(|X_{ni} - c_i| \leq \varepsilon) \to 1$. Note that

$$P(|X_{ni} - c_i| \leq \varepsilon) \;=\; P(-\varepsilon \leq X_{ni} - c_i \leq \varepsilon) \;\leq\; P(X_{ni} - c_i \geq -\varepsilon),$$

thus we have $P(X_{ni} - c_i + \varepsilon \geq 0) \to 1$. By definition this means that, for any $\delta > 0$ there exists a $n_0 \in \mathbb{N}$ such that for any $n > n_0$,

$$P(X_{ni} - c_i + \varepsilon \geq 0) \;>\; 1 - \frac{\delta}{L}, \;\; i = 1, \ldots, L.$$

Thus by Lemma 3.1 we have

$$P\left(\bigcap_{i=1}^{L}\{X_{ni} - c_i + \varepsilon \geq 0\}\right) > 1 - \delta$$

or equivalently

$$P\left(\min_i\{X_{ni} - c_i\} + \varepsilon \geq 0\right) > 1 - \delta,$$

which implies $P\left(\min_i\{X_{ni} - c_i\} + \varepsilon \geq 0\right) \to 1$. This together with (3.11) gives $P\left(\min_i\{X_{ni}\} - \min_i\{c_i\} \geq -\varepsilon\right) \to 1$, i.e. $P\left(\min_i\{X_{ni}\} - \min_i\{c_i\} < -\varepsilon\right) \to 0$.

Let $c_{i_0} = \min_i \{c_i\}$. Then by definition $P(|X_{ni_0} - c_{i_0}| \leq \varepsilon) \to 1$, and further $P(X_{ni_0} - c_{i_0} \leq \varepsilon) \to 1$. Note that $P(X_{ni_0} - c_{i_0} \leq \varepsilon) \leq P(\min_i \{X_{ni}\} - c_{i_0} \leq \varepsilon)$. Thus we have $P(\min_i \{X_{ni}\} - c_{i_0} \leq \varepsilon) \to 1$, i.e. $P\left(\min_i \{X_{ni}\} - \min_i \{c_i\} > \varepsilon\right) \to 0$.

□

Theorem 3.1. $\hat{\lambda}_L \xrightarrow{\mathscr{P}} \lambda - \lambda \min_{\substack{l=1,\ldots,L \\ p_{1l} \neq 0}} \left\{ \frac{p_{2l}}{p_{1l}} \right\}$ *as* $m, n \to \infty$.

Proof. By WLLN, we have $m_l/m \xrightarrow{\mathscr{P}} p_{1l}$ and $n_l/n \xrightarrow{\mathscr{P}} (1-\lambda)p_{1l} + \lambda p_{2l}$ as $m, n \to \infty$. Then for any $p_{1l} \neq 0$, we have

$$\frac{n_l/n}{m_l/m} \xrightarrow{\mathscr{P}} (1-\lambda) + \lambda \frac{p_{2l}}{p_{1l}}.$$

By Lemma 3.2 we have

$$\min_{\substack{l=1,\ldots,L \\ m_l \neq 0}} \left\{ \frac{n_l/n}{m_l/m} \right\} \xrightarrow{\mathscr{P}} 1 - \lambda + \lambda \min_{\substack{l=1,\ldots,L \\ p_{1l} \neq 0}} \left\{ \frac{p_{2l}}{p_{1l}} \right\},$$

and hence the result. □

From Theorem 3.1 we obtain immediately a sufficient condition for $\hat{\lambda}_L$ being consistent.

Corollary 3.1. *If* $p_{2l} = 0$ *for some* l *such that* $p_{1l} \neq 0$, *then* $\hat{\lambda}_L \xrightarrow{\mathscr{P}} \lambda$ *as* $m, n \to \infty$.

Proof. The result follows directly from Theorem 3.1. □

For fixed L, the sufficient condition of consistency in Corollary 3.1 holds only when f and g have different support. When f and g have common finite support, obviously the condition is not satisfied. But when the lower limit of the common support $D_f = -\infty$ and we allow $L \to \infty$, then the estimator $\hat{\lambda}_L$ is consistent under the condition of identifiability given in Theorem 2.1. This result is given in the following corollary.

Corollary 3.2. *Let the maximum length of the intervals* $R_1, \ldots, R_L$ *go to zero when* $L \to \infty$. *If* $g(x)/f(x) \to 0$ *as* $x \to D_f^+$, *then* $\hat{\lambda}_L \to \lambda$ *as* $m, n, L \to \infty$.

Proof. From Theorem 3.1 we have, as $m,n \to \infty$ and then $L \to \infty$, that

$$\hat{\lambda}_L \xrightarrow{\mathscr{P}} \lambda - \lambda \inf_{\{x:f(x)\neq 0\}} \left\{ \frac{g(x)}{f(x)} \right\} = \lambda.$$

□

Remark 3.1. *The stochastic dominance constraint $F \geq G$ in the multinomial approximation is reduced to*

$$\sum_{i=1}^{k} p_{1i} \geq \sum_{i=1}^{k} p_{2i}, \ \ k = 1, \ldots, L. \tag{3.12}$$

We don't impose this dominance constraint on the estimating equations (3.3)-(3.7) when the asymptotic properties of MLE are discussed, simply to avoid technical difficulties. The study of the asymptotic properties of MLE under the constraint (3.12) is planned for future research. On the other hand, the MLE under (3.12) is examined in the simulation studies in the next section.

3.3 Simulation studies

To evaluate the finite-sample performance of our proposed estimator $\hat{\lambda}_L$ given in (3.10), we consider the same mixture models listed in Table 2.1 of Section 2.4, with the same varying mixing proportion values. We also use the same sample sizes $m = n = 30$ and $m = n = 100$ and the same number of replications $N = 1000$. Since the mixture of normals, i.e. models M1 and M2, satisfy the conditions in Corollary 3.2, we will let the number of partitioned intervals L increases when sample sizes increases from $m = n = 30$ to $m = n = 100$. Specifically, we tried $L = 3$ and $L = 5$ for $m = n = 30$ and $L = 10$, $L = 15$ and $L = 20$ for $m = n = 100$. Note that, large L results in small bins with mostly zero observations and small L results in very large bins with most of the observations clumped in one bin. We chose $L = 3$ and $L = 10$ for our simulation study as they perform better than the rest of the choice for L.

As discussed in Remark 3.1, we examine the MLE of λ under the stochastic dominance constraint (3.12). We still use $\hat{\lambda}_L$ to denote the MLE under (3.12) without confusion. Even though the MLE of λ without this constraint does not exist theoretically, our simulation results below show that the MLE with this constraint does exist. The MLE under the dominance constraint is calculated by optimizing the likelihood function (3.2) subject to (3.6), (3.7), (3.12) and all $p_{il} \geq 0$, $i = 1,2$ and $l = 1,\ldots,L$. We use function "optim" in R statistical software for this optimization. For simplicity, we use λ_+ given in (2.12) as the initial estimate of λ. We use the relative frequency m_l/m as the initial estimate of p_{1l}. To give an initial estimate of p_{2l}, we use the relationship $g(x) = \frac{h(x)}{\lambda} - \frac{1-\lambda}{\lambda} f(x)$ from (1.2) with λ replaced with its initial estimate and f and h replaced with the relative frequency m_l/m and n_l/n respectively.

As in Section 2.4, for each model we calculated the bias and MSE of $\hat{\lambda}_L$ and the misclassification rate MR when the same classification rule as in Section 2.4 is used. The simulation results are presented in Table 3.1. From Table 3.1, we can see that $\hat{\lambda}_L$ performs very well in terms of bias and MSE in most cases. Not surprisingly, the estimation accuracy is higher when the two components are well separated (M2 & M4) than when they are not (M1 & M3). These observations indicate that even though we group the data and thus lose some information, we can still estimate the mixing proportion quite well. However, the MRs are much higher than the OMRs in most cases. This is expected since with use of discretization, all the observations falling into the same interval will be classified as from the same component. When the interval is relatively wide, for example $L = 3$ or even $L = 10$, the discretization will generate a higher misclassification rate.

When the estimator $\hat{\lambda}_L$ we proposed in this chapter is compared with $\hat{\lambda}$ in (2.1) we proposed in Chapter 2, we observe that both perform competitively in terms of bias and MSE while $\hat{\lambda}$ has better performance in terms of MR than $\hat{\lambda}_L$. Therefore, if our interest is in λ only, then either method should work well. But if we are interested in classification, then method in Chapter 2 works much better.

Table 3.1: Bias and MSE of $\hat{\lambda}_L$ and MR (%) of a classification rule based on $\hat{p}_L$.

Model	λ	$m=n=30$ $Bias(\hat{\lambda}_L)(MSE(\hat{\lambda}_L))$	$m=n=30$ MR	$m=n=100$ $Bias(\hat{\lambda}_L)(MSE(\hat{\lambda}_L))$	$m=n=100$ MR	OMR
M1	0.05	0.020 (0.031)	5.60	0.022 (0.009)	5.00	4.99
	0.20	0.032 (0.068)	22.13	0.011 (0.016)	20.16	18.61
	0.50	-0.039 (0.078)	46.77	-0.048 (0.026)	44.37	30.85
	0.80	-0.144 (0.089)	41.57	-0.121 (0.033)	27.39	18.61
	0.95	-0.181 (0.090)	25.17	-0.149 (0.035)	11.73	4.99
M2	0.05	0.022 (0.031)	4.97	0.026 (0.009)	5.00	0.24
	0.20	0.055 (0.071)	20.57	0.033 (0.016)	19.90	0.48
	0.50	0.029 (0.066)	39.13	0.017 (0.019)	25.61	0.62
	0.80	-0.009 (0.036)	19.97	0.000 (0.010)	4.56	0.48
	0.95	-0.011 (0.009)	4.53	-0.002 (0.003)	2.21	0.24
M3	0.05	0.002 (0.015)	5.40	-0.000 (0.004)	5.01	4.76
	0.20	-0.017 (0.046)	22.03	-0.022 (0.013)	20.09	13.90
	0.50	-0.076 (0.010)	43.67	-0.073 (0.029)	44.03	19.05
	0.80	-0.128 (0.101)	35.53	-0.121 (0.038)	29.74	13.35
	0.95	-0.163 (0.102)	22.57	-0.143 (0.034)	14.27	6.06
M4	0.05	0.003 (0.016)	5.43	0.000 (0.004)	5.01	3.14
	0.20	-0.007 (0.047)	21.60	-0.000 (0.012)	20.08	7.03
	0.50	-0.028 (0.084)	37.97	-0.009 (0.021)	35.49	10.19
	0.80	-0.033 (0.059)	24.20	-0.018 (0.013)	14.68	7.82
	0.95	-0.048 (0.037)	10.03	-0.023 (0.005)	5.41	3.21
M5	0.05	0.026 (0.032)	22.07	0.033 (0.009)	5.01	2.50
	0.20	0.049 (0.069)	20.80	0.035 (0.014)	20.13	10.00
	0.50	0.025 (0.065)	45.10	0.011 (0.019)	37.34	25.00
	0.80	-0.001 (0.029)	23.07	-0.001 (0.008)	13.38	10.00
	0.95	-0.012 (0.010)	5.70	-0.005 (0.002)	3.62	2.50

Chapter 4

Semiparametric Estimation I: MLE

Due to the unidentifiability in general of the nonparametric mixture model (1.2), it is difficult to obtain an estimator with good asymptotic properties such as asymptotic normality. In this chapter, we impose a semiparametric structure on (1.2) and construct the MLE of the resulting semiparametric mixture model. In Section 4.1 we introduce this structure and explain why it successfully accommodates the stochastic dominance condition. In Section 4.2, we construct the MLE for the resulting semiparametric mixture model. Asymptotic properties of the proposed MLE are discussed in Section 4.3 while the simulation studies are presented in Section 4.4.

4.1 A semiparametric mixture model

Let Z denote a binary response variable and Y the associated covariate. Then the logistic regression model is given by

$$P(Z=1|Y=y) \;=\; \frac{\exp[\alpha^* + \beta^\top r(y)]}{1+\exp[\alpha^* + \beta^\top r(y)]}, \tag{4.1}$$

where $r(y) = (r_1(y), \ldots, r_p(y))^\top$ is a given $p \times 1$ vector of functions of y, α^* is the intercept parameter and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is the $p \times 1$ coefficient parameter vector. In case-control studies data are collected retrospectively. For example, a random sample of subjects with disease $Z=1$ ('case') and a separate random sample of subjects without disease $Z=0$ ('control') are selected with Y observed for each subject. Let $\pi = P(Z=1) = 1-P(Z=0)$. Let $f(y)$ and $g(y)$ denote the conditional p.d.f.s of Y given $Z=0$ and $Z=1$ respectively, then it follows from (4.1) and Bayes' rule that

$$g(y) \;=\; \exp[\alpha + \beta^\top r(y)] f(y), \tag{4.2}$$

where $\alpha = \alpha^* + \log[(1-\pi)/\pi]$. Now model (1.2) is reduced to the semiparametric mixture model

$$
\begin{aligned}
X_1, \ldots, X_m &\overset{\text{i.i.d.}}{\sim} f(x), \\
Y_1, \ldots, Y_n &\overset{\text{i.i.d.}}{\sim} h_\theta(x) := h(x) = \left\{(1-\lambda) + \lambda \exp\left[\alpha + \beta^\top r(x)\right]\right\} f(x)
\end{aligned}
\tag{4.3}
$$

where $\theta = (\lambda, \alpha, \beta^\top)^\top$ is the parameter vector of interest.

The relationship (4.2) between two p.d.f.s was first proposed by Anderson (1972). It essentially assumes that the log-likelihood ratio of the two p.d.f.s is linear in the observations. With $r(x) = x$ or $r(x) = (x, x^2)^\top$, it has wide applications in logistic discriminant analysis (Anderson, 1972 & 1979) and case-control studies (Prentice and Pyke, 1979; Breslow and Day, 1980). For $r(x) = x$, (4.2) encompasses many common distributions, including two exponential distributions with different means and two normal distributions with common variance but different means. Model (4.2) with $r(x) = (x, x^2)^\top$ also coincides with the exponential family of densities considered in Efron and Tibshirani (1996). Moreover, model (4.2) can be viewed as a biased sampling model with the 'tilt' weight function $\exp[\alpha + \beta^\top r(x)]$ depending on the unknown parameters α and β. Note that the test of equality of f and g can be regarded as a special case of model (4.3) with $\beta = 0$.

Qin and Zhang (1997) discussed a goodness-of-fit test for logistic regression based on case-control data where the first sample comes from the control group f and independently the second sample comes from the case group g. They proposed a Kolmogorov-Smirnov type statistic to test the validity of (4.2) with $r(y) = y$. When data from both the mixture and the two individual components satisfying (4.2) are available, Qin (1999) developed an empirical likelihood ratio based statistic for constructing confidence intervals of the mixing proportion. For the same model and data structure, Zhang (2002) proposed an EM algorithm to calculate the MLE while Zhang (2006) proposed a score statistic to test the mixing proportion. Chen and Wu (2013) employed (4.2) to model differentially expressed genes of acute lymphoblastic leukemia patients and acute myeloid leukemia patients.

For model (4.3) with $r(y) = y$, if $\beta > 0$ then we can easily check that $p(y)$ in (1.3), the probability of y being from g, is a monotonic increasing function. Further we can prove below that if $\beta > 0$, then the stochastic dominance constraint $F \geq G$ is implied by (4.2).

Theorem 4.1. *Model (4.3) with $r(y) = y$ is identifiable. If further $\beta > 0$ and m is sufficiently large, then $F \geq G$.*

Proof. Since we have a sample from f, so when look at identifiability we can equivalently assume f is known. Then if $h_{\theta_1} = h_{\theta_2}$, i.e.

$$\left\{1 - \lambda_1 + \lambda_1 \exp\left[\alpha_1 + \beta_1^\top r(x)\right]\right\} f(x) = \left\{1 - \lambda_2 + \lambda_2 \exp\left[\alpha_2 + \beta_2^\top r(x)\right]\right\} f(x) \text{ for all } x,$$

then we must have $\lambda_1 = \lambda_2$, $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. Thus h_θ is identifiable.

Let $s(x) = f(x) - g(x) = f(x)\left[1 - \exp(\alpha + \beta x)\right]$ and $S(x) = \int_{-\infty}^{x} s(t)dt = F(x) - G(x)$. Let x_0 denote the solution to $1 - \exp(\alpha + \beta x) = 0$. Then $s(x) > 0$ when $x < x_0$ and $s(x) \leq 0$ when $x \geq x_0$, and hence $S(x)$ increases for $x < x_0$ and decreases for $x \geq x_0$. If $F(x') < G(x')$ for some x', i.e. $S(x') < 0$, then $x' \geq x_0$ since $S(x) > 0$ for all $x < x_0$. Since $S(x)$ decreases when $x \geq x_0$, we have $S(x) \leq S(x') < 0$ for all $x > x'$ and thus $S(\infty) < 0$. However $S(\infty) = F(\infty) - G(\infty) = 1 - 1 = 0$, a contradiction. Therefore $F \geq G$. □

Even though Theorem 4.1 tells us that the condition (4.2) is stronger than the original stochastic dominance constraint, the thus resulted semiparametric mixture model (4.3) is identifiable and has better interpretation than the nonparametric mixture model (1.2). In addition, the estimation of (4.3) may possess better asymptotic properties, such as normality, than those of (1.2). So from now on, we will focus on model (4.3) with $r(y) = y$ and $\beta > 0$.

4.2 Construction of MLE

In this section, we construct the semiparametric empirical MLE of θ in (4.3). Let $(T_1, \ldots, T_{m+n}) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ be the pooled data and $p_i = dF(T_i)$. Then the empirical likelihood function is

$$L(\lambda,\alpha,\beta) \;=\; \prod_{i=1}^{m} dF(X_i)\prod_{j=1}^{n} dH(Y_j) \;=\; \prod_{i=1}^{m+n} p_i \prod_{j=1}^{n}\left[(1-\lambda)+\lambda e^{\alpha+\beta Y_j}\right],$$

subject to constraints $\beta \geq 0$, $0 \leq \lambda \leq 1$, $p_i \geq 0$, $\sum_{i=1}^{m+n} p_i = 1$, and $\sum_{i=1}^{m+n} p_i e^{\alpha+\beta T_i} = 1$. To find the MLE, we use the Lagrange multipliers and maximize

$$\sum_{i=1}^{m+n}\log p_i + \sum_{j=1}^{n}\log\left[(1-\lambda)+\lambda e^{\alpha+\beta Y_j}\right] - t_1\left[\sum_{i=1}^{m+n} p_i - 1\right] - t_2\left[\sum_{i=1}^{m+n} p_i e^{\alpha+\beta T_i} - 1\right].$$

Taking partial derivatives gives the estimating equation system

$$\frac{1}{p_i} - t_1 - t_2 e^{\alpha+\beta T_i} \;=\; 0, \quad i = 1,\ldots,m+n, \tag{4.4}$$

$$\sum_{j=1}^{n}\frac{e^{\alpha+\beta Y_j}-1}{(1-\lambda)+\lambda e^{\alpha+\beta Y_j}} \;=\; 0, \tag{4.5}$$

$$\sum_{j=1}^{n}\frac{\lambda e^{\alpha+\beta Y_j}}{(1-\lambda)+\lambda e^{\alpha+\beta Y_j}} - t_2\sum_{i=1}^{m+n} p_i e^{\alpha+\beta T_i} \;=\; 0, \tag{4.6}$$

$$\sum_{j=1}^{n}\frac{Y_j\lambda e^{\alpha+\beta Y_j}}{(1-\lambda)+\lambda e^{\alpha+\beta Y_j}} - t_2\sum_{i=1}^{m+n} p_i T_i e^{\alpha+\beta T_i} \;=\; 0, \tag{4.7}$$

$$\sum_{i=1}^{m+n} p_i \;=\; 1, \tag{4.8}$$

$$\sum_{i=1}^{m+n} p_i e^{\alpha+\beta T_i} \;=\; 1. \tag{4.9}$$

From (4.5) and $\sum_{j=1}^{n}\frac{(1-\lambda)+\lambda e^{\alpha+\beta Y_j}}{(1-\lambda)+\lambda e^{\alpha+\beta Y_j}} = n$, we have

$$\sum_{j=1}^{n}\frac{e^{\alpha+\beta Y_j}}{(1-\lambda)+\lambda e^{\alpha+\beta Y_j}} \;=\; \sum_{j=1}^{n}\frac{1}{(1-\lambda)+\lambda e^{\alpha+\beta Y_j}} \;=\; n,$$

and plugging it into (4.6) gives

$$n\lambda - t_2\sum_{i=1}^{m+n} p_i e^{\alpha+\beta T_i} \;=\; 0. \tag{4.10}$$

From (4.4) we get

$$(m+n) - t_1\sum_{i=1}^{m+n} p_i - t_2\sum_{i=1}^{m+n} p_i e^{\alpha+\beta T_i} \;=\; 0.$$

This together with (4.9) and (4.10) gives $t_2 = n\lambda$ and $t_1 = m+n-n\lambda$. Then by (4.4) again we have

$$p_i = \frac{1}{(m+n)\left[1+\rho_N\lambda\left(e^{\alpha+\beta T_i}-1\right)\right]},$$

where $\rho_N = n/(m+n)$ with $N = m+n$. Therefore ignoring a constant, the log-likelihood function is

$$l(\lambda,\alpha,\beta) \propto \sum_{j=1}^{n}\log[(1-\lambda)+\lambda e^{\alpha+\beta Y_j}] - \sum_{i=1}^{m+n}\log[1+\rho_N\lambda(e^{\alpha+\beta T_i}-1)]. \tag{4.11}$$

Maximizing (4.11) over (λ,α,β) gives the system of score functions

$$\frac{\partial l(\lambda,\alpha,\beta)}{\partial\lambda} = \sum_{j=1}^{n}\frac{e^{\alpha+\beta Y_j}-1}{1-\lambda+\lambda e^{\alpha+\beta Y_j}} - \sum_{i=1}^{m+n}\frac{\rho_N(e^{\alpha+\beta T_i}-1)}{1+\rho_N\lambda(e^{\alpha+\beta T_i}-1)} = 0, \tag{4.12}$$

$$\frac{\partial l(\lambda,\alpha,\beta)}{\partial\alpha} = \sum_{j=1}^{n}\frac{\lambda e^{\alpha+\beta Y_j}}{1-\lambda+\lambda e^{\alpha+\beta Y_j}} - \sum_{i=1}^{m+n}\frac{\rho_N\lambda e^{\alpha+\beta T_i}}{1+\rho_N\lambda(e^{\alpha+\beta T_i}-1)} = 0, \tag{4.13}$$

$$\frac{\partial l(\lambda,\alpha,\beta)}{\partial\beta} = \sum_{j=1}^{n}\frac{\lambda Y_j e^{\alpha+\beta Y_j}}{1-\lambda+\lambda e^{\alpha+\beta Y_j}} - \sum_{i=1}^{m+n}\frac{\rho_N\lambda T_i e^{\alpha+\beta T_i}}{1+\rho_N\lambda(e^{\alpha+\beta T_i}-1)} = 0. \tag{4.14}$$

Let $\hat{\theta}_{MLE} = (\hat{\lambda}_{MLE}, \hat{\alpha}_{MLE}, \hat{\beta}_{MLE})^\top$ be the MLE of θ, i.e. the maximizer of the log-likelihood function l in (4.11). Then the MLE of $p(y)$ in (1.3) is

$$\hat{p}_{MLE}(y) = \frac{\hat{\lambda}_{MLE}\exp[\hat{\alpha}_{MLE}+\hat{\beta}_{MLE}y]}{(1-\hat{\lambda}_{MLE})+\hat{\lambda}_{MLE}\exp[\hat{\alpha}_{MLE}+\hat{\beta}_{MLE}y]}. \tag{4.15}$$

Note that the system (4.12)-(4.14) does not yield an explicit solution for the semiparametric MLE $(\hat{\lambda},\hat{\alpha},\hat{\beta})$. Thus one has to compute iteratively by using the Newton-Raphson method or some variant.

4.3 Asymptotic properties

In this section, we develop asymptotic properties of the proposed MLE $\hat{\theta}_{MLE}$. The proofs of results in this section are very similar to Qin and Zhang (1997) but for a different model. We first present a lemma used to prove the asymptotic normality of $\hat{\theta}_{MLE}$. Let

$$S_N \;=\; -\frac{1}{N}\begin{pmatrix} \frac{\partial^2 l}{\partial\lambda^2} & \frac{\partial^2 l}{\partial\lambda\partial\alpha} & \frac{\partial^2 l}{\partial\lambda\partial\beta} \\ \\ \frac{\partial^2 l}{\partial\alpha\partial\lambda} & \frac{\partial^2 l}{\partial\alpha^2} & \frac{\partial^2 l}{\partial\alpha\partial\beta} \\ \\ \frac{\partial^2 l}{\partial\beta\partial\lambda} & \frac{\partial^2 l}{\partial\alpha\partial\beta} & \frac{\partial^2 l}{\partial\beta^2} \end{pmatrix} \tag{4.16}$$

with l defined in (4.11), and

$$S \;=\; \int \left(\frac{\partial w_1(y)}{\partial\theta}\right)\left(\frac{\partial w_1(y)}{\partial\theta}\right)^{\top} \frac{f}{w_1 w_2}(y)dy \;=\; \begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{12} & S_{22} & S_{23} \\ S_{13} & S_{23} & S_{33} \end{pmatrix}, \tag{4.17}$$

where

$$\begin{aligned} S_{11} \;&=\; \int (e^{\alpha+\beta y}-1)^2 \cdot \frac{f}{w_1 w_2}(y)\,dy, \\ S_{22} \;&=\; \lambda^2 \int e^{2\alpha+2\beta y} \cdot \frac{f}{w_1 w_2}(y)\,dy, \\ S_{33} \;&=\; \lambda^2 \int y^2 e^{2\alpha+2\beta y} \cdot \frac{f}{w_1 w_2}(y)\,dy, \\ S_{12} \;&=\; \lambda \int e^{\alpha+\beta y}(e^{\alpha+\beta y}-1) \cdot \frac{f}{w_1 w_2}(y)\,dy, \\ S_{13} \;&=\; \lambda \int y e^{\alpha+\beta y}(e^{\alpha+\beta y}-1) \cdot \frac{f}{w_1 w_2}(y)\,dy, \\ S_{23} \;&=\; \lambda^2 \int y e^{2\alpha+2\beta y} \cdot \frac{f}{w_1 w_2}(y)\,dy \end{aligned} \tag{4.18}$$

with

$$w_1(y) \;=\; 1-\lambda+\lambda e^{\alpha+\beta y}, \tag{4.19}$$

$$w_2(y) \;=\; 1-\rho\lambda+\rho\lambda e^{\alpha+\beta y}. \tag{4.20}$$

Lemma 4.1. *Assume $\rho_N \to \rho$ as $N \to \infty$. Then $S_N \xrightarrow{\mathscr{P}} \rho(1-\rho)S$ as $N \to \infty$, where S_N and S are defined in (4.16) and (4.17) respectively.*

Proof. Define

$$w_{2N}(y) \;=\; 1-\rho_N\lambda+\rho_N\lambda e^{\alpha+\beta y}. \tag{4.21}$$

The second-order partial derivatives of the log-likelihood function l in (4.11) are

$$\begin{aligned}
\frac{\partial^2 l}{\partial\lambda^2} &= -\sum_{j=1}^{n}\frac{(e^{\alpha+\beta Y_j}-1)^2}{w_1^2(Y_j)}+\sum_{i=1}^{m+n}\frac{\rho_N^2(e^{\alpha+\beta T_i}-1)^2}{w_{2N}^2(T_i)},\\
\frac{\partial^2 l}{\partial\alpha^2} &= \sum_{j=1}^{n}\frac{\lambda(1-\lambda)e^{\alpha+\beta Y_j}}{w_1^2(Y_j)}-\sum_{i=1}^{m+n}\frac{\rho_N\lambda(1-\rho_N\lambda)e^{\alpha+\beta T_i}}{w_{2N}^2(T_i)},\\
\frac{\partial^2 l}{\partial\beta^2} &= \sum_{j=1}^{n}\frac{\lambda(1-\lambda)Y_j^2e^{\alpha+\beta Y_j}}{w_1^2(Y_j)}-\sum_{i=1}^{m+n}\frac{\rho_N\lambda(1-\rho_N\lambda)T_i^2e^{\alpha+\beta T_i}}{w_{2N}^2(T_i)},\\
\frac{\partial^2 l}{\partial\lambda\partial\alpha} &= \sum_{j=1}^{n}\frac{e^{\alpha+\beta Y_j}}{w_1^2(Y_j)}-\sum_{i=1}^{m+n}\frac{\rho_N e^{\alpha+\beta T_i}}{w_{2N}^2(T_i)},\\
\frac{\partial^2 l}{\partial\lambda\partial\beta} &= \sum_{j=1}^{n}\frac{Y_je^{\alpha+\beta Y_j}}{w_1^2(Y_j)}-\sum_{i=1}^{m+n}\frac{\rho_N T_i e^{\alpha+\beta T_i}}{w_{2N}^2(T_i)},\\
\frac{\partial^2 l}{\partial\alpha\partial\beta} &= \sum_{j=1}^{n}\frac{\lambda(1-\lambda)Y_je^{\alpha+\beta Y_j}}{w_1^2(Y_j)}-\sum_{i=1}^{m+n}\frac{\rho_N\lambda(1-\rho_N\lambda)T_ie^{\alpha+\beta T_i}}{w_{2N}^2(T_i)}.
\end{aligned}$$

Straight calculation gives

$$E\left[-\frac{1}{N}\cdot\frac{\partial^2 l}{\partial\lambda^2}\right] \;=\; \rho_N(1-\rho_N)\int(e^{\alpha+\beta y}-1)^2\frac{f}{w_1w_{2N}}(y)\,dy \longrightarrow\; \rho(1-\rho)S_{11}.$$

By WLLN, $-\frac{1}{N}\cdot\frac{\partial^2 l}{\partial\lambda^2}\xrightarrow{\mathscr{P}}\rho(1-\rho)S_{11}$. Similarly we have the convergence of other components of the matrix S_N. $\square$

The following theorem gives the asymptotic normality of the MLE $\hat\theta_{MLE}$ that maximizes l in (4.11). Let

$$\begin{aligned}
V &= \int\left(\frac{\partial w_1(y)}{\partial\theta}\right)\left(\frac{\partial w_1(y)}{\partial\theta}\right)^{\top}\frac{f}{w_1w_2}(y)dy-\int\frac{\partial w_1(y)}{\partial\theta}\frac{f}{w_2}(y)dy\int\left(\frac{\partial w_1(y)}{\partial\theta}\right)^{\top}\frac{f}{w_2}(y)dy\\
&= S-\int\frac{\partial w_1(y)}{\partial\theta}\frac{f}{w_2}(y)dy\int\left(\frac{\partial w_1(y)}{\partial\theta}\right)^{\top}\frac{f}{w_2}(y)dy\\
&= \begin{pmatrix} V_{11} & V_{12} & V_{13}\\ V_{12} & V_{22} & V_{23}\\ V_{13} & V_{23} & V_{33}\end{pmatrix},
\end{aligned} \tag{4.22}$$

where

$$V_{11} = \int (e^{\alpha+\beta y}-1)^2 \frac{f}{w_1 w_2}(y)\,dy - \left[\int (e^{\alpha+\beta y}-1)\frac{f}{w_2}(y)\,dy\right]^2,$$

$$V_{22} = \lambda^2 \int e^{2\alpha+2\beta y} \frac{f}{w_1 w_2}(y)\,dy - \lambda^2 \left[\int e^{\alpha+\beta y}\frac{f}{w_2}(y)\,dy\right]^2,$$

$$V_{33} = \lambda^2 \int y^2 e^{2\alpha+2\beta y} \frac{f}{w_1 w_2}(y)\,dy - \lambda^2 \left[\int y e^{\alpha+\beta y}\frac{f}{w_2}(y)\,dy\right]^2,$$

$$V_{12} = \lambda \int e^{\alpha+\beta y}(e^{\alpha+\beta y}-1) \frac{f}{w_1 w_2}(y)\,dy - \lambda \int (e^{\alpha+\beta y}-1)\frac{f}{w_2}(y)\,dy \int e^{\alpha+\beta y}\frac{f}{w_2}(y)\,dy,$$

$$V_{13} = \lambda \int y e^{\alpha+\beta y}(e^{\alpha+\beta y}-1) \frac{f}{w_1 w_2}(y)\,dy - \lambda \int (e^{\alpha+\beta y}-1)\frac{f}{w_2}(y)\,dy \int y e^{\alpha+\beta y}\frac{f}{w_2}(y)\,dy,$$

$$V_{23} = \lambda^2 \int y e^{2\alpha+2\beta y} \frac{f}{w_1 w_2}(y)\,dy - \lambda^2 \int e^{\alpha+\beta y}\frac{f}{w_2}(y)\,dy \int y e^{\alpha+\beta y}\frac{f}{w_2}(y)\,dy$$

with w_1 and w_2 defied in (4.19) and (4.20) respectively.

Theorem 4.2. *Assume $\rho_N \to \rho$ as $N \to \infty$. Then under some regularity conditions (for MLE in general),*

$$\sqrt{N}\begin{pmatrix} \hat{\lambda}_{MLE}-\lambda \\ \hat{\alpha}_{MLE}-\alpha \\ \hat{\beta}_{MLE}-\beta \end{pmatrix} \xrightarrow{\mathscr{L}} N(0,\Sigma),$$

where $\Sigma = \frac{1}{\rho(1-\rho)} S^{-1} V S^{-1}$ with S and V defined in (4.17) and (4.22) respectively.

Proof. Let $Q_N = \frac{1}{N}\left(\frac{\partial l}{\partial \lambda}, \frac{\partial l}{\partial \alpha}, \frac{\partial l}{\partial \beta}\right)^\top$, where $\frac{\partial l}{\partial \lambda}$, $\frac{\partial l}{\partial \alpha}$ and $\frac{\partial l}{\partial \beta}$ are given in (4.12), (4.13) and

(4.14) respectively. Then $E[Q_N]=0$. From (4.12) we have, as $N\to\infty$,

$$\begin{aligned}
&\frac{1}{N}Var\left[\frac{\partial l}{\partial\lambda}\right]\\
=\ &\frac{1}{N}Var\left[\sum_{j=1}^{n}\left(\frac{e^{\alpha+\beta Y_j}-1}{w_1(Y_j)}-\frac{\rho_N(e^{\alpha+\beta Y_j}-1)}{w_{2N}(Y_j)}\right)-\sum_{i=1}^{m}\frac{\rho_N(e^{\alpha+\beta X_i}-1)}{w_{2N}(X_i)}\right]\\
=\ &\rho_N Var\left[\frac{(1-\rho_N)(e^{\alpha+\beta Y_1}-1)}{w_1(Y_1)w_{2N}(Y_1)}\right]+(1-\rho_N)Var\left[\frac{\rho_N(e^{\alpha+\beta X_1}-1)}{w_{2N}(X_1)}\right]\\
=\ &\rho_N(1-\rho_N)^2\left\{\int\frac{(e^{\alpha+\beta y}-1)^2}{w_1^2(y)w_{2N}^2(y)}w_1(y)f(y)dy-\left[\int\frac{e^{\alpha+\beta y}-1}{w_1(y)w_{2N}(y)}w_1(y)f(y)dy\right]^2\right\}\\
&+\rho_N^2(1-\rho_N)\left\{\int\frac{(e^{\alpha+\beta y}-1)^2}{w_{2N}^2(y)}f(y)dy-\left[\int\frac{e^{\alpha+\beta y}-1}{w_{2N}(y)}f(y)dy\right]^2\right\}\\
=\ &\rho_N(1-\rho_N)\left[\int(e^{\alpha+\beta y}-1)^2\frac{f}{w_1w_{2N}}(y)dy-\left[\int(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}}(y)dy\right]^2\right]\\
\to\ &\rho(1-\rho)V_{11}.
\end{aligned}$$

Similarly we have $\frac{1}{N}Var\left[\frac{\partial l}{\partial\alpha}\right]\to\rho(1-\rho)V_{22}$ and $\frac{1}{N}Var\left[\frac{\partial l}{\partial\beta}\right]\to\rho(1-\rho)V_{33}$ as $N\to\infty$.
From (4.12) and (4.13) we have

$$\begin{aligned}
&\frac{1}{N}Cov\left[\frac{\partial l}{\partial\lambda},\frac{\partial l}{\partial\alpha}\right]\\
=\ &\frac{1}{N}E\left[\frac{\partial l}{\partial\lambda}\cdot\frac{\partial l}{\partial\alpha}\right]\\
=\ &\frac{1}{N}E\left[\left\{\sum_{j=1}^{n}\left(\frac{e^{\alpha+\beta Y_j}-1}{w_1(Y_j)}-\frac{\rho_N(e^{\alpha+\beta Y_j}-1)}{w_{2N}(Y_j)}\right)-\sum_{i=1}^{m}\frac{\rho_N(e^{\alpha+\beta X_i}-1)}{w_{2N}(X_i)}\right\}\right.\\
&\qquad\left.\cdot\left\{\sum_{j=1}^{n}\left(\frac{\lambda e^{\alpha+\beta Y_j}}{w_1(Y_j)}-\frac{\rho_N\lambda e^{\alpha+\beta Y_j}}{w_{2N}(Y_j)}\right)-\sum_{i=1}^{m}\frac{\rho_N\lambda e^{\alpha+\beta X_i}}{w_{2N}(X_i)}\right\}\right]\\
=\ &\frac{1}{N}E\left[\left\{(1-\rho_N)\sum_{j=1}^{n}\frac{e^{\alpha+\beta Y_j}-1}{w_1(Y_j)w_{2N}(Y_j)}-\rho_N\sum_{i=1}^{m}\frac{e^{\alpha+\beta X_i}-1}{w_{2N}(X_i)}\right\}\right.\\
&\qquad\left.\cdot\left\{(1-\rho_N)\lambda\sum_{j=1}^{n}\frac{e^{\alpha+\beta Y_j}}{w_1(Y_j)w_{2N}(Y_j)}-\rho_N\lambda\sum_{i=1}^{m}\frac{e^{\alpha+\beta X_i}}{w_{2N}(X_i)}\right\}\right]\\
=\ &\frac{1}{N}E\left[(A-B)(C-D)\right],\ \text{say}\\
=\ &\frac{1}{N}\left\{E[AC]+E[BD]-E[A]E[D]-E[B]E[C]\right\},
\end{aligned}$$

where

$$
\begin{aligned}
E[AC] &= (1-\rho_N)^2\lambda\left\{nE\left[\frac{e^{\alpha+\beta Y_1}(e^{\alpha+\beta Y_1}-1)}{w_1^2(Y_1)w_{2N}^2(Y_1)}\right]\right. \\
&\qquad\left. + n(n-1)E\left[\frac{e^{\alpha+\beta Y_1}-1}{w_1(Y_1)w_{2N}(Y_1)}\right]E\left[\frac{e^{\alpha+\beta Y_1}}{w_1(Y_1)w_{2N}(Y_1)}\right]\right\} \\
&= n(1-\rho_N)^2\lambda\left\{\int e^{\alpha+\beta y}(e^{\alpha+\beta y}-1)\frac{f}{w_1 w_{2N}^2}(y)\,dy\right. \\
&\qquad\left. + (n-1)\int(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}}(y)\,dy\int e^{\alpha+\beta y}\frac{f}{w_{2N}}(y)\,dy\right\},
\end{aligned}
$$

$$
\begin{aligned}
E[BD] &= m\rho_N^2\lambda\left\{\int e^{\alpha+\beta y}(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}^2}(y)\,dy\right. \\
&\qquad\left. + (m-1)\int(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}}(y)\,dy\int e^{\alpha+\beta y}\frac{f}{w_{2N}}(y)\,dy\right\},
\end{aligned}
$$

$$
E[A]E[D] = E[B]E[C] = mn\rho_N(1-\rho_N)\lambda\int(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}}(y)\,dy\int e^{\alpha+\beta y}\frac{f}{w_{2N}}(y)\,dy,
$$

and thus as $N\to\infty$,

$$
\begin{aligned}
\frac{1}{N}Cov\left[\frac{\partial l}{\partial\lambda},\frac{\partial l}{\partial\alpha}\right] &= \rho_N(1-\rho_N)^2\lambda\int e^{\alpha+\beta y}(e^{\alpha+\beta y}-1)\frac{f}{w_1 w_{2N}^2}(y)\,dy \\
&\quad + (n-1)\rho_N(1-\rho_N)^2\lambda\int(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}}(y)\,dy\int e^{\alpha+\beta y}\frac{f}{w_{2N}}(y)\,dy \\
&\quad + \rho_N^2(1-\rho_N)\lambda\int e^{\alpha+\beta y}(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}^2}(y)\,dy \\
&\quad + (m-1)\rho_N^2(1-\rho_N)\lambda\int(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}}(y)\,dy\int e^{\alpha+\beta y}\frac{f}{w_{2N}}(y)\,dy \\
&\quad - 2m\rho_N^2(1-\rho_N)\lambda\int(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}}(y)\,dy\int e^{\alpha+\beta y}\frac{f}{w_{2N}}(y)\,dy \\
&= \rho_N(1-\rho_N)\lambda\left[\int e^{\alpha+\beta y}(e^{\alpha+\beta y}-1)\frac{f}{w_1 w_{2N}^2}(y)\,dy\right. \\
&\qquad\left. - \int(e^{\alpha+\beta y}-1)\frac{f}{w_{2N}}(y)\,dy\int e^{\alpha+\beta y}\frac{f}{w_{2N}}(y)\,dy\right] \\
&\to \rho(1-\rho)V_{12}.
\end{aligned}
$$

Similarly we have $\frac{1}{N}Cov\left[\frac{\partial l}{\partial\lambda},\frac{\partial l}{\partial\beta}\right] \to \rho(1-\rho)V_{13}$ and $\frac{1}{N}Cov\left[\frac{\partial l}{\partial\alpha},\frac{\partial l}{\partial\beta}\right] \to \rho(1-\rho)V_{23}$ as $N\to\infty$. Thus by CLT

$$
\sqrt{N}Q_N \xrightarrow{\mathscr{L}} N(0,\rho(1-\rho)V).
$$

From Lemma 4.1 along with Slutsky's theorem, hence the result. □

4.4 Simulation studies

In this section, we examine the finite-sample performance of our proposed MLE $\hat{\theta}_{MLE}$ through a Monte Carlo simulation study. We consider the same mixture models listed in Table 2.1 of Section 2.4. For each model, the true values of α and β are derived under (4.2). For example, if we consider a mixture of two normals $f \sim N(0,1)$ and $g \sim N(\mu,\sigma)$, then $\alpha = -\frac{1}{2}(\log\sigma^2 + \frac{\mu}{\sigma^2})$ and $\beta = \frac{\mu}{\sigma^2}$. Similarly for mixture of two Poisson distributions $f \sim Po(\mu_1)$ and $g \sim Po(\mu_2)$, the values of α and β are $\alpha = \mu_1 - \mu_2$ and $\beta = \log\frac{\mu_2}{\mu_1}$ respectively. Note that model M5 does not satisfy (4.3) since the two components have different support. The true values of α and β for each model are given in Table 4.1. From Table 4.1 we can see that always $\beta > 0$ for each of those four models, M1-M4, for which the two components have common support. This demonstrate that $\beta > 0$ generally implies the stochastic dominance (Theorem 4.1), even for discrete populations.

Table 4.1: Mixture models considered in simulation study.

Model	Form	α	β
M1	$(1-\lambda)N(0,1)+\lambda N(1,1)$	-0.5	1
M2	$(1-\lambda)N(0,1)+\lambda N(5,1)$	-12.5	5
M3	$(1-\lambda)Po(2)+\lambda Po(4)$	-2	0.693
M4	$(1-\lambda)Po(2)+\lambda Po(6)$	-4	1.099
M5	$(1-\lambda)U(0,4)+\lambda U(2,6)$	NA	NA

For each model we consider the same varying mixing proportion values as in Sections 2.4 and 3.3. We also use the same sample sizes $m = n = 30$ and $m = n = 100$ and the same number of replications $N = 1000$. To find the MLE $\hat{\theta}_{MLE}$ that maximizes the log-likelihood function l in (4.11), we use package“optim” in `R` statistical software. For simplicity, we use λ_+ given in (2.12) as the initial estimate of λ. Initial value for α and β are calculated by exploiting the relationship in (4.3). Specifically, (4.3) indicates

$$\log\frac{h(x)/f(x)-(1-\lambda)}{\lambda} \;=\; \alpha+\beta x.$$

Thus for each T_i in the pooled sample, we generate the pair (T_i, R_i), where $R_i = \log \frac{h_n(T_i)/f_m(T_i)-(1-\lambda_+)}{\lambda_+}$ with f_m, h_n and λ_+ defined in (2.3), (2.4) and (2.12) respectively. Finally we use (T_i, R_i), $i = 1, \ldots, N$, to fit a least-squares regression line and the fitted coefficients will be used as the initials of α and β. As in Sections 2.4 and 3.3, for each model we calculate the bias and MSE of $\hat{\lambda}_{MLE}$ and the misclassification rate MR when the same classification rule as in Sections 2.4 and 3.3 is used. We also calculate the coverage probability (CP) of the 95% confidence interval constructed based on $\hat{\lambda}_{MLE}$ using the asymptotic variance given in Theorem 4.2. We plug in the MLE $\hat{\lambda}_{MLE}$, $\hat{\alpha}_{MLE}$, $\hat{\beta}_{MLE}$ and the kernel estimator f_m into the expression to calculate the matrix Σ. Since λ is our main interest we don't give the CP for $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$. The simulation results are presented in Table 4.2.

From Table 4.2 we can see that, as expected, $\hat{\theta}_{MLE}$ has smaller bias, MSE and MR for larger sample sizes than for smaller sample sizes. The $\hat{\lambda}_{MLE}$ always gives small bias and MSE, especially for larger sample sizes, while $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$ generally give relatively large bias and MSE even for larger sample sizes. Nevertheless, the MR is reasonably close to OMR, the optimal misclassification rate assuming the probability function p in (1.3) is known, regardless of sample size. Even for M5 where the assumption (4.2) doesn't hold, the MLE of λ based on (4.2) performs surprisingly well and the MR doesn't deviate from OMR too much for large sample sizes. We also observe that when the two components are close to each other in terms of location (M1 and M3), the estimated mixing proportion $\hat{\lambda}_{MLE}$ has larger bias and MSE than the cases when the two components are far apart (M2 and M4). However, the bias and MSE of $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$ for M2 are larger than those for M1. This could be explained by the larger magnitude of the true α and β values for M2 than for M1. The CP of the confidence interval based on $\hat{\lambda}_{MLE}$ is close to the nominal level of 95% for most of the cases.

When the MLE $\hat{\lambda}_{MLE}$ we proposed in this chapter based on the assumption (4.2) is compared with the two nonparametric estimators, $\hat{\lambda}$ in Chapter 2 and $\hat{\lambda}_L$ in Chapter 3, without this assumption, we observe that the three perform quite competitively while $\hat{\lambda}_{MLE}$ and $\hat{\lambda}$ have a little better MR than $\hat{\lambda}_L$.

Table 4.2: Bias and MSE of $\hat{\theta}_{MLE}$, CP (%) of $\hat{\lambda}_{MLE}$ and MR (%) of a classification rule based on $\hat{p}_{MLE}$.

Model	λ	$m=n=30$					$m=n=100$					
		$Bias(\hat{\lambda}_{MLE})$ $(MSE(\hat{\lambda}_{MLE}))$	$Bias(\hat{\alpha}_{MLE})$ $(MSE(\hat{\alpha}_{MLE}))$	$Bias(\hat{\beta}_{MLE})$ $(MSE(\hat{\beta}_{MLE}))$	CP	MR	$Bias(\hat{\lambda}_{MLE})$ $(MSE(\hat{\lambda}_{MLE}))$	$Bias(\hat{\alpha}_{MLE})$ $(MSE(\hat{\alpha}_{MLE}))$	$Bias(\hat{\beta}_{MLE})$ $(MSE(\hat{\beta}_{MLE}))$	CP	MR	OMR
	0.05	0.139 (0.091)	-0.204 (0.198)	0.006 (0.436)	100.0	18.53	0.066 (0.036)	-0.133 (0.184)	0.025 (0.427)	96.2	11.04	4.99
M1	0.20	0.093 (0.106)	-0.156 (0.183)	0.174 (0.469)	87.4	31.20	0.053 (0.059)	-0.072 (0.159)	0.174 (0.407)	88.9	27.04	18.61
$\alpha=-0.5$	0.50	0.035 (0.102)	-0.118 (0.154)	0.225 (0.376)	100.0	38.04	0.051 (0.062)	-0.079 (0.123)	0.094 (0.198)	88.1	37.29	30.85
$\beta=1$	0.80	-0.024 (0.056)	-0.125 (0.116)	0.200 (0.245)	96.1	24.50	-0.006 (0.029)	-0.086 (0.079)	0.095 (0.113)	97.8	22.23	18.61
	0.95	-0.062 (0.032)	-0.147 (0.096)	0.219 (0.205)	94.3	10.53	-0.041 (0.014)	-0.098 (0.054)	0.116 (0.082)	98.3	8.14	4.99
	0.05	0.029 (0.005)	0.123 (0.943)	0.255 (1.125)	90.3	1.13	0.008 (0.001)	0.053 (0.865)	0.594 (0.654)	92.2	0.69	0.24
M2	0.20	0.019 (0.007)	-0.073 (0.908)	0.581 (0.648)	95.3	1.24	-0.003 (0.004)	-0.330 (0.873)	0.564 (0.560)	93.7	0.90	0.48
$\alpha=-12.5$	0.50	0.004 (0.014)	-0.197 (0.889)	0.564 (0.594)	95.9	1.27	0.001 (0.005)	-0.409 (0.889)	0.411 (0.441)	94.8	0.91	0.62
$\beta=5$	0.80	-0.008 (0.014)	-0.352 (0.900)	0.500 (0.532)	95.7	0.53	0.002 (0.002)	-0.383 (0.899)	0.316 (0.386)	90.6	0.61	0.48
	0.95	0.002 (0.002)	-0.558 (0.901)	0.428 (0.471)	95.3	0.37	0.001 (0.001)	-0.366 (0.908)	0.255 (0.331)	94.8	0.29	0.24
	0.05	0.058 (0.055)	0.544 (1.104)	0.063 (0.641)	89.6	13.63	0.019 (0.019)	0.564 (1.082)	0.078 (0.536)	97.9	7.61	4.76
M3	0.20	-0.019 (0.089)	0.089 (1.296)	0.427 (0.734)	89.9	25.47	-0.026 (0.045)	0.084 (1.292)	0.399 (0.529)	93.6	21.75	13.90
$\alpha=-2$	0.50	-0.091 (0.148)	-0.392 (1.698)	0.504 (0.599)	80.4	36.13	-0.022 (0.078)	-0.319 (1.29)	0.246 (0.296)	92.5	34.89	19.05
$\beta=0.693$	0.80	-0.109 (0.119)	-0.682 (1.686)	0.335 (0.331)	90.2	26.97	-0.023 (0.036)	-0.353 (0.851)	0.119 (0.091)	98.4	22.87	13.35
	0.95	-0.139 (0.093)	-0.774 (1.529)	0.301 (0.241)	95.3	15.73	-0.057 (0.018)	-0.392 (0.622)	0.130 (0.055)	100.0	9.55	6.06
	0.05	-0.003 (0.012)	0.651 (0.889)	0.029 (0.705)	84.5	6.70	-0.024 (0.004)	0.648 (0.765)	0.201 (0.565)	81.9	4.41	3.14
M4	0.20	-0.084 (0.037)	0.255 (0.844)	0.512 (0.492)	88.0	15.53	-0.080 (0.025)	0.189 (0.789)	0.450 (0.389)	94.2	11.14	7.03
$\alpha=-4$	0.50	-0.092 (0.087)	-0.179 (0.849)	0.337 (0.271)	82.9	19.00	-0.018 (0.025)	-0.148 (0.743)	0.115 (0.097)	96.2	16.49	10.19
$\beta=1.099$	0.80	-0.027 (0.051)	-0.263 (0.791)	0.144 (0.109)	98.4	14.23	0.004 (0.007)	-0.163 (0.618)	0.053 (0.039)	94.0	13.00	7.82
	0.95	-0.027 (0.021)	-0.375 (0.673)	0.131 (0.075)	99.4	6.07	-0.007 (0.002)	-0.223 (0.468)	0.066 (0.033)	96.1	5.18	3.21
	0.05	0.166 (0.177)	NA	NA	81.3	25.97	0.082 (0.099)	NA	NA	86.2	15.31	2.5
	0.20	-0.002 (0.114)	NA	NA	88.0	23.04	-0.093 (0.040)	NA	NA	94.9	16.59	10
M5	0.50	-0.121 (0.135)	NA	NA	84.4	29.83	-0.132 (0.057)	NA	NA	84.6	26.38	25
	0.80	-0.055 (0.076)	NA	NA	94.4	22.47	-0.048 (0.032)	NA	NA	93.0	20.16	10
	0.95	-0.022 (0.018)	NA	NA	99.7	8.23	-0.006 (0.005)	NA	NA	95.2	5.92	2.5

Chapter 5

Semiparametric Estimation II: MHDE

In this chapter we still assume (4.2) or equivalently the semiparametric mixture model (4.3) with $r(x) = x$ and $\beta > 0$. In last chapter we construct the MLE for (4.3) which is asymptotically normally distributed and performs well for finite sample sizes. However, MLE is generally not robust against outliers and model misspecification. To achieve robustness, we propose in this chapter a minimum Hellinger distance estimation (MHDE). In Section 5.1 we review the general definition of MHDE and construct the MHDE specifically for the two-sample semiparametric mixture model (4.3). In Section 5.2 we present its asymptotic properties, such as consistency and asymptotic normality. Through simulation studies, Section 5.3 gives the efficiency study of the proposed MHDE while Section 5.4 is devoted to its robustness study and comparison with other estimators. Finally Section 5.5 gives detailed proofs of the asymptotic properties presented in Section 5.2.

5.1 Construction of MHDE

Due to its excellent robustness properties and simultaneous efficiency, MHDE has been popular in practice. The Hellinger distance between two functions f_1 and f_2 is defined as $\|f_1^{1/2} - f_2^{1/2}\|$, the L^2-norm of root functions. For a fully parametric model $\{h_\theta : \theta \in \Theta\}$ with Θ the parameter space, the MHDE of θ is defined as the value $\hat{\theta}_{MHDE}$ that minimizes the Hellinger distance between the parametric model and an appropriate nonparametric density estimator, say, $\hat{h}$ based on data, i.e.

$$\hat{\theta}_{MHDE} = \arg\min_{t \in \Theta} \left\| h_t^{1/2} - \hat{h}^{1/2} \right\|. \tag{5.1}$$

MHDE was first introduced by Beran (1977) for this fully parametric model of general form. Beran (1977) showed that the MHDE for parametric model has both full efficiency and good robustness properties. Lindsay (1994) outlined the comparison between MHDE and MLE in terms of robustness and efficiency and showed that MHDE and MLE are members of a larger class of efficient estimators with various second-order efficiency properties. However, the literature on MHDE for mixture models is not redundant. Lu, Hui and Lee (2003) considered the MHDE for mixture of Poisson regression models. MHDE of mixture complexity for finite mixture models was investigated by Woo and Sriram (2006 & 2007). Recently, MHDE has been extended from parametric models to semiparametric models. Wu, Karunamuni and Zhang (2010) proposed a MHDE for two-sample case-control data under model (4.2) and investigated the asymptotic properties and robustness of the proposed estimator. Xiang, Yao and Wu (2014) proposed a minimum profile Hellinger distance estimation (MPHDE) for two-component semiparamtric mixture models studied by Bordes, Delmas and Vandekerkhove (2006a) where one component is known and the other is an unknown symmetric function with unknown location parameter. Wu, Yao and Xiang (2017) proposed an algorithm for the MPHDE in two-component semiparametric location-shifted mixture models. Inspired by these works, we propose in this chapter to use the MHDE to estimate the parameters in (4.3).

In model (4.3), even though α and β can possibly take any value on real line, we can essentially use intervals that are large enough to cover their true values. So for practical purpose, without loss of generality we can assume that $\theta \in \Theta$ with Θ a compact subset of $\mathbb{R}^3$. To give the MHDE for model (4.3), note that the MHDE defined in (5.1) is not available in practice since the f in $h_t(x) = (1 - t_1 + t_1 e^{t_2 + t_3 x}) f(x)$, with $t = (t_1, t_2, t_3)^\top$, is unknown. Intuitively, we can use the kernel density estimator f_m given in (2.3) to replace f and apply the plug-in rule to give an estimated parametric model

$$\hat{h}_t(x) \;=\; (1 - t_1 + t_1 e^{t_2 + t_3 x}) f_m(x), \tag{5.2}$$

We can use the kernel density estimator h_n given in (2.4) as the nonparametric estimator of h. Now we define the MHDE of $\theta = (\lambda, \alpha, \beta)^\top$ as

$$\hat{\theta}_{MHDE} \;=\; T(f_m, h_n) \;=\; \arg\min_{t \in \Theta} \left\| \hat{h}_t^{1/2} - h_n^{1/2} \right\|. \tag{5.3}$$

That is, $\hat{\theta}_{MHDE}$ is the minimizer t of the Hellinger distance between the estimated parametric model $\hat{h}_t$ and the nonparametric density estimator h_n. Then the MHDE of $p(y)$ in (1.3) is given by

$$\hat{p}_{MHDE}(y) = \frac{\hat{\lambda}_{MHDE}\exp[\hat{\alpha}_{MHDE} + \hat{\beta}_{MHDE}y]}{(1-\hat{\lambda}_{MHDE}) + \hat{\lambda}_{MHDE}\exp[\hat{\alpha}_{MHDE} + \hat{\beta}_{MHDE}y]}.$$

This MHDE defined in (5.3) is similar to Beran's (1977) original mechanism of obtaining MHDE for fully parametric models. Thus, we would expect $\hat{\theta}_{MHDE}$ to have good robustness and asymptotic efficiency properties. Note that in (5.3) we do not impose any restriction on $\hat{h}_t$ to make it a density function, i.e. $\int \hat{h}_t(x)dx = 1$. The reason behind that is, even though for a particular $t \in \Theta$ such that $\hat{h}_t$ is not a density, it could make h_t a density. The true parameter value θ may not make $\hat{h}_\theta$ a density, but it is not reasonable to exclude θ as the estimate $\hat{\theta}_{MHDE}$ of itself. As the explicit expression of $\hat{\theta}_{MHDE}$ does not exist, one needs to use iterative methods such as Newton-Raphson to numerically calculate it. Karunamuni and Wu (2011) has shown that with appropriate initial value, even a one-step iteration will work well and give a quite accurate approximation of MHDE.

5.2 Asymptotic properties

In this section we investigate the asymptotic properties of our proposed $\hat{\theta}_{MHDE}$ given in (5.3). The results and their proofs in this section are very similar to Wu, Karunamuni and Zhang (2010) but for a different model. For completeness we still present them here even though we follow exactly the same lines as in Wu, Karunamuni and Zhang (2010).

Let $\mathscr{H}$ be the set of all c.d.f.s with respect to Lebesgue measure on the real line. We decompose the parameter vector θ into two parts

$$\theta = (\lambda, \theta_r^\top)^\top,$$

where $\theta_r = (\alpha, \beta)^\top$ represents the regression coefficient parameters in (4.3). Note that $g_{\theta_r}(x) = e^{\alpha+\beta x} f(x)$ is essentially the g in (4.2). Parallelly for each $t \in \Theta$ we write $t = (t_1, t_r^\top)^\top$ with

$t_r = (t_2, t_3)^\top$ and $g_{t_r}(x) = e^{t_2+t_3x} f(x)$. We first list some conditions that will be used for later proof of asymptotics.

(D1) There exists an ε-neighbourhood $B(\theta_r, \varepsilon)$ of θ_r for some $\varepsilon \geq 0$ such that $g_{t_r} - g_{\theta_r}$ is bounded by an integrable function for any $t_r \in B(\theta_r, \varepsilon)$.

(D2) f and K_0 in (4.3) and (2.3) respectively have compact supports.

(D3) f in (4.3) has infinite support, K_0 in (2.3) is a bounded symmetric density with support $[-a_0, a_0]$ for some $0 < a_0 < \infty$, and there exists a sequence α_m of positive numbers such that as $m \to \infty$, $\alpha_m \to \infty$ and

$$\sup_{\theta \in \Theta} \int I_{\{|x|>\alpha_m\}} h_\theta(x) dx \;\longrightarrow\; 0, \tag{5.4}$$

$$b_m^2 \sup_{\theta \in \Theta} \int I_{\{|x|>\alpha_m\}} h_\theta(x) \sup_{|t| \leq a_0} \frac{|f^{(2)}(x) + t b_m|}{f(x)} dx \;\to\; 0, \tag{5.5}$$

$$m^{-1} b_m^{-1} \sup_{\theta \in \Theta} \int I_{\{|x|\leq\alpha_m\}} h_\theta(x) \sup_{|t| \leq a_0} \frac{f(x + t b_m)}{f^2(x)} dx \;\to\; 0, \tag{5.6}$$

$$b_m^4 \sup_{\theta \in \Theta} \int I_{\{|x|\leq\alpha_m\}} h_\theta(x) \sup_{|t| \leq a_0} \left[\frac{f^{(2)}(x + t b_m)}{f(x)}\right]^2 dx \;\to\; 0, \tag{5.7}$$

where $f^{(k)}$ denotes that k^{th} derivative of f.

Lemma 5.1. *If (D1) holds for $\theta \in \Theta$, then $d(t) = \|h_t^{1/2} - \varphi^{1/2}\|$ is continuous at point $t = \theta$, for any $\varphi \in \mathscr{H}$.*

Proof. Suppose $\theta_k \to \theta$ as $k \to \infty$. From Minkowski's inequality,

$$|d(\theta_k) - d(\theta)| \;\leq\; \left\| h_{\theta_k}^{1/2} - h_\theta^{1/2} \right\| \;\leq\; \left[\int \left| h_{\theta_k}(x) - h_\theta(x) \right| dx \right]^{1/2}. \tag{5.8}$$

Note that

$$\begin{aligned} \left| h_{\theta_k}(x) - h_\theta(x) \right| &= \left| -(\lambda_k - \lambda) f(x) + \lambda_k f(x)(e^{\alpha_k+\beta_k x} - e^{\alpha+\beta x}) + (\lambda_k - \lambda) e^{\alpha+\beta x} f(x) \right| \\ &\leq f(x) + \left| f(x)(e^{\alpha_k+\beta_k x} - e^{\alpha+\beta x}) \right| + e^{\alpha+\beta x} f(x). \end{aligned}$$

By (D1), $f(x)(e^{\alpha_k+\beta_k x}-e^{\alpha+\beta x})$ is bounded by an integrable function, say $B(x)$, and as a result $\left|h_{\theta_k}(x)-h_\theta(x)\right|$ is bounded by integrable function $f(x)+B(x)+g(x)$. Therefore by the DCT we have $\int |h_{\theta_k}(x)-h_\theta(x)|dx \to 0$ as $k\to\infty$, i.e., $d(\theta_k)\to d(\theta)$ as $k\to\infty$ and $d(t)$ is continuous at point $t=\theta$. □

Theorem 5.1. *Suppose* $(D1)$ *holds for all* $t\in\Theta$. *Then*

(i) For every $\varphi\in\mathscr{H}$, *there exist* $T(f,\varphi)$ *and* $T(f_m,\varphi)$ *in* Θ *satisfying (5.3), where* f_m *is defined in (2.3) with the kernel* K_0 *compactly supported.*

(ii) Suppose that $m,n\to\infty$ *as* $N\to\infty$ *and* $\theta=T(f,\varphi)$ *is unique. Then* $\theta_N=T(f_m,\varphi_n)\to\theta$ *as* $N\to\infty$ *for any density sequences* f_m *and* φ_n *such that* $\|\varphi_n^{1/2}-\varphi^{1/2}\|\to 0$ *and* $\sup_{t\in\Theta}\|\hat{h}_t^{1/2}-h_t^{1/2}\|\to 0$ *as* $N\to\infty$ *with* $\hat{h}_t$ *given in (5.2).*

(iii) $T(f,h_\theta)=\theta$ *uniquely for any* $\theta\in\Theta$.

Proof. (i) Let $d_m(t)=\|\hat{h}_t^{1/2}-\varphi^{1/2}\|$. Suppose sequence $\{t_k\}\subset\Theta$ such that $t_k\to t$ as $k\to\infty$. Since Θ is compact, $t\in\Theta$. Similar to (5.8), we have

$$|d_m(t_k)-d_m(t)| \;\le\; \left[\int\left|\lambda_k-\lambda-\lambda_k e^{\alpha_k+\beta_k x}+\lambda e^{\alpha+\beta x}\right| f_m(x)dx\right]^{1/2}.$$

Since f_m is compactly supported, we have by the DCT that $d_m(t_k)\to d_m(t)$ as $k\to\infty$, i.e. $d_m(t)$ is continuous and achieves a minimum over $t\in\Theta$. Let $d(t)=\|h_t^{1/2}-\varphi^{1/2}\|$. By Lemma 5.1, $d(t)$ is continuous in t and therefore achieves a minimum over $t\in\Theta$.

(ii) Suppose $\|\varphi_n^{1/2}-\varphi^{1/2}\|\to 0$ and $\sup_{t\in\Theta}\|\hat{h}_t^{1/2}-h_t^{1/2}\|\to 0$ as $N\to\infty$. Let $d_N(t)=\|\hat{h}_t^{1/2}-\varphi_n^{1/2}\|$ and $d(t)=\|h_t^{1/2}-\varphi^{1/2}(x)\|$. By Minkowski's inequality

$$\begin{aligned}|d_N(t)-d(t)| &\le \left\{\int\left[\hat{h}_t^{1/2}(x)-\varphi_n^{1/2}(x)-h_t^{1/2}(x)+\varphi^{1/2}(x)\right]^2dx\right\}^{1/2}\\ &\le \left\{2\int\left[\hat{h}_t^{1/2}(x)-h_t^{1/2}\right]^2dx+2\int\left[\varphi_n^{1/2}-\varphi^{1/2}(x)\right]^2dx\right\}^{1/2}\end{aligned}$$

and consequently $\sup_{t\in\Theta}|d_N(t)-d(t)|\to 0$ as $N\to\infty$. Thus as $N\to\infty$, $d_N(\theta)\to d(\theta)$ and $d_N(\theta_N)-d(\theta_N)\to 0$. If $\theta_N\nrightarrow\theta$, then there exists a subsequence $\{\theta_{N_i}\}\subseteq\{\theta_N\}$ such that, $\theta_{N_i}\to\theta'\neq\theta$. Since Θ is compact, $\theta'\in\Theta$. Lemma 5.1 yields that $d(\theta_{N_i})\to d(\theta')$. From the above results we have $d_{N_i}(\theta_{N_i}) - d_{N_i}(\theta) \to d(\theta^{'}) - d(\theta)$. By the definition of θ_{N_i}, $d_{N_i}(\theta_{N_i}) - d_{N_i}(\theta) \le 0$. Hence, $d(\theta^{'}) - d(\theta) \le 0$. But by the definition and uniqueness of θ, $d(\theta^{'}) > d(\theta)$. This is a contradiction. Therefore $\theta_N \to \theta$.

(iii) Since by Theorem 4.1 $\{h_t\}_{t\in\Theta}$ is identifiable, we have $T(f, h_\theta) = \theta$ uniquely for any $\theta \in \Theta$. □

In order to prove the consistency of our proposed $\hat{\theta}_{MHDE}$ given in (5.3), we need the following lemma.

Lemma 5.2. *Suppose (D3) holds. Then as $m \to \infty$,*

$$\sup_{\theta\in\Theta} \int w_1(x) \left[f_m^{1/2}(x) - f^{1/2}(x) \right]^2 dx \xrightarrow{\mathscr{P}} 0.$$

Proof. By the continuity of the function w_1 in θ and the compactness of Θ, there exists a $\theta_m \in \Theta$ which maximizes $\int w_1(x)[f_m^{1/2}(x) - f^{1/2}(x)]^2 dx$. By (5.4), (5.5) and a Taylor expansion, one has

$$\begin{aligned}
& E\left|\int I_{\{|x|>\alpha_m\}} w_1(x) f_m(x) dx\right| \\
= & \int\int I_{\{|x|>\alpha_m\}} w_1(x) \frac{1}{b_m} K_0\left(\frac{y-x}{b_m}\right) f(y) dy\, dx \\
= & \int I_{\{|x|>\alpha_m\}} w_1(x) \int K_0(t) f(x+tb_m) dt\, dx \\
= & \int I_{\{|x|>\alpha_m\}} w_1(x) \int K_0(t) \left[f(x) + f^{(1)}(x) t b_m + \frac{1}{2} f^{(2)}(\xi) t^2 b_m^2 \right] dt\, dx \\
\le & \int I_{\{|x|>\alpha_m\}} h_\theta(x) dx + \frac{1}{2} b_m^2 \int I_{\{|x|>\alpha_m\}} h_\theta(x) \sup_{|t|\le a_0} \frac{f^{(2)}(x+tb_m)}{f(x)} dx \int t^2 K_0(t) dt \\
\le & \sup_{\theta\in\Theta} \int I_{\{|x|>\alpha_m\}} h_\theta(x) dx + \frac{1}{2} b_m^2 \int t^2 K_0(t) dt \sup_{\theta\in\Theta} \int I_{\{|x|>\alpha_m\}} h_\theta(x) \sup_{|t|\le a_0} \frac{f^{(2)}(x+tb_m)}{f(x)} dx \\
\to & \; 0.
\end{aligned}$$

Thus as $m \to \infty$,

$$\int I_{\{|x|>\alpha_m\}} w_1(x) f_m(x) dx \xrightarrow{\mathscr{P}} 0$$

and

$$
\begin{aligned}
& \int I_{\{|x|>\alpha_m\}} w_1(x) \left[f_m^{1/2}(x) - f^{1/2}(x)\right]^2 dx \\
\leq \quad & 2\int I_{\{|x|>\alpha_m\}} w_1(x) \left[f_m(x) + f(x)\right] dx \\
\leq \quad & 2\int I_{\{|x|>\alpha_m\}} w_1(x) f_m(x) dx + 2\int I_{\{|x|>\alpha_m\}} h_\theta(x) dx \\
\stackrel{\mathscr{P}}{\to} \quad & 0.
\end{aligned} \tag{5.9}
$$

On the other hand,

$$
\begin{aligned}
\left|\int I_{\{|x|\leq\alpha_m\}} w_1(x) \left[f_m^{1/2}(x) - f^{1/2}(x)\right]^2 dx\right| \quad = \quad & \int I_{\{|x|\leq\alpha_m\}} w_1(x) \frac{[f_m(x) - f(x)]^2}{\left[f_m^{1/2}(x) + f^{1/2}(x)\right]^2} dx \\
\leq \quad & \int I_{\{|x|\leq\alpha_m\}} w_1(x) f^{-1}(x) [f_m(x) - f(x)]^2 dx \\
\leq \quad & 2\int I_{\{|x|\leq\alpha_m\}} w_1(x) f^{-1}(x) [f_m(x) - E[f_m(x)]]^2 dx \\
& +2\int I_{\{|x|\leq\alpha_m\}} w_1(x) f^{-1}(x) \left[E[f_m(x)] - f(x)\right]^2 dx \\
= \quad & 2(A_{1m} + A_{2m}), \text{ say.}
\end{aligned}
$$

Now by (5.6) as $m \to \infty$

$$
\begin{aligned}
E[A_{1m}] \quad = \quad & \int I_{\{|x|\leq\alpha_m\}} w_1(x) f^{-1}(x) E[f_m(x) - E[f_m(x)]]^2 dx \\
\leq \quad & \int I_{\{|x|\leq\alpha_m\}} w_1(x) f^{-1}(x) \frac{1}{mb_m^2} \int K_0^2\left(\frac{y-x}{b_m}\right) f(y) dy dx \\
= \quad & m^{-1} b_m^{-1} \int I_{\{|x|\leq\alpha_m\}} w_1(x) \int_{-a_0}^{a_0} K_0^2(t) f(x+tb_m) f^{-1}(x) dt dx \\
\leq \quad & m^{-1} b_m^{-1} \int_{-a_0}^{a_0} K_0^2(t) dt \sup_{\theta\in\Theta} \int I_{\{|x|\leq\alpha_m\}} h_\theta(x) \sup_{|t|\leq a_0} \frac{f(x+tb_m)}{f^2(x)} dx \\
\to \quad & 0,
\end{aligned}
$$

i.e, $A_{1m} \stackrel{\mathscr{P}}{\to} 0$ as $m \to \infty$. By a Taylor expansion and (5.7),

$$
\begin{aligned}
|A_{2m}| \quad = \quad & \int I_{\{|x|\leq\alpha_m\}} w_1(x) f^{-1}(x) \left[\int_{-a_0}^{a_0} K_0(t)(f(x+tb_m) - f(x)) dt\right]^2 dx \\
\leq \quad & \frac{1}{4} b_m^4 \int I_{\{|x|\leq\alpha_m\}} w_1(x) f^{-1}(x) \left[\sup_{|t|\leq a_0} |f^{(2)}(x+tb_m)| \int_{-a_0}^{a_0} t^2 K_0(t)\right]^2 dx \\
\leq \quad & \frac{1}{4} b_m^4 \left[\int_{-a_0}^{a_0} K_0(t) t^2 dt\right]^2 \sup_{\theta\in\Theta} \int I_{\{|x|\leq\alpha_m\}} h_\theta(x) \sup_{|t|\leq a_0} \left[\frac{f^{(2)}(x+tb_m)}{f(x)}\right]^2 dx \\
\to \quad & 0
\end{aligned}
$$

Therefore, $\int I_{\{|x|\leq\alpha_m\}} w_1(x)[f_m^{1/2}(x)-f^{1/2}(x)]^2dx \xrightarrow{\mathscr{P}} 0$ as $m\to\infty$. This combined with (5.9) gives $\int w_1(x)[f_m^{1/2}(x)-f^{1/2}(x)]^2dx \xrightarrow{\mathscr{P}} 0$ for any $\theta\in\Theta$. By the continuity of the function in θ and the compactness of Θ, hence the result. □

Following the results above, the theorem below presents the consistency of $\hat{\theta}_{MHDE}$ defined in (5.3).

Theorem 5.2. *Let $m,n\to\infty$ as $N\to\infty$. Suppose that (D1) holds for any $\theta\in\Theta$ and the bandwidths b_m and b_n in (2.3) and (2.4) respectively satisfy $b_m,b_n\to 0$ and $mb_m, nb_n\to\infty$ as $N\to\infty$. Further suppose that either (D2) or (D3) holds. Then $\|f_m^{1/2}-f^{1/2}\|\xrightarrow{\mathscr{P}} 0$, $\|h_n^{1/2}-h_\theta^{1/2}\|\xrightarrow{\mathscr{P}} 0$ and $\sup_{t\in\Theta}\|\hat{h}_t^{1/2}-h_t^{1/2}\|\xrightarrow{\mathscr{P}} 0$ as $N\to\infty$. Furthermore, $\hat{\theta}_{MHDE}\xrightarrow{\mathscr{P}}\theta$ as $N\to\infty$, where $\hat{\theta}_{MHDE}$ is defined in (5.3) with f_m, h_n and $\hat{h}_t$ given by (2.3), (2.4) and (5.2) respectively.*

Proof. If we can prove that $\|h_n^{1/2}-h_\theta^{1/2}\|\xrightarrow{\mathscr{P}} 0$ and $\sup_{t\in\Theta}\|\hat{h}_t^{1/2}-h_t^{1/2}\|\xrightarrow{\mathscr{P}} 0$ as $N\to\infty$, then by Theorem 5.1 (iii) and then (ii) we have $\hat{\theta}_{MHDE}\xrightarrow{\mathscr{P}}\theta$ as $N\to\infty$.

It is known that $f_m\xrightarrow{\mathscr{P}} f$ and $h_n\xrightarrow{\mathscr{P}} h$ as $N\to\infty$ (see Rao, 1983). Since $\int h_\theta(x)dx = \int h_n(x)dx = 1$, $\int[h_\theta(x)-h_n(x)]^+dx = \int[h_\theta(x)-h_n(x)]^-dx$ and $\|h_n^{1/2}-h_\theta^{1/2}\|^2 \leq \int|h_\theta(x)-h_n(x)|dx = 2\int[h_\theta(x)-h_n(x)]^+dx$. Since, $[h_\theta(x)-h_n(x)]^+ < h_\theta(x)$, by the DCT it follows that $\|h_n^{1/2}-h_\theta^{1/2}\|\xrightarrow{\mathscr{P}} 0$ as $n\to\infty$. Similarly $\|f_m^{1/2}-f^{1/2}\|\xrightarrow{\mathscr{P}} 0$ as $m\to\infty$.

Note that $\int[\hat{h}_\theta^{1/2}(x)-h_\theta^{1/2}(x)]^2dx = \int w_1(x)[f_m^{1/2}(x)-f^{1/2}(x)]^2dx \leq \int w_1(x)|f_m(x)-f(x)|dx$. If (D2) holds then f_m-f will have a compact support on which $w_1(x)$ is bounded. Therefore, $\int[\hat{h}_\theta^{1/2}(x)-h_\theta^{1/2}(x)]^2dx \leq C_1\int|f_m(x)-f(x)|dx = 2C_1\int[f(x)-f_m(x)]^+dx$ for some positive number C_1. Since $f_m\xrightarrow{\mathscr{P}} f$, by the DCT we have $\sup_{\theta\in\Theta}\|\hat{h}_\theta^{1/2}-h_\theta^{1/2}\|\xrightarrow{\mathscr{P}} 0$. If (D3) holds then Lemma 5.2 gives $\sup_{\theta\in\Theta}\|\hat{h}_\theta^{1/2}-h_\theta^{1/2}\|\xrightarrow{\mathscr{P}} 0$. □

Next we prove the asymptotic normality of $\hat{\theta}_{MHDE}$. Under condition (D2) we derive in the next theorem an expression of the bias term $\hat{\theta}_{MHDE}-\theta$. Note that the first-order partial derivatives of w_1 given in (4.19) are

$$\frac{\partial w_1(x)}{\partial\lambda} = e^{\alpha+\beta x}-1,$$

$$\frac{\partial w_1(x)}{\partial \alpha} = \lambda e^{\alpha+\beta x},$$
$$\frac{\partial w_1(x)}{\partial \beta} = \lambda x e^{\alpha+\beta x}.$$

Define symmetric matrix

$$\Delta(\theta) = \int \left(\frac{\partial w_1(x)}{\partial \theta}\right)\left(\frac{\partial w_1(x)}{\partial \theta}\right)^\top \frac{f}{w_1}(x)dx = \begin{pmatrix} \Delta_{11}(\theta) & \Delta_{12}(\theta) & \Delta_{13}(\theta) \\ \Delta_{12}(\theta) & \Delta_{22}(\theta) & \Delta_{23}(\theta) \\ \Delta_{13}(\theta) & \Delta_{23}(\theta) & \Delta_{33}(\theta) \end{pmatrix}, \quad (5.10)$$

where

$$\Delta_{11}(\theta) = \int \left(\frac{\partial w_1(x)}{\partial \lambda}\right)^2 \frac{f}{w_1}(x)dx = \int (e^{\alpha+\beta x}-1)^2 \frac{f}{w_1}(x)dx,$$
$$\Delta_{22}(\theta) = \int \left(\frac{\partial w_1(x)}{\partial \alpha}\right)^2 \frac{f}{w_1}(x)dx = \lambda^2 \int e^{2\alpha+2\beta x} \frac{f}{w_1}(x)dx,$$
$$\Delta_{33}(\theta) = \int \left(\frac{\partial w_1(x)}{\partial \beta}\right)^2 \frac{f}{w_1}(x)dx = \lambda^2 \int x^2 e^{2\alpha+2\beta x} \frac{f}{w_1}(x)dx,$$
$$\Delta_{12}(\theta) = \int \frac{\partial w_1(x)}{\partial \lambda}\frac{\partial w_1(x)}{\partial \alpha} \frac{f}{w_1}(x)dx = \lambda \int e^{\alpha+\beta x}(e^{\alpha+\beta x}-1) \frac{f}{w_1}(x)dx,$$
$$\Delta_{13}(\theta) = \int \frac{\partial w_1(x)}{\partial \lambda}\frac{\partial w_1(x)}{\partial \beta} \frac{f}{w_1}(x)dx = \lambda \int x e^{\alpha+\beta x}(e^{\alpha+\beta x}-1) \frac{f}{w_1}(x)dx,$$
$$\Delta_{23}(\theta) = \int \frac{\partial w_1(x)}{\partial \alpha}\frac{\partial w_1(x)}{\partial \beta} \frac{f}{w_1}(x)dx = \lambda^2 \int x e^{2\alpha+2\beta x} \frac{f}{w_1}(x)dx.$$

Let

$$A_N(\theta) = \int \frac{\partial w_1}{\partial \theta}(x)\left[\frac{f_m^{1/2}h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right] dx = \begin{pmatrix} A_{N1}(\theta) \\ A_{N2}(\theta) \\ A_{N3}(\theta) \end{pmatrix}, \quad (5.11)$$

where

$$A_{N1}(\theta) = \int (e^{\alpha+\beta x}-1)\left[\frac{f_m^{1/2}h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right] dx,$$
$$A_{N2}(\theta) = \int \lambda e^{\alpha+\beta x}\left[\frac{f_m^{1/2}h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right] dx,$$
$$A_{N3}(\theta) = \int \lambda x e^{\alpha+\beta x}\left[\frac{f_m^{1/2}h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right] dx.$$

Theorem 5.3. *Suppose that $\theta \in int(\Theta)$, K_0 in (2.3) has compact support, and assumptions in Theorem 5.2 hold. Further suppose (D2) holds. Then it follows that*

$$\hat{\theta}_{MHDE} - \theta \;=\; 2\left[\Delta^{-1}(\theta) + R_N\right] A_N(\theta), \tag{5.12}$$

where $\hat{\theta}_{MHDE}$ is defined by (5.3) and R_N is a 3×3 matrix with elements tending to zero in probability as $N \to \infty$.

Proof. From Theorem 5.2 we have $\hat{\theta}_{MHDE} \xrightarrow{\mathscr{P}} \theta$ as $N \to \infty$. Since $t = \hat{\theta}_{MHDE} \in \Theta$ minimizes the Hellinger distance between $\hat{h}_t$ and h_n, $\hat{\theta}_{MHDE}$ maximizes $2\int \hat{h}_t^{1/2}(x) h_n^{1/2}(x)dx - \int \hat{h}_t(x)dx$. Also since K_0 has compact support, we have

$$\int \frac{\partial}{\partial t}\left[2\hat{h}_t^{1/2}(x) h_n^{1/2}(x)dx - \hat{h}_t(x)\right]\bigg|_{t=\hat{\theta}_{MHDE}} dx \;=\; 0.$$

For notation simplicity we use $\hat{\theta}$ to denote $\hat{\theta}_{MHDE}$ and use $\hat{w}_1$ to denote w_1 in (4.19) with θ replaced by $\hat{\theta}_{MHDE}$. Let

$$M_\theta(x) \;=\; 2\hat{h}_\theta^{1/2}(x) h_n^{1/2}(x)dx - \hat{h}_\theta(x),$$

then by a Taylor expansion of $\hat{\theta}$ at θ it follows that

$$\int \frac{\partial M_\theta(x)}{\partial \theta}dx + \left[\int \frac{\partial^2 M_\theta(x)}{\partial\theta\partial\theta^\top}dx + R_N\right] \cdot \left(\hat{\lambda} - \lambda, \hat{\alpha} - \alpha, \hat{\beta} - \beta\right)^\top \;=\; 0, \tag{5.13}$$

where, by (D2), R_N is a 3×3 matrix with elements tending to zero in probability as $N \to \infty$. Direct calculation gives

$$\frac{\partial M_\theta(x)}{\partial \lambda} \;=\; (e^{\alpha+\beta x} - 1)\left[\frac{f_m^{1/2} h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right], \tag{5.14}$$

$$\frac{\partial M_\theta(x)}{\partial \alpha} \;=\; \lambda e^{\alpha+\beta x}\left[\frac{f_m^{1/2} h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right], \tag{5.15}$$

$$\frac{\partial M_\theta(x)}{\partial \beta} \;=\; \lambda x e^{\alpha+\beta x}\left[\frac{f_m^{1/2} h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right], \tag{5.16}$$

$$\frac{\partial^2 M_\theta(x)}{\partial \lambda^2} \;=\; -\frac{(e^{\alpha+\beta x} - 1)^2}{2w_1^{3/2}(x)} f_m^{1/2}(x) h_n^{1/2}(x), \tag{5.17}$$

$$\frac{\partial^2 M_\theta(x)}{\partial\lambda\partial\alpha} = \frac{e^{\alpha+\beta x}(w_1(x)+1)}{2w_1^{3/2}} f_m^{1/2}(x)h_n^{1/2}(x) - e^{\alpha+\beta x} f_m(x), \tag{5.18}$$

$$\frac{\partial^2 M_\theta(x)}{\partial\lambda\partial\beta} = \frac{xe^{\alpha+\beta x}(w_1(x)+1)}{2w_1^{3/2}(x)} f_m^{1/2}(x)h_n^{1/2}(x) - xe^{\alpha+\beta x} f_m(x), \tag{5.19}$$

$$\frac{\partial^2 M_\theta(x)}{\partial\alpha^2} = \frac{\lambda e^{\alpha+\beta x}(w_1(x)+1-\lambda)}{2w_1^{3/2}(x)} f_m^{1/2}(x)h_n^{1/2}(x) - \lambda e^{\alpha+\beta x} f_m(x), \tag{5.20}$$

$$\frac{\partial^2 M_\theta(x)}{\partial\alpha\partial\beta} = \frac{\lambda xe^{\alpha+\beta x}(w_1(x)+1-\lambda)}{2w_1^{3/2}(x)} f_m^{1/2}(x)h_n^{1/2}(x) - \lambda xe^{\alpha+\beta x} f_m(x), \tag{5.21}$$

$$\frac{\partial^2 M_\theta(x)}{\partial\beta^2} = \frac{\lambda x^2 e^{\alpha+\beta x}(w_1(x)+1-\lambda)}{2w_1^{3/2}(x)} f_m^{1/2}(x)h_n^{1/2}(x) - \lambda x^2 e^{\alpha+\beta x} f_m(x). \tag{5.22}$$

Since (D2) holds, for (5.17) we have by Theorem 5.2

$$\begin{aligned}
&\left|\int \frac{(e^{\alpha+\beta x}-1)^2}{2w_1^{3/2}(x)}\left[f_m^{1/2}(x)h_n^{1/2}(x) - f^{1/2}(x)h_\theta^{1/2}(x)\right]dx\right| \\
\leq\ & C\left[\int f_m^{1/2}(x)\left|h_n^{1/2}(x)-h_\theta^{1/2}(x)\right|dx + \int h_\theta^{1/2}(x)\left|f_m^{1/2}(x)-f^{1/2}(x)\right|dx\right] \\
\leq\ & C\left[\left\|h_n^{1/2}-h_\theta^{1/2}\right\| + \left\|f_m^{1/2}-f^{1/2}\right\|\right] \\
\stackrel{\mathscr{P}}{\to}\ & 0.
\end{aligned}$$

Thus for (5.17),

$$\begin{aligned}
-\int \frac{(e^{\alpha+\beta x}-1)^2}{2w_1^{3/2}(x)} f_m^{1/2}(x)h_n^{1/2}(x) \quad &\stackrel{\mathscr{P}}{\longrightarrow} \quad -\int \frac{(e^{\alpha+\beta x}-1)^2}{2w_1^{3/2}(x)} f^{1/2}(x)h_\theta^{1/2}(x)dx \\
&= \quad -\frac{1}{2}\int (e^{\alpha+\beta x}-1)^2 \frac{f}{w_1}(x)dx \\
&= \quad -\frac{1}{2}\Delta_{11}(\theta).
\end{aligned} \tag{5.23}$$

For (5.18), similarly we have

$$\begin{aligned}
\int \frac{e^{\alpha+\beta x}(w_1(x)+1)}{2w_1^{3/2}(x)} f_m^{1/2}(x)h_n^{1/2}(x)dx \quad &\stackrel{\mathscr{P}}{\longrightarrow} \quad \int \frac{e^{\alpha+\beta x}(w_1(x)+1)}{2w_1^{3/2}(x)} f^{1/2}(x)h_\theta^{1/2}(x)dx \\
&= \quad \int \frac{g(x)(w_1(x)+1)}{2w_1(x)}dx \\
&= \quad \frac{1}{2} + \frac{1}{2}\int \frac{e^{\alpha+\beta x}}{w_1(x)} f(x)dx
\end{aligned}$$

and

$$\begin{aligned}
\left|\int e^{\alpha+\beta x}\left[f_m(x)-f(x)\right]dx\right| &\leq C\int\left|\left[f_m^{1/2}(x)-f^{1/2}(x)\right]\left[f_m^{1/2}(x)+f^{1/2}(x)\right]\right|dx \\
&\leq C\left\|f_m^{1/2}-f^{1/2}\right\|\cdot\left\|f_m^{1/2}+f^{1/2}\right\| \\
&\leq 2C\left\|f_m^{1/2}-f^{1/2}\right\| \\
&\stackrel{\mathscr{P}}{\to} 0,
\end{aligned}$$

i.e. $\int e^{\alpha+\beta x}f_m(x)dx \stackrel{\mathscr{P}}{\to} \int e^{\alpha+\beta x}f(x)dx$. Thus for (5.18),

$$\begin{aligned}
&\int \frac{e^{\alpha+\beta x}(w_1(x)+1)}{2w_1^{3/2}(x)}f_m^{1/2}(x)h_n^{1/2}(x)-e^{\alpha+\beta x}f_m(x)dx \\
&\qquad \stackrel{\mathscr{P}}{\longrightarrow} -\frac{1}{2}\lambda\int e^{\alpha+\beta x}(e^{\alpha+\beta x}-1)\frac{f}{w_1}(x)dx \;=\; -\frac{1}{2}\Delta_{12}(\theta).
\end{aligned} \tag{5.24}$$

Similarly for (5.19)-(5.22),

$$\begin{aligned}
&\int \frac{xe^{\alpha+\beta x}(w_1(x)+1)}{2w_1^{3/2}(x)}f_m^{1/2}(x)h_n^{1/2}(x)-xe^{\alpha+\beta x}f_m(x)dx \\
&\qquad \stackrel{\mathscr{P}}{\longrightarrow} -\frac{1}{2}\lambda\int xe^{\alpha+\beta x}(e^{\alpha+\beta x}-1)\frac{f}{w_1}(x)dx \;=\; -\frac{1}{2}\Delta_{13}(\theta),
\end{aligned} \tag{5.25}$$

$$\begin{aligned}
&\int \frac{\lambda e^{\alpha+\beta x}(w_1(x)+1-\lambda)}{2w_1^{3/2}(x)}f_m^{1/2}(x)h_n^{1/2}(x)-\lambda e^{\alpha+\beta x}f_m(x)dx \\
&\qquad \stackrel{\mathscr{P}}{\longrightarrow} -\frac{1}{2}\lambda^2\int e^{2\alpha+2\beta x}\frac{f}{w_1}(x)dx \;=\; -\frac{1}{2}\Delta_{22}(\theta),
\end{aligned} \tag{5.26}$$

$$\begin{aligned}
&\int \frac{\lambda xe^{\alpha+\beta x}(w_1(x)+1-\lambda)}{2w_1^{3/2}(x)}f_m^{1/2}(x)h_n^{1/2}(x)-\lambda xe^{\alpha+\beta x}f_m(x)dx \\
&\qquad \stackrel{\mathscr{P}}{\longrightarrow} -\frac{1}{2}\lambda^2\int xe^{2\alpha+2\beta x}\frac{f}{w_1}(x)dx \;=\; -\frac{1}{2}\Delta_{23}(\theta),
\end{aligned} \tag{5.27}$$

$$\begin{aligned}
&\int \frac{\lambda x^2e^{\alpha+\beta x}(w_1(x)+1-\lambda)}{2w_1^{3/2}(x)}f_m^{1/2}(x)h_n^{1/2}(x)-\lambda x^2e^{\alpha+\beta x}f_m(x)dx \\
&\qquad \stackrel{\mathscr{P}}{\longrightarrow} -\frac{1}{2}\lambda^2\int x^2e^{2\alpha+2\beta x}\frac{f}{w_1}(x)dx \;=\; -\frac{1}{2}\Delta_{33}(\theta).
\end{aligned} \tag{5.28}$$

Now together with (5.17)-(5.28), (5.13) is reduced to

$$A_N(\theta)+\left[-\frac{1}{2}\Delta(\theta)+R_N\right]\left(\hat{\theta}-\theta\right)=0,$$

where, $\Delta(\theta)$ and $A_N(\theta)$ are given in (5.10) and (5.11) respectively. Hence the result. □

We now state the asymptotic distribution of the proposed MHDE $\hat{\theta}_{MHDE}$. The following conditions are made in the next theorem.

Let $\{\alpha_N\}$ be a sequence of positive numbers such as $\alpha_N \to \alpha$ as $N \to \infty$ and

(C0) f has infinite support $(-\infty, \infty)$.

(C1) The second derivative of f exists.

(C2) $\frac{n}{N} \to \rho \in (0,1)$ as $N \to \infty$.

(C3) K_0 and K_1 in (2.3) and (2.4) respectively are bounded symmetric densities with support $[-a_0, a_0]$ and $[-a_1, a_1]$ respectively, where $0 < a_0, a_1 < \infty$.

(C4) All the elements in both $\Delta(\theta)$ and $\bar{\Delta}(\theta)$ are finite, where $\Delta(\theta)$ is defined in (5.10) and $\bar{\Delta}(\theta) = \int \left(\frac{\partial w_1(x)}{\partial \theta}\right)\left(\frac{\partial w_1(x)}{\partial \theta}\right)^{\top} f(x)dx$.

(C5) The second derivative of f exists and satisfies for $i = 1,2,3$ that as $N \to \infty$,

$$b_m^2 \int \varepsilon_{Ni}^2(x) \frac{f}{w_1}(x) \sup_{|t| \le a_0} \frac{f^{(2)}(x+tb_m)}{f(x)} dx = O(1),$$

where $\varepsilon_N(x) = \frac{\partial w_1(x)}{\partial \theta} I_{\{|x| > \alpha_N\}}$.

(C6)

$$N \cdot P(|X_1| > \alpha_N - a_0 b_m) \to 0 \quad \text{as } N \to \infty,$$

$$N \cdot P(|Y_1| > \alpha_N - a_1 b_n) \to 0 \quad \text{as } N \to \infty.$$

(C7) The second derivatives of f exists and satisfies

$$N^{-1/2} b_n^{-1} \int |\delta_N(x)| w_1^{-1}(x) \sup_{|t| \le a_1} \frac{h_\theta(x+tb_n)}{h_\theta(x)} dx \to 0 \quad \text{as } N \to \infty,$$

$$N^{1/2} b_n^4 \int |\delta_N(x)| f(x) \sup_{|t| \le a_1} \left[\frac{h_\theta^{(2)}(x+tb_n)}{h_\theta(x)}\right]^2 dx \to 0 \quad \text{as } N \to \infty,$$

$$N^{-1/2} b_m^{-1} \int |\delta_N(x)| \sup_{|t| \le a_0} \frac{f(x+tb_m)}{f(x)} dx \to 0 \quad \text{as } N \to \infty,$$

$$N^{1/2}b_m^4\int|\delta_N(x)|f(x)\sup_{|t|\le a_0}\left[\frac{f^{(2)}(x+tb_m)}{f(x)}\right]^2dx\to 0 \quad \text{as } N\to\infty,$$

where $\delta_N(x)=\frac{\partial w_1(x)}{\partial\theta}I_{\{|x|\le\alpha_N\}}$.

(C8) The second derivatives of f exists and satisfies

$$N^{1/2}b_n^2\int|\delta_N(x)|f(x)\sup_{|t|\le a_1}\frac{|h_\theta^{(2)}(x+tb_n)|}{h_\theta(x)}dx\to 0 \quad \text{as } N\to\infty,$$

$$N^{1/2}b_m^2\int|\delta_N(x)|f(x)\sup_{|t|\le a_0}\frac{|f^{(2)}(x+tb_m)|}{f(x)}dx\to 0 \quad \text{as } N\to\infty.$$

(C9)

$$\begin{aligned}\sup_{|x|\le\alpha_N}\sup_{|t|\le a_1}\frac{h_\theta(x+tb_n)}{h_\theta(x)} &= O(1) \quad \text{as } N\to\infty,\\ \sup_{|x|\le\alpha_N}\sup_{|t|\le a_0}\frac{f(x+tb_m)}{f(x)} &= O(1) \quad \text{as } N\to\infty.\end{aligned}$$

(C10)

$$b_n^2\int I_{\{|x|\le\alpha_N\}}h_\theta(x)\sup_{|t|\le a_1}\left[\frac{\partial^2\log w_1(y)}{\partial\theta\partial y}|_{y=x+tb_n}\right]^2dx\to 0 \quad \text{as } N\to\infty,$$

$$b_m^2\int I_{\{|X|\le\alpha_N\}}f(x)\sup_{|t|\le a_0}\left[\frac{\partial^2 w_1(y)}{\partial\theta\partial y}|_{y=x+tb_m}\right]^2dx\to 0 \quad \text{as } N\to\infty.$$

(C11) The second derivative of f exists and satisfies

$$N^{1/2}b_m^2\int|\varepsilon_N(x)|f(x)\sup_{|t|\le a_0}\frac{|f^{(2)}(x+tb_m)|}{f(x)}dx=o(1) \quad \text{as } N\to\infty.$$

Theorem 5.4. *Suppose that $\hat\theta_{MHDE}$ defined in (5.3) satisfies (5.12). Further suppose that conditions $(C0)-(C9)$ hold. Then the asymptotic distribution of $\sqrt{N}(\hat\theta_{MHDE}-\theta)$ is $N(0,\Sigma)$, where Σ is defined by*

$$\begin{aligned}\Sigma &= \Delta^{-1}(\theta)\left[\frac{1}{1-\rho}\bar\Delta(\theta)+\frac{1}{\rho}\Delta(\theta)\right]\Delta^{-1}(\theta)\\ &= \frac{1}{\rho(1-\rho)}\Delta^{-1}(\theta)\left[\Delta(\theta)-\rho\left(\Delta(\theta)-\bar\Delta(\theta)\right)\right]\Delta^{-1}(\theta)\end{aligned}$$

with

$$\bar{\Delta}(\theta) = \int \frac{\partial w_1}{\partial \theta}(x) \left[\frac{\partial w_1}{\partial \theta}(x)\right]^\top f(x)dx, \tag{5.29}$$

$$\Delta(\theta) = \int \frac{\partial w_1}{\partial \theta}(x) \left[\frac{\partial w_1}{\partial \theta}(x)\right]^\top \frac{f}{w_1}(x)dx. \tag{5.30}$$

Proof. We give the sketch of the proof here. In order to find the asymptotic distribution of $\hat{\theta}_{MHDE} - \theta$, by (5.12) we only need to find the asymptotic distribution of $\sqrt{N}A_N(\theta)$. Note that by (5.11),

$$\begin{aligned}
A_N(\theta) &= \int \frac{\partial w_1}{\partial \theta}(x) \left[\frac{f_m^{1/2} h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right] dx \\
&= \int \frac{\partial w_1}{\partial \theta}(x) \frac{f_m^{1/2}}{w_1^{1/2}}(x) \left[h_n^{1/2}(x) - h_\theta^{1/2}(x)\right] dx - \int \frac{\partial w_1}{\partial \theta}(x) f_m^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x)\right] dx \\
&= \int \frac{\partial w_1}{\partial \theta}(x) \frac{f^{1/2}}{w_1^{1/2}}(x) \left[h_n^{1/2}(x) - h_\theta^{1/2}(x)\right] dx \\
&\quad + \int \frac{\partial w_1}{\partial \theta}(x) \frac{1}{w_1^{1/2}(x)} \left[f_m^{1/2}(x) - f^{1/2}(x)\right] \left[h_n^{1/2}(x) - h_\theta^{1/2}(x)\right] dx \\
&\quad - \int \frac{\partial w_1}{\partial \theta}(x) f^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x)\right] dx \\
&\quad - \int \frac{\partial w_1}{\partial \theta}(x) \left[f_m^{1/2}(x) - f^{1/2}(x)\right] \left[f_m^{1/2}(x) - f^{1/2}(x)\right] dx.
\end{aligned}$$

We can prove that as $N \to \infty$,

$$\sqrt{N} \int \frac{\partial w_1}{\partial \theta}(x) \frac{1}{w_1^{1/2}(x)} \left[f_m^{1/2}(x) - f^{1/2}(x)\right] \left[h_n^{1/2}(x) - h_\theta^{1/2}(x)\right] dx \xrightarrow{\mathscr{P}} 0$$

and

$$\sqrt{N} \int \frac{\partial w_1}{\partial \theta}(x) \left[f_m^{1/2}(x) - f^{1/2}(x)\right] \left[f_m^{1/2}(x) - f^{1/2}(x)\right] dx \xrightarrow{\mathscr{P}} 0.$$

Thus we only need to give the asymptotic distribution of

$$\int \frac{\partial w_1}{\partial \theta}(x) \frac{f^{1/2}}{w_1^{1/2}}(x) \left[h_n^{1/2}(x) - h_\theta^{1/2}(x)\right] dx$$

and

$$\int \frac{\partial w_1}{\partial \theta}(x) f^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x)\right] dx$$

separately as they are independent. Details of the proof is given in Section 5.5. □

5.3 Simulation studies

In this section we carry out a simulation study to examine the finite-sample performance of the proposed MHDE $\hat{\theta}_{MHDE}$. We consider the same mixture models as in Table 4.1 with the true parameter values listed there. For each model we consider the same varying mixing proportion values $\lambda = 0.05, 0.2, 0.5, 0.8, 0.95$ as in previous chapters. We also use the same sample sizes $m = n = 30$ and $m = n = 100$ and the same number of replications $N = 1000$.

In the two kernel density estimators f_m of f and h_n of h in (2.3) and (2.4) respectively, we use the truncated standard normal function for both K_0 and K_1. Specifically, we truncate the standard normal curve at ± 2 and rescale it to have total area 1 under the curve, i.e. we use

$$K(u) = \frac{1}{2\Phi(2)-1} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} I_{\{|u| \le 2\}}$$

for both K_0 and K_1. The bandwidths b_m and b_n defined in (2.8) are used here as in Chapter 2. In order to find the $\hat{\theta}_{MHDE}$ defined in (5.3), we use function "optim" in `R` statistical software. We use the same initial estimates of λ, α and β as in Section 4.4. In other words, we use λ_+ as the initial estimate of λ and use the least-squares regression estimate based on model assumption (4.2) as initial values for α and β. As in previous chapters, for each model we calculate the bias and MSE of $\hat{\lambda}_{MHDE}$, coverage probability of the 95% confidence interval constructed based on $\hat{\lambda}_{MHDE}$ using the asymptotic variance defined in Theorem 5.4 and the misclassification rate MR when the same classification rule as in Section 4.4 is used. The simulation results are presented in Table 5.1.

From Table 5.1 we can see that, as expected, $\hat{\theta}_{MHDE}$ has smaller bias, MSE and MR for larger sample sizes than for smaller sample sizes. The $\hat{\lambda}_{MHDE}$ always gives small bias and MSE, especially for larger sample sizes, while $\hat{\alpha}_{MHDE}$ and $\hat{\beta}_{MHDE}$ generally give relatively large bias and MSE even for larger sample sizes. Nevertheless, the MR is reasonably close to OMR, the optimal misclassification rate assuming the probability function p in (1.3) is known, regardless of sample size. Even for M5 where the assumption (4.2) doesn't hold, the MHDE of λ based on (4.2) performs surprisingly well and the MR doesn't deviate from OMR too much for large sample sizes. We also observe that when the two components are close to each other in terms of location (M1 and M3), the estimated mixing proportion $\hat{\lambda}_{MHDE}$ has larger bias and MSE than the cases when the two components are far apart (M2 and M4). However, the bias and MSE of $\hat{\alpha}_{MHDE}$ and $\hat{\beta}_{MHDE}$ for M2 are larger than those for M1. This could be explained by the larger magnitude of the true α and β values for M2 than for M1. The coverage probability of the 95% confidence interval based on $\hat{\lambda}_{MHDE}$ is higher than the nominal level 95% for most of the cases.

When the MHDE $\hat{\theta}_{MHDE}$ we proposed in this chapter is compared with the MLE $\hat{\theta}_{MLE}$ proposed in Chapter 4 based on the same assumption (4.2), they have similar performance for estimating λ, MHDE has a bit better performance than MLE for estimating α while MHDE has much better performance than MLE for estimating β. The MHDE and MLE give comparable misclassification rate MR. Even though MHDE and MLE have similar efficiency in terms of bias, MSE and MR, MLE generally suffers from lack of robustness in the presence of outliers. We investigate the robustness properties of the MHDE in the next section.

Table 5.1: Bias and MSE of $\hat{\theta}_{MHDE}$, CP (%) for $\hat{\lambda}_{MHDE}$ and MR (%) of a classification rule based on $\hat{p}_{MHDE}$.

Model	λ	$m=n=30$					$m=n=100$					OMR
		$Bias(\hat{\lambda})$ $(MSE(\hat{\lambda}))$	$Bias(\hat{\alpha})$ $(MSE(\hat{\alpha}))$	$Bias(\hat{\beta})$ $(MSE(\hat{\beta}))$	CP	MR	$Bias(\hat{\lambda})$ $(MSE(\hat{\lambda}))$	$Bias(\hat{\alpha})$ $(MSE(\hat{\alpha}))$	$Bias(\hat{\beta})$ $(MSE(\hat{\beta}))$	CP	MR	
	0.05	0.126 (0.066)	-0.287 (0.197)	-0.117 (0.359)	82.4	14.90	0.059 (0.026)	-0.249 (0.184)	-0.127 (0.349)	84.3	8.81	4.99
M1	0.20	0.069 (0.041)	-0.239 (0.183)	0.023 (0.387)	82.4	28.13	0.039 (0.044)	-0.120 (0.149)	0.034 (0.342)	100.0	24.97	18.61
$\alpha=-0.5$	0.5	-0.052 (0.087)	-0.165 (0.151)	0.110 (0.328)	100.0	37.93	0.039 (0.057)	-0.065 (0.119)	-0.007 (0.193)	85.5	37.98	30.85
$\beta=1$	0.80	-0.059 (0.059)	-0.132 (0.120)	0.062 (0.223)	100.0	24.03	-0.018 (0.031)	-0.064 (0.085)	-0.011 (0.121)	100.0	23.04	18.61
	0.95	-0.099 (0.040)	-0.130 (0.102)	0.059 (0.119)	100.0	12.40	-0.058 (0.019)	-0.069 (0.062)	-0.002 (0.092)	100.0	9.04	4.99
	0.05	0.054 (0.007)	0.302 (1.072)	-1.504 (5.266)	100.0	2.37	0.013 (0.001)	0.079 (0.717)	-0.789 (3.278)	100.0	1.67	0.24
M2	0.20	0.047 (0.017)	0.221 (1.240)	0.456 (1.009)	100.0	1.70	0.001 (0.006)	-0.219 (0.477)	0.534 (0.598)	100.0	0.97	0.48
$\alpha=-12.5$	0.50	0.055 (0.017)	0.213 (1.158)	0.632 (0.641)	100.0	1.33	0.029 (0.004)	-0.242 (0.397)	0.535 (0.519)	100.0	0.89	0.62
$\beta=5$	0.80	0.009 (0.009)	-0.062 (0.697)	0.552 (0.592)	99.7	1.03	0.005 (0.002)	-0.381 (0.277)	0.463 (0.434)	100.0	0.61	0.48
	0.95	0.004 (0.001)	-0.238 (0.458)	0.468 (0.599)	100.0	0.63	0.004 (0.001)	-0.438 (0.254)	0.423 (0.385)	98.8	0.29	0.24
	0.05	0.142 (0.054)	-0.188 (0.767)	-0.299 (0.265)	85.8	12.27	0.062 (0.018)	-0.131 (0.808)	-0.229 (0.219)	98.2	6.89	4.76
M3	0.20	0.091 (0.060)	-0.349 (1.195)	-0.094 (0.168)	100.0	25.00	0.049 (0.029)	-0.215 (1.182)	-0.025 (0.104)	98.6	20.73	13.90
$\alpha=-2$	0.50	0.026 (0.062)	-0.429 (1.319)	0.044 (0.097)	99.0	34.50	0.029 (0.039)	-0.199 (0.954)	0.016 (0.068)	98.0	34.06	19.05
$\beta=0.693$	0.80	-0.059 (0.043)	-0.433 (1.076)	0.072 (0.077)	100.0	25.00	-0.043 (0.028)	-0.237 (0.705)	0.035 (0.052)	100.0	23.68	13.35
	0.95	-0.116 (0.039)	-0.455 (0.937)	0.085 (0.068)	89.8	14.73	-0.088 (0.024)	-0.241 (0.557)	0.034 (0.043)	100.0	11.27	6.06
	0.05	0.067 (0.016)	-0.140 (0.544)	-0.242 (0.610)	100.0	6.30	0.003 (0.004)	-0.542 (0.915)	-0.155 (0.579)	95.8	4.51	3.14
M4	0.20	0.027 (0.024)	-0.346 (0.809)	0.059 (0.292)	95.2	12.30	-0.005 (0.012)	-0.284 (0.901)	0.093 (0.187)	95.2	10.87	7.03
$\alpha=-4$	0.50	0.016 (0.021)	-0.471 (0.922)	0.098 (0.085)	100.0	16.97	0.002 (0.011)	-0.248 (0.813)	0.042 (0.066)	97.6	15.85	10.19
$\beta=1.099$	0.80	-0.032 (0.014)	-0.472 (0.878)	0.090 (0.073)	100.0	13.17	-0.028 (0.006)	-0.281 (0.750)	0.039 (0.049)	100.0	12.94	7.82
	0.95	-0.054 (0.009)	-0.471 (0.835)	0.089 (0.070)	98.4	7.03	-0.037 (0.004)	-0.241 (0.651)	0.033 (0.048)	99.6	5.96	3.21
	0.05	0.237 (0.190)	NA	NA	99.3	25.53	0.186 (0.170)	NA	NA	97.7	22.18	2.5
	0.20	0.178 (0.127)	NA	NA	99.8	30.87	0.093 (0.056)	NA	NA	100.0	21.61	10
M5	0.50	0.076 (0.063)	NA	NA	100.0	31.90	0.037 (0.028)	NA	NA	100.0	28.77	25
	0.80	0.046 (0.029)	NA	NA	100.0	18.70	0.063 (0.012)	NA	NA	100.0	15.77	10
	0.95	0.014 (0.005)	NA	NA	100.0	5.13	0.033 (0.002)	NA	NA	100.0	4.84	2.5

5.4 Robustness study and comparison

An estimator being robust implies that the estimator is resistant to outlying observations and model misspecification. Estimators with good robustness properties are not heavily affected by small departures from model assumptions (presence of outliers is one type) while estimators with poor robustness are badly affected. For example, the sample median is considered to be more robust than the sample mean because outliers have much less impact on the sample median than on the sample mean. In statistics, many classical estimation methods, such as MLE and least-squares estimation, depend heavily on model assumption. However in real life sometimes these model assumptions are not met, especially in the presence of outliers. Therefore, it is important to provide robust statistics which can tolerate outliers and deviations from model assumption.

In this section, we investigate the robustness properties of MHDE. We also compare the robustness of $\hat{\lambda}_{MHDE}$ with that of all other estimators we proposed in previous chapters, i.e. $\hat{\lambda}$ in Chapter 2, $\hat{\lambda}_L$ in Chapter 3 and $\hat{\lambda}_{MLE}$ in Chapter 4. We can only look at the estimation of λ but not α and β in model (4.3), since this model is not assumed for the two estimations in Chapters 2 and 3. Specifically, we examine the behaviour of all the proposed estimators when data are contaminated by a single outlying observation. Presence of several outliers will be similar and thus omitted here. Note that the outlying observation can be in either the first sample from $f(x)$ or in the second sample from the mixture $h(x)$. Here we only consider the case when the outlier comes from $h(x)$, similar results apply to the other case as well. We look at the change in estimate before and after data contamination. A small change in estimate indicates that the estimator is not influenced much by outliers and thus is considered robust. For this purpose, the α-influence function (α-IF) given in Beran (1997) is an appropriate measure of the change in estimate. However its application in mixture context is very difficult, as discussed in Karlis and Xekalaki (1998). Therefore, we use an adaptive version of α-IF as in Lu et al. (2003) which uses the change in estimate, before and after outlying observations are included, divided by contamination rate (proportion of outlying observations).

In our simulation we consider the same mixture models listed in Table 2.1 or Table 4.1. For each mixture model, we consider varying λ values and two sets of sample sizes $m = n = 30$ and $m = n = 100$ as in previous chapters. Take model M1 for example, after drawing two independent samples with one from $N(0,1)$ and the other from the mixture $(1-\lambda)N(0,1) + \lambda N(1,1)$, we replace the last observation generated from the mixture with a single outlier, an integer with range $[-30,20]$. Thus the contamination rate is $1/60$ for $m = n = 30$ and $1/200$ for $m = n = 100$. Then the α-IF is calculated by

$$IF(x) = \frac{W((X_i)_{i=1}^{m}, (x, Y_i)_{i=1}^{n-1}) - W((X_i)_{i=1}^{m}, (Y_i)_{i=1}^{n})}{1/N}$$

over 100 replications, where $N = 60$ or 200 and W is any estimator of λ based on the samples. In our study, W is either $\hat{\lambda}$, $\hat{\lambda}_L$, $\hat{\lambda}_{MLE}$ or $\hat{\lambda}_{MHDE}$. Similar procedure is used for mixture of Poisson components with outliers varying over the range $[0,20]$. The simulation results are presented in Figures 5.1-5.3 for both $m = n = 30$ and $m = n = 100$. Figure 5.1 is for model M1 with $\lambda = 0.15$ and 0.55, Figure 5.2 is for M2 with $\lambda = 0.25$ and 0.75, and Figure 5.3 is for M3 $\lambda = 0.25$ and 0.75. The results for other models and λ values are very similar and thus omitted.

From Figures 5.1-5.3 we can see that no matter for which model and what sample size, $\hat{\lambda}_{MLE}$ always performs the worst, $\hat{\lambda}_{MHDE}$ performs the best and the behavior of $\hat{\lambda}$, $\hat{\lambda}_L$ and $\hat{\lambda}_{MHDE}$ are quite similar. The α-IF of $\hat{\lambda}_{MLE}$ is generally unbounded while that of $\hat{\lambda}$, $\hat{\lambda}_L$ and $\hat{\lambda}_{MHDE}$ seems bounded when the outlying observation increases in both directions for mixture of normals and in the right direction for mixture of Poissons. This indicates that $\hat{\lambda}_{MLE}$ is generally not resistant to outliers while $\hat{\lambda}$, $\hat{\lambda}_L$ and $\hat{\lambda}_{MHDE}$ are. The bad performance of $\hat{\lambda}_{MLE}$ is mostly for when the outlying observation is bigger than 10. When the outlying observation is less than 10, the performance of $\hat{\lambda}_{MLE}$ is generally ok and is similar to that of other three estimators. The α-IFs of $\hat{\lambda}$, $\hat{\lambda}_L$ and $\hat{\lambda}_{MHDE}$ are almost constants for outliers beyond the range $[-10,5]$ for mixture of normals and $[0,5]$ for mixture of Poissons, though the constants are of different magnitude, and they fluctuate within the ranges. When $\hat{\lambda}$, $\hat{\lambda}_L$ and $\hat{\lambda}_{MHDE}$ are compared, $\hat{\lambda}_L$ behaves the worst in terms of having largest α-IF for mixture of normals and $\hat{\lambda}$ behaves the worst for mixture of Poissons. The performance of $\hat{\lambda}$ follows closely with $\hat{\lambda}_{MHDE}$ except for possible large spikes at around -2 for mixture of normals and possible large α-IF for mixture of Poissons. In summary, $\hat{\lambda}_{MHDE}$ has the best robustness, followed by $\hat{\lambda}$ and then $\hat{\lambda}_L$, and $\hat{\lambda}_{MLE}$ doesn't have robustness against outliers.

5.5 Proof of asymptotic normality

To prove the asymptotic normality of $\hat{\theta}_{MHDE}$ stated in Theorem 5.4, we need a series of lemmas presented below.

Lemma 5.3. *Suppose that (C3)-(C6) hold. Then as $N \to \infty$,*

$$N^{1/2} \int \varepsilon_N(x) w_1^{-1/2}(x) f_m^{1/2}(x) h_n^{1/2}(x) dx \xrightarrow{\mathscr{P}} 0, \tag{5.31}$$

$$N^{1/2} \int \varepsilon_N(x) f^{1/2}(x) f_m^{1/2}(x) dx \xrightarrow{\mathscr{P}} 0. \tag{5.32}$$

Proof. By the Cauchy-Schwarz Inequality,

$$\begin{aligned} & N \cdot E\left[\int \varepsilon_{Ni}(x) w_1^{-1/2} f_m^{1/2}(x) h_n^{1/2}(x) dx\right]^2 \\ \leq \; & N \cdot E\left[\int \varepsilon_{Ni}^2(x) w_1^{-1}(x) f_m(x) dx\right] \cdot E\left[\int I_{\{|x|>\alpha_N\}} h_n(x) dx\right] \\ = \; & N \cdot \Delta_1 \cdot \Delta_2, \;\; \text{say}. \end{aligned}$$

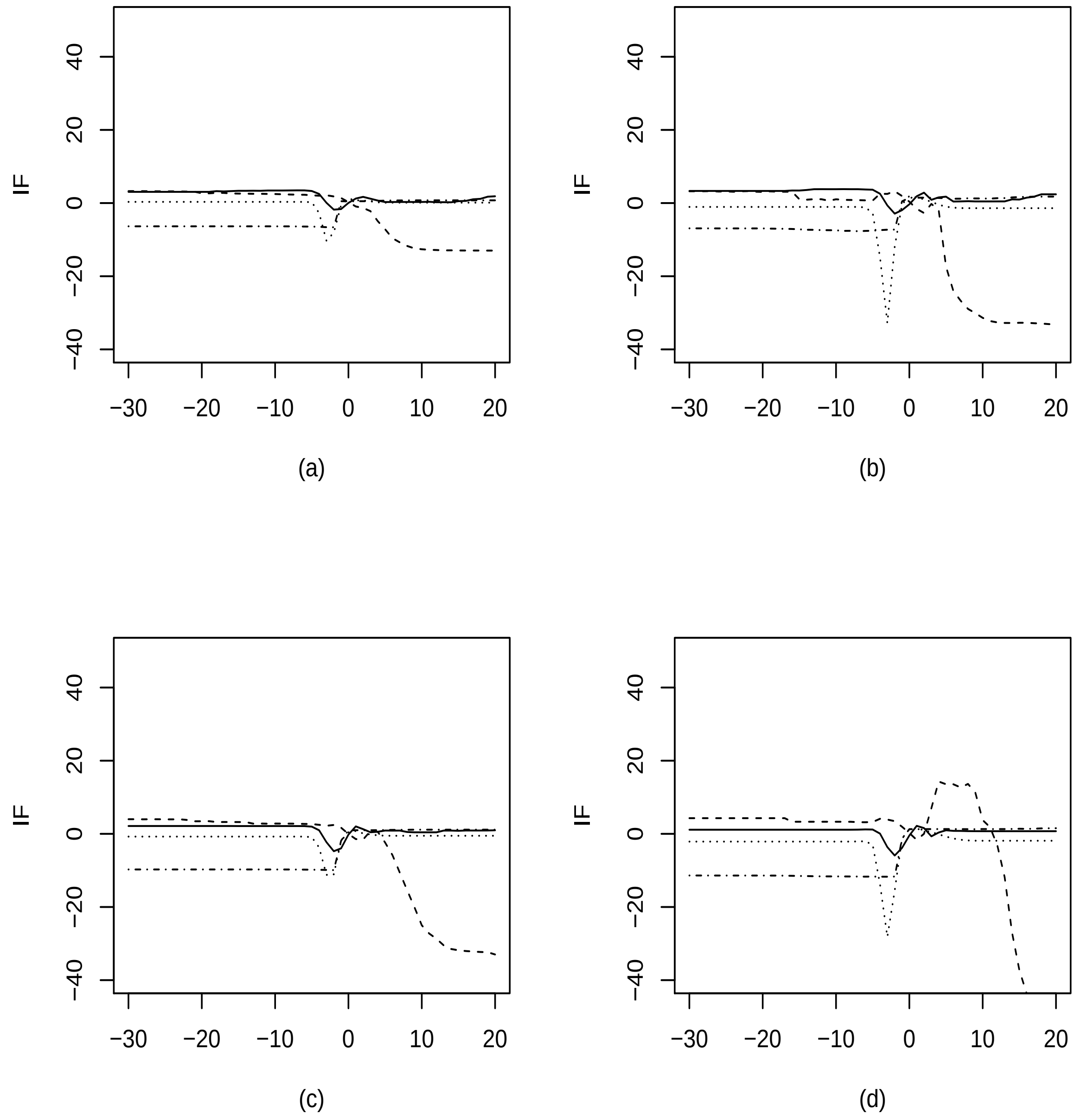

Figure 5.1: The α-IFs of $\hat{\lambda}$ (dotted), $\hat{\lambda}_L$ (dot-dashed), $\hat{\lambda}_{MLE}$ (dashed) and $\hat{\lambda}_{MHDE}$ (solid) for mixture model M1 $(1-\lambda)N(0,1)+\lambda N(1,1)$: (a) $\lambda = 0.15$ and $m = n = 30$; (b) $\lambda = 0.15$ and $m = n = 100$; (c) $\lambda = 0.55$ and $m = n = 30$; (d) $\lambda = 0.55$ and $m = n = 100$.

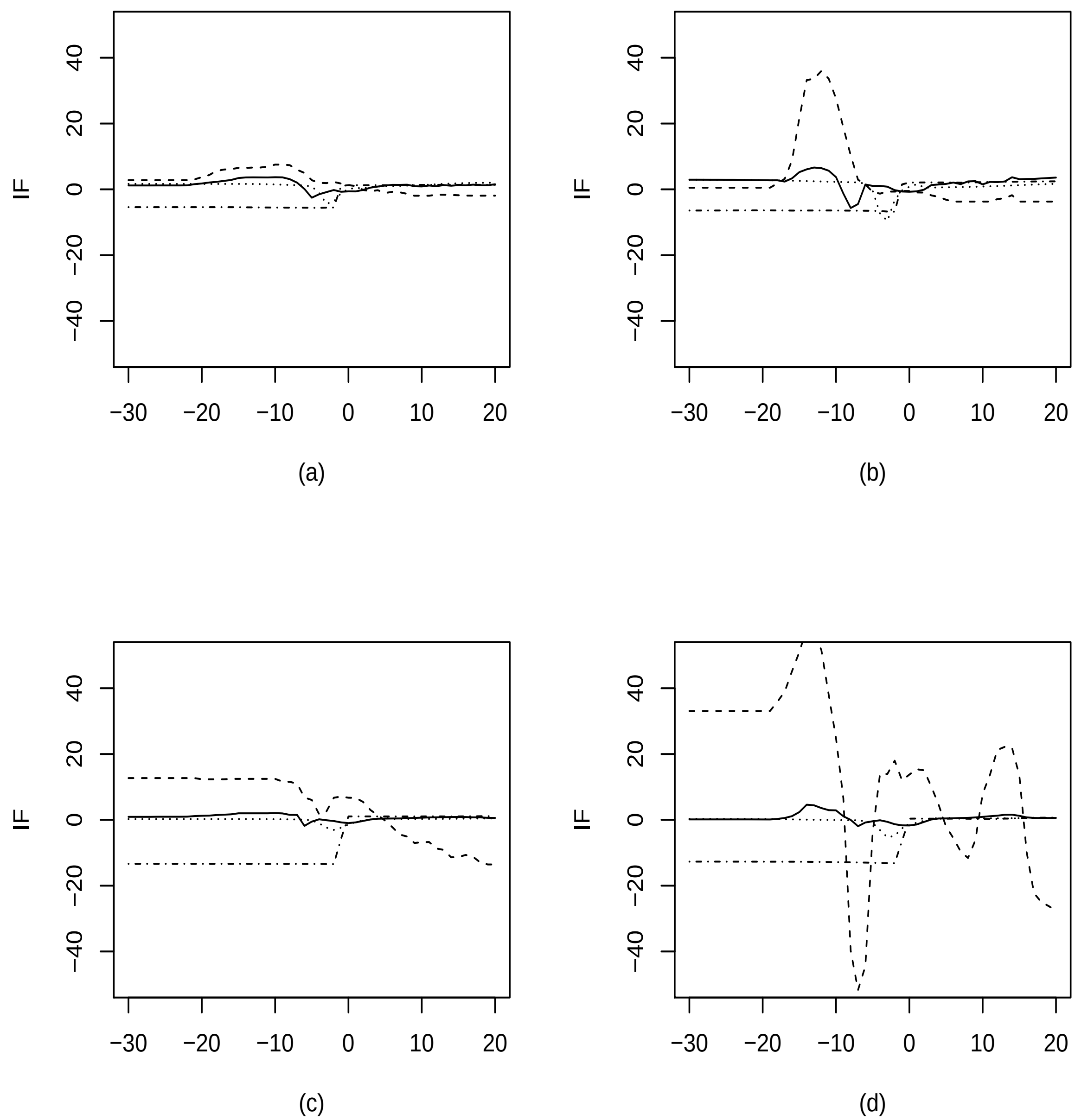

Figure 5.2: The α-IFs of $\hat{\lambda}$ (dotted), $\hat{\lambda}_L$ (dot-dashed), $\hat{\lambda}_{MLE}$ (dashed) and $\hat{\lambda}_{MHDE}$ (solid) for mixture model M2 $(1-\lambda)N(0,1)+\lambda N(5,1)$: (a) $\lambda = 0.25$ and $m = n = 30$; (b) $\lambda = 0.25$ and $m = n = 100$; (c) $\lambda = 0.75$ and $m = n = 30$; (d) $\lambda = 0.75$ and $m = n = 100$.

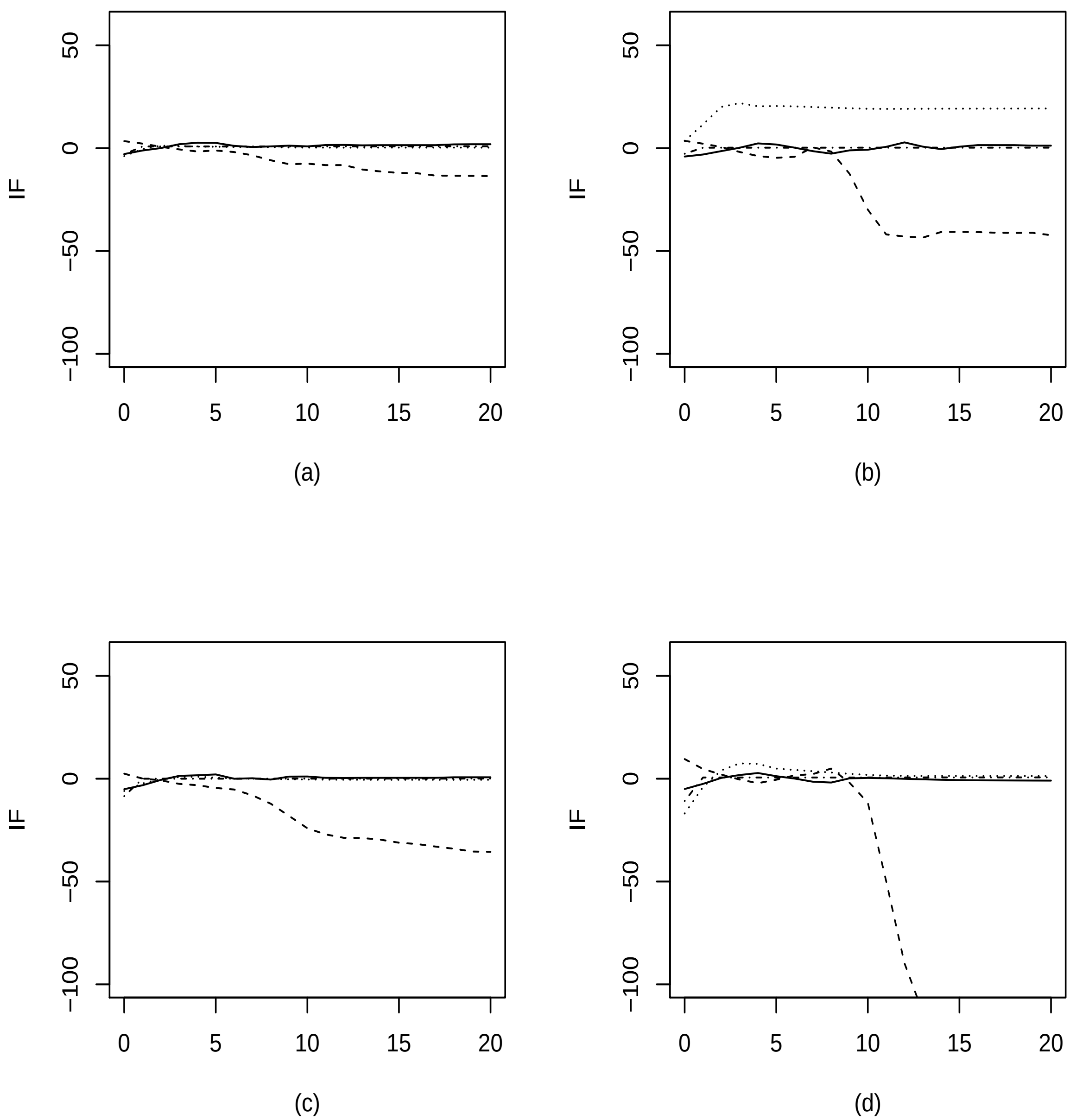

Figure 5.3: The α-IFs of $\hat{\lambda}$ (dotted), $\hat{\lambda}_L$ (dot-dashed), $\hat{\lambda}_{MLE}$ (dashed) and $\hat{\lambda}_{MHDE}$ (solid) for mixture model M3 $(1-\lambda)Po(2)+\lambda Po(4)$: (a) $\lambda = 0.25$ and $m = n = 30$; (b) $\lambda = 0.25$ and $m = n = 100$; (c) $\lambda = 0.75$ and $m = n = 30$; (d) $\lambda = 0.75$ and $m = n = 100$.

By a Taylor expansion and assumptions (C4) and (C5), it follows that

$$
\begin{aligned}
|\Delta_1| &= \int\int \varepsilon_{Ni}^2(x) w_1^{-1}(x) \frac{1}{b_m} K_0(\frac{y-x}{b_m}) f(y) dy dx \\
&= \int \varepsilon_{Ni}^2(x) w_1^{-1}(x) \int_{-a_0}^{a_0} K_0(t) f(x+tb_m) dt dx \\
&= \int \varepsilon_{Ni}^2(x) w_1^{-1}(x) \int_{-a_0}^{a_0} K_0(t) \left[f(x) + f^{(1)}(x) t b_m + \frac{f^{(2)}}{2}(\xi) t^2 b_m^2 \right] dt dx \\
&\leq \int \left[\frac{\partial w_1(x)}{\partial \theta} \right]_i^2 \frac{f}{w_1}(x) dx + \frac{1}{2} b_m^2 \int \varepsilon_{Ni}^2(x) w_1^{-1}(x) |f^{(2)}(x+tb_m)| dx \int_{-a_0}^{a_0} t^2 K_0(t) dt \\
&\leq \Delta_{ii}(\theta) + \frac{1}{2} b_m^2 \int \varepsilon_{Ni}^2(x) \frac{f}{w_1}(x) \sup_{|t|\leq a_0} \frac{|f^{(2)}(x+tb_m)|}{f(x)} dx \int_{-a_0}^{a_0} t^2 K_0(t) dt \\
&= O(1),
\end{aligned}
$$

i.e., Δ_1 is bounded. On the other hand,

$$
\begin{aligned}
|\Delta_2| &= \int\int I_{\{|x|>\alpha_N\}} \frac{1}{b_n} K_1(\frac{y-x}{b_n}) h_\theta(y) dy dx \\
&= \int\int I_{\{|x|>\alpha_N\}} K_1(t) h_\theta(x+tb_n) dt dx \\
&= \int_{-a_1}^{a_1} K_1(t) \int_{|y-tb_n|>\alpha_N} h_\theta(y) dy dt \\
&\leq \int_{-a_1}^{a_1} K_1(t) dt \int_{|y|>\alpha_N - a_1 b_n} h_\theta(y) dy \\
&= P(|Y_1| > \alpha_N - a_1 b_n).
\end{aligned}
$$

From (C6) we have $N \cdot E[\int \varepsilon_{Ni}(x) w_1^{-1/2}(x) f_m^{1/2}(x) h_n^{1/2}(x) dx]^2 \to 0$ i.e., (5.31) holds.

By the Cauchy-Schwarz Inequality and a similar argument we have

$$
\begin{aligned}
& N \cdot E\left[\int \varepsilon_{Ni}(x) f^{1/2}(x) f_m^{1/2}(x) \right]^2 \\
\leq\ & N \cdot \int \left[\frac{\partial w_1(x)}{\partial \theta} \right]^2 f(x) dx \cdot E\left[\int I_{\{|x|>\alpha_N\}} f_m(x) dx \right] \\
=\ & N \cdot \int \left[\frac{\partial w_1(x)}{\partial \theta} \right]^2 f(x) dx \int\int I_{\{|x|>\alpha_N\}} \frac{1}{b_m} K_0(\frac{y-x}{b_m}) f(y) dy dx \\
\leq\ & N \cdot \int \left[\frac{\partial w_1(x)}{\partial \theta} \right]^2 f(x) dx \cdot P\left(|X_1| > \alpha_N - a_0 b_m \right),
\end{aligned}
$$

and (5.32) follows by assumptions (C4) and (C6). □

Lemma 5.4. *Suppose (C0)-(C3) and (C7) hold. Then as $N \to \infty$,*

$$
N^{1/2} \int |\delta_N(x)| w_1^{-1}(x) \left[h_n^{1/2}(x) - h_\theta^{1/2}(x) \right]^2 dx \xrightarrow{\mathscr{P}} 0, \tag{5.33}
$$

$$N^{1/2}\int|\delta_N(x)|\left[f_m^{1/2}(x)-f^{1/2}(x)\right]^2dx \xrightarrow{\mathscr{P}} 0. \tag{5.34}$$

Proof. Note that

$$\begin{aligned}
& N^{1/2}\int|\delta_N(x)|w_1^{-1}(x)\left[h_n^{1/2}(x)-h_\theta^{1/2}(x)\right]^2dx \\
\leq\ & N^{1/2}\int|\delta_N(x)|w_1^{-1}(x)h_\theta^{-1}(x)\left[h_n(x)-h_\theta(x)\right]^2 \\
\leq\ & 2\left\{N^{1/2}\int|\delta_N(x)|w_1^{-1}(x)h_\theta^{-1}(x)\left[h_n(x)-E(h_n(x))\right]^2dx\right. \\
& \left.+N^{1/2}\int|\delta_N(x)|w_1^{-1}(x)h_\theta^{-1}(x)\left[E(h_n(x))-h_\theta(x)\right]^2dx\right\} \\
=\ & 2(B_{1N}+B_{2N}),\ \text{say.}
\end{aligned}$$

By conditions (C0), (C2), (C3) and (C7), we have, as $N\to\infty$,

$$\begin{aligned}
E[B_{1N}] &= N^{1/2}\int|\delta_N(x)|w_1^{-1}(x)h_\theta^{-1}(x)Var[h_n(x)]dx \\
&\leq N^{1/2}\int|\delta_N(x)|w_1^{-1}(x)h_\theta^{-1}(x)\frac{1}{nb_n^2}\int K_1^2\left(\frac{y-x}{b_n}\right)h_\theta(y)dydx \\
&= N^{1/2}n^{-1}b_n^{-1}\int|\delta_N(x)|w_1^{-1}(x)\int_{-a_1}^{a_1}K_1^2(t)h_\theta(x+tb_n)h_\theta^{-1}(x)dtdx \\
&= N^{1/2}n^{-1}b_n^{-1}\int|\delta_N(x)|w_1^{-1}(x)\sup_{|t|\leq a_1}\frac{h_\theta(x+tb_n)}{h_\theta(x)}dx\int_{-a_1}^{a_1}K_1^2(t)dt \\
&\to 0
\end{aligned}$$

i.e., $B_{1N}\xrightarrow{\mathscr{P}}0$ as $N\to\infty$. Using a Taylor expansion and conditions (C1) and (C7) we have, as $N\to\infty$,

$$\begin{aligned}
|B_{2N}| &= N^{1/2}\int|\delta_N(x)|w_1^{-1}(x)h_\theta^{-1}(x)\left[\int_{-a_1}^{a_1}K_1(t)\left[h_\theta(x+tb_n)-h_\theta(x)\right]dt\right]^2dx \\
&\leq \frac{1}{4}N^{1/2}b_n^4\int|\delta_N(x)|w_1^{-1}(x)h_\theta^{-1}(x)\left[\sup_{|t|\leq a_1}|h_\theta^{(2)}(x+tb_n)|\int_{-a_1}^{a_1}t^2K_1(t)dt\right]^2dx \\
&\leq \frac{1}{4}N^{1/2}b_n^4\int|\delta_N(x)|f(x)\sup_{|t|\leq a_1}\left[\frac{h_\theta^{(2)}(x+tb_n)}{h_\theta(x)}\right]^2dx\left(\int_{-a_1}^{a_1}t^2K_1(t)dt\right)^2 \\
&\to 0
\end{aligned}$$

Hence (5.33) holds. Using similar idea one can prove (5.34). □

Lemma 5.5. *Suppose (C0)-(C8) hold. Then the asymptotic distribution of*

$$N^{1/2}\int\frac{\partial w_1(x)}{\partial\theta}\frac{f_m^{1/2}(x)}{w_1^{1/2}(x)}\left[h_n^{1/2}(x)-h_\theta^{1/2}(x)\right]dx \tag{5.35}$$

is the same as that of

$$N^{1/2}\int \delta_N(x)\frac{f^{1/2}(x)}{w_1^{1/2}(x)}\left[h_n^{1/2}(x)-h_\theta^{1/2}(x)\right]dx.$$

Proof. Lemma 5.4 gives

$$N^{1/2}\int \varepsilon_N(x)w_1^{-1/2}(x)f_m^{1/2}(x)\left[h_n^{1/2}(x)-h_\theta^{1/2}(x)\right]dx \to 0,$$

thus the asymptotic distribution of (5.35) is the same as that of

$$N^{1/2}\int \delta_N(x)w_1^{-1/2}(x)f_m^{1/2}(x)\left[h_n^{1/2}(x)-h_\theta^{1/2}(x)\right]dx \to 0.$$

If we can prove $N^{1/2}\int \delta_N(x)w_1^{-1/2}(x)\left[f_m^{1/2}(x)-f^{1/2}(x)\right]\left[h_n^{1/2}(x)-h_\theta^{1/2}(x)\right]dx \xrightarrow{\mathscr{P}} 0$, then the result. By Cauchy-Schwarz inequality and Lemma 5.4,

$$\begin{aligned}
&\left\{N^{1/2}\int \delta_N(x)w_1^{-1/2}(x)\left[f_m^{1/2}(x)-f^{1/2}(x)\right]\left[h_n^{1/2}(x)-h_\theta^{1/2}(x)\right]dx\right\}^2\\
\leq\ & N^{1/2}\int |\delta_N(x)|\left[f_m^{1/2}(x)-f^{1/2}(x)\right]^2 dx\cdot N^{1/2}\int |\delta_N(x)|w_1^{-1}(x)\left[h_n^{1/2}(x)-h_\theta^{1/2}(x)\right]^2 dx\\
\xrightarrow{\mathscr{P}}\ & 0.
\end{aligned}$$

□

Lemma 5.6. *Suppose (C4) and (C6) hold. Then as $N\to\infty$,*

$$N^{1/2}\int |\varepsilon_N(x)|f(x)dx \longrightarrow 0,$$

$$N^{1/2}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_N(Y_i)w_1^{-1}(Y_i) \xrightarrow{\mathscr{P}} 0,$$

$$N^{1/2}\frac{1}{m}\sum_{i=1}^{m}\varepsilon_N(X_i) \xrightarrow{\mathscr{P}} 0.$$

Proof. By the Cauchy-Schwarz inequality and conditions (C4) and (C6),

$$\begin{aligned}
N^{1/2}\int |\varepsilon_N(x)|f(x)dx &\leq \left[N\int I_{\{|x|>\alpha_N\}}h_\theta(x)dx\right]^{1/2}\left[\int\left(\frac{\partial w_1(x)}{\partial\theta}\right)^2\frac{f}{w_1}(x)dx\right]^{1/2}\\
&= [N\cdot P(|Y_1|>\alpha_N)]^{1/2}\left[\int\left(\frac{\partial w_1(x)}{\partial\theta}\right)^2\frac{f}{w_1}(x)dx\right]^{1/2}\\
&\to 0.
\end{aligned}$$

It follows that

$$\begin{aligned} E\left|N^{1/2}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_N(Y_i)w_1^{-1}(Y_i)\right| &\leq E\left[N^{1/2}\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_N(Y_i)|w_1^{-1}(Y_i)\right] \\ &= N^{1/2}\int|\varepsilon_N(x)|f(x)dx \\ &\to 0, \end{aligned}$$

i.e. $N^{1/2}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_N(Y_i)w_1^{-1}(Y_i) \stackrel{\mathscr{P}}{\longrightarrow} 0$. Similarly,

$$\begin{aligned} E\left|N^{1/2}\frac{1}{m}\sum_{i=1}^{m}\varepsilon_N(X_i)\right| &\leq E\left[N^{1/2}\frac{1}{m}\sum_{i=1}^{m}|\varepsilon_N(X_i)|\right] \\ &= N^{1/2}\int|\varepsilon_N(x)|\,f(x)dx \\ &\to 0, \end{aligned}$$

i.e. $N^{1/2}\frac{1}{m}\sum_{i=1}^{m}\varepsilon_N(X_i) \stackrel{\mathscr{P}}{\longrightarrow} 0$. □

Lemma 5.7. *Suppose (C0)-(C4) and (C8)-(C10) hold. Then as $N\to\infty$,*

$$N^{1/2}\left[\int\delta_N(x)w_1^{-1}(x)h_n(x)dx-\frac{1}{n}\sum_{i=1}^{n}\delta_N(Y_i)w_1^{-1}(Y_i)\right] \stackrel{\mathscr{P}}{\longrightarrow} 0,$$

$$N^{1/2}\left[\int\delta_N(x)f_m(x)dx-\frac{1}{m}\sum_{i=1}^{m}\delta_N(X_i)\right] \stackrel{\mathscr{P}}{\longrightarrow} 0.$$

Proof. We will only give the proof of the second convergence and the proof of the first convergence is similar. Let

$$D_{Ni} \;=\; N^{1/2}\left[\int\delta_{Ni}(x)f_m(x)dx-\frac{1}{m}\sum_{i=1}^{m}\delta_{Ni}(X_i)\right],$$

then by (C8),

$$\begin{aligned} |E(D_{Ni})| &= N^{1/2}\left|\int\delta_{Ni}(x)E[f_m(x)]dx-\int\delta_{Ni}(x)f(x)dx\right| \\ &= N^{1/2}\left|\int\delta_{Ni}(x)\int_{-a_0}^{a_0}K_0(t)[f(x+tb_m)-f(x)]dtdx\right| \\ &\leq N^{1/2}b_m^2\int|\delta_{Ni}(x)|f(x)\sup_{|t|\leq a_0}\frac{|f^{(2)}(x+tb_m)|}{f(x)}dx\int_{-a_0}^{a_0}t^2K_0(t)dt \\ &\to 0. \end{aligned}$$

On the other hand,

$$\begin{aligned}
Var(D_{Ni}) &= \frac{N}{m}Var\left[\int \delta_{Ni}(x)\frac{1}{b_m}K_0(\frac{x-X_1}{b_m})dx - \delta_{Ni}(X_1)\right] \\
&\le \frac{N}{m}E\left[\int \delta_{Ni}(x)\frac{1}{b_m}K_0(\frac{x-X_1}{b_m})dx - \delta_{Ni}(X_1)\right]^2 \\
&= \frac{N}{m}E\left[\int_{-a_0}^{a_0} K_0(t)\left(\delta_{Ni}(X_1+tb_m) - \delta_{Ni}(X_1)\right)dt\right]^2 \\
&= \frac{N}{m}E\left[\int_{-a_0}^{a_0} K_0(t)\left(\frac{\partial w_1(X_1+tb_m)}{\partial\theta}\right)_i\left(I_{\{|X_1+tb_m|\le\alpha_N\}} - I_{\{|X_1|\le\alpha_N\}}\right)dt\right. \\
&\qquad \left. + \int_{-a_0}^{a_0} K_0(t)I_{\{|X_1|\le\alpha_N\}}\left(\frac{\partial w_1(X_1+tb_m)}{\partial\theta} - \frac{\partial w_1(X_1)}{\partial\theta}\right)_i\right]^2 dt \\
&\le \frac{2N}{m}\left\{E\left[\int_{-a_0}^{a_0} K_0(t)\left(\frac{\partial w_1(X_1+tb_m)}{\partial\theta}\right)_i\left(I_{\{|X_1+tb_m|\le\alpha_N\}} - I_{\{|X_1|\le\alpha_N\}}\right)dt\right]^2\right. \\
&\qquad \left. + E\left[\int_{-a_0}^{a_0} K_0(t)I_{\{|X_1|\le\alpha_N\}}\left(\frac{\partial w_1(X_1+tb_m)}{\partial\theta} - \frac{\partial w_1(X_1)}{\partial\theta}\right)_i\right]^2 dt\right\} \\
&= \frac{2N}{m}(B_{Ni}+C_{Ni}), \text{ say.}
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
B_{Ni} &\le E\left[\int_{-a_0}^{a_0} K_0(t)\left(\frac{\partial w_1(X_1+tb_m)}{\partial\theta}\right)_i^2\left(I_{\{|X_1+tb_m|\le\alpha_N\}} - I_{\{|X_1|\le\alpha_N\}}\right)^2 dt\right] \\
&= \int_{-a_0}^{a_0} K_0(t)\int\left(\frac{\partial w_1(x+tb_m)}{\partial\theta}\right)_i^2\left(I_{\{|x+tb_m|\le\alpha_N\}} - I_{\{|x|\le\alpha_N\}}\right)^2 f(x)dxdt \\
&= \int_0^{a_0} K_0(t)\left\{\int_{-\alpha_N-tb_m}^{-\alpha_N}\left(\frac{\partial w_1(x+tb_m)}{\partial\theta}\right)_i^2 f(x)dx + \int_{\alpha_N-tb_m}^{\alpha_N}\left(\frac{\partial w_1(x+tb_m)}{\partial\theta}\right)_i^2 f(x)dx\right\}dt \\
&\quad + \int_{-a_0}^{0} K_0(t)\left\{\int_{-\alpha_N}^{-\alpha_N-tb_m}\left(\frac{\partial w_1(x+tb_m)}{\partial\theta}\right)_i^2 f(x)dx + \int_{\alpha_N}^{\alpha_N-tb_m}\left(\frac{\partial w_1(x+tb_m)}{\partial\theta}\right)_i^2 f(x)dx\right\}dt.
\end{aligned} \tag{5.36}$$

Note that $\left(\frac{\partial w_1(x)}{\partial\theta}\right)_i^2 f(x)$ is bounded by (C0) and (C4), and therefore by (C9)

$$\begin{aligned}
& \int_0^{a_0} K_0(t)\int_{-\alpha_N-tb_m}^{-\alpha_N}\left(\frac{\partial w_1(x+tb_m)}{\partial\theta}\right)_i^2 f(x)dxdt \\
&= \int_0^{a_0} K_0(t)\int_{-\alpha_N}^{-\alpha_N+tb_m}\left(\frac{\partial w_1(x)}{\partial\theta}\right)_i^2 f(x-tb_m)dxdt \\
&\le \sup_{|x|\le\alpha_N}\sup_{|t|\le a_0}\frac{f(x+tb_m)}{f(x)}\int_0^{a_0} K_0(t)\int_{-\alpha_N}^{-\alpha_N+tb_m}\left(\frac{\partial w_1(x)}{\partial\theta}\right)_i^2 f(x)dxdt \\
&= O\left(b_m\int_0^{a_0} tK_0(t)dt\right) \\
&\to 0.
\end{aligned}$$

The other three terms on the right hand side of equation (5.36) go to zero using similar argument. Therefore, $B_{Ni} \to 0$ as $N \to \infty$.

For C_{Ni}, using the Cauchy-Schwarz inequality and condition (C10) we have

$$\begin{aligned}
C_{Ni} &\leq E\left[\int_{-a_0}^{a_0} K_0(t) I_{\{|X_1|\leq \alpha_N\}} \left(\frac{\partial w_1(X_1+tb_m)}{\partial\theta} - \frac{\partial w_1(X_1)}{\partial\theta}\right)_i^2\right] dt \\
&= \int_{-a_0}^{a_0} K_0(t) \int I_{\{|x|\leq \alpha_N\}} \left(\frac{\partial w_1(x+tb_m)}{\partial\theta} - \frac{\partial w_1(x)}{\partial\theta}\right)_i^2 f(x)dxdt \\
&\leq b_m^2 \int I_{\{|X|\leq \alpha_N\}} f(x) \sup_{|t|\leq a_0} \left[\frac{\partial^2 w_1(y)}{\partial\theta\partial y}|_{y=x+tb_m}\right]_i^2 dx \int_{-a_0}^{a_0} t^2 K_0(t) dt \\
&\to 0.
\end{aligned}$$

Thus $Var(D_{Ni}) \to 0$ as $N \to \infty$ and $E(D_{Ni}^2) = Var(D_{Ni}) + (E(D_{Ni}))^2 \to 0$. Therefore $D_{Ni} \xrightarrow{\mathscr{P}} 0$ as $N \to \infty$. $\square$

Corollary 5.1. *Suppose that (C0)-(C10) hold. Then the asymptotic distribution of (5.35) is* $N(0, \frac{1}{4\rho}\Delta(\theta))$ *with* $\Delta(\theta)$ *defined in (5.30).*

Proof. In view of Lemma 5.5, we only need to give the asymptotic distribution of $N^{1/2}\int \delta_N(x)\frac{f^{1/2}}{w_1^{1/2}}(x)\left[h_n^{1/2}(x) - h_\theta^{1/2}(x)\right]dx$. By using the following algebraic expression, with > 0, $b \geq 0$,

$$b^{1/2} - a^{1/2} = \frac{b-a}{2a^{1/2}} - \frac{(b^{1/2}-a^{1/2})^2}{2a^{1/2}}, \tag{5.37}$$

we have that, as $N \to \infty$,

$$
\begin{aligned}
& N^{1/2} \int \delta_N(x) \frac{f^{1/2}}{w_1^{1/2}}(x) \left[h_n^{1/2}(x) - h_\theta^{1/2}(x) \right] dx \\
= \; & \frac{1}{2} N^{1/2} \int \delta_N(x) w_1^{-1}(x) \left[h_n(x) - h_\theta(x) \right] dx - \frac{1}{2} N^{1/2} \int \delta_N(x) w_1^{-1}(x) \left[h_n^{1/2}(x) - h_\theta^{1/2}(x) \right]^2 dx \\
= \; & \frac{1}{2} N^{1/2} \int \delta_N(x) w_1^{-1}(x) \left[h_n(x) - h_\theta(x) \right] dx + o_p(1) \quad \text{(by Lemma 5.4)} \\
= \; & \frac{1}{2} N^{1/2} \left[\frac{1}{n} \sum_{i=1}^n \delta_N(Y_i) w_1^{-1}(Y_i) - \int \delta_N(x) f(x) dx \right] \\
& \quad + \frac{1}{2} N^{1/2} \left[\int \delta_N(x) w_1^{-1}(x) h_n(x) dx - \frac{1}{n} \sum_{i=1}^n \delta_N(Y_i) w_1^{-1}(Y_i) \right] + o_p(1) \\
= \; & \frac{1}{2} N^{1/2} \left[\frac{1}{n} \sum_{i=1}^n \delta_N(Y_i) w_1^{-1}(Y_i) - \int \delta_N(x) f(x) dx \right] + o_p(1) \quad \text{(by Lemma 5.7)} \\
= \; & \frac{1}{2} N^{1/2} \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial w_1(Y_i)}{\partial \theta} w_1^{-1}(Y_i) - \int \frac{\partial w_1(x)}{\partial \theta} f(x) dx \right] + o_p(1) \quad \text{(by Lemma 5.6).}
\end{aligned}
$$

Now by CLT the asymptotic distribution of $n^{1/2}[\frac{1}{n}\sum_{i=1}^n \frac{\partial w_1(Y_i)}{\partial \theta} w_1^{-1}(Y_i) - \int \frac{\partial w_1(x)}{\partial \theta} f(x)dx]$ is $N(0, \Delta(\theta))$. □

Lemma 5.8. *Suppose that (C0)-(C7) and (C11) hold. Then the asymptotic distribution of*

$$
N^{1/2} \int \frac{\partial w_1(x)}{\partial \theta} f_m^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x) \right] dx \tag{5.38}
$$

is the same as that of

$$
N^{1/2} \int \delta_N(x) f^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x) \right] dx.
$$

Proof. By a Taylor expansion, condition (C11) and Lemma 5.6, for $i = 1, 2, 3$,

$$
\begin{aligned}
& E \left| N^{1/2} \int \varepsilon_{Ni}(x) f_m(x) dx \right| \\
\leq \; & N^{1/2} \int |\varepsilon_{Ni}(x)| \int_{-a_0}^{a_0} K_0(t) f(x + t b_m) dt dx \\
\leq \; & N^{1/2} \int |\varepsilon_{Ni}(x)| \int_{-a_0}^{a_0} K_0(t) \left[f(x) + f^{(1)}(x) t b_m + \frac{1}{2} t^2 b_m^2 \sup_{|t| \leq a_0} |f^{(2)}(x + t b_m)| \right] dt dx \\
\leq \; & N^{1/2} \int |\varepsilon_{Ni}(x)| f(x) dx + \frac{1}{2} N^{1/2} b_m^2 \int |\varepsilon_{Ni}(x)| f(x) \sup_{|t| \leq a_0} \frac{|f^{(2)}(x + t b_m)|}{f(x)} dx \int_{-a_0}^{a_0} t^2 K_0(t) dt \\
\to \; & 0.
\end{aligned}
$$

Thus $N^{1/2} \int \varepsilon_N(x) f_m(x) \xrightarrow{\mathscr{P}} 0$. Combined with (5.32) from Lemma 5.3, we have

$$N^{1/2} \int \varepsilon_N(x) f_m^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x) \right] dx \xrightarrow{\mathscr{P}} 0.$$

So the asymptotic distribution of (5.38) is same as that of

$$N^{1/2} \int \delta_N(x) f_m^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x) \right] dx \xrightarrow{\mathscr{P}} 0.$$

The result follows from (5.34) in Lemma 5.4. □

Corollary 5.2. *Suppose that (C0)-(C11) hold. Then the asymptotic distribution of (5.38) in* $N(0, \frac{1}{4(1-\rho)} \bar{\Delta}(\theta))$ *with* $\bar{\Delta}(\theta)$ *defined in (5.29).*

Proof. The proof is very similar to that of Corollary 5.1. In view of Lemma 5.8, we only need to give the asymptotic distribution of $N^{1/2} \int \delta_N(x) f^{1/2}(x) [f_m^{1/2}(x) - f^{1/2}(x)] dx$. Applying the same algebraic expression (5.37) we have that, as $N \to \infty$,

$$\begin{aligned}
& N^{1/2} \int \delta_N(x) f^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x) \right] dx \\
= \; & \frac{1}{2} N^{1/2} \int \delta_N(x) \left[f_m(x) - f(x) \right] dx + \frac{1}{2} N^{1/2} \int \delta_N(x) \left[f_m^{1/2}(x) - f^{1/2}(x) \right]^2 dx \\
= \; & \frac{1}{2} N^{1/2} \int \delta_N(x) [f_m(x) - f(x)] dx + o_p(1) \quad \text{(By Lemma 5.4)} \\
= \; & \frac{1}{2} N^{1/2} \left[\frac{1}{m} \sum_{i=1}^{m} \delta_N(X_i) - \int \delta_N(x) f(x) dx \right] + \frac{1}{2} N^{1/2} \left[\int \delta_N(x) f_m(x) dx - \frac{1}{m} \sum_{i=1}^{m} \delta_N(X_i) \right] + o_p(1) \\
= \; & \frac{1}{2} N^{1/2} \left[\frac{1}{m} \sum_{i=1}^{m} \delta_N(X_i) - \int \delta_N(x) f(x) dx \right] + o_p(1) \quad \text{(By Lemma 5.7)} \\
= \; & \frac{1}{2} N^{1/2} \left[\frac{1}{m} \sum_{i=1}^{m} \frac{\partial w_1(X_i)}{\partial \theta} - \int \frac{\partial w_1(x)}{\partial \theta} f(x) dx \right] + o_p(1) \quad \text{(By Lemma 5.6).}
\end{aligned}$$

Note that by CLT the asymptotic distribution of $m^{1/2} [\frac{1}{m} \sum_{i=1}^{m} \frac{\partial w_1(X_i)}{\partial \theta} - \int \frac{\partial w_1(x)}{\partial \theta} f(x) dx]$ is $N(0, \bar{\Delta}(\theta))$. Hence the result. □

Proof of Theorem 5.4. By (5.12), we only need to find the asymptotic distribution of $A_N(\theta)$. From (5.11) we have,

$$\begin{aligned}
& N^{1/2} \int \frac{\partial w_1(x)}{\partial \theta} w_1^{-1/2}(x) f_m^{1/2}(x) h_n^{1/2}(x) dx - N^{1/2} \int \frac{\partial w_1(x)}{\partial \theta} f_m(x) dx \\
= \; & N^{1/2} \int \frac{\partial w_1(x)}{\partial \theta} \frac{f_m^{1/2}(x)}{w_1^{1/2}(x)} \left[h_n^{1/2}(x) - h_\theta^{1/2}(x) \right] dx - N^{1/2} \int \frac{\partial w_1(x)}{\partial \theta} f_m^{1/2}(x) \left[f_m^{1/2}(x) - f^{1/2}(x) \right] dx.
\end{aligned}$$

Because the two terms on the right hand side of the preceding expression are independent, we only need to find their individual asymptotic distributions. Then by Corollaries 5.1 and Corollary 5.2 and Slutsky's theorem, we have the result. $\square$

Chapter 6

Test of the Semiparametric Model

In this chapter we discuss the validity of the semiparametric mixture model (4.3), or equivalently the model (4.2), with $r(x) = x$ assumed for both Chapters 4 and 5. In Section 6.1, we construct two Kolmogorov-Smirnov (K-S) type test statistics based on MLE and MHDE that we proposed in Chapters 4 and 5 respectively. We also discuss how to use bootstrap method to find the approximate distribution of the constructed test statistics. In Section 6.2 we use simulation studies to demonstrate the performance of the two tests.

6.1 Kolmogorov-Smirnov tests based on MLE and MHDE

Several goodness-of-fit test statistics for testing the model (4.2) in case-control studies are available in literature; see, for example, Qin and Zhang (1997), Zhang (1999, 2001 & 2006), and Deng, Wan and Zhang (2009). Zhang (1999) considered a chi-squared statistic to test the validity of (4.2) by adapting the goodness-of-fit test of Nikulin-Rao-Robson-Moore. Zhang (2001) suggested a test based on information matrix which requires high-dimensional matrix inversion. For a semiparametric finite mixture model where a sample is available from each component as well as from the mixture, a test based on score statistics is discussed by Zhang (2006). Deng, Wan and Zhang (2009) proposed an improved goodness-of-fit test introduced by Zhang (1999) by randomly partitioning the case-control data. Qin and Zhang (1997) proposed a K-S type statistic based on MLE to test the validity of (4.2) and used a bootstrap sampling technique to find critical values of the test statistic. In this chapter, we propose similar K-S type statistics but for our special model (4.3) and based on both MLE and MHDE.

The idea of K-S test statistic is to use the discrepancy between two c.d.f. estimates, one with the model assumption and the other without, to assess the validity of a model. For our model (4.3) with $r(x) = x$, we can use the empirical c.d.f. based on the first sample X_i's as the first estimation and the MLE or MHDE based on both samples X_i's and Y_i's exploiting (4.3) as the second. We first look at the special case of model (4.3) when $\beta = 0$. Note that α is only a standardization parameter and $\alpha = 0$ whenever $\beta = 0$. In model (4.3), $\beta = 0$ implies the equality of the two components F and G in the mixture, and thus the equality of F and H. For testing the equality of two populations, a commonly used test statistic is the K-S statistic. The two-sample K-S statistic for testing the equality of F and H is given by

$$\sup_t |\hat{F}(t) - \hat{H}(t)| \; = \; \frac{N}{n} \sup_t |\hat{F}(t) - \tilde{F}_0(t)|, \tag{6.1}$$

where $N = m+n$, $(T_1, \ldots, T_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is the pooled sample, and

$$\begin{aligned} \hat{F}(t) \; &= \; \frac{1}{m} \sum_{i=1}^{m} I(X_i \le t), \qquad (6.2) \\ \hat{H}(t) \; &= \; \frac{1}{n} \sum_{i=1}^{n} I(Y_i \le t), \\ \tilde{F}_0(t) \; &= \; \frac{1}{N} \sum_{i=1}^{N} I(T_i \le t). \end{aligned}$$

Note that $\hat{F}$ and $\hat{H}$ are, respectively, the nonparametric MLE of F and H without the assumption of $F = H$, whereas $\tilde{F}_0$ is the nonparametric MLE of F with the assumption of $F = H$. Now consider the general case of $\beta \ne 0$. Motivated by the construction of K-S statistic 6.1, to test the validity of model (4.3) with $r(x) = x$, we propose to use the test statistic

$$KS \; = \; N^{1/2} \sup_t \left| \hat{F}(t) - \tilde{F}(t) \right|, \tag{6.3}$$

where the empirical distribution $\hat{F}$ is given in (6.2) and $\tilde{F}$ is either the MLE or an estimator of F based on MHDE of $\theta = (\lambda, \alpha, \beta)^\top$ with model assumption (4.3). Recall in Chapter 4, with $p_i = dF(T_i)$ and $\rho = n/N$, the likelihood function under (4.3) is given by

$$L(\lambda, \alpha, \beta) \; = \; \prod_{i=1}^{m} dF(x_i) \prod_{j=1}^{n} dH(y_j) \; = \; \prod_{i=1}^{N} p_i \prod_{j=1}^{n} \left[(1-\lambda) + \lambda e^{\alpha + \beta y_j} \right]$$

and the MLE of p_i is given by

$$\hat{p}_i \;=\; \frac{1}{N\left[1+\rho\hat{\lambda}\left(e^{\hat{\alpha}+\hat{\beta}T_i}-1\right)\right]},\;\; i=,1,\ldots,N. \tag{6.4}$$

Now the estimator $\tilde{F}$ of F under model (4.3) is given by

$$\tilde{F}(t) \;=\; \sum_{i=1}^{N}\hat{p}_i I(T_i\leq t) \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{I(T_i\leq t)}{1-\rho\hat{\lambda}+\rho\hat{\lambda}e^(\hat{\alpha}+\hat{\beta}T_i)}. \tag{6.5}$$

If the $\hat{\theta}=(\hat{\lambda},\hat{\alpha},\hat{\beta})^\top$ in (6.4) and (6.5) is the MLE $\hat{\theta}_{MLE}$ we constructed in Chapter 4, then the resulting $\tilde{F}_{MLE}$ is the actual MLE of F under (4.3) and we denote the corresponding test statistic in (6.3) as KS_{MLE}. This test statistic is essentially the same as that in Qin and Zhang (1997), but they used it for case-control data instead of our more complicated mixture model (4.3). Intuitively, we can also use the MHDE $\hat{\theta}_{MHDE}$ we proposed in (5.3) of Chapter 5 for $\hat{\theta}$, then we denote the resulting $\tilde{F}$ in (6.5) and KS in (6.3) as $\tilde{F}_{MHDE}$ and KS_{MHDE} respectively.

To use the two test statistics KS_{MLE} and KS_{MHDE} to test the validity of model (4.3), we need to give the distribution of them or at least their approximated distributions. Following the same idea as in Qin and Zhang (1997), we use bootstrap procedure to find the approximated distributions and critical values for hypothesis testing. To generate bootstrapping data, we randomly select independent samples X_i^*'s from $d\tilde{F}(x)$ and Y_i^*'s from $(1-\hat{\lambda}+\hat{\lambda}e^{\hat{\alpha}+\hat{\beta}x})d\tilde{F}(x)$, where $\hat{\theta}$ and $\tilde{F}$ are either the MLEs $\hat{\theta}_{MLE}$ and $\tilde{F}_{MLE}$ or the MHDEs $\hat{\theta}_{MHDE}$ and $\tilde{F}_{MHDE}$ respectively. Note that both X_i^*'s and Y_i^*'s are selected from the pooled data $(T_1,\ldots,T_N)$ but with different probability distribution function. This means that some of the selected X_i^*'s could be values in the original second sample Y_i's and some of the selected Y_i^*'s could be values in the original first sample X_i's. Let $(T_1^*,\ldots,T_N^*)$ denote the combined bootstrapping sample and $\hat{\theta}^*=(\hat{\lambda}^*,\hat{\alpha}^*,\hat{\beta}^*)^\top$ be either the MLE or the MHDE based on the bootstrapping samples X_i^*'s and Y_i^*'s. Then we can calculate the function in (6.2) based on X_i^*'s, the quantities in (6.4) and the function in (6.5) based on T_i^*'s and $\hat{\theta}^*$, with results denoted by $\hat{F}^*$, $\hat{p}_i^*$ and $\tilde{F}^*$ respectively. Then finally the bootstrapping KS test statistic is

$$KS^* \;=\; N^{1/2}\sup_t\left|\hat{F}^*(t)-\tilde{F}^*(t)\right|.$$

We can generate 1000 bootstrapping samples to give 1000 bootstrapping KS statistic values for both KS_{MLE} and KS_{MHDE} at the same time. Then the distributions, and thus the critical values, of KS_{MLE} and KS_{MHDE} can be estimated by these respective 1000 statistic values.

6.2 Simulation study

In our simulation study, we consider model (4.3) with $r(x) = (x, x^2)^\top$ as the collection of all possible models under consideration. Then we test whether the reduced model (4.3) with $r(x) = x$ is the actual true model or not. For demonstration purpose, we only consider mixture of normals $H(x) = (1-\lambda)F(x) + \lambda G(x)$ with $F \sim N(0,1)$ and $G \sim N(\mu, \sigma^2)$. Then $f(x)$ and $h(x)$ are related by

$$h_\theta(x) \;=:\; h(x) \;=\; \left(1 - \lambda + \lambda e^{\alpha + \beta x + \gamma x^2}\right) f(x), \tag{6.6}$$

where

$$\alpha = -\frac{1}{2}\left(\log \sigma^2 + \frac{\mu^2}{\sigma^2}\right), \quad \beta = \frac{\mu}{\sigma^2}, \quad \gamma = \frac{1}{2}\left(1 - \frac{1}{\sigma^2}\right). \tag{6.7}$$

Note that (6.6) is a special case of (4.3) when $r(x) = (x, x^2)^\top$. If $\sigma = 1$, then $\gamma = 0$ and thus model (4.3) holds with $r(x) = x$. So testing the validity of model (4.3) with $r(x) = x$ is equivalent to testing the null hypothesis $H_0 : \gamma = 0$ under model (6.6). In our simulation study, we consider $\gamma = 0$, -0.9 and -1.5, $\lambda = 0.35$ and 0.65, and sample sizes $m = n = 30$ and $m = n = 100$. For simplicity, we just fix $\mu = 1$ and as a result $\sigma = 1, 0.6$ and 0.5 for $\gamma = 0, -0.9$ and -1.5 respectively. For each λ, γ and sample size considered, we use 500 total number of replications for our calculation. Within each replication, we use totally 1000 bootstrapping samples to estimate the distribution and critical value of the test statistics KS_{MLE} and KS_{MHDE}. We choose different level of significance $a = 0.10$, 0.05 and 0.01. The simulation results are presented in Table 6.1. Note that $\gamma = 0$ means model (4.3) with $r(x) = x$ is correct and thus the correspondingly calculated values in Table 6.1 are the estimated significance levels. When $\gamma \neq 0$, model (4.3) with $r(x) = x$ is not correct and thus the correspondingly calculated values in Table 6.1 are the estimated powers at that value of γ.

Table 6.1: Estimated significance level and power of KS_{MLE} and KS_{MHDE}.

			$m=n=30$		$m=n=100$	
λ	γ	Significance level	KS_{MLE}	KS_{MHDE}	KS_{MLE}	KS_{MHDE}
0.35	0	0.10	0.040	0.104	0.156	0.186
		0.05	0.030	0.014	0.122	0.084
		0.01	0.002	0.000	0.002	0.002
	-0.9	0.10	0.950	0.860	0.956	0.870
		0.05	0.904	0.802	0.910	0.710
		0.01	0.734	0.410	0.578	0.184
	-1.5	0.10	0.948	0.966	0.958	0.998
		0.05	0.898	0.912	0.910	0.984
		0.01	0.716	0.580	0.536	0.846
0.65	0	0.10	0.036	0.388	0.096	0.136
		0.05	0.030	0.170	0.122	0.056
		0.01	0.008	0.010	0.002	0.006
	-0.9	0.10	0.970	0.910	0.894	0.928
		0.05	0.888	0.818	0.708	0.758
		0.01	0.464	0.282	0.158	0.120
	-1.5	0.10	0.956	0.876	0.990	0.990
		0.05	0.858	0.762	0.908	0.944
		0.01	0.424	0.302	0.174	0.396

From Table 6.1 we can see that, the two test statistics KS_{MLE} and KS_{MHDE} are quite competitive in terms of achieved significance level and power. The achieved levels of significance are quite close to the true levels for most of the cases except for the case of KS_{MHDE} with $\lambda = 0.65$ and $m = n = 30$. The powers of KS_{MHDE} become larger when γ is away from 0 except for the case with $\lambda = 0.65$ and $m = n = 30$. Surprisingly, the powers of KS_{MLE} become smaller when γ is away from 0 except for the case with $\lambda = 0.65$ and $m = n = 100$. As expected, when the significance level a decrease, both the observed significance level and power decrease. For both KS_{MLE} and KS_{MHDE}, the powers are generally high for significance levels $a = 0.10$ and 0.05.

Chapter 7

Data Examples and Discussion

In Section 7.1 we consider two real life data examples and demonstrate the application of our proposed estimators. In Section 7.2 we summarize the whole thesis and present some discussion of future work.

7.1 Two real data examples

Example 1: Grain data.

Smith and Vounatsou (1997) analyzed a data where an autoradiography assay was used to determine the intracellular transfer of small molecules in mouse cells in culture. The assay was used to determine the proportion of cells in the test population which were exposed to radio active materials. The cells in control group were not exposed to radioactivity, but otherwise were similar in nature. Autoradiograph of the cells can determine the amount of radio active material in the cell by counting the number of grains, X. Now grains can appear in autoradiograph due to the presence of radioactive material or due to background fogging. Hence the proportion of cells with radio active material can only be revealed by comparing the distribution of grain counts in test sample and that in control sample. This data set is given in Table 7.1.

This data were originally analyzed by Smith, Smith and Hooper (1986) by fitting a parametric Poisson mixture model to density ratio. Later Smith and Vounatsou (1997) proposed several estimation methods including odds ratio (using two-by-two table), logistic power model, nonparametric monotone regression, and latent class model. We apply all the four estimation methods we proposed in Chapters 2-5 for this data and compare them with the estimators in Smith and Vounatsou (1997) and Smith, Smith and Hooper (1986). To calculate

Table 7.1: Frequency distribution for the test group and control group in the grain data.

Number of grains (X)	Frequency in recipients (test sample from mixture h)	Frequency in controls (control sample from f)
0	2	3
1	2	6
2	2	12
3	3	16
4	4	8
5	3	11
6	1	9
7	2	5
8	4	9
9	2	5
10	4	5
11	3	1
12	4	3
14	3	0
15	1	0
16	2	0
17	1	1
18	2	0
$>$19	49	0
Total	94	94

a 95% confidence interval, we use the bootstrap method with 1000 bootstrapping samples to estimate the standard deviation of an estimator. Both the point and interval estimation results, for the proportion of cells in the test population which were exposed to radio active materials, are given in Table 7.2. From Table 7.2 we observe that our proposed methods give very similar point estimate of λ in comparison with the method by Smith et al. (1986) and the several estimators presented in Smith and Vounatsou (1997). In addition, all of the four estimators we proposed give reasonable confidence intervals strictly within the range $[0,1]$ and with relatively smaller widths. Comparatively, the methods based on Poisson mixture, two-by-two table and monotone logistic give very wide confidence intervals with bound either 0 or 1. Also in Table 7.2 the confidence interval in the parentheses for $\hat{\lambda}_{MLE}$ and $\hat{\lambda}_{MHDE}$ are calculated using the asymptotic covariance matrices given in Theorems 4.2 and 5.4 respectively.

From the results we see that bootstrap approximation is quite accurate.

Table 7.2: Point and interval estimation of the proportion for the grain data.

Method	Estimate	95% confidence interval
Poisson mixture (Smith, Smith and Hooper, 1986)	0.77	0.00 – 0.91
Two-by-two table (Smith and Vounatsou, 1997)	0.20	0.00 – 1.00
Logistic power (Smith and Vounatsou, 1997)	0.61	0.58 – 0.64
Monotone logistic (Smith and Vounatsou, 1997)	0.74	0.61 – 1.00
Latent class (Smith and Vounatsou, 1997)	0.73	0.63 – 0.83
$\hat{\lambda}$ based on c.d.f.s	0.78	0.68 – 0.87
$\hat{\lambda}_L$ based on multinomial approximation	0.79	0.58 – 0.88
$\hat{\lambda}_{MLE}$ based on semiparametric MLE	0.75	0.61 – 0.88(0.60 – 0.89)
$\hat{\lambda}_{MHDE}$ based on semiparametric MHDE	0.76	0.64 – 0.92(0.65 – 0.88)

Example 2: Malaria data.

We also study a clinical malaria data set discussed in Vounatsou, Smith and Smith (1998). Vounatsou, Smith and Smith (1998) considered a Bayesian approach to estimate the probabilities of children with different level of parasitaemia having fever due to malaria. Clinical malaria is diagnosed by measuring the parasite densities in a child's body who has fever. They formulated the parasite densities in children with fever using a two-component mixture model, where one component represents the parasite densities in children without clinical malaria (f) and the other with clinical malaria (g). Parasite levels in children from the community are available and are used as a training sample, i.e. a sample that comes from the component of the mixture corresponding to children without clinical malaria (f) but who may have parasites. The mixing proportion λ represents the proportion of children whose fever is attributable to malaria.

This data were first described in Kitua et al. (1996). The data arose from repeated cross-sectional surveys of parasitaemia and fever among 426 children up to one year old resided in a village in Kilombero district in Tanzania. A subset of this data was analyzed by Vounatsou, Smith and Smith (1998) where they considered children aged between 6 and 9 months and two seasons: the wet season (January-June) during which the mosquito population, and hence exposure to malaria infection, is high, and the dry season (July-December) during which the mosquito population is lower. The original data were grouped into 10 categories and the parasite level refers to the midpoint of each category. The data is given in Table 7.3.

Table 7.3: Frequency distribution of parasite density for children aged between 6 and 9 months in the malaria data.

Category	Wet Season	Frequency		Dry Season	Frequency	
	Parasite level	f	h	Parasite level	f	h
1	0	43	60	0	43	42
2	3251	40	58	11370	68	116
3	9673	3	14	34029	8	30
4	16095	3	13	56689	2	16
5	22518	2	10	79348	0	7
6	28940	1	8	102008	0	7
7	35362	0	7	124668	0	6
8	41785	1	6	147327	0	2
9	48207	1	6	169987	0	3
10	225685	0	69	290634	0	16
Total		94	251		122	245

We apply our proposed methods, $\hat{\lambda}_L$ based on multinomial approximation and the semi-parametric MLE $\hat{\lambda}_{MLE}$ based on model (4.3), to this data and compare them with the Bayesian approach proposed by Vounatsou, Smith and Smith (1998). Note that this is a discretized data, so kernel smoothing is not appropriate and as a result the $\hat{\lambda}$ based on c.d.f.s and the semiparametric MHDE $\hat{\lambda}_{MHDE}$ are not appropriate and thus not applied to this data as they use kernel density estimations. The estimation results are given in Table 7.4. The numbers in parentheses are the estimated standard errors of the corresponding estimates based on 500 bootstrapping samples. From Table 7.4 we can see that both $\hat{\lambda}_L$ and $\hat{\lambda}_{MLE}$ give consistent estimates with that of the Bayesian approach in Vounatsou, Smith and Smith (1998).

Table 7.4: The estimates of $\hat{\lambda}_L$, $\hat{\lambda}_{MLE}$ and the Bayesian method for the malaria data.

Method	Wet Season	Dry Season
$\hat{\lambda}_L$	0.435 (0.083)	0.349 (0.102)
$\hat{\lambda}_{MLE}$	0.461 (0.093)	0.330 (0.191)
Bayesian	0.444 (0.054)	0.305 (0.118)

7.2 Summary and discussion

In this thesis we studied a two-component mixture model (1.2) with a stochastic dominance constraint. We first proposed two nonparametric estimators with one based on kernel c.d.f. estimations ($\hat{\lambda}$) and the other the MLE based on multinomial approximation ($\hat{\lambda}_L$). The $\hat{\lambda}$ can be easily calculated and gives relatively small bias and MSE for large sample sizes. However for small sample sizes it doesn't show good efficiency for mixtures of discrete distributions or distributions with different support. We also computed the misclassification rate (MR) based on a simple classification rule (0.5 threshhold) and compared it with the optimal misclassification rate (OMR) when the classification likelihood function p in (1.3) is assumed known completely. It turns out that the MR of $\hat{\lambda}$ is quite close to the OMR. We next proposed $\hat{\lambda}_L$, the MLE based on multinomial approximation. This estimator gives slightly improved performance in terms of smaller bias and MSE over $\hat{\lambda}$. We lose some information in multinomial approximation as the original data is grouped. One disadvantage of this method is that the discretization is somewhat arbitrary and results in higher MR than $\hat{\lambda}$. Some theoretical results, such as model identifiability and estimation consistency, are discussed for both estimators.

As nonparametric mixture model generally suffers from identifiability problem, we consider a semiparametric structure (4.2) for the two components. For the resulting two-sample semiparametric mixture model (4.3), we constructed the MLE $\hat{\lambda}_{MLE}$ for which we derived the asymptotic distribution. This estimator gives smaller bias, MSE and MR than the two nonparametric estimators $\hat{\lambda}$ and $\hat{\lambda}_L$. MLE is known to have good efficiency, but it lacks in robustness. Thus we proposed a more robust MHDE $\hat{\lambda}_{MHDE}$ under the same model (4.3). We proved that

$\hat{\lambda}_{MHDE}$ is consistent and asymptotically normally distributed. The MHDE performs competitively with MLE in terms of bias, MSE and MR. In order to compare the robustness of the four estimators we proposed, we used α-IF to show that the two estimators $\hat{\lambda}$ and $\hat{\lambda}_{MHDE}$ based on kernel smoothing technique are more robust than the two MLEs $\hat{\lambda}_L$ and $\hat{\lambda}_{MLE}$ in the presence of outlying observations. Since both $\hat{\lambda}_{MLE}$ and $\hat{\lambda}_{MHDE}$ are based on the assumption of the semiparametric structure (4.2), we constructed two Kolmogorov-Smirnov type test statistics, one based on $\hat{\lambda}_{MLE}$ and the other based on $\hat{\lambda}_{MHDE}$, to test the validity of (4.2). Bootstrap technique was used to approximate the distributions of the test statistics. Both test statistics show promising results in our simulation study.

For future work, I may consider to use minimum profile Hellinger distance estimation (MPHDE) for the semiparametric mixture model (4.3). Wu and Karunamuni (2015) first introduced the profile Hellinger distance particularly for semiparametric models and investigated the MPHDE for semiparametric model of general form. Wu and Karunamuni (2015) proved that the MPHDE is as robust as MHDE and achieves full efficiency at the true model. Wu, Yao and Xiang (2017) applied this MPHDE for a two-component semiparametric location-shifted mixture model. Xiang, Yao and Wu (2014) proposed the MPHDE for different semiparametric mixture model where one component is known up to some unknown parameters while the other component is unspecified with unknown location parameter. Another direction to approach the estimation problem of the two-component mixture model could be Bayesian method. Vounatsou, Smith and Smith (1998) applied Gibbs sampling approach to estimate the unknown mixing proportion in a two-component mixture model with discretized data.

Bibliography

[1] Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19-35.

[2] Anderson, J.A. (1979). Multivariate logistic compounds. *Biometrika* **66**, 17-26.

[3] Berslow, N. and Day, N.E. (1980). *Statistical Methods in Cancer Research, Vol. 1: The analysis of Case-Control Studies.* Lyon:IARC.

[4] Benaglia, T., Chauveau, D. and Hunter D.R. (2009). An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *J. Comput. Graph. Statist.* **18**, 505-526.

[5] Beran, R. (1977). Minimum Hellinger distance estimators for parametric models. *Ann. Statist.* **5**, 445-463.

[6] Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications*. Chapman and Hall/CRC, Boca Raton, FL.

[7] Bordes, L., Chauveau, D. and Vandekerkhove, P. (2007). An EM algorithm for a semi-parametric mixture model. *Comput. Statist. Data Anal.* **51**, 5429-5443.

[8] Bordes, L., Delmas, C. and Vandekerkhove, P. (2006a). Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. statist.* **33**, 733-752.

[9] Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006b). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34**, 1204-1232.

[10] Bordes, L. and Vandekerkhove, P. (2010). Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Math. Methods Statist.* **19**, 22-41.

[11] Chauveau, D. and Hoang, V.T.L. (2016). Nonparametric mixture models with conditionally independent multivariate component densities. *Comput. Statist. Data Anal.* **103**, 1-16.

[12] Chen, H., Chen, J. and Kalbfleisch, J.D. (2001). A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models. *J. Roy. Statist. Soc. Ser. B* **63**, 19-29.

[13] Chen, H., Chen, J. and Kalbfleisch, J.D. (2004). Testing for a finite mixture model with two components. *J. Roy. Statist. Soc. Ser. B* **66**, 95-115.

[14] Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Ann. Statist.* **37**, 2523-2542.

[15] Chen, G. and Wu, J. (2013). Molecular classification of acute leukemia". *Some Recent Advances in Mathematics and Statistics* - Proceedings of Statistics 2011 Canada/IMST 2011-FIM XX, 60-74.

[16] Cohen, A.C. (1967). Estimation in mixture of two normal distributions. *Technometrics* **9**, 15-28.

[17] Cruz-Medina, I.R. (2001). *Almost nonparametric and nonparametric estimation in mixture models*. Ph.D. Dissertation, Pennsylvania State University.

[18] Cruz-Medina, I.R. and Hettmansperger, T.P. (2004). Nonparametric estimation in semiparametric univariate mixture models. *J. Statist. Comput. Simul.* **74**, 513-524.

[19] Cruz-Medina, I.R., Hettmansperger, T.P. and Thomas, H. (2004). Semiparametric mixture models and repeated measures: the multinomial cut point model. *J. Roy. Statist. Soc. Ser. C* **53**, 463-474.

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

[21] Deng, X., Wan, S. and Zhang, B. An improved Goodness-of-Fit test for logistic regression models based on case-control data by random partition (2009). *Comm. Statist. Simulation Comput.* **38**, 233-243.

[22] Di, C-Z., Chan, K.C.G., Zheng, C. and Liang, K-Y. (2017). Testing homogeneity in semiparametric mixture case-control models. *Comm. Statist. Theory Methods* **46**, 9092-9100.

[23] Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56**, 363-375.

[24] Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.* **24**, 2431-2461.

[25] Elmore, R.T., Hettmansperger, T.P. and Thomas, H. (2004). Estimating component cumulative distribution functions in finite mixture models. *Comm. Statist. Theory Methods* **33**, 2075-2086.

[26] Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577-588.

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

[28] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Verlag, New York.

[29] Hall, P. (1981). On the nonparametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **43**, 147-156.

[30] Hall, P. (1983). Orthogonal series distribution function estimation, with application. *J. Roy. Statist. Soc. Ser. B* **45**, 81-88.

[31] Hall, P., Neeman, A., Pakyari, R. and Elmore, R. (2005). Nonparametric inference in multivariate mixtures. *Biometrika* **92**, 667-678.

[32] Hall, P. and Titterington, D.M. (1984). Efficient nonparametric estimation of mixing proportions. *J. Roy. Statist. Soc. B* **46**, 465-473.

[33] Hall, P. and Zhou, X. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31**, 201-224.

[34] Hammel T.A.W. (2010). *Semiparametric estimation for finite mixture models using an exponential tilt*. Ph.D. Dissertation, Pennsylvania State University.

[35] Hettmansperger, T.P. and Thomas, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *J. Roy. Statist. Soc. Ser. B* **62**, 811-825.

[36] Hohmann, D. and Holzmann, H. (2013). Two component mixtures with independent coordinates as conditional mixtures: Nonparametric identification and estimation. *Electron. J. Stat.* **7**, 859-880.

[37] Hosmer, D.W., Jr. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics* **29**, 761-770.

[38] Hunter, D.R., Wang, S. and Hettmansperger, T.P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35**, 224-251.

[39] Irion, A., Beck, H.P. and Smith, T. (2002). Assessment of positivity in immuno-assays with variability in background measurements: a new approach applied to the antibody response to Plasmodium falciparum MSP2. *J. Immunological Methods* **259**, 111-118.

[40] Jochmans, K., Henry, M. and Salanie, B. (2016). Inference on two component mixtures under tail restrictions. *Econometric Theory* **33**, 610-635.

[41] Karlis, D. and Xekalaki, E. (1998). Minimum Hellinger distance estimation for Poisson mixtures. *Comput. Statist. Data Anal.* **29**, 81-103.

[42] Karunamuni, R.J. and Wu, J. (2009). Minimum Hellinger distance estimation in a non-parametric mixture model. *J. Statist. Plann. Inference* **139**, 1118-1133.

[43] Karunamuni, R.J. and Wu, J. (2011). One-step minimum Hellinger distance estimation. *Comput. Statist. Data Anal* **55**, 3148-3164.

[44] Kitua, A.Y., Smith, T., Alonso, P.L., Masanja, H., Urassa, H., Menendez, C., Kimario, J. and Tanner, M. (1996). *Plasmodium falciparum* malaria in the first year of life in an area of intense and perennial transmission. *Trop. Med. Int. Hlth* **1**, 475-484.

[45] Leung, D.H.-Y., and Qin, J. (2006). Semi-parametric inference in a bivariate (multivariate) mixture model. *Statist. Sinica* **16**, 153-163.

[46] Li, P., Liu, Y. and Qin, J. (2017). Semiparametric inference in a genetic mixture model. *J. Amer. Statist. Assoc.* **112**, 1250-1260.

[47] Lindsay,B.G. (1983a). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11**, 86-94.

[48] Lindsay,B.G. (1983b). The geometry of mixture likelihoods, part II: the exponential family. *Ann. Statist.* **11**, 783-792.

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

[50] Lindsay, B.G. (1995). *Mixture models: theory, geometry and applications*. In: NSF-CBMS Regional Conference Series in Probability and Statistics, IMS, Hayward, CA.

[51] Lindsay B.G. and Basak, P. (1993). Multivariate normal mixtures: a fast consistent method of moments. *J. Amer. Statist. Assoc.* **88**, 468-476.

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

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

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

[55] Murray, G.D. and Titterington, D.M. (1978). Estimation problems with data from a mixture. *J. Roy. Statist. Soc. Ser. C* **27**, 325-334.

[56] Nagelkerke, N.J.D., Boshuizen1, H.C., de Melker1, H.E., Schellekens, J.F.P., Peeters, M.F. and Spaendonck, M.C. (2003). Estimating the incidence of subclinical infections with Legionella Pneumonia using data augmentation: analysis of an outbreak in The Netherlands. *Stat. Med.* **22**, 3713-3724.

[57] Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411.

[58] Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann. Statist.* **27**, 1368-1384.

[59] Qin, J. and Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609-618.

[60] Qin, J. and Liang, K-Y. (2011). Hypothesis testing in mixture case-control data. *Biometrics* **67**, 182-193.

[61] Quandt, R.E. and Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer Statist. Assoc.* **73**, 730-738.

[62] Rao, B.L.S.P. (1983). *Nonparametric Functional Estimation*. Academic Press, Orlando, FL.

[63] Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26**, 195-239.

[64] Robin, S.A., Bar-Hen, A., Daudin, J-J. and Pierre, L. (2007). A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Comput. Statist. Data Anal.* **51**, 5483-5493.

[65] Roeder, K. (1994). A Graphical Technique for Determining the Number of Components in a Mixture of Normals. *J. Amer. Statist. Assoc.* **89**, 487-500.

[66] Smith, T., Schellenberg, J.A., and Hayes, R. (l994). Attributable fraction estimates and case definitions for malaria in endemic areas. *Stat. Med.* **13**, 2345-2358.

[67] Smith, T.A., Smith, A.G. and Hooper, M.L. (1986). Selection of a mouse embryonal carcinoma clone resistant to the inhibition of metabolic cooperation by retinoic acid. *Exp. Cell Res.* **165**, 417-430.

[68] Smith, T. and Vounatsou, P. (1997). Logistic regression and latent class models for estimating positives in diagnostic assays with poor resolution *Comm. Statist. Theory methods* **27**, 1677-1700.

[69] Song, S., Nicolae, D.L. and Song, J. (2010). Estimating the mixing proportion in a semiparametric mixture model. *Comput. Statist. Data Anal.* **54**, 2276-2283.

[70] Thomas, H. and Hettmansperger T.P. (2001). Modelling change in cognitive understanding with finite mixtures. *Appl. Statist.* **50**, 435-448.

[71] Titterington, D.M. (1983). Minimum distance nonparametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **45**, 37-46.

[72] Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of finite Mixture Distributions*. Wiley, Chichester.

[73] Vounatsou, P., Smith, T. and Smith, A.F.M. (1998). Bayesian analysis of two- component mixture distributions applied to estimating malaria attributable fractions. *J. Roy. Statist. Soc. Ser. C* **47**, 575-587.

[74] Woo, M-J. and Sriram, T.N. (2006). Robust estimation of mixture complexity. *J. Amer. Statist. Assoc.* **101**, 1475-1486.

[75] Woo, M-J. and Sriram, T.N. (2007). Robust estimation of mixture complexity for count data. *Comput. Statist. Data Anal.* **51**, 4379-4392.

[76] Wu, J. and Karunamuni, R.J. (2015). Profile Hellinger distance estimation. *Statistics* **49**, 711-740.

[77] Wu, J., Karunamuni, R.J. and Zhang, B. (2010). Minimum Hellinger distance estimation in a two-sample semiparametric model. *J. Multivariate Anal.* **101**, 1102-1122.

[78] Wu, J., Yao, W. and Xiang, S. (2017). Computation of an efficient and robust estimator in a semiparametric mixture model. *J. Stat. Comput. Simul.* **87**, 2128-2137.

[79] Xiang, S., Yao, W. and Wu, J. (2014). Minimum profile Hellinger distance estimation for a semiparametric mixture model. *Canad. J. Statist.* **42**, 246-267.

[80] Zhang, B. (1999). A chi-squared goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **86**, 531-539.

[81] Zhang, B. (2001). An information matrix test for logistic regression models based on case-control data. *Biometrika* **88**, 921-932.

[82] Zhang, B. (2002). An EM algorithm for a semiparametric finite mixture model. *J. Stat. Comput. Simul.* **72**, 791-802.

[83] Zhang, B. (2006). A score test under a semiparametric finite mixture model. *J. Stat. Comput. Simul.* **76**, 691-703.

[84] Zhang, S. (2005). *Statistical inferences under a semiparametric finite mixture model*. Ph.D. dissertation, University of Toledo.

[85] Zou, F., Fine, J.P. and Yandell, B.S. (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika* **89**, 61-75.