University of Alberta

Minimum Hellinger Distance Estimation in Semiparametric Models

by

Jingjing Wu

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences

Edmonton, Alberta
Spring 2008

Your file Votre référence
ISBN: 978-0-494-45629-3
Our file Notre référence
ISBN: 978-0-494-45629-3

ABSTRACT

We examine the use of Hellinger distance method to obtain robust statistics in a variety of problems in statistical inference. Minimum Hellinger distance (MHD) estimators are proposed and investigated for the two-component mixture model, a two-sample semiparametric model, and semiparametric models of general form. We demonstrate that the proposed MHD estimators have excellent robustness and efficiency properties for semiparametric models.

In Chapter 2, we consider the problem of estimating the mixture proportion in the two-component mixture model. We propose a MHD estimator of the mixture proportion which is strongly consistent, asymptotically normally distributed, and asymptotically efficient at a special case. Furthermore, the proposed MHD estimator is robust, a property that is not generally shared by the classical estimators such as the maximum likelihood estimator (MLE). Using a Monte Carlo study, the proposed estimator is shown to have good robustness properties with respect to a single outlier. A real data set is also analyzed to estimate the proportion of male halibut.

In Chapter 3, we consider a two-sample semiparametric model, which includes the two-sample location-scale model as a special case. We construct a

MHD estimator of regression parameters and examine the asymptotic properties of the proposed estimator. We show good robustness properties of the proposed estimator through a simulation study. A real data set is analyzed to investigate the relationship between age and coronary disease status.

In Chapter 4, we consider the semiparametric models of general form. We construct MHD and minimum profile Hellinger distance (MPHD) estimators of the parametric component. We investigate asymptotic properties of the proposed estimators such as consistency, asymptotic normality, efficiency and adaptivity. We show the robustness and good small sample properties of the proposed estimators using Monte Carlo studies. This chapter demonstrates that both MHD and MPHD estimators in semiparametric models are generally efficient and robust.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my supervisor, Dr. Rohana Karunamuni, for his guidance, encouragement and careful reading of this thesis; and for his constant support during my studies at the University of Alberta.

I also would like to express sincere appreciation to Dr. Anton Schick, Department of Mathematical Sciences, Binghamton University, for his inspiring discussions and guidance on my thesis, in particular on the materials related to Chapter 4.

I also wish to thank Dr. Peter Hooper and Dr. Edit Gombay for their support during my studies at the University of Alberta.

I also express my great gratitude to other members of my committee, Dr. Biao Huang, Dr. Biao Zhang and Dr. Peng Zhang, for their guidance and assistance.

I would certainly like to give my greatest thanks to my mother Zhongxia Zhao and my father Zhonghua Wu for their support throughout my studies and in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

$\|(a_1, \ldots, a_p)\|$	$\|a_1\| + \|a_2\| + \cdots + \|a_p\|$
$\|\|f\|\|$	$[\int f^2(x)dx]^{1/2},\ f \in L_2$
$f^{(k)}(x)$	the $k-$th derivative of $f(x)$
$f \perp g$	$\int f(x)g(x)dx = 0,\ f, g \in L_2$
$< f,\ g >$	$\int f(x)g(x)dx$
int(Θ)	interior of the parameter space Θ
I_A	indicator function of a set A
L_1	$\{f : \int \|f(x)\|dx < \infty\}$
L_2	$\{f : \int f^2(x)dx < \infty\}$
$N(\mu, \sigma)$	normal distribution with mean μ and standard deviation σ
$\mathbb{N}$	the set of natural numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^p$	$\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ (p copies)
$x_n \xrightarrow{P} 0$	random sequence x_n converges to zero in probability as $n \to \infty$

$x_n \xrightarrow{\mathcal{L}} \mathcal{F}$	the distribution of random sequence x_n converges to $\mathcal{F}$ as $n \to \infty$
$x_n = o(r_n)$	$x_n/r_n \to 0$ as $n \to \infty$
$x_n = o_P(r_n)$	$x_n/r_n \xrightarrow{P} 0$ as $n \to \infty$
$x_n = O(r_n)$	$\{x_n/r_n\}$ is a bounded sequence
$x_n = O_P(r_n)$	$\{x_n/r_n\}$ is a sequence bounded in probability
c.d.f.	cumulative distribution function
CLT	Central Limit Theorem
i.i.d.	independent and identically distributed
IF	influence function
l.h.s. (r.h.s.)	left(right) – hand side
MHD	minimum Hellinger distance
MPHD	minimum profile Hellinger distance
MLE	maximum likelihood estimator
MSE	mean squared error
r.v.	random variable
w.p.1	with probability one
w.r.t	with respect to

CHAPTER ONE: INTRODUCTION

1.1 Background of This Research

Statistical inference is based on statistical models for data. During most of the history of the subject, these have been parametric: the mechanism generating the data could be identified by specifying a few real parameters. However, during the last thirty years nonparametric and semiparametric models have flourished. The main reason has of course been the rise of computing power permitting of such models to large data sets showing the inadequacy of parametric models. The deficiency in interpretability of nonparametric models was filled by the development of semiparametric models. The main focus of research in this area has been the construction of such models and corresponding statistical procedures in response to particular types of data arising in various disciplines, primarily in biostatistics and econometrics. The well-known semiparametric models include the Cox proportional hazard model in survival analysis, econometric index models, regression models and errors-in-variables models, among many others. In this thesis, I mainly focus on semiparametric models.

Many authors have considered efficient and adaptive estimation in semiparametric models for the past twenty years; see, for example, Bickel (1982), Schick (1986) and Forrester et al. (2003) for most references. However, the robustness in semiparametric models has been paid little attention. The efficiency when the model has been appropriately chosen and the robustness when it has not are two fundamental ideas in parametric estimation. It was long thought that there was an inherent contradiction between the aims of achieving robustness and efficiency; i.e., a robust estimator could not be efficient and vice versa. Some of the practical deficiencies of maximum likelihood estimators (MLEs) are the lack of resistance to outliers and the general non-robustness with respect to model misspecification. The need for robust statistics in statistical inference has been widely recognized now. Many different approaches for finding robust statistics for parametric models have been proposed, see Huber (1980) and Maronna et al. (2007) for summaries of most important methods. Such methods have had varying degree of success in dealing with "bad" data, but they may suffer from a loss of efficiency if the postulated model distribution is the true one. This is, however, not the case with minimum Hellinger distance (MHD) estimators. Lindsay (1994) has shown that MLE and MHD estimators are members of a

larger class of efficient estimators with various second-order efficiency properties. MHD estimators have been shown to have excellent robustness properties in parametric models such as the resistance to outliers and robustness with respect to model misspecification, see Beran (1977) and Donoho and Liu (1988). [In fact, Donoho and Liu (1988) have shown a much stronger result that all minimum distance estimators are automatically robust with respect to the stability of the quantity being estimated.] Efficiency combined with excellent robustness properties make MHD estimators appealing in practice. Furthermore, Hellinger distance has the special attraction that it is dimensionless. For a comparison between MHD estimators with the MLEs and the balance between robustness and efficiency of estimators see the articles of Lindsay (1994) and Karlis and Xekalaki (1998, 2001). The literature on MHD estimation has been dominated by MHD estimation in fully parametric models. There appears to be very little research has been done on application of the MHD methodology to semiparametric models. In this thesis, I extend the use of MHD approach to the semiparametric models to obtain robust efficient estimators.

1.2 MHD Estimation

Consider the situation where we observe a sequence of independent and identically distributed (i.i.d.) random variables (r.v.) $X_1, X_2, \ldots, X_n$ from a distribution with density function f. If f belongs to a specified parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ then θ may be estimated using well-known likelihood procedures. However, assuming f belongs strictly to the family $\mathcal{F}$ ignores the possibility of departures from the parametric model. In practice, data contamination, lack of information, and other factors beyond our control can make the parametric model incorrect for the data at hand. Instead, we assume that f is either in $\mathcal{F}$ or close to a member of $\mathcal{F}$, and use a minimum distance estimation procedure. We use the minimum Hellinger distance approach as our estimation procedure, in which the estimate is chosen to minimize the Hellinger distance between the parametric model and a nonprametric density estimator of f. In other words, the MHD estimator of θ is defined as the value of the parameter that minimizes the Hellinger distance between a density estimator and the parametric density. If we use $\widehat{\theta}$ to denote the MHD estimator, then $\widehat{\theta}$ is defined by

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \| f_\theta^{1/2} - f_n^{1/2} \|,$$

where $\|\cdot\|$ denotes the L_2-norm and f_n is a nonparametric density estimator of f based on the observations $X_1, X_2, \ldots, X_n$.

It is interesting to note that this estimator $\widehat{\theta}$ is related heuristically to the MLE of θ. For n sufficiently large, the MLE should be close to the true parameter value θ and the density estimator f_n should be close to f_θ. Finding the MLE amounts to maximizing $\int \log f_t(x) dF_n(x)$ over $t \in \Theta$, where F_n is the

empirical distribution function of the data. Arguing formally, we expect that this procedure is nearly the same as maximizing over t near θ the quantity

$$\begin{aligned}\int f_n(x)\log\Big[\frac{f_t(x)}{f_n(x)}\Big]\,dx &= 2\int f_n(x)\log\Big[1+\Big(\frac{f_t^{1/2}(x)}{f_n^{1/2}(x)}-1\Big)\Big]\,dx \\ &\approx 2\int f_n(x)\Big[\Big(\frac{f_t^{1/2}(x)}{f_n^{1/2}(x)}-1\Big)-\frac{1}{2}\Big(\frac{f_t^{1/2}(x)}{f_n^{1/2}(x)}-1\Big)^2\Big]\,dx \\ &= -2\|f_t^{1/2}-f_n^{1/2}\|^2.\end{aligned}$$

Thus, it is not unreasonable to expect that the MHD estimator $\widehat{\theta}$ is asymptotically efficient under f_θ. On the other hand, simple calculation shows that

$$\|f_t^{1/2}-f_n^{1/2}\|^2 \;\le\; \int |f_t(x)-f_n(x)|dx \;\le\; 2\|f_t^{1/2}-f_n^{1/2}\|,$$

so the topology induced on the space of probability measures by the Hellinger metric is the same as that induced by the L_1-norm. It is known that the L_1-norm induces a robust topology. Thus, the MHD estimator could be expected to be robust as well. In fact, various asymptotic and robustness properties of $\widehat{\theta}$ have been studied under some regularity conditions in Beran (1977), Stather (1981) and Tamura and Boos (1986), among others.

Now assume that f belongs to a class of general semiparametric models of the form

$$\{f_{\theta,\eta}:\theta\in\Theta\subseteq\mathbb{R}^p,\eta\in\mathcal{H}\}, \tag{1.1}$$

where Θ is a compact subset of $\mathbb{R}^p$ and $\mathcal{H}$ is an arbitrary set, typically of infinite dimension. The problem is to estimate the parameter θ assuming that η as a *nuisance parameter.* If $\mathcal{H}$ is finite dimensional, then (1.1) is a fully parametric model. If we still use $\widehat{\theta}$ to denote the MHD estimator, then

$$\widehat{\theta}=\underset{\theta\in\Theta}{\operatorname{argmin}}\underset{\eta\in\mathcal{H}}{\inf}\|f_{\theta,\eta}^{1/2}-f_n^{1/2}\|. \tag{1.2}$$

For semiparametric models, i.e., $\mathcal{H}$ is infinite dimensional, we can define a MHD estimator θ_n of θ in a natural way as

$$\theta_n=\arg\min_{t\in\Theta}\|f_{t,\eta_n}^{1/2}-f_n^{1/2}\|,$$

where η_n is a suitable estimator of η.

The literature on MHD estimation has been dominated by MHD estimation in fully parametric models. Beran (1977) has shown that the MHD estimator $\widehat{\theta}$ defined in (1.2) has excellent robustness and efficiency. Tamura and Boos (1986) extended the work of Beran (1977) to a multivariate setting, while the

corresponding MHD estimation for count data can be found in Simpson (1987). Yang (1991) and Ying (1992) investigated MHD estimation for censored data. Sriram and Vidyashankar (2000) and Woo and Sriram (2006, 2007) have studied MHD estimates for branching processes and the mixture complexity in a finite mixture model, respectively. However, there seems to be relatively very few attempts to apply the MHD approach to semiparametric models. The Hellinger deviance test was introduced in Karlis and Xekalaki (1998) for testing a semiparametric Poisson mixture. The only notable work reported in the literature on MHD estimation in semiparametric models appears to be that of the work by Lu et al. (2003). The preceding authors have investigated a MHD estimator for finite mixtures of Poisson regression models with the distribution of the covariate variable unknown.

1.3 Summary of Results

In Chapter 2, we consider the problem of estimating the mixture proportion in the two-component mixture model $\theta F+(1-\theta)G$, where F and G are two different distribution functions. Specifically, suppose we observe three independent samples

$$\begin{aligned} X_1, \ldots, X_{n_0} &\overset{\text{i.i.d.}}{\sim} F \\ Y_1, \ldots, Y_{n_1} &\overset{\text{i.i.d.}}{\sim} G \\ Z_1, \ldots, Z_{n_2} &\overset{\text{i.i.d.}}{\sim} \theta F + (1-\theta)G \end{aligned}$$

with density functions f, g and $h_\theta = \theta f + (1-\theta)g$, respectively. Here θ is called the mixture proportion, where $\theta \in [0,1]$. The problem is to estimate the mixture parameter θ, treating f and g as nuisance parameters. We propose to estimate θ using the MHD approach. Let $n = n_0 + n_1 + n_2$. We define a MHD estimator θ_n of θ as follows:

$$\theta_n = \arg\min_{t\in[0,1]} \|(t\widehat{f} + (1-t)\widehat{g})^{1/2} - \widehat{h}^{1/2}\|,$$

where $\widehat{f}$, $\widehat{g}$ and $\widehat{h}$ are kernel-type density estimators of f, g and h_θ, respectively, based on the samples X_i's, Y_i's and Z_i's, respectively. In other words, we minimize the Hellinger distance between a totally nonparametric kernel density estimator and a parameterized convolution of estimated component densities.

In Theorem 2.1 we show the existence and the continuity of θ_n as a functional. Theorem 2.2 shows that θ_n is consistent, while the asymptotic distribution of θ_n is established in Theorem 2.4 which is a consequence of Theorem 2.3. To see the performance of θ_n, in Section 2.3 we obtain a MLE of θ with the asymptotic distribution established in Theorem 2.5. The proposed MHD estimator is compared with the MLE and asymptotic efficiency properties of θ_n are examined in Section 2.4. This is done by constructing a Cramér-Rao type lower

bound in Theorem 2.6 for nonparametric estimators of the mixture proportion. The full efficiency is achieved by the MHD estimator θ_n at a special case as shown in Corollary 2.1, which is a simple consequence of Theorems 2.4 and 2.6. The robustness properties of our proposed MHD estimator θ_n are studied using a Monte Carlo study in Section 2.5. Theoretical results on the robustness of MHD estimator seem difficult in the present context. We study four different mixtures of normal distributions in the simulation. The α-influence functions (IFs) demonstrate that the MHD estimator is very robust in the presence of outliers, a property that is not generally possessed by the classical estimators such as the MLEs. When compared with two MLEs constructed in Section 2.5, our proposed MHD estimator θ_n shows good efficiency properties. In Section 2.6, a real data set is analyzed to estimate the proportion of male halibut.

In Chapter 3, we consider a two-sample semiparametric model, where the log ratio of the two underlying density functions is of a regression model, i.e., $h_\theta(x) = g(x)\exp[\alpha + r(x)\beta]$ with $\theta = (\alpha, \beta)$. This setup includes the two-sample location-scale model as a special case. This model is also closely related to the logistic regression model. We construct a MHD estimator of regression parameters in a quite nature way and examine the asymptotic properties of the proposed estimator. The existence and continuity of the proposed MHD estimator are shown in Theorem 3.1. Theorem 3.2 shows that the proposed MHD estimator is consistent for both finite and infinite support cases of g. Due to the fact that techniques developed in Chapter 2 could be used to derive the asymptotic distribution of the proposed MHD estimator for the finite support case, we concentrate on developing the asymptotic distribution of the proposed MHD estimator for the infinite support case of g. Theorem 3.4 establishes the asymptotic normality of the proposed MHD estimator, which is a consequence of Theorem 3.3. Similar techniques as in Stather (1981) are used to prove the theorems. However, we extend his results developed for parametric models to semiparametric models. For the case that g has infinite support, we need to prove several technical results to control the effect of the tails. These require some conditions on the underlying densities g and h_θ, which are satisfied by a variety of families, such as the location-scale families as shown in Sections 3.2 and 3.3. To see the performance of the proposed MHD estimator, in Section 3.4 we compare the proposed MHD estimator with the semiparametric likelihood estimator developed in Zhang (2000), assuming that g and h_θ are normal distributions $N(0,1)$ and $N(\mu,1)$, respectively. We observe that the proposed MHD estimator has comparative asymptotic variance when compared with the semiparametric likelihood estimator, especially when μ is close to zero; see Remark 3.9. To see the small sample properties, a Monte Carlo simulation is conducted. While the estimated bias and MSE of the proposed MHD estimator of α are higher than those of the semiparametric likelihood estimator of α, our proposed MHD estimator of β performs uniformly better than the semiparametric likelihood estimator of β in the sense of having smaller estimated bias and MSE.

Note that β plays a more important role than α in most applications. To investigate the robustness, the α-IFs are calculated for a single outlying observation. The α-IFs of the proposed MHD estimators are bounded while those of the semiparametric likelihood estimator seem to increases dramatically in absolute value when the outlying observation moves to the left from -1. This shows that our proposed MHD estimator has good robustness properties. A real data set is also analyzed in Section 3.5 to investigate the relationship between age and coronary disease status.

In Chapter 4, we consider the semiparametric models of general form: $\{f_{\theta,\eta} : \theta \in \Theta \subseteq \mathbb{R}^p, \eta \in \mathcal{H}\}$, where Θ is a compact subset of $\mathbb{R}^p$ and $\mathcal{H}$ is an arbitrary set of infinite dimension. The problem is to estimate the parameter θ assuming that η as a nuisance parameter. Theorem 4.1 generalizes similar result of Beran (1977) on the efficiency of MHD estimator of a fully parametric model. For semiparametric models, a MHD estimator is constructed using a plug-in rule. This estimator is shown to be adaptive under certain assumptions, see Theorem 4.2. An efficient (in the semiparametric sense) MHD estimator is also investigated in Section 4.3. This estimator was studied by Huang (1982), who has left the consistency of the estimator an open problem. The consistency is established in Theorem 4.3, solving the preceding problem. We construct a minimum profile Hellinger distance (MPHD) estimator in Section 4.4 and it is shown to be efficient under certain conditions, see Theorems 4.5 and 4.6 and Remark 4.10. It is also shown in Section 4.5 that the proposed MHD estimator of Theorem 4.2 is still asymptotically normally distributed even though the underlying density function is not strictly from the semiparametric model described above. In some sense, this shows the robustness of the MHD estimator proposed in Section 4.2. A special form of contamination is considered and it also shows that the MHD estimator proposed in Section 4.2 is robust. A Monte Carlo study is designed to demonstrate the efficiency and robustness of the MHD estimator proposed in Section 4.2. In the simulation study, we consider the mixture of two normal distributions and the mixture proportion is considered as the parameter of interest. For comparison purposes, two MLEs of the mixture proportion are also constructed. When compared with the two MLEs, the proposed MHD estimator is observed to be more robust. In fact, the α-IF of the MHD estimator, with respect to a single outlying observation, is almost a constant valued around zero, while those of the two MLEs have big jumps when the outlying observation is further away from zero. This means that the MHD estimator is not much affected by a single outlying observation, while the MLEs are affected by the outlying observation. We also show that the breakdown point for the MHD estimator is about 0.5 (the best possible value), while that for one of the MLEs is around 0.25. In other words, MHD estimator shows more robust behavior than the MLEs analyzed. Furthermore, the MHD estimator has competitive efficiency when compared with the MLEs in the sense of having smaller estimated bias and MSE, under the true model

(without contamination). Under the contaminated model, the MHD estimator performs generally better than the MLEs. When the contamination rate is high, the MHD estimator has much smaller estimated bias and MSE than the MLEs. As an example, a symmetric location model is investigated in Section 4.7 and adaptive MHD and MPHD estimators are constructed for this model.

In summary, we show that the MHD approach in parametric model can be extended successfully to semiparametric models, either for particular models or for a general model. The proposed MHD estimators in semiparametric models have been shown to have good efficiency and robustness properties. The success of this approach in the problems considered of this thesis could encourage its further development in many other problems. We consider the following problems, among others, to be worthy candidates for future study: theoretical development of the robustness, application to semiparametric regression models, robust hypothesis testing, and classification.

CHAPTER TWO: MHD ESTIMATION IN THE TWO-COMPONENT MIXTURE MODEL

2.1 Introduction

Let F and G be two probability distributions and θ be a positive real number between 0 and 1. Then $\theta F + (1-\theta)G$ defines a two-component mixture distribution with mixture weights θ and $(1-\theta)$. When component distributions F and G are known to have some specific forms, then $\theta F + (1-\theta)G$ is called a *parametric mixture*. On the other hand, if F and G are completely unspecified but are different distributions then $\theta F + (1-\theta)G$ is known as a *nonparametric mixture*. A great deal of work has been done in parametric mixture models; see, e.g., Titterington et al. (1985), Lindsay (1995), Chen (1995, 1998), McLachlan and Peel (2000), and Scott (2001), among others for examples, applications and theory. The estimation problem of the mixture parameter θ in a nonparametric mixture model, however, is faced with the lack of identifiability of θ. One way of overcoming this difficulty is to take training samples from each component distribution as in Hall (1981). More specifically, suppose we observe three independent samples

$$\begin{aligned} X_1, \ldots, X_{n_0} &\overset{iid}{\sim} F \\ Y_1, \ldots, Y_{n_1} &\overset{iid}{\sim} G \\ Z_1, \ldots, Z_{n_2} &\overset{iid}{\sim} \theta F + (1-\theta)G, \end{aligned} \tag{2.1}$$

then the problem is to estimate the mixture parameter θ, treating F and G as nuisance parameters. For model (2.1), Hall (1981, 1983) described minimum distance estimators based on empirical distribution functions, Titterington (1983) considered minimum distance estimators based on density estimators, and Hall and Titterington (1984) constructed a sequence of multinomial approximations and related MLE estimators of θ by grouping data for a similar model to (2.1). Qin (1999) developed a confidence interval for θ using an empirical likelihood ratio based statistic assuming the log-likelihood ratio of densities of F and G is linear in observations. Hosmer (1973) used the model (2.1) to estimate the proportions of male and female fish in a population of halibut from some univariate data provided by International Halibut Commission in Seattle, Washington. More applications can be found in the papers of the specific issue of *Communi-*

cations in Statistics on Remote Sensing (1976).

Robust methods such as M-estimation are not easily adapted for nonparametric mixtures (Cutler and Cordero-Braña, 1996). Minimum distance estimation is an alternative approach that produces robust estimators. The model (2.1) has not been fully investigated using the preceding approach. In this chapter, we propose to estimate the mixture parameter θ using the MHD approach. The Hellinger distance has the special attraction that it is dimensionless. Furthermore, MHD estimators have been shown to have excellent robustness properties such as resistance to outliers and robustness with respect to model misspecification (Beran, 1977 and Donoho and Liu, 1988). Many robust estimators achieve robustness at some cost in first-order efficiency. This is, however, not the case with MHD estimators. Lindsay (1994) has shown that MLE and MHD estimators are members of a larger class of efficient estimators with various robustness and second-order efficiency properties.

The setup of Beran (1977) assumes that the observed random variables are i.i.d. with some unknown density g which is close in the Hellinger metric to a member of some specified parametric class $\{f_\theta : \theta \in [0,1]\}$. The model at (2.1) is not parametric, however. Thus, the results in this chapter exhibit an extension of Beran's (1977) MHD technique to a semiparametric model. Furthermore, the combined data set of (2.1), $X_1, \ldots, X_{n_0}, Y_1, \ldots, Y_{n_1}, Z_1, \ldots, Z_{n_2}$, is a collection of independent observations, but not necessarily identically distributed. This feature also adds a degree of complexity to the development of asymptotic theoretical results of the proposed MHD estimator of θ.

There have been very few attempts to estimate the parameters in a mixture problem with the MHD method or similar minimum distance approaches. The only work on MHD estimation for mixtures appears to be that of Woodward et al. (1995), Cordero-Braña (1994), Cutler and Cordero-Braña (1996) and Lu et al. (2003). However, their results are for the case that F and G are fully parametric models. More specifically, Woodward et al. (1995) have concentrated on estimating the mixture proportions $(\pi_1, \ldots, \pi_{k-1})$ in a fully parametric model of the form $\sum_{i=1}^{k} \pi_i f(x|\phi_i)$, whereas Cordero-Braña (1994) and Cutler and Cordero-Braña (1996) have assumed that all the mixture parameters $(\pi_1, \ldots, \pi_{k-1}, \phi_1, \ldots, \phi_k)$ are of interest, extending the work of Woodward et al. (1995), where $f(\cdot|\phi_1), \ldots, f(\cdot|\phi_k)$ are density functions on the real line and $\phi_i \in \Phi \subseteq \mathbb{R}^s$, $i = 1, \ldots, k$. Lu et al. (2003) have examined MHD estimation for finite mixtures of Poisson regression models. The present work thus shows a further extension of above papers to the case where the distributions F and G in model (2.1) are completely unknown.

MHD estimation has been applied in many other settings. For example, Tamura and Boos (1986) extended the work of Beran (1977) to a multivariate setting, while the corresponding MHD estimation for count data can be found in Simpson (1987). Yang (1991) and Ying (1992) investigated MHD estimation

for censored data. Sriram and Vidyashankar (2000) and Woo and Sriram (2006, 2007) have studied MHD estimators for branching processes and the mixture complexity in a finite mixture model, respectively.

In Section 2.2, our proposed MHD estimator of θ is given. Our approach is very natural. We minimize the Hellinger distance between a totally nonparametric adaptive kernel density estimator and a parameterized convolution of estimated component densities. We study asymptotic theoretical properties such as strong consistency and asymptotic normality of the proposed estimator. In Section 2.3, we obtain a MLE of θ using the approach of Hall and Titterington (1984). Asymptotic efficiency properties of the proposed MHD estimator are examined in Section 2.4. This is done by constructing a Cramér-Rao type lower bound for nonparametric estimators of the mixture proportion. In Section 2.5, robustness properties of the proposed MHD estimator are studied using a Monte Carlo study. It is observed that the MHD estimator is very robust in the presence of outliers. Examples and concluding remarks are given in Sections 2.6 and 2.7, respectively. All the proofs are deferred to Section 2.8.

2.2 MHD Estimator of Mixture Proportion

In this section, we assume the setup of model (2.1). In order to employ the MHD technique of Beran (1977), we first define a parametric family of densities

$$h_\theta(x) = \theta f(x) + (1-\theta)g(x) \tag{2.2}$$

where f and g denote two different densities of F and G, respectively; i.e., we suppose that $\int |f(x) - g(x)|dx > 0$. Next we define following adaptive kernel density estimators (see, e.g., Scott, 1992) of f and g, respectively, based on data $X_1, \ldots, X_{n_0}$ and $Y_1, \ldots, Y_{n_1}$ of (2.1):

$$\widehat{f}(x) = \frac{1}{n_0 S_{n_0} b_{n_0}} \sum_{i=1}^{n_0} K_0\Big(\frac{x - X_i}{S_{n_0} b_{n_0}}\Big), \tag{2.3}$$

$$\widehat{g}(x) = \frac{1}{n_1 S_{n_1} b_{n_1}} \sum_{j=1}^{n_1} K_1\Big(\frac{x - Y_j}{S_{n_1} b_{n_1}}\Big), \tag{2.4}$$

where K_0 and K_1 are two smooth density functions, bandwidths b_{n_0} and b_{n_1} are positive constants such that $b_{n_i} \to 0$ as $n_i \to \infty$, $i = 0, 1$, and $S_{n_0} = S_{n_0}(X_1, \ldots, X_{n_0})$ and $S_{n_1} = S_{n_1}(Y_1, \ldots, Y_{n_1})$ are robust scale statistics (these statistics generally estimate the scale parameters of respective distributions). In a realistic situation, the bandwidths usually take the form $b_{n_i} = n_i^{-r}$ with $0 < r < 1$ for $i = 0, 1$. Estimators (2.3) and (2.4) are similar to the ones used in Beran (1977) for density estimation. For any $t \in [0, 1]$ define

$$\widetilde{h}_t(x) = t\widehat{f}(x) + (1-t)\widehat{g}(x). \tag{2.5}$$

Note that $\widetilde{h}_\theta$ is a parametric density function with the only unknown parameter being θ. Furthermore, θ is identifiable from (2.5) since $\theta_1 \neq \theta_2$ implies $\widetilde{h}_{\theta_1} \neq \widetilde{h}_{\theta_2}$ (Titterington et al. (1985, Section 3.1)). Next we define a kernel density estimator based on the Z_i's as follows:

$$\widehat{h}(x) = \frac{1}{n_2 S_{n_2} b_{n_2}} \sum_{i=1}^{n_2} K_2\left(\frac{x - Z_i}{S_{n_2} b_{n_2}}\right), \tag{2.6}$$

where again K_2 is a smooth density function, bandwidth b_{n_2} is a positive constant such that $b_{n_2} \to 0$ as $n_2 \to \infty$, and $S_{n_2} = S_{n_2}(Z_1, \ldots, Z_{n_2})$ is a robust scale statistic.

Let $\mathcal{H}$ be the set of all densities w.r.t. Lebesgue measure on the real line. Following Beran (1977), we first define a MHD functional $T_0 : \mathcal{H} \longrightarrow [0,1]$ such that

$$T_0(\phi) = \arg\min_{t\in[0,1]} \| h_t^{1/2} - \phi^{1/2} \|, \tag{2.7}$$

where $\| \cdot \|$ denotes the L_2-norm. When h_t is known, the MHD estimator of $T_0(\phi)$ is defined as $T_0(\widehat{\phi})$, where $\widehat{\phi}$ is a nonparametric density estimator of ϕ. Since h_t is unknown in our model (2.1), we propose to replace h_t with $\widetilde{h}_t$, the parameterized convolution of estimated component densities defined by (2.5). Then a MHD estimator of $T_0(\phi)$ is defined as functional $\widehat{T}(\phi)$ at $\widehat{\phi}$, where

$$\widehat{T}(\phi) = \arg\min_{t\in[0,1]} \| \widetilde{h}_t^{1/2} - \phi^{1/2} \| . \tag{2.8}$$

Since the parameter space $[0,1]$ is compact, $\widehat{T}(\widehat{\phi})$ is attained. However, $\widehat{T}(\widehat{\phi})$ may be multiple valued and so we shall use the notation $\widehat{T}(\widehat{\phi})$ to indicate any one of the possible values chosen arbitrarily (cf., Beran, 1977). In our situation, $\phi = h_\theta$ and $\widehat{\phi} = \widehat{h}$. Therefore, our proposed MHD estimator of θ is defined as

$$\theta_n = \widehat{T}(\widehat{h}), \tag{2.9}$$

where $\widehat{h}$ is given by (2.6) and where $n = n_0 + n_1 + n_2$ is the total sample size. That is, θ_n is the minimizer of the Hellinger distance between $\theta\widehat{f} + (1-\theta)\widehat{g}$ and $\widehat{h}$ with $\widehat{f}$ and $\widehat{g}$ defined by (2.3) and (2.4), respectively. We are interested in both the asymptotic and local properties of θ_n. So we let $n \to \infty$ and at the same time suppose that $n_i/n \to \rho_i$ for some positive constants ρ_i as $n \to \infty$, $i = 0, 1, 2$.

We now discuss asymptotic properties of the proposed MHD estimator. First, we give some results on the existence, consistency and asymptotic unique-

ness of the MHD estimator of θ. The next theorem, which gives conditions for the existence of θ_n and the continuity of the functionals is analogous to Theorem 1 of Beran (1977).

Theorem 2.1. *Suppose that T_0 and $\widehat{T}$ are defined by (2.7) and (2.8), respectively. Then,*

(i) For every $\phi \in \mathcal{H}$, there exists $\widehat{T}(\phi) \in [0,1]$ satisfying (2.8).

(ii) If $T_0(\phi)$ is unique, then $\widehat{T}(\phi_n) \to T_0(\phi)$ for any sequences $\{\phi_n\}_{n\in\mathbb{N}}$ and $\{\widetilde{h}_t\}_{n\in\mathbb{N}}$ such that $\| \phi_n^{1/2} - \phi^{1/2} \| \to 0$ and $\sup_{t\in[0,1]} \| \widetilde{h}_t^{1/2} - h_t^{1/2} \| \to 0$ as $n \to \infty$.

(iii) $T_0(h_t) = t$ uniquely for any $t \in [0,1]$, where $h_t(x) = tf(x) + (1-t)g(x)$.

Remark 2.1. Beran (1977) gave conditions for the Hellinger consistency of the density estimator. Devroye and Wagner (1979) have proved the L_1 convergence of such estimators under weaker conditions. In view of the equivalence of the Hellinger and L_1 topologies (see, Devroye and Györfi, 1985), the Hellinger consistency of ϕ_n is equivalent to $\int |\phi_n - \phi| dx \xrightarrow{P} 0$ as $n \to \infty$.

With further assumptions on the bandwidths and kernels in (2.3), (2.4) and (2.6), consistency of the MHD estimator θ_n of θ follows from the continuity of the functionals in the Hellinger topology. This result is given next. We first list the assumptions made in the theorems of this section:

C1. The kernels K_0, K_1 and K_2 in (2.3), (2.4) and (2.6), respectively, are absolutely continuous on their compact support, and the first derivatives $K_0^{(1)}$, $K_1^{(1)}$ and $K_2^{(1)}$ are bounded.

C2. f and g are uniformly continuous on their support.

C3. The positive constants b_{n_0}, b_{n_1}, b_{n_2} in (2.3), (2.4) and (2.6), respectively, satisfy $b_{n_i} \to 0$ and $n_i^{1/2} b_{n_i} \to \infty$ as $n_i \to \infty$, $i = 0, 1, 2$.

C4. $S_{n_i} \xrightarrow{P} S_i$, as $n_i \to \infty$, $i = 0, 1, 2$.

C5. The sequences of densities $\{\widehat{h}\}_{n\in\mathbb{N}}$ and $\{\widetilde{h}_t\}_{n\in\mathbb{N}}$ converge to h_θ and h_t, respectively, in the sense that $\| \widehat{h}^{1/2} - h_\theta^{1/2} \| \to 0$ and $\sup_{t\in[0,1]} \| \widetilde{h}_t^{1/2} - h_t^{1/2} \| \to 0$ as $n \to \infty$, where $\theta \in (0,1)$ and $\widetilde{h}_t = t\widehat{f} + (1-t)\widehat{g}$ with $\widehat{f}$ and $\widehat{g}$ converging to f and g uniformly.

C6. f and g have the same compact support, say W, on which $h_t(x) > 0$ for any $t \in [0,1]$; and f, g, $\widehat{f}$, $\widehat{g}$ and $\widehat{h}$ are piecewise continuous.

C7. K_0, K_1 and K_2 are symmetric about zero and have compact support, and the second derivatives $K_0^{(2)}$, $K_1^{(2)}$ and $K_2^{(2)}$ exist and are bounded.

C8. $\dot{S}_\theta = \frac{\partial}{\partial\theta} h_\theta^{1/2}$ has compact support W on which it is continuous, where h_θ is given by (2.2).

C9. $f, g > 0$ on W and the second derivatives $f^{(2)}$ and $g^{(2)}$ exist and are bounded.

C10. $b_{n_i} \to 0$, $n_i^{1/2} b_{n_i} \to \infty$ and $n_i^{1/2} b_{n_i}^2 \to 0$ as $n_i \to \infty$, $i = 0, 1, 2$.

C11. There exist positive finite constants S_0, S_1 and S_2 depending on f and g such that $n_i^{1/2}(S_{n_i} - S_i) = O_P(1)$ as $n \to \infty$, $i = 0, 1, 2$.

Theorem 2.2. *Suppose that $n_i/n \to \rho_i$ for some positive constants ρ_i as $n \to \infty$, $i = 0, 1, 2$. Further suppose that assumptions C1 to C4 hold with $\widetilde{h}_t$, $\widehat{h}$ and θ_n given by (2.5), (2.6) and (2.9) respectively. Then C5 holds and it follows that $\theta_n \xrightarrow{P} \theta$ as $n \to \infty$.*

Remark 2.2. If S_{n_i} satisfies a stronger condition that $S_{n_i} \to S_i$ as $n \to \infty$, w.p.1, and $\sum_{n_i=1}^{\infty} \exp\big(-\gamma n_i b_{n_i}^2\big) < \infty$ for any $\gamma > 0$ and $i = 0, 1, 2$, then the convergence in probability result given in Theorem 2.2 above can be changed to almost surely. For example, if one takes $b_{n_i} = [\frac{\log n_i}{n_i^{\epsilon}}]^{1/2}$ for some $0 < \epsilon < 1$, then $\sum_{n_i=1}^{\infty} \exp\big(-\gamma n_i b_{n_i}^2\big) < \infty$ for any $\gamma > 0$ and C3 is also satisfied.

We now state results on the asymptotic distribution of the proposed MHD estimator θ_n. The next theorem gives an expression for the difference $\theta_n - \theta$, which is fundamental for further developments of theory.

Theorem 2.3. *Suppose that densities $\widetilde{h}_t$ defined in (2.5) and $\widehat{h}$ in (2.6) satisfy assumptions C5 and C6. Define functional $T(\{h_t\}_{t\in[0,1]}, \phi) = \arg\min_{t\in[0,1]} \| h_t^{1/2} - \phi^{1/2} \|$ and suppose that the functional T is continuous at $(\{h_t\}_{t\in[0,1]}, h_\theta)$ in the sense of Theorem 2.1 (ii). Then, it follows that*

$$
\begin{aligned}
\theta_n - \theta &= T\big(\{\widetilde{h}_t\}_{t\in[0,1]}, \widehat{h}\big) - T\big(\{h_t\}_{t\in[0,1]}, h_\theta\big) \\
&= \Big\{\Big[\int \frac{(f-g)^2}{2(\theta f + (1-\theta)g)^{3/2}} h_\theta^{1/2} dx\Big]^{-1} + \gamma_n\Big\} \times \\
&\quad \Big\{\int \frac{f-g}{(\theta f + (1-\theta)g)^{1/2}} (\widehat{h}^{1/2} - h_\theta^{1/2}) dx \\
&\qquad + \int \frac{\frac{\theta}{2}(f-g) + g}{(\theta f + (1-\theta)g)^{3/2}} (\widehat{f} - f)\widehat{h}^{1/2} dx \\
&\qquad - \int \frac{\frac{1}{2}(1+\theta)(f-g) + g}{(\theta f + (1-\theta)g)^{3/2}} (\widehat{g} - g)\widehat{h}^{1/2} dx \\
&\qquad + \alpha_n \int (\widehat{f} - f)^2 dx + \beta_n \int (\widehat{g} - g)^2 dx\Big\},
\end{aligned}
\tag{2.10}
$$

where $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\gamma_n\}$ are bounded sequences of real numbers and $\gamma_n \to 0$ as $n \to \infty$.

Under further conditions on the parametric family h_θ at (2.2) and the kernels, the next theorem shows that $\theta_n = \widehat{T}(\widehat{h})$ is asymptotically normally distributed about $\theta = T_0(h_\theta)$.

Theorem 2.4. *Suppose that $n_i/n \to \rho_i$ for some positive constants ρ_i as $n \to \infty$, $i = 0, 1, 2$. Suppose that $\widetilde{h}_t$, $\widehat{h}$ and θ_n given by (2.5), (2.6) and (2.9), respectively, satisfy assumptions C7 to C11. Then the asymptotic distribution of $n^{1/2}(\theta_n - \theta)$ is $N(0, \sigma^2)$, where σ^2 is defined by*

$$\Big\{Var\big[\frac{\partial \log h_\theta(Z_1)}{\partial \theta}\big]\Big\}^{-2}\Big\{\frac{\theta^2}{\rho_0}Var\big[\frac{\partial \log h_\theta(X_1)}{\partial \theta}\big] + \frac{(1-\theta)^2}{\rho_1}Var\big[\frac{\partial \log h_\theta(Y_1)}{\partial \theta}\big] + \frac{1}{\rho_2}Var\big[\frac{\partial \log h_\theta(Z_1)}{\partial \theta}\big]\Big\}.$$

Remark 2.3. The regularity conditions assumed in Theorems 2.1 to 2.4 above are typical in MHD estimation context; see, e.g., Beran (1977) and Cordero-Braña (1994). From Theorem 2.2 we observe that the proposed MHD estimator of θ is consistent without the compact support requirement on the densities f and g. However, this assumption is critical for the asymptotic normality results established in Theorem 2.4. Therefore, in order to prove the asymptotic normality for the infinite support case of f and g, we must employ a different technique. Note that the asymptotic normality of an estimator is related to the differentiability of the functional T_0 defined at (2.7). One way to achieve such a goal is to concentrate on the Hadamard (or compact) differentiability of the functional T_0; see Fernholz (1983). It is known that Hadamard differentiability will yield the asymptotic normality. Hadamard differentiability is weaker than Fréchet differentiability that we rarely have for functionals. Fernholz (1983) has built up the Hadamard differentiability of three important estimators, M-, L- and R-estimators, and hence has obtained their asymptotic normality. The norm chosen on the domain of the functional is a crucial factor for the differentiability, and moreover, it is desirable to have a topology which suggests "robustness" according to Hampel (1971). The weak topology, uniform topology and the topology induced by the Hellinger metric are all "robust". Fernholz (1983) adopted the uniform topology, which is stronger than the weak topology but weaker than the topology induced by the Hellinger metric. Thus, what we may need to do is to set up the Hadamard differentiability under the Hellinger norm. Another way to obtain the asymptotic normality for the infinite support case is to consider the technique used in Stather (1981). We apply similar technique in Chapter 3 for a two-sample semiparametric model.

2.3 MLE of Mixture Proportion

In this section, we construct a MLE of the mixture proportion θ of the model (2.1). We follow the approach of Hall and Titterington (1984), where they have obtained a MLE for Hosmer's (1973) model M2. Our model (2.1) with the assumptions made in this chapter is similar to the model M1 described in

Hosmer (1973), and there is no formal development of MLEs available for the mixture parameter θ under model M1 in the literature. In Hosmer (1973), three models were investigated. The case when there are data only from the mixed distribution is referred to as the M0 model. The data from the M0 model is called mixed data. A sample where the component of origin of each observation is known with certainty is called known data. Two types of known data are possible according to whether or not the known data contains information about the mixture proportion. A sample which contains both mixed and known data and where the known data contains no information about the mixture proportion is called the M1 model. An M2 model refers to the case where the sample contains both mixed and known data, and information about the mixture proportion is contained in the relative number of observations from the two components in the known data. In model (2.1), we do not assume any information contained in the known data. It is appropriate and safe to use model (2.1) for the following three cases: (1) the relative number of observations from the two components contains no information about the mixture proportion; (2) we are not sure whether or not it contains any information and (3) we do not know in which way it contains the information. In this sense, model (2.1) is more robust than the model considered in Hall and Titterington (1984).

As in Hall and Titterington (1984), we first partition the support of h_θ into L regions $R_1, R_2, \ldots, R_L$ so that each observation may be assigned uniquely to a single region. Define

$$\alpha_{0l} = \int_{R_l} f(x)dx,$$

$$\alpha_{1l} = \int_{R_l} g(x)dx,$$

$$\alpha_{2l} = \int_{R_l} h_\theta(x)dx = \theta\alpha_{0l} + (1-\theta)\alpha_{1l},$$

where f, g and h_θ denote densities of F, G and $\theta F + (1-\theta)G$ of model (2.1). Note that $\sum_{l=1}^{L}\alpha_{il} = 1$, $i = 0, 1$. Let n_{il} denote the number out of the n_i which come from region R_l, $i = 0, 1, 2$. The likelihood from the data sets is then proportional to

$$\prod_{l=1}^{L}(\alpha_{0l})^{n_{0l}}(\alpha_{1l})^{n_{1l}}(\theta\alpha_{0l} + (1-\theta)\alpha_{1l})^{n_{2l}}. \tag{2.11}$$

Let θ_{nL} denote the MLE which maximizes (2.11). Unlike in Hall and Titterington (1984), an explicit solution which maximizes the likelihood function (2.11) is not easily available. Instead, we obtain a MLE from its implicit form. Taking the derivatives of the log likelihood and equating them to zero yields the following estimating system :

$$\frac{n_{0l}}{\widehat{\alpha}_{0l}}-\frac{n_{0L}}{\widehat{\alpha}_{0L}}+\frac{n_{2l}\theta_{nL}}{\theta_{nL}\widehat{\alpha}_{0l}+(1-\theta_{nL})\widehat{\alpha}_{1l}}-\frac{n_{2L}\theta_{nL}}{\theta_{nL}\widehat{\alpha}_{0L}+(1-\theta_{nL})\widehat{\alpha}_{1L}}=0,\ l=1,\ldots,L-1 \tag{2.12}$$

$$\frac{n_{1l}}{\widehat{\alpha}_{1l}}-\frac{n_{1L}}{\widehat{\alpha}_{1L}}+\frac{n_{2l}(1-\theta_{nL})}{\theta_{nL}\widehat{\alpha}_{0l}+(1-\theta_{nL})\widehat{\alpha}_{1l}}-\frac{n_{2L}(1-\theta_{nL})}{\theta_{nL}\widehat{\alpha}_{0L}+(1-\theta_{nL})\widehat{\alpha}_{1L}}=0,\ l=1,\ldots,L-1 \tag{2.13}$$

$$\sum_{l=1}^{L}\frac{n_{2l}(\widehat{\alpha}_{0l}-\widehat{\alpha}_{1l})}{\theta_{nL}\widehat{\alpha}_{0l}+(1-\theta_{nL})\widehat{\alpha}_{1l}}=0 \tag{2.14}$$

with constraints

$$\begin{aligned}&\sum_{l=1}^{L}\widehat{\alpha}_{il}=1,\quad i=0,1\\ &\widehat{\alpha}_{il}\geq 0,\quad i=0,1,\ l=1,\ldots,L.\end{aligned} \tag{2.15}$$

The consistency and asymptotic normality of θ_{nL} obtained via equations (2.12)-(2.15) are established in the next theorem.

Theorem 2.5. *Suppose that $\theta\neq 0,1$ and that $n_i/n\to\rho_i$ as $n\to\infty$, $i=0,1,2$. There exist consistent MLEs of θ. Furthermore, if $\rho_0/\rho_1=\theta/(1-\theta)$ and $\sqrt{n}(n_i/n-\rho_i)\to 0$, $i=0,1,2$, then the consistent MLE θ_{nL} is asymptotically normally distributed with mean θ and variance Δ_L, where*

$$\Delta_L=\frac{1}{\Delta^{(2)}}\Big[\frac{(1-\theta)^4}{\rho_0}\Delta^{(0)}+\frac{(1-\theta)^4}{\rho_1}\Delta^{(1)}+\frac{(1-\theta)^2}{\rho_2}\Delta^{(2)}\Big], \tag{2.16}$$

with $\Delta^{(0)}=\sum_{l=1}^{L}\frac{\alpha_{1l}^2}{\alpha_{2l}^2}\alpha_{0l}-\big(\sum_{l=1}^{L}\frac{\alpha_{1l}}{\alpha_{2l}}\alpha_{0l}\big)^2$, $\Delta^{(1)}=\sum_{l=1}^{L}\frac{\alpha_{0l}^2}{\alpha_{2l}^2}\alpha_{1l}-\big(\sum_{l=1}^{L}\frac{\alpha_{0l}}{\alpha_{2l}}\alpha_{1l}\big)^2$ *and* $\Delta^{(2)}=\sum_{l=1}^{L}\frac{\alpha_{0l}^2}{\alpha_{2l}}-1$.

Remark 2.4. As stated above, we assume that the known data (learning samples) may not contain information about λ since our model is similar to model M1 of Hosmer (1973). In Theorem 2.5 it is shown that the MLE θ_{nL} is asymptotically normal when the learning samples contain some information about θ, i.e., when $\rho_0/\rho_1=\theta/(1-\theta)$ holds. For the case that $\rho_0/\rho_1\neq\theta/(1-\theta)$ the method used to prove Theorem 2.5 does not seem to work very well and it needs further study. On the other hand, the MHD estimator θ_n defined in (2.9) is asymptotically normal whether $\rho_0/\rho_1=\theta/(1-\theta)$ holds or not, see Theorem 2.4 above.

Remark 2.5. The MLE obtained using the likelihood function (2.11) is a function of L, the number of regions. In other words, we have a sequence of MLEs depending on L. In fact, the likelihood (2.11) is not the true likelihood

of the original data set given in (2.1). If $\sup_{l\in\{1,2,\ldots,L\}} \{\alpha_{0l}, \alpha_{1l}\} \to 0$ as $L \to \infty$, then for large L, (2.11) is a good approximation to the true likelihood of data in (2.1). As L increases, the number of unknown parameters α_{il}'s ($i = 0, 1$) in (2.11) increases, which in turn makes the maximization process of (2.11) tedious.

2.4 *Asymptotic Efficiency of MHD Estimator*

In this section, we discuss asymptotic efficiency properties of the proposed MHD estimator given in Section 2.2. In particular, we ask the question, "Are the MHD and MLE estimators optimal in some sense?" Asymptotic efficiencies of MHD estimators and MLEs are well-known in parametric models (Beran, 1977 and Lindsay, 1994). However, such properties in nonparametric or semiparametric settings have been less studied. Hall and Titterington (1984) have derived a Cramér-Rao type lower bound for nonparametric estimators of the mixture proportions and thereby characterize asymptotically optimal procedures for the case of sampling model M2 of Hosmer (1973). Furthermore, they have constructed a sequence of maximum likelihood estimators that attain the above mentioned lower bound and are therefore asymptotically optimal in this sense. Following the ideas of Hall and Titterington (1984), we also obtain a Cramér-Rao type lower bound for nonparametric estimators of the mixture proportion θ. Then we show that the proposed MHD estimator attains this lower bound under certain regularity conditions, showing an asymptotically optimal property of the proposed MHD estimator.

Theorem 2.6. *Let θ_n' denote a nonparametric estimator of θ such that $n^{1/2}(\theta_n' - \theta) \to N(0, V(\theta, f, g, h_\theta))$ and $nVar(\theta_n' - \theta) \to V(\theta, f, g, h_\theta)$ as $n \to \infty$, where f, g and h_θ denote the densities of distributions F, G and $\theta F + (1-\theta)G$, respectively, of (2.1). Suppose that $n_i/n - \rho_i \to 0$, $i = 0, 1, 2$, and $\rho_0/\rho_1 = \theta/(1-\theta)$. If $V(\theta, f_n, g_n, h_n) \to V(\theta, f, g, h_\theta)$ whenever $f_n \to f$, $g_n \to g$ and $h_n \to h_\theta$ in the class of uniformly piecewise continuous densities, then*

$$V(\theta, f, g, h_\theta) \geq \Delta(\theta, f, g, h_\theta)$$

for any $f \neq g$ and $\theta \in (0, 1)$, where

$$\Delta(\theta, f, g, h_\theta) = \frac{1}{\Delta_2^2}\Big[\frac{(1-\theta)^4}{\rho_0}\Delta_0 + \frac{(1-\theta)^4}{\rho_1}\Delta_1 + \frac{(1-\theta)^2}{\rho_2}\Delta_2\Big] \tag{2.17}$$

with $\Delta_0 = \int \frac{g^2}{h_\theta^2} f dx - (\int \frac{g}{h_\theta} f dx)^2$, $\Delta_1 = \int \frac{f^2}{h_\theta^2} g dx - (\int \frac{f}{h_\theta} g dx)^2$ and $\Delta_2 = \int \frac{f^2}{h_\theta} dx - 1$.

Corollary 2.1. *Assume that conditions of Theorem 2.4 hold. Then the asymptotic variance of the MHD estimator θ_n of (2.9) is equal to $\Delta(\theta, f, g, h_\theta)$, where $\Delta(\theta, f, g, h_\theta)$ is given in (2.17). In this sense, θ_n is asymptotically efficient.*

Remark 2.6. Theorem 2.5 only gives the asymptotic distribution when $\rho_0/\rho_1 = \theta/(1-\theta)$. For other cases, the method adopted in Section 2.3 does not work well and one may need to seek different ways to find a lower bound of the asymptotic variance. Therefore, for the case $\rho_0/\rho_1 \neq \theta/(1-\theta)$, the full efficiency of the MHD estimator θ_n of (2.9) is unknown and it needs further research. Nevertheless, we have shown in Theorem 2.4 that θ_n is $n^{1/2}$-consistent, i.e. $n^{1/2}(\theta_n - \theta) = O_P(1)$, which demonstrates that θ_n has good efficiency properties whether $\rho_0/\rho_1 = \theta/(1-\theta)$ holds or not. One can also see this behavior from the numerical studies in Section 2.5.

2.5 Robustness and Simulation Studies

It is difficult to establish any theoretical results on the robustness of our MHD estimator because of the inherent complexity of this problem. Thus to study robustness properties of our estimator we relied on Monte Carlo methods. We considered a mixture of two normal distributions in this numerical study. Specifically, we studied the following four mixture models:

$$\begin{array}{lrcl} \text{Model I}: & h_\theta & = & 0.25N(0,1) + 0.75N(3.60,1), \\ \text{Model II}: & h_\theta & = & 0.25N(0,1) + 0.75N(2.32,1), \\ \text{Model III}: & h_\theta & = & 0.5N(0,1) + 0.5N(3.76,1), \\ \text{Model IV}: & h_\theta & = & 0.5N(0,1) + 0.5N(2.56,1). \end{array} \tag{2.18}$$

That is, we set the distributions F and G of (2.1) as $N(0,1)$ and $N(\mu,1)$, respectively, where $\mu \neq 0$ depends on the Model. Note that Models I and III have an overlap of 0.03, whereas Models II and IV have an overlap of 0.1. Here the overlap is defined as the probability of misclassification using the rule: classify an observation x as being from population F if $x < x_c$ and from population G if $x \geq x_c$, where x_c is the unique point between 0 and μ such that $\theta f(x_c) = (1-\theta)g(x_c)$. We examined the resistance of our MHD estimator defined at (2.9) to a single outlying observation. For this purpose, the α-IF given in Beran (1977) is a suitable measure of the change in the estimator. It has been observed, however, that analytical evaluation of the α-IF is almost impossible in the mixture context (Karlis and Xekalaki, 1998). For this reason, adapted versions of the α-IF have been employed by many authors in the mixture context; see, e.g., Lu et al. (2003). In this study, we have used the adapted α-IF defined in the preceding paper.

First, we considered the case that $\rho_0/\rho_1 = \theta/(1-\theta)$. For Models I and II, we chose sample sizes $n_0 = 50$, $n_1 = 150$ and $n_2 = 300$, and for Models III and IV,

$n_0 = n_1 = 100$ and $n_2 = 300$ were chosen. Note that the outlying observation could come from any one of the three distributions. That is, for example, for Model I, the outlier may be from distributions $N(0,1)$, or $N(3.60,1)$ or from the mixture distribution $0.25N(0,1)+0.75N(3.60,1)$. Thus, after drawing data sets of the specified sizes, 147 alternate versions of the data were created by replacing the last observation in the first data set, the last observation in the second data set, or the last observation in the third data set by an integer from -24 to 24. Here we have chosen a moderate sample size of $n = 500$ in our study, and we have done 1000 replications and averaged the results over the 1000 replications. The contamination rate is then $1/500$ and the three α-IFs are given by

$$IF_0(x) = \frac{W\big((x, X_i)_{i=1}^{n_0-1}, (Y_i)_{i=1}^{n_1}, (Z_i)_{i=1}^{n_2}\big) - W\big((X_i)_{i=1}^{n_0}, (Y_i)_{i=1}^{n_1}, (Z_i)_{i=1}^{n_2}\big)}{1/500},$$

$$IF_1(x) = \frac{W\big((X_i)_{i=1}^{n_0}, (x, Y_i)_{i=1}^{n_1-1}, (Z_i)_{i=1}^{n_2}\big) - W\big((X_i)_{i=1}^{n_0}, (Y_i)_{i=1}^{n_1}, (Z_i)_{i=1}^{n_2}\big)}{1/500},$$

$$IF_2(x) = \frac{W\big((X_i)_{i=1}^{n_0}, (Y_i)_{i=1}^{n_1}, (x, Z_i)_{i=1}^{n_2-1}\big) - W\big((X_i)_{i=1}^{n_0}, (Y_i)_{i=1}^{n_1}, (Z_i)_{i=1}^{n_2}\big)}{1/500},$$

where W could be any functional (estimator of θ) based on three data sets from f, g and h_θ, respectively. In our case, W is given by functional $\widehat{T}(\widehat{h})$ defined in (2.9) (which is also based on three data sets from f, g and h_θ, respectively). We used the compact-supported Epanechnikov kernel function

$$K(x) = \frac{3}{4}\left(1 - x^2\right) I_{[-1,1]}(x),$$

for all three kernels K_0, K_1 and K_2 in (2.3), (2.4) and (2.6), respectively. The positive constants b_{n_0}, b_{n_1} and b_{n_2} in (2.3), (2.4) and (2.6), respectively, were taken to be $b_{n_0} = n_0^{-1/3}$, $b_{n_1} = n_1^{-1/3}$ and $b_{n_2} = n_2^{-1/3}$. This selection satisfies the bandwidth assumptions in the theorems of Section 2.2. For scale statistics S_{n_0}, S_{n_1} and S_{n_2} in (2.3), (2.4) and (2.6), respectively, we used the following robust scale estimator proposed by Rousseeuw and Croux (1993),

$$S_n = 1.1926 \ \mathrm{med}_i\big(\mathrm{med}_j(|X_i - X_j|)\big).$$

The choices of kernel function, bandwidth and scale estimator satisfy conditions C1, C3 and C4. Thus C5 is satisfied by Theorem 2.2. For the average of the 1000 replications, the α-IFs under the four models are graphically displayed in Figure 2.1. From Figure 2.1, we can see that as the outlier approaches $\pm\infty$, the α-IF appears to converge to a constant, i.e., $\lim_{x\to\infty} IF_i(x) = \lim_{x\to-\infty} IF_i(x)$, $i = 0, 1, 2$. In fact, the α-IFs outside the interval $[-3, 7]$ seem to be constant,

while they take varying values inside the interval $[-3, 7]$. Specifically, IF_0 has a higher value inside the interval $[-3, 7]$ than outside the interval, whereas IF_1 has a lower value inside the interval $[-3, 7]$ than outside the interval.

Next, we considered the case that $\rho_0/\rho_1 \neq \theta/(1-\theta)$. We have used the same four models as above but with $n_0/n_1 \neq \theta/(1-\theta)$. We observed that the resulting α-IFs were similar to those in the case that $n_0/n_1 = \theta/(1-\theta)$ considered above. Two typical examples are given in Figure 2.2, in which figure (a) is under Model I with sample sizes $n_0 = n_1 = 100$ and $n_2 = 300$, and figure (b) is under Model IV with sample sizes $n_0 = 50$, $n_1 = 150$ and $n_2 = 300$. Robustness of the MHD estimator is evident from Figures 2.1 and 2.2 by the fact that the α-IFs are bounded.

We also compared our MHD estimator with two MLEs. For the reasons stated in Remark 2.5, the MLE constructed in Section 2.3 was not used in our comparison. Instead, we examined two ML estimators based on following likelihood functions combined with the data $(Z_1, \ldots, Z_{n_2})$:

$$L = \prod_{i=1}^{n_2} [\theta f(Z_i) + (1-\theta) g(Z_i)]$$

and

$$\widetilde{L} = \prod_{i=1}^{n_2} [\theta \widehat{f}(Z_i) + (1-\theta) \widehat{g}(Z_i)],$$

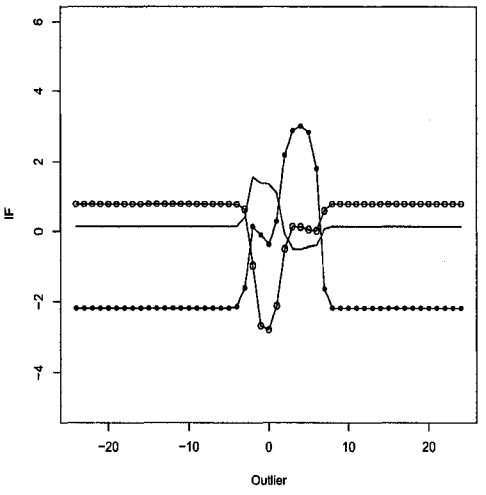
where $\widehat{f}$ and $\widehat{g}$ are the kernel density estimators of f and g defined by (2.3) and (2.4), respectively, with f and g as in model (2.2). In other words, the likelihood L is constructed assuming that density functions f and g are completely known, whereas $\widetilde{L}$ is obtained by replacing f and g by their estimators. Thus, L and $\widetilde{L}$ are rather naturally constructed for simulation purposes. We define

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta \in [0,1]} L \tag{2.19}$$

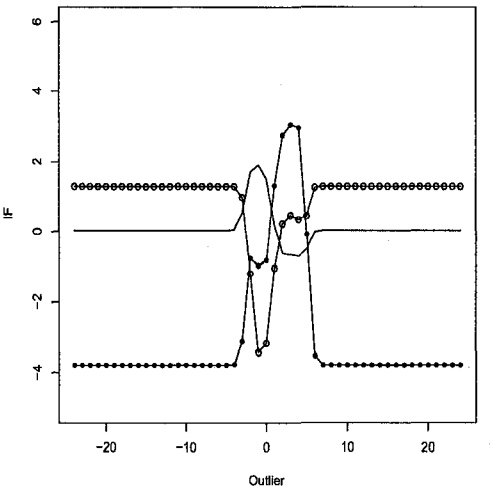
and

$$\widetilde{\theta}_{\text{MLE}} = \arg\max_{\theta \in [0,1]} \widetilde{L} \tag{2.20}$$

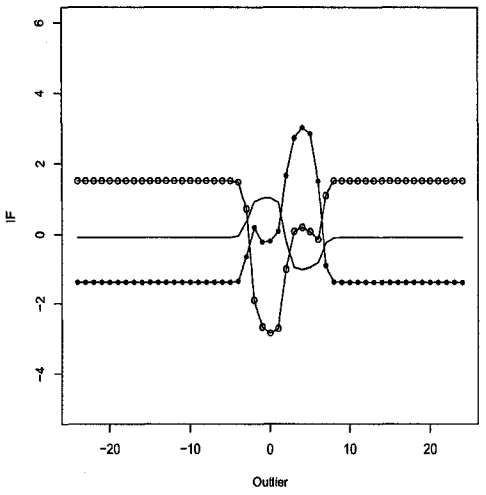
as the MLEs of θ based on L and $\widetilde{L}$, respectively. In our simulation, the data were again generated from the models defined in (2.18). For each model, 500 samples with $n_0 = n_1 = 30$ and $n_2 = 100$ were obtained from the corresponding distributions. For instance, for Model I, samples of size $n_0 = 30$ and $n_1 = 30$ were obtained from the distributions $N(0, 1)$ and $N(3.60, 1)$, respectively, while a sample of size $n_2 = 100$ was obtained from the mixture distribution $0.25N(0, 1) + 0.75N(3.60, 1)$. In each of the distributional situations considered, we obtained estimates of the bias and mean squared error (MSE) as follows:



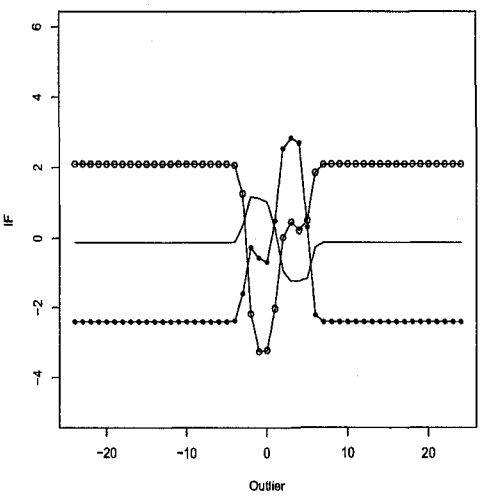
(a) Model I



(b) Model II

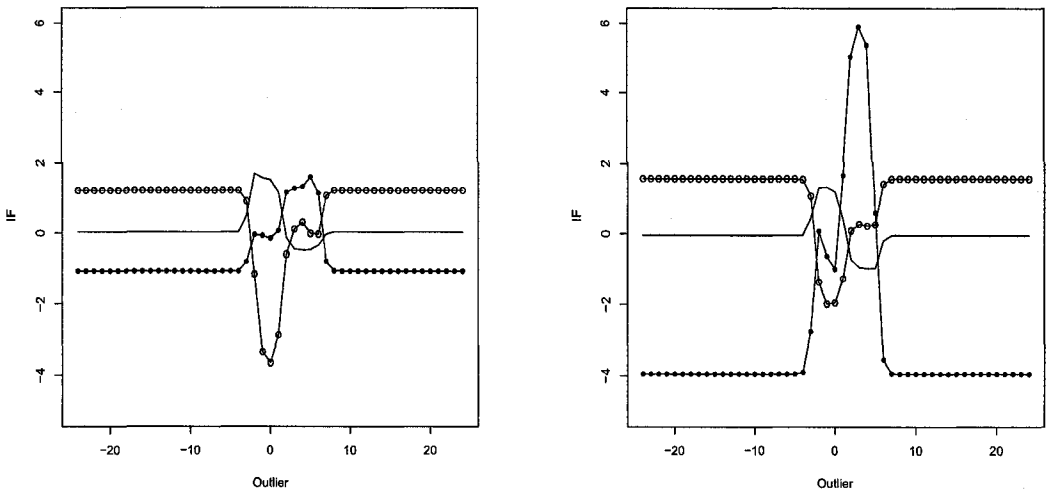


(c) Model III



(d) Model IV

Fig. 2.1: The α-influence function of MHD estimator θ_n with respect to single outlier under Model I-IV and $\rho_0/\rho_1 = \theta/(1-\theta)$, with $\bullet$ - IF_0, $\circ$ - IF_1 and $--$ - IF_2.



(a) Model I: $n_0 = n_1 = 100$ and $n_2 = 300$

(b) Model IV: $n_0 = 50$, $n_1 = 150$ and $n_2 = 300$

Fig. 2.2: The α-influence function of MHD estimator θ_n with respect to single outlier under Model I and IV and $\rho_0/\rho_1 \neq \theta/(1-\theta)$, with $\bullet$ - IF_0, $\circ$ - IF_1 and $--$ - IF_2.

$$\widehat{\text{Bias}} = \frac{1}{N_s}\sum_{i=1}^{N_s}(\widehat{\mu}_i - \mu)$$

and

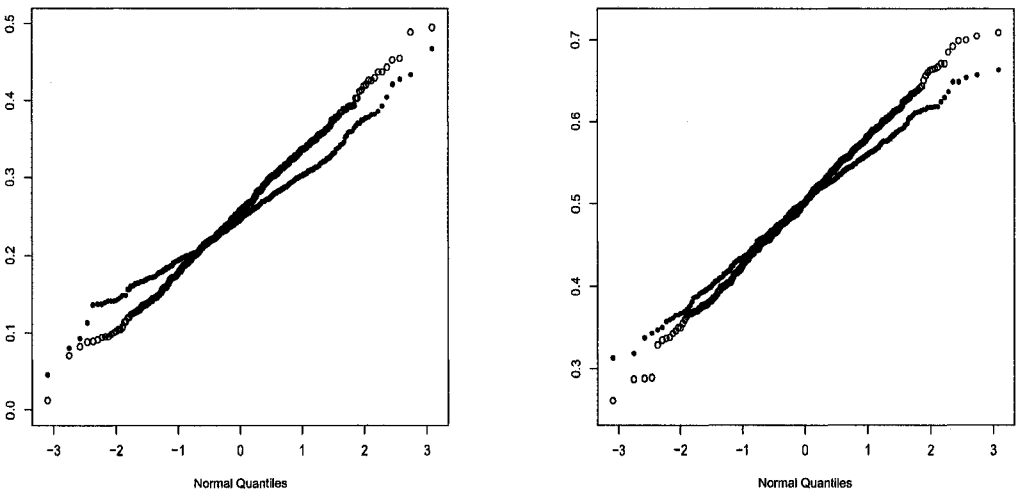
$$\widehat{\text{MSE}} = \frac{1}{N_s}\sum_{i=1}^{N_s}(\widehat{\mu}_i - \mu)^2,$$

where N_s is the number of replications ($N_s = 500$ in our case), and $\widehat{\mu}_i$ denotes an estimate of μ for the ith replication. Here $\mu = \theta$ and $\widehat{\mu}$ denotes either the proposed MHD estimator θ_n, $\widehat{\theta}_{\text{MLE}}$ or $\widetilde{\theta}_{\text{MLE}}$. Kernel estimators $\widehat{f}$ and $\widehat{g}$ are the same as those employed in the robustness study above. Simulation results are summarized in Table 2.1.

Tab. 2.1: Estimates of the biases and MSEs of θ_n, $\widehat{\theta}_{\text{MLE}}$ and $\widetilde{\theta}_{\text{MLE}}$.

Model	$\widehat{\text{Bias}}(\theta_n)$	$\widehat{\text{MSE}}(\theta_n)$	$\widehat{\text{Bias}}(\widehat{\theta}_{\text{MLE}})$	$\widehat{\text{MSE}}(\widehat{\theta}_{\text{MLE}})$	$\widehat{\text{Bias}}(\widetilde{\theta}_{\text{MLE}})$	$\widehat{\text{MSE}}(\widetilde{\theta}_{\text{MLE}})$
I	-0.0021	0.0033	-0.0023	0.0021	-0.0098	0.0028
II	0.0071	0.0061	-0.0028	0.0029	-0.0853	0.0119
III	-0.0006	0.0039	-0.0018	0.0029	-0.0295	0.0052
IV	0.0031	0.0060	-0.0024	0.0036	-0.1399	0.0281

We found that the MHD estimator θ_n performed better than the MLE $\widetilde{\theta}_{\text{MLE}}$ for models II, III and IV, and both were comparable for model I. On the other hand, the MLE $\widehat{\theta}_{\text{MLE}}$, which is based on assuming f and g are known, showed the best performance among the three estimators considered for all four models. However, this behavior can be expected here since $\widehat{\theta}_{\text{MLE}}$ employs more information (i.e., knowing f and g, or in other words $n_0 = \infty$ and $n_1 = \infty$) than either $\widetilde{\theta}_{\text{MLE}}$ or θ_n. Note that $\widehat{\theta}_{\text{MLE}}$ is not available in practice and the sole purpose of analyzing it here is to examine the amount of loss in performance when f and g are unknown. The bias and MSE of θ_n were less affected by the preceding fact compared to those of $\widetilde{\theta}_{\text{MLE}}$. Note that $\widehat{\theta}_{\text{MLE}}$ uses only the mixture sample of size $n_2 = 100$, whereas θ_n and $\widetilde{\theta}_{\text{MLE}}$ are based on all three samples of sizes $n_0 = 30$, $n_1 = 30$ and $n_2 = 100$. (Data from f and g are not required for $\widehat{\theta}_{\text{MLE}}$ since it is based on the fact that f and g are known.) Thus, one might argue that a direct comparison between θ_n and $\widehat{\theta}_{\text{MLE}}$ may not be fair. In Figure 2.3, we have also given the normal probability plots of the proposed MHD estimator θ_n based on the 500 replications for all the four models. Figure 2.3 demonstrates that the sampling distribution of θ_n closely approximates a normal curve for each model, no matter $\rho_0/\rho_1 = \theta/(1-\theta)$ holds or not.



(a) Model I and II

(b) Model III and IV

Fig. 2.3: Normal probability plots of MHD estimator θ_n for sample sizes $n_0 = n_1 = 30$ and $n_2 = 100$, with • - Model I and III and ∘ - Model II and IV.

2.6 Examples

In this section, we applied the proposed MHD estimator constructed in Section 2.2 to a real data set used in Hosmer (1973). The International Halibut Commission in Seattle, Washington wanted to estimate the proportions of male or female halibut. They provided the lengths of 74 eleven year old male halibut and 134 eleven year old female halibut caught on one of their research cruises. A summary of the data is given in Table 2.2.

Tab. 2.2: Frequency distribution of the lengths in centimeters of 11 year old male and female halibut caught on Western Trip I, April 1957.

Sex	75	80	85	90	95	100	105	110	115	120	125	130	135
Males	2	7	8	6	7	11	10	9	9	3	2	0	0
Females	0	1	0	0	4	2	7	18	22	29	28	13	10

The sample proportion of males in this example is $74/208 \approx 0.3558$. To illustrate computation of the MHD estimator, we randomly selected 14 male lengths and 26 female lengths from 74 and 134, respectively, so that the remaining male proportion was about the same as 0.3558 (60/168). These samples formed the first, second and the mixture samples, respectively. That is, $n_0 = 14$, $n_1 = 26$ and $n_2 = 168$. This idea of selection of samples is similar to model M1 sampling mechanism described in Section 2.3. Based on above sample sizes, we carried out a simulation with 10 and 100 replications and averaged the results. The resulting MHD estimates were 0.2755 and 0.3144, respectively, for the male proportion. The average squared errors from the sample proportion 0.3558 were 0.0112 and 0.0125, respectively. Based on one replication, Hosmer (1973) obtained a MLE of male proportion of 0.465 with a squared error of 0.012 from the sample proportion. Thus, our results are similar to those in Hosmer (1973). However, our estimator is constructed without the normality assumption on the densities f and g, whereas Hosmer (1973) assumed the two component distributions were normal. The kernels K_0, K_1, K_2, bandwidths and robust scale estimators of (2.3), (2.4) and (2.6), respectively, were chosen the same way as in Section 2.5 in the above simulation.

Another example is given in Anderson (1979). Anderson (1979) generated samples $(X_1, \ldots, X_{n_0}) = (1.15, 0.25, 2.31, 2.44, 3.28, 3.34)$ from the $N(2,1)$ distribution with $n_0 = 6$, $(Y_1, \ldots, Y_{n_1}) = (0.74, -0.50, 1.08, 1.34, -0.74, 0.15)$ from the $N(0,1)$ distribution with $n_1 = 6$, and $(Z_1, \ldots, Z_{n_2}) = (-0.23, 0.71, 0.92, -0.53, -0.68, 1.04, 0.61, -0.88, -0.61, 0.59, 2.96, 2.59)$ from the mixture $\theta N(2,1) + (1-\theta)N(0,1)$ distribution with $n_2 = 12$ and $\theta = 0.25$. Using the assumption that the log ratio of the two component density functions is linear (this is the case here), Anderson (1979) obtained a MLE of θ as 0.19. Later Zhang (2002)

proposed an EM algorithm based argument on the same log linear model to calculate a MLE, and he gave his estimate of θ to be 0.1890. Using the same data set, we obtained the MHD estimate of θ defined in Section 2.2. Our estimate of θ came to be 0.2045. Note that our estimate is much closer to the actual value of θ than both Anderson (1979) and Zhang (2002) estimates, even though we made no assumptions about the relationship of the two component distributions while they made an extra assumption that the log ratio of the two component densities is linear.

2.7 Concluding Remarks

In this chapter, we have considered the problem of estimating the mixture proportion in a general two-population mixture, when samples of sizes n_0 and n_1 are available from the two individual populations while a sample of size n_2 is available from the mixture population. There have been very few attempts in the literature to estimate the parameters in a mixture problem under the preceding setup using the minimum distance approaches or by the method of maximum likelihood. Here we have constructed a MHD estimator of the mixture proportion. The proposed MHD estimator has been shown to have good efficiency and robustness properties. By constructing a sequence of multinomial approximations, we have also obtained a sequence of asymptotically normal MLE estimator of the mixture proportion. Furthermore, we have derived a Cramér-Rao type lower bound for nonparametric estimators of the mixture proportion and thereby characterized asymptotically efficient estimators.

The results in this chapter could be extended to the more general nonparametric mixture model studied in Hall and Titterington (1984) of the form $\sum_{i=1}^{k} p_i f_i$, where $0 \leq p_i \leq 1$ and $\sum_{i=1}^{k} p_i = 1$. We believe that results similar to those in this chapter can be established for the semiparametric model proposed in Anderson (1979) as well. He assumed that the log ratio of the densities f and g is linear of the form $\log\big(g(x)/f(x)\big) = \beta_0 + \beta_1 x$, or equivalently $g(x) = f(x)\exp(\beta_0 + \beta_1 x)$. The three data sets in (2.1) then would come from the distributions $f(x)$, $f(x)\exp(\beta_0 + \beta_1 x)$ and $\{\theta + (1-\theta)\exp(\beta_0 + \beta_1 x)\}f(x)$, respectively; and MHD estimators of θ, β_0 and β_1 may be developed along arguments similar to those given in Section 2.2 above. A more general two-sample semiparametric model than the one considered in Anderson (1979) is investigated in Chapter 3.

2.8 Proofs

Proof of Theorem 2.1.

The method of proof is similar to that of Theorem 2.1 of Beran (1977). For completeness, we give the proof below.

(i) Let $d_n(t) = \| \widetilde{h}_t^{1/2}(x) - g^{1/2}(x) \|$. For any sequence $\{t_k : t_k \to t, t_k, t \in [0,1]\}$,

$$\begin{aligned} | d_n^2(t_k) - d_n^2(t) | &= | \int [\widetilde{h}_{t_k}^{1/2}(x) - g^{1/2}(x)]^2 dx - \int [\widetilde{h}_t^{1/2}(x) - g^{1/2}(x)]^2 dx | \\ &= 2 | \int [\widetilde{h}_{t_k}^{1/2}(x) - \widetilde{h}_t^{1/2}(x)] g^{1/2}(x) dx | \\ &\leq 2 \| \widetilde{h}_{t_k}^{1/2}(x) - \widetilde{h}_t^{1/2}(x) \| . \end{aligned}$$

Since $\int \widetilde{h}_t(x)dx = \int \widetilde{h}_{t_k}(x)dx = 1$, $\int [\widetilde{h}_t(x) - \widetilde{h}_{t_k}(x)]^+ dx = \int [\widetilde{h}_t(x) - \widetilde{h}_{t_k}(x)]^- dx$. Thus, $\| \widetilde{h}_{t_k}^{1/2}(x) - \widetilde{h}_t^{1/2}(x) \|^2 \leq \int | \widetilde{h}_t(x) - \widetilde{h}_{t_k}(x) | dx = 2 \int [\widetilde{h}_t(x) - \widetilde{h}_{t_k}(x)]^+ dx$. Also, $[\widetilde{h}_t(x) - \widetilde{h}_{t_k}(x)]^+ \leq \widetilde{h}_t(x)$ and, for every x, $\widetilde{h}_t(x)$ is continuous in t. Thus, by the Dominated Convergence Theorem, $\| \widetilde{h}_{t_k}^{1/2}(x) - \widetilde{h}_t^{1/2}(x) \| \to 0$ as $k \to \infty$. So, $d_n(t_k) \to d_n(t)$ as $k \to \infty$, i.e., d_n is continuous on $[0,1]$ and achieves a minimum over $t \in [0,1]$. Similarly, $d(t) = \| h_t^{1/2}(x) - g^{1/2}(x) \|$ is continuous on $[0,1]$.

(ii) Suppose $\| \phi_n^{1/2} - \phi^{1/2} \| \to 0$ and $\sup_{t \in [0,1]} \| \widetilde{h}_t^{1/2} - h_t^{1/2} \| \to 0$ as $n \to \infty$. Put $d_n(t) = \| \widetilde{h}_t^{1/2}(x) - \phi_n^{1/2}(x) \|$ and $d(t) = \| h_t^{1/2}(x) - \phi^{1/2}(x) \|$. By Minkowski's inequality,

$$\begin{aligned} &| d_n(t) - d(t) | \\ &\qquad \leq \left\{ \int \left[\widetilde{h}_t^{1/2}(x) - \phi_n^{1/2}(x) - h_t^{1/2}(x) + \phi^{1/2}(x) \right]^2 dx \right\}^{1/2} \\ &\qquad \leq \left\{ 2 \int \left[\widetilde{h}_t^{1/2}(x) - h_t^{1/2}(x) \right]^2 dx + 2 \int \left[\phi_n^{1/2}(x) - \phi^{1/2}(x) \right]^2 dx \right\}^{1/2}. \end{aligned}$$

Consequently, $\sup_{t \in [0,1]} | d_n(t) - d(t) | \leq \{ 2 \sup_{t \in [0,1]} \int [\widetilde{h}_t^{1/2}(x) - h_t^{1/2}(x)]^2 dx + 2 \int [\phi_n^{1/2}(x) - \phi^{1/2}(x)]^2 dx \}^{1/2}$, and the r.h.s. of the preceding expression goes to zero as $n \to \infty$ by assumptions. Therefore, we have, as $n \to \infty$, $d_n(\theta_0) \to d(\theta_0)$ and $d_n(\theta_n) - d(\theta_n) \to 0$, with $\theta_0 = T_0(\phi)$ and $\theta_n = \widehat{T}(\phi_n)$. If $\theta_n \not\to \theta_0$, then there exists a subsequence $\{\theta_m\} \subseteq \{\theta_n\}$ such that $\theta_m \to \theta' \neq \theta_0$, implying $\theta' \in [0,1]$ and $d(\theta_m) \to d(\theta')$ by continuity of d. From above results, we have $d_m(\theta_m) - d_m(\theta_0) \to d(\theta') - d(\theta_0)$. By the definition of θ_m, $d_m(\theta_m) - d_m(\theta_0) \leq 0$. Hence, $d(\theta') - d(\theta_0) \leq 0$. But by the definition of θ_0 and the uniqueness of it, $d(\theta') > d(\theta_0)$. This is a contradiction. Therefore, $\theta_n \to \theta_0$.

(iii) For fixed f and g, $t_1 \neq t_2$ implies $h_{t_1} \neq h_{t_2}$. So $\{h_t\}_{t \in [0,1]}$ is identifiable. Immediately, we have $T_0(h_t) = t$ uniquely. □

Proof of Theorem 2.2.

If we can prove that as $n \to \infty$, $\| \widehat{h}^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$ and $\sup_{t\in[0,1]} \| \widetilde{h}_t^{1/2} - h_t^{1/2} \| \xrightarrow{P} 0$, then by Theorem 2.1, $\theta_n - \theta = \widehat{T}(\widehat{h}) - T_0(h_\theta) \xrightarrow{P} 0$ as $n \to \infty$. It is easy to show that $\sup_x |\widehat{f}(x) - f(x)| \xrightarrow{P} 0$, $\sup_x |\widehat{g}(x) - g(x)| \xrightarrow{P} 0$, $\sup_x |\widehat{h}(x) - h_\theta(x)| \xrightarrow{P} 0$ and $\sup_{t\in[0,1]} \sup_x |\widetilde{h}_t(x) - h_t(x)| \xrightarrow{P} 0$, see below. From an argument similar to the proof of Theorem 2.1 (i), we have $\| \widehat{h}^{1/2}(x) - h_\theta^{1/2}(x) \|^2 \leq \int \mid h_\theta(x) - \widehat{h}(x) \mid dx = 2\int [h_\theta(x) - \widehat{h}(x)]^+ dx$ and $[h_\theta(x) - \widehat{h}(x)]^+ < f(x) + g(x)$. Then by the Dominated Convergence Theorem, it follows that $\| \widehat{h}^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$ as $n \to \infty$. On the other hand, $\sup_x |\widetilde{h}_t(x) - h_t(x)| \leq \sup_x |\widehat{f}(x) - f(x)| + \sup_x |\widehat{g}(x) - g(x)| \xrightarrow{P} 0$. Since $\sup_{t\in[0,1]} \int [\widetilde{h}_t^{1/2}(x) - h_t^{1/2}(x)]^2 dx \leq \sup_{t\in[0,1]} \int \mid \widetilde{h}_t(x) - h_t(x) \mid dx \leq \int |f(x) - \widehat{f}(x)| dx + \int |g(x) - \widehat{g}(x)| dx$, it follows that $\sup_{t\in[0,1]} \| \widetilde{h}_t^{1/2} - h_t^{1/2} \| \xrightarrow{P} 0$.

Finally we prove that $\sup_x |\widehat{f}(x) - f(x)| \xrightarrow{P} 0$ as $n \to \infty$. Define

$$\overline{E}\widehat{f}(x) = (b_{n_0} S_{n_0})^{-1} \int K_0\Big(\frac{x-y}{b_{n_0} S_{n_0}}\Big) dF(y)$$

and $B_{n_0}(x) = n_0^{1/2}[D_{n_0}(x) - F(x)]$, where D_{n_0} denote the empirical c.d.f. of $(X_1, X_2, \ldots, X_{n_0})$. We have $\sup_x |B_{n_0}(x)| = O_P(1)$ (see Kiefer and Wolfowitz (1958)) and then

$$\sup_x |\widehat{f}(x) - \overline{E}\widehat{f}(x)| \leq n_0^{-1/2} (b_{n_0} S_{n_0})^{-1} \sup_x |B_{n_0}(x)| \cdot \int |K_0^{(1)}(x)| dx \xrightarrow{P} 0. \quad (2.21)$$

Suppose K_0 has compact support $[a_0, b_0]$, then

$$\begin{aligned} \sup_x |\overline{E}\widehat{f}(x) - f(x)| &= \sup_x \Big| \int_{a_0}^{b_0} K_0(t) f(x - b_{n_0} S_{n_0} t) dt - f(x) \Big| \\ &= \sup_x \Big| \int_{a_0}^{b_0} K_0(t) dt f(x - b_{n_0} S_{n_0} \xi_{n_0}) - f(x) \Big|, \\ &\qquad \text{with } \xi_{n_0} \in [a_0, b_0] \\ &\leq \sup_x \sup_{t\in[a_0,b_0]} |f(x - b_{n_0} S_{n_0} t) - f(x)| \\ &\xrightarrow{P} 0. \end{aligned} \quad (2.22)$$

From (2.21) and (2.22), one has $\sup_x |\widehat{f}(x) - f(x)| \xrightarrow{P} 0$. Similarly, $\sup_x |\widehat{g}(x) - g(x)| \xrightarrow{P} 0$, $\sup_x |\widehat{h}(x) - h_\theta(x)| \xrightarrow{P} 0$, and $\sup_{t\in[0,1]} \sup_x |\widetilde{h}_t(x) - h_t(x)| \leq \sup_x |\widehat{f}(x) -$

$f(x)| + \sup_x |\widehat{g}(x) - g(x)| \xrightarrow{P} 0.$ □

Proof of Theorem 2.3.

Since T is continuous at $(\{h_t\}_{t\in[0,1]}, h_\theta)$ in the sense of Theorem 2.1 (ii) and $\| \widehat{h}^{1/2} - h_\theta^{1/2} \| \to 0$ and $\sup_{t\in[0,1]} \| \widetilde{h}_t^{1/2} - h_t^{1/2} \| \to 0$ as $n \to \infty$, we obtain that $T\big(\{\widetilde{h}_t\}_{t\in[0,1]}, \widehat{h}\big) \to T\big(\{h_t\}_{t\in[0,1]}, h_\theta\big)$ as $n \to \infty$. That is, $\theta_n \to \theta$ as $n \to \infty$. Thus, for large n, $\theta_n \in (0,1)$ since $\theta \in (0,1)$. Denote $S_t = \widetilde{h}_t^{1/2}$. We claim that for any $t \in (0,1)$

$$S_{t+\alpha}(x) = S_t(x) + \alpha \dot{S}_t(x) + \alpha\mu_\alpha(x), \tag{2.23}$$

$$\dot{S}_{t+\alpha}(x) = \dot{S}_t(x) + \alpha \ddot{S}_t(x) + \alpha\nu_\alpha(x), \tag{2.24}$$

where $\dot{S}_t(x) = \frac{\partial S_t(x)}{\partial t}$ and $\ddot{S}_t(x) = \frac{\partial^2 S_t(x)}{\partial t^2}$ are in L_2, and $\mu_\alpha(x)$ and $\nu_\alpha(x)$ tend to zero in L_2 as $\alpha \to 0$. The proof of this statement is shown at the end of this proof. Since $\theta_n \in (0,1)$ minimizes the Hellinger distance between $\widetilde{h}_t$ and $\widehat{h}$, or in other words θ_n maximizes $\int \widetilde{h}_t^{1/2}(x)\widehat{h}^{1/2}(x)dx$, (2.23) yields that

$$\begin{aligned}
&\lim_{\alpha\to 0} \alpha^{-1} \int [\widetilde{h}_{t+\alpha}^{1/2}(x) - \widetilde{h}_t^{1/2}(x)]\widehat{h}^{1/2}(x)dx \\
&= \int \frac{\partial \widetilde{h}_t^{1/2}(x)}{\partial t}\widehat{h}^{1/2}(x)dx + \lim_{\alpha\to 0}\int \mu_\alpha(x)\widehat{h}^{1/2}(x)dx \\
&= \int \frac{\partial \widetilde{h}_t^{1/2}(x)}{\partial t}\widehat{h}^{1/2}(x)dx,
\end{aligned}$$

and so we have $0 = \int \frac{\partial \widetilde{h}_{\theta_n}^{1/2}(x)}{\partial \theta_n}\widehat{h}^{1/2}(x)dx$. Since $\widehat{f} \to f$ and $\widehat{g} \to g$ uniformly, by a Taylor expansion one obtains

$$\begin{aligned}
2\frac{\partial}{\partial\theta_n}\widetilde{h}_{\theta_n}^{1/2} &= \frac{f-g}{[\theta_n f + (1-\theta_n)g]^{1/2}} + \frac{\frac{\theta_n}{2}(f-g)+g}{[\theta_n f + (1-\theta_n)g]^{3/2}} \times (\widehat{f} - f) \\
&\quad - \frac{\frac{1}{2}(1+\theta_n)(f-g)+g}{[\theta_n f + (1-\theta_n)g]^{3/2}} \times (\widehat{g} - g) \\
&\quad - \frac{\frac{\theta_n^2}{4}(f_r - g_r) + \theta_n g_r}{2[\theta_n f_r + (1-\theta_n)g_r]^{5/2}} \times (\widehat{f} - f)^2 \\
&\quad + \frac{\frac{1}{4}(1-\theta_n)(3+\theta_n)(f_r - g_r) + (1-\theta_n)g_r}{2[\theta_n f_r + (1-\theta_n)g_r]^{5/2}} \times (\widehat{g} - g)^2 \\
&\quad + \frac{\frac{1}{4}\theta_n(1+\theta_n)(f_r - g_r) + (\theta_n - \frac{1}{2})g_r}{[\theta_n f_r + (1-\theta_n)g_r]^{5/2}} \times (\widehat{f} - f)(\widehat{g} - g),
\end{aligned} \tag{2.25}$$

where $f_r(x) = r(x)f(x) + [1 - r(x)]\widehat{f}(x)$, $g_r(x) = r(x)g(x) + [1 - r(x)]\widehat{g}(x)$ and $r(x) \in [0,1]$. Since $f_r(x) \to f(x)$ and $g_r(x) \to g(x)$ uniformly as $n \to \infty$ and $h_t \geq C$ for all $t \in [0,1]$ and some $C > 0$, we have

$$|\int \frac{\frac{\theta_n^2}{4}(f_r-g_r)+\theta_n g_r}{2[\theta_n f_r+(1-\theta_n)g_r]^{5/2}}\,(\widehat{f}-f)^2\,\widehat{h}^{1/2}dx\,|\le \tau_1\int(\widehat{f}-f)^2dx,$$

$$|\int \frac{\frac{1}{4}(1-\theta_n)(3+\theta_n)(f_r-g_r)+(1-\theta_n)g_r}{2[\theta_n f_r+(1-\theta_n)g_r]^{5/2}}\,(\widehat{g}-g)^2\,\widehat{h}^{1/2}dx\,|\le \tau_2\int(\widehat{g}-g)^2dx,$$

$$|\int \frac{\frac{1}{4}\theta_n(1+\theta_n)(f_r-g_r)+(\theta_n-\frac{1}{2})g_r}{[\theta_n f_r+(1-\theta_n)g_r]^{5/2}}\,(\widehat{f}-f)(\widehat{g}-g)\,\widehat{h}^{1/2}dx\,|$$
$$\le\ \tau_3\int(\widehat{f}-f)^2dx+\tau_4\int(\widehat{g}-g)^2dx$$

for some positive constants τ_i $(i=1,2,3,4)$. Then from (2.25) and above three inequalities we have

$$\begin{aligned}
0 &= 2\int\frac{\partial\widetilde{h}_{\theta_n}^{1/2}(x)}{\partial\theta_n}\,\widehat{h}^{1/2}(x)dx\\
&= \int\frac{f-g}{[\theta_n f+(1-\theta_n)g]^{1/2}}\,\widehat{h}^{1/2}dx+\int\frac{\frac{\theta_n}{2}(f-g)+g}{[\theta_n f+(1-\theta_n)g]^{3/2}}\,(\widehat{f}-f)\,\widehat{h}^{1/2}dx\\
&\quad-\int\frac{\frac{1}{2}(1+\theta_n)(f-g)+g}{[\theta_n f+(1-\theta_n)g]^{3/2}}\,(\widehat{g}-g)\,\widehat{h}^{1/2}dx\\
&\quad+\alpha_n\int(\widehat{f}-f)^2dx+\beta_n\int(\widehat{g}-g)^2dx,
\end{aligned}$$

where $|\alpha_n|\le\tau_5$ and $|\beta_n|\le\tau_5$ for some positive constant $\tau_5>0$. Again since $h_t\ge C>0$ for all $t\in[0,1]$ and f and g are continuous on compact set W, (2.23) holds for $S_t=\frac{f-g}{[tf+(1-t)g]^{1/2}}$, $S_t=\frac{\frac{t}{2}(f-g)+g}{[tf+(1-t)g]^{3/2}}$ and $S_t=\frac{\frac{1}{2}(1+t)(f-g)+g}{[tf+(1-t)g]^{3/2}}$ on W. Applying (2.23) to preceding expressions, we obtain

$$\begin{aligned}
0 &= \Big\{\int\frac{f-g}{[\theta f+(1-\theta)g]^{1/2}}\,\widehat{h}^{1/2}dx+\int\frac{\frac{\theta}{2}(f-g)+g}{[\theta f+(1-\theta)g]^{3/2}}\,(\widehat{f}-f)\,\widehat{h}^{1/2}dx\\
&\qquad-\int\frac{\frac{1}{2}(1+\theta)(f-g)+g}{[\theta f+(1-\theta)g]^{3/2}}\,(\widehat{g}-g)\,\widehat{h}^{1/2}dx\Big\}\\
&\quad-(\theta_n-\theta)\Big\{\int\frac{(f-g)^2}{2[\theta f+(1-\theta)g]^{3/2}}\,\widehat{h}^{1/2}dx\\
&\qquad+\int\frac{\frac{\theta}{4}(f-g)^2+g(f-g)}{[\theta f+(1-\theta)g]^{5/2}}\,(\widehat{f}-f)\,\widehat{h}^{1/2}dx\\
&\qquad-\int\frac{\frac{1}{4}(3+\theta)(f-g)^2+g(f-g)}{[\theta f+(1-\theta)g]^{5/2}}\,(\widehat{g}-g)\,\widehat{h}^{1/2}dx\Big\}\\
&\quad+(\theta_n-\theta)\Big\{\int\mu_n\widehat{h}^{1/2}dx+\int\nu_n(\widehat{f}-f)\widehat{h}^{1/2}dx+\int\omega_n(\widehat{g}-g)\widehat{h}^{1/2}dx\Big\}\\
&\quad+\Big\{\alpha_n\int(\widehat{f}-f)^2dx+\beta_n\int(\widehat{g}-g)^2dx\Big\}
\end{aligned}$$

$$= (A_1 + A_2 - A_3) - (\theta_n - \theta)(B_1 + B_2 - B_3) + (\theta_n - \theta)(C_1 + C_2 + C_3) + (D_1 + D_2), \quad \text{say,}$$

where $\mu_n(x)$, $\nu_n(x)$ and $\omega_n(x)$ tend to zero in L_2 as $n \to \infty$. Then it follows that $\theta_n - \theta = [(B_1 + B_2 - B_3) - (C_1 + C_2 + C_3)]^{-1}(A_1 + A_2 - A_3 + D_1 + D_2)$. It is easy to show that $C_i \to 0$ $(i = 1, 2, 3)$, $B_i \to 0$ $(i = 2, 3)$ and that $|\int \frac{(f-g)^2}{2[\theta f + (1-\theta)g]^{3/2}} \cdot (\widehat{h}^{1/2} - h_\theta^{1/2})dx| \le C[\int (\widehat{h}^{1/2} - h_\theta^{1/2})^2 dx]^{1/2} \to 0$. Hence the result.

Finally we prove (2.23) and (2.24) hold for $S_t = \widetilde{h}_t^{1/2}$. By a Taylor expansion,

$$S_{t+\alpha} = S_t + \alpha \frac{\widehat{f} - \widehat{g}}{2S_t} + \frac{\alpha}{2}\Big[\frac{\widehat{f} - \widehat{g}}{S_{t+r\alpha}} - \frac{\widehat{f} - \widehat{g}}{S_t}\Big]$$

with $r = r(x) \in [0, 1]$. Note that

$$\begin{aligned}
\Big\|\frac{\widehat{f} - \widehat{g}}{S_{t+r\alpha}} - \frac{\widehat{f} - \widehat{g}}{S_t}\Big\| &\le \int \Big|\frac{(\widehat{f}(x) - \widehat{g}(x))^2}{\widetilde{h}_{t+r\alpha}(x)} - \frac{(\widehat{f}(x) - \widehat{g}(x))^2}{\widetilde{h}_t(x)}\Big| dx \\
&\le |\alpha| \int \frac{|\widehat{f}(x) - \widehat{g}(x)|^3}{\widetilde{h}_{t+r\alpha}(x)\widetilde{h}_t(x)} dx \\
&\le |\alpha| \int_{\widehat{f} > \widehat{g}} \frac{|\widehat{f}(x) - \widehat{g}(x)|^3}{t(t + r\alpha)(\widehat{f}(x) - \widehat{g}(x))^2} dx \\
&\quad + |\alpha| \int_{\widehat{f} \le \widehat{g}} \frac{|\widehat{f}(x) - \widehat{g}(x)|^3}{(1-t)(1-t-r\alpha)(\widehat{f}(x) - \widehat{g}(x))^2} dx \\
&\le |\alpha|\Big[\frac{1}{t(t - |\alpha|)} + \frac{1}{(1-t)(1-t-|\alpha|)}\Big] \int |\widehat{f}(x) - \widehat{g}(x)| dx \\
&\le 2|\alpha|\Big[\frac{1}{t(t - |\alpha|)} + \frac{1}{(1-t)(1-t-|\alpha|)}\Big] \\
&\to 0
\end{aligned}$$

as $\alpha \to 0$. Similarly one can prove that $\dot{S}_t \in L_2$ and (2.24) holds. □

Proof of Theorem 2.4.

Since the proof of Theorem 2.2 gives that $\sup_x |\widehat{f}(x) - f(x)| \xrightarrow{P} 0$, $\sup_x |\widehat{g}(x) - g(x)| \xrightarrow{P} 0$, $\| \widehat{h}^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$ and $\sup_{t \in [0,1]} \| \widetilde{h}_t^{1/2} - h_t^{1/2} \| \xrightarrow{P} 0$, (2.10) holds w.p.1 for some versions by Skorokhod's representation theorem. So it suffices to give the asymptotic distribution of $n^{1/2} \int \sigma_1(x)[\widehat{h}^{1/2}(x) - h_\theta^{1/2}(x)]dx + n^{1/2} \int \sigma_2(x)[\widehat{f}(x) - f(x)]\widehat{h}^{1/2}(x)dx - n^{1/2} \int \sigma_3(x)[\widehat{g}(x) - g(x)]\widehat{h}^{1/2}(x)dx + n^{1/2}\alpha_n \int [\widehat{f}(x) - f(x)]^2 dx +$

$n^{1/2}\beta_n \int [\widehat{g}(x) - g(x)]^2 dx$, where $\sigma_1 = \frac{f-g}{(\theta f+(1-\theta)g)^{1/2}}$, $\sigma_2 = \frac{\frac{\theta}{2}(f-g)+g}{(\theta f+(1-\theta)g)^{3/2}}$ and $\sigma_3 = \frac{\frac{1}{2}(1+\theta)(f-g)+g}{(\theta f+(1-\theta)g)^{3/2}}$. Denote H_{n_2} the empirical c.d.f. of $(Z_1, Z_2, \ldots, Z_{n_2})$ and H the c.d.f. of h_θ. Using the algebraic identity

$$b^{1/2} - a^{1/2} = \frac{b-a}{2a^{1/2}} - \frac{(b-a)^2}{2a^{1/2}(b^{1/2}+a^{1/2})^2}, \qquad b \geq 0,\ a > 0,$$

we have

$$n^{1/2} \int \sigma_1(x)[\widehat{h}^{1/2}(x) - h_\theta^{1/2}(x)]dx = n^{1/2} \int_W \sigma_1(x) \frac{\widehat{h}(x) - h_\theta(x)}{2h_\theta^{1/2}(x)} dx + R_n,$$

where, for $\delta = \min_{x \in W} h_\theta(x) > 0$,

$$\begin{aligned} |R_n| &\leq n_2^{1/2} \int |\sigma_1(x)| \frac{(\widehat{h}(x) - h_\theta(x))^2}{2h_\theta^{3/2}(x)} dx \\ &\leq \delta^{-3/2} \Big\{ n_2^{1/2} \int |\sigma_1(x)| [\widehat{h}(x) - \overline{E\widehat{h}}(x)]^2 dx \\ &\qquad + n_2^{1/2} \int |\sigma_1(x)| [\overline{E\widehat{h}}(x) - h_\theta(x)]^2 dx \Big\} \\ &= \delta^{-3/2}(W_{1n} + W_{2n}), \qquad \text{say,} \end{aligned} \tag{2.26}$$

where $\overline{E\widehat{h}}(x)$ is defined by

$$\overline{E\widehat{h}}(x) = (b_{n_2} S_{n_2})^{-1} \int K_2\Big(\frac{x-y}{b_{n_2} S_{n_2}}\Big) dH(y).$$

By denoting $B_{n_2}(x) = n_2^{1/2}[H_{n_2}(x) - H(x)]$, we have

$$\widehat{h}(x) - \overline{E\widehat{h}}(x) = n_2^{-1/2}(b_{n_2} S_{n_2})^{-1} \int K_2\Big(\frac{x-y}{b_{n_2} S_{n_2}}\Big) dB_{n_2}(y) = T_{1n}(x) + T_{2n}(x), \tag{2.27}$$

where

$$T_{1n}(x) = n_2^{-1/2}(b_{n_2} S_2)^{-1} \int K_2\Big(\frac{x-y}{b_{n_2} S_2}\Big) dB_{n_2}(y),$$

and

$$\begin{aligned} T_{2n}(x) &= -n_2{}^{-1/2} \int \int_{b_{n_2} S_2}^{b_{n_2} S_{n_2}} t^{-2} \Big[K_2\Big(\frac{x-y}{t}\Big) + \frac{x-y}{t} K_2^{(1)}\Big(\frac{x-y}{t}\Big) \Big] dt dB_{n_2}(y) \\ &= n_2{}^{-1/2} \int_{b_{n_2} S_2}^{b_{n_2} S_{n_2}} t^{-2} \int B_{n_2}(x - tz) \big[2K_2^{(1)}(z) + z K_2^{(2)}(z) \big] dz dt. \end{aligned}$$

By direct calculation,

$$
\begin{aligned}
E[T_{1n}^2(x)] &= E\Big[\frac{1}{n_2 b_{n_2} S_2}\sum_{i=1}^{n_2} K_2(\frac{x-Z_i}{b_{n_2}S_2}) - \frac{1}{b_{n_2}S_2}\int K_2(\frac{x-y}{b_{n_2}S_2})dD(y)\Big]^2 \\
&= \frac{1}{n_2} Var\Big[\frac{1}{b_{n_2}S_2}K_2(\frac{x-Z_1}{b_{n_2}S_2})\Big] \\
&\le \frac{1}{n_2 b_{n_2} S_2}\int K_2^2(z) h_\theta(x - b_{n_2}S_2 z)dz
\end{aligned}
$$

and

$$
\begin{aligned}
\sup_x T_{2n}(x) &= \sup_x n_2^{-1/2}\int_{b_{n_2}S_2}^{b_{n_2}S_{n_2}} t^{-2}dt \cdot O_p(1)\cdot O(1) \\
&= n_2^{-1/2} b_{n_2}(S_{n_2} - S_2) b_{n_2}^{-2} O_p(1) \\
&= O_p\big((n_2 b_{n_2})^{-1}\big)
\end{aligned}
\tag{2.28}
$$

since $n_2^{1/2}(S_{n_2} - S_2) = O_p(1)$. By CLT, $T_{1n}(x) = O_p((n_2 h_{n_2})^{-1/2})$. Then by (2.27), $\widehat{h}(x) - \overline{E\widehat{h}}(x) = O_p((n_2 b_{n_2})^{-1/2})$ and thus $W_{1n} = n_2^{1/2} O_p((n_2 b_{n_2})^{-1}) = O_p(n_2^{-1/2} b_{n_2}^{-1}) \xrightarrow{P} 0$ by (2.26) and $n_2^{1/2} b_{n_2} \to \infty$. Further since

$$
\begin{aligned}
\sup_x |\overline{E\widehat{h}}(x) - h_\theta(x)| &= \sup_x \Big|\int K_2(t)[h_\theta(x - b_{n_2}S_{n_2}t) - h_\theta(x)]dt\Big| \\
&\le 2^{-1} b_{n_2}^2 S_{n_2}^2 \sup_x |h_\theta^{(2)}(x)| \int x^2 K_2(x)dx,
\end{aligned}
\tag{2.29}
$$

$n_2^{1/2} b_{n_2}^2 \to 0$ and $\sqrt{n}(n_i/n - \rho_i) \to 0$, we have $W_{2n} \xrightarrow{P} 0$ as well. Consequently, $R_n \xrightarrow{P} 0$ as $n \to \infty$. Using a similar argument as for $R_n \xrightarrow{P} 0$, it can be shown that $n_0^{1/2}\int(\widehat{f}(x) - f(x))^2 dx \xrightarrow{P} 0$ and $n_1^{1/2}\int(\widehat{g}(x) - g(x))^2 dx \xrightarrow{P} 0$. So it suffices to give the asymptotic distribution of $n^{1/2}\int \sigma_1(x)\frac{\widehat{h}(x) - h_\theta(x)}{2h_\theta^{1/2}(x)}dx + n^{1/2}\int \sigma_2(x)[\widehat{f}(x) - f(x)]\widehat{h}^{1/2}(x)dx - n^{1/2}\int \sigma_3(x)[\widehat{g}(x) - g(x)]\widehat{h}^{1/2}(x)dx$. Since for large n,

$$
\int \sigma_2(x)[\widehat{f}(x) - f(x)][\widehat{h}^{1/2}(x) - h_\theta^{1/2}(x)]dx = o\Big(\int \sigma_1(x)[\widehat{h}^{1/2}(x) - h_\theta^{1/2}(x)]dx\Big)
$$

and

$$
\int \sigma_3(x)[\widehat{g}(x) - g(x)][\widehat{h}^{1/2}(x) - h_\theta^{1/2}(x)]dx = o\Big(\int \sigma_1(x)[\widehat{h}^{1/2}(x) - h_\theta^{1/2}(x)]dx\Big),
$$

we only need to find the asymptotic distribution of $n^{1/2}\int \sigma_1(x)\frac{\widehat{h}(x) - h_\theta(x)}{2h_\theta^{1/2}(x)}dx + n^{1/2}\int \sigma_2(x)[\widehat{f}(x) - f(x)]h_\theta^{1/2}(x)dx - n^{1/2}\int \sigma_3(x)[\widehat{g}(x) - g(x)]h_\theta^{1/2}(x)dx$. Let $A(x) = \sigma_1(x)/[2h_\theta^{1/2}(x)]$. Then by (2.27)

$$n_2^{1/2}\int \sigma_1(x)\frac{\widehat{h}(x)-h_\theta(x)}{2h_\theta^{1/2}(x)}dx = n_2^{1/2}\int A(x)T_{1n}(x)dx + n_2^{1/2}\int A(x)T_{2n}(x)dx + n_2^{1/2}\int A(x)[\overline{E\widehat{h}}(x)-h_\theta(x)]dx. \tag{2.30}$$

Also from (2.28), (2.29) and assumption C10 in the theorem, we have $n_2^{1/2}\int A(x)T_{2n}(x)dx \xrightarrow{P} 0$ and $n_2^{1/2}\int A(x)[\overline{E\widehat{h}}(x)-h_\theta(x)]dx \xrightarrow{P} 0$. The first term in (2.30) can be expressed as

$$\begin{aligned} n_2^{1/2}\int A(x)T_{1n}(x)dx &= \int A(x)\frac{1}{b_{n_2}S_2}\int K_2(\frac{x-y}{b_{n_2}S_2})dB_{n_2}(y)dx \\ &= \int \frac{1}{b_{n_2}S_2}\int A(x)K_2(\frac{x-y}{b_{n_2}S_2})dxdB_{n_2}(y) \\ &= \int\int A(y+b_{n_2}S_2z)K_2(z)dzdB_{n_2}(y). \end{aligned}$$

Thus straightforward calculations give

$$\begin{aligned} & E\big[n_2^{1/2}\int A(x)T_{1n}(x)dx - \int A(y)dB_{n_2}(y)\big]^2 \\ = \; & E\{\int K_2(z)\int [A(y+b_{n_2}S_2z)-A(y)]dB_{n_2}(y)dz\}^2 \\ \le \; & E\{\int K_2^2(z)[\int (A(y+b_{n_2}S_2z)-A(y))dB_{n_2}(y)]^2dz\} \\ = \; & \int K_2^2(z)E[\int (A(y+b_{n_2}S_2z)-A(y))dB_{n_2}(y)]^2dz \\ = \; & \int K_2^2(z)Var[A(Z_1+b_{n_2}S_2z)-A(Z_1)]dz \\ \le \; & \int K_2^2(z)\int [A(x+b_{n_2}S_2z)-A(x)]^2h_\theta(x)dxdz, \end{aligned}$$

which goes to zero as $n\to\infty$. Therefore, $n_2^{1/2}\int A(x)T_{1n}(x)dx - \int A(y)dB_{n_2}(y) \xrightarrow{P} 0$ as $n_2\to\infty$, and the asymptotic distribution of $n_2^{1/2}\int \sigma_1(x)\frac{\widehat{h}(x)-h_\theta(x)}{2h_\theta^{1/2}(x)}dx$ is the same as that of $\int \frac{\sigma_1(y)}{2h_\theta^{1/2}(y)}dB_{n_2}(y) = \sqrt{n_2}\big[\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{\sigma_1(Z_i)}{2h_\theta^{1/2}(Z_i)} - \int \frac{\sigma_1(x)}{2h_\theta^{1/2}(x)}h_\theta(x)dx\big]$. Applying a similar argument to $n_0^{1/2}\int \sigma_2(x)[\widehat{f}(x)-f(x)]h_\theta^{1/2}(x)dx$ and $n_1^{1/2}\int \sigma_3(x)\cdot[\widehat{g}(x)-g(x)]h_\theta^{1/2}(x)dx$, it is enough to find the asymptotic distribution of

$$\sqrt{n}\Big\{\Big[\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{\sigma_1(Z_i)}{2h_\theta^{1/2}(Z_i)}-\int\frac{\sigma_1(x)}{2h_\theta^{1/2}(x)}h_\theta(x)dx\Big]+\Big[\frac{1}{n_0}\sum_{i=1}^{n_0}\sigma_2(X_i)h_\theta^{1/2}(X_i)-$$
$$\int\sigma_2(x)h_\theta^{1/2}(x)f(x)dx\Big]-\Big[\frac{1}{n_1}\sum_{i=1}^{n_1}\sigma_3(Y_i)h_\theta^{1/2}(Y_i)-\int\sigma_3(x)h_\theta^{1/2}(x)g(x)dx\Big]\Big\},$$

i.e.,

$$\sqrt{n}\Big\{\Big[\frac{1}{n_0}\sum_{i=1}^{n_0}\sigma_2(X_i)h_\theta^{1/2}(X_i)-\frac{1}{n_1}\sum_{i=1}^{n_1}\sigma_3(Y_i)h_\theta^{1/2}(Y_i)+\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{\sigma_1(Z_i)}{2h_\theta^{1/2}(Z_i)}\Big]$$
$$-\Big[\int\frac{\sigma_1(x)}{2h_\theta^{1/2}(x)}h_\theta(x)dx+\int\sigma_2(x)h_\theta^{1/2}(x)f(x)dx-\int\sigma_3(x)h_\theta^{1/2}(x)g(x)dx\Big]\Big\}. \tag{2.31}$$

By Liapounov's theorem, (2.31) is asymptotically normally distributed with mean zero and asymptotic variance

$$\begin{aligned}
&\frac{1}{\rho_0}Var\big[\sigma_2(X_1)h_\theta^{1/2}(X_1)\big]+\frac{1}{\rho_1}Var\big[\sigma_3(Y_1)h_\theta^{1/2}(Y_1)\big]+\frac{1}{\rho_2}Var\big[\frac{\sigma_1(Z_1)}{2h_\theta^{1/2}(Z_1)}\big]\\
=\;&\frac{1}{4\rho_0}Var\big[\frac{g(X_1)}{h_\theta(X_1)}\big]+\frac{(1-\theta)^2}{4\rho_1\theta^2}Var\big[\frac{g(Y_1)}{h_\theta(Y_1)}\big]+\frac{1}{4\rho_2\theta^2}Var\big[\frac{g(Z_1)}{h_\theta(Z_1)}\big]\\
=\;&\frac{1}{4}\Big\{\frac{\theta^2}{\rho_0}Var\big[\frac{\partial\log h_\theta(X_1)}{\partial\theta}\big]+\frac{(1-\theta)^2}{\rho_1}Var\big[\frac{\partial\log h_\theta(Y_1)}{\partial\theta}\big]\\
&\qquad+\frac{1}{\rho_2}Var\big[\frac{\partial\log h_\theta(Z_1)}{\partial\theta}\big]\Big\}.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 2.5.

By simple calculations, for large n, the estimating equations (2.12)-(2.15) are equivalent to the following equations:

$$A(1-\theta_{nL})^2\widehat{\alpha}_{1l}^2+B_l(1-\theta_{nL})\widehat{\alpha}_{1l}+C_l=0,\ l=1,\ldots,L, \tag{2.32}$$

$$\widehat{\alpha}_{0l}=\frac{1}{\theta_{nL}}\times\frac{1}{n_2+\frac{n_0}{\theta_{nL}}}\times\Big[(n_{0l}+n_{1l}+n_{2l})-\big(n_2+\frac{n_1}{1-\theta_{nL}}\big)(1-\theta_{nL})\widehat{\alpha}_{1l}\Big],\ l=1,\ldots,L, \tag{2.33}$$

and

$$\sum_{l=1}^{L}\widehat{\alpha}_{1l}=1, \tag{2.34}$$

where $A=-\big(n_2+\frac{n_1}{1-\theta_{nL}}\big)\big(\frac{n_0}{\theta_{nL}}-\frac{n_1}{1-\theta_{nL}}\big)$, $B_l=(n_{1l}+n_{2l})\big(\frac{n_0}{\theta_{nL}}-\frac{n_1}{1-\theta_{nL}}\big)-(n_{0l}+n_{1l})\big(n_2+\frac{n_1}{1-\theta_{nL}}\big)$ and $C_l=n_{1l}(n_{0l}+n_{1l}+n_{2l})$. Note that $n_{il}/n_i\to\alpha_{il}$ and $n_i/n\to$

ρ_i as $n \to \infty$. If we let $n \to \infty$ and plug these limits into equations (2.32)-(2.34), then we observe that $\{\alpha_{0l}, \alpha_{1l}, \theta\}$ is a solution. This means that there exists a consistent sequence of roots to the likelihood equation. For notational convenience, we use $(\widehat{\alpha}_{01}, \ldots, \widehat{\alpha}_{0L}, \widehat{\alpha}_{11}, \ldots, \widehat{\alpha}_{1L}, \theta_{nL})$ to denote the consistent sequence, i.e., $(\widehat{\alpha}_{01}, \ldots, \widehat{\alpha}_{0L}, \widehat{\alpha}_{11}, \ldots, \widehat{\alpha}_{1L}, \theta_{nL}) \xrightarrow{P} (\alpha_{01}, \ldots, \alpha_{0L}, \alpha_{11}, \ldots, \alpha_{1L}, \theta)$. One can easily see that the consistent solution to (2.32) is

$$\widehat{\alpha}_{1l} = -\frac{1}{1-\theta_{nL}} \times \frac{B_l + \sqrt{B_l^2 - 4AC_l}}{2A}, \; l = 1, \ldots, L.$$

By substituting above equation into (2.34), we have that

$$\sum_{l=1}^{L} \sqrt{B_l^2 - 4AC_l} \;=\; 2A(\theta_{nL} - 1) - \sum_{l=1}^{L} B_l. \tag{2.35}$$

Note that (2.35) has θ_{nL} as the only unknown parameter, so we can use (2.35) to investigate the asymptotic properties of the MLE θ_{nL}. Applying Taylor expansion to the left hand side (l.h.s.) of (2.35), we have $\sum_{l=1}^{L} \sqrt{B_l^2 - 4AC_l} = \sum_{l=1}^{L} \big[- B_l + 2AC_l/B_l + 2A^2C_l^2/B_l^3 + o_p(A^2/n) \big]$, and then $\theta_{nL} - 1 = \sum_{l=1}^{L} C_l/B_l + A\sum_{l=1}^{L} C_l^2/B_l^3 + o_p(A/n)$. Applying Taylor expansion again, we obtain

$$\begin{aligned}
&\theta_{nL} - 1 \\
=\; & -\sum_{l=1}^{L} \frac{n_{1l}(n_{0l} + n_{1l} + n_{2l})}{(n_{0l} + n_{1l})(n_2 + \frac{n_1}{1-\theta})} \\
& -\sum_{l=1}^{L} \frac{n_{1l}(n_{0l} + n_{1l} + n_{2l})}{(n_{0l} + n_{1l})^2(n_2 + \frac{n_1}{1-\theta})^2}(n_{1l} + n_{2l})\Big(\frac{n_0}{\theta} - \frac{n_1}{1-\theta}\Big) \\
& +(\theta_{nL} - \theta) \times \Big\{ \sum_{l=1}^{L} \frac{n_{1l}(n_{0l} + n_{1l} + n_{2l})}{(n_{0l} + n_{1l})(n_2 + \frac{n_1}{1-\theta})^2} \frac{n_1}{(1-\theta)^2} \\
& \qquad + \sum_{l=1}^{L} \frac{n_{1l}(n_{0l} + n_{1l} + n_{2l})}{(n_{0l} + n_{1l})^2(n_2 + \frac{n_1}{1-\theta})^2}(n_{1l} + n_{2l})\Big(\frac{n_0}{\theta^2} + \frac{n_1}{(1-\theta)^2}\Big) \Big\} \\
& +(n_2 + \frac{n_1}{1-\theta})\Big(\frac{n_0}{\theta} - \frac{n_1}{1-\theta}\Big) \sum_{l=1}^{L} \frac{[n_{1l}(n_{0l} + n_{1l} + n_{2l})]^2}{[(n_{0l} + n_{1l})(n_2 + \frac{n_1}{1-\theta})]^3} \\
& +(\theta_{nL} - \theta)\Big[\frac{(2\theta - 1)n_0 n_1}{\theta^2(1-\theta)^2} - \frac{2n_1^2}{(1-\theta)^3} - \frac{n_0 n_2}{\theta^2} - \frac{n_1 n_2}{(1-\theta)^2} \Big] \times \\
& \qquad \sum_{l=1}^{L} \frac{[n_{1l}(n_{0l} + n_{1l} + n_{2l})]^2}{[(n_{0l} + n_{1l})(n_2 + \frac{n_1}{1-\theta})]^3} + o_p(\theta_{nL} - \theta) + o_p\Big(\frac{n_0}{n\theta} - \frac{n_1}{n(1-\theta)}\Big).
\end{aligned}$$

Since $\sqrt{n}(n_i/n - \rho_i) \to 0$, $i = 0, 1, 2$, further calculations then show that

$$\begin{aligned}
\theta_{nL} - \theta &= 1 - \theta - \sum_{l=1}^{L} \frac{n_{1l}(n_{0l} + n_{1l} + n_{2l})}{(n_{0l} + n_{1l})(n_2 + \frac{n_1}{1-\theta})} \\
&\quad -(\frac{n_0}{\theta} - \frac{n_1}{1-\theta})(n_2 + \frac{n_1}{1-\theta})^{-2} \sum_{l=1}^{L} \frac{n_{1l}(n_{0l} + n_{1l} + n_{2l}) n_{0l} n_{2l}}{(n_{0l} + n_{1l})^3} \\
&\quad +(\theta_{nL} - \theta) \times \Big\{ \sum_{l=1}^{L} \frac{n_{1l}(n_{0l} + n_{1l} + n_{2l})}{(n_{0l} + n_{1l})(n_2 + \frac{n_1}{1-\theta})^2} \frac{n_1}{(1-\theta)^2} \\
&\qquad +\sum_{l=1}^{L} \frac{n_{1l}(n_{0l} + n_{1l} + n_{2l})}{(n_{0l} + n_{1l})^2 (n_2 + \frac{n_1}{1-\theta})^2} (n_{1l} + n_{2l}) (\frac{n_0}{\theta^2} + \frac{n_1}{(1-\theta)^2}) \\
&\qquad +\Big[\frac{(2\theta - 1) n_0 n_1}{\theta^2 (1-\theta)^2} - \frac{2 n_1^2}{(1-\theta)^3} - \frac{n_0 n_2}{\theta^2} - \frac{n_1 n_2}{(1-\theta)^2} \Big] \times \\
&\qquad \sum_{l=1}^{L} \frac{[n_{1l}(n_{0l} + n_{1l} + n_{2l})]^2}{[(n_{0l} + n_{1l})(n_2 + \frac{n_1}{1-\theta})]^3} \Big\} \\
&\quad +o_p(\theta_{nL} - \theta) + o_p(\frac{n_0}{n\theta} - \frac{n_1}{n(1-\theta)}) \\
&= \frac{1}{(n_2 + \frac{n_1}{1-\theta})} \sum_{l=1}^{L} \frac{(1-\theta) n_{0l} n_{2l} - \theta n_{1l} n_{2l}}{n_{0l} + n_{1l}} \\
&\quad +\Big\{ \rho_0 + \rho_1 + \frac{\rho_2}{\theta} - \frac{\rho_2 (1-\theta)}{\theta} \sum_{l=1}^{L} \frac{\alpha_{1l}^2}{\theta \alpha_{0l} + (1-\theta)\alpha_{1l}} \Big\} (\theta_{nL} - \theta) \\
&\quad +o_p(n^{-1/2}) + o_p(\theta_{nL} - \theta),
\end{aligned}$$

equivalently,

$$\begin{aligned}
&\theta(1-\theta)\rho_2 \sum_{l=1}^{L} \frac{(\alpha_{0l} - \alpha_{1l})^2}{\theta \alpha_{0l} + (1-\theta)\alpha_{1l}} (\theta_{nL} - \theta) \\
&\qquad = \frac{1}{(n_2 + \frac{n_1}{1-\theta})} \sum_{l=1}^{L} \frac{(1-\theta) n_{0l} n_{2l} - \theta n_{1l} n_{2l}}{n_{0l} + n_{1l}} + o_p(n^{-1/2}) + o_p(\theta_{nL} - \theta),
\end{aligned}$$

or

$$\begin{aligned}
&\sqrt{n}(\theta_{nL} - \theta) \\
&= \Big(\theta(1-\theta) \sum_{l=1}^{L} \frac{(\alpha_{0l} - \alpha_{1l})^2}{\theta \alpha_{0l} + (1-\theta)\alpha_{1l}} \Big)^{-1} \times \sqrt{n} \Big(\sum_{l=1}^{L} \frac{n_{0l}}{n_{0l} + n_{1l}} \cdot \frac{n_{2l}}{n_2} - \theta \Big) + o_p(1) \\
&= \frac{1-\theta}{\theta} (\sum_{l=1}^{L} \frac{\alpha_{0l}^2}{\alpha_{2l}} - 1)^{-1} \times \sqrt{n} \Big(\sum_{l=1}^{L} \frac{n_{0l}}{n_{0l} + n_{1l}} \cdot \frac{n_{2l}}{n_2} - \theta \Big) + o_p(1),
\end{aligned} \tag{2.36}$$

since $n_{0l}/n_0 \xrightarrow{P} \alpha_{0l}$, $n_{1l}/n_1 \xrightarrow{P} \alpha_{1l}$ and $n_{2l}/n_2 \xrightarrow{P} \theta\alpha_{0l} + (1-\theta)\alpha_{1l}$, $l = 1, \ldots, L$.

We only need to find the asymptotic distribution of $\sqrt{n}\{\sum_{l=1}^{L}\frac{n_{0l}}{n_{0l}+n_{1l}}\cdot\frac{n_{2l}}{n_2}-\theta\}$. Observe that

$$\sum_{l=1}^{L}\frac{n_{0l}}{n_{0l}+n_{1l}}\cdot\frac{n_{2l}}{n_2}-\theta=\sum_{l=1}^{L}\Big(\frac{n_{0l}}{n_{0l}+n_{1l}}-\frac{\theta\alpha_{0l}}{\alpha_{2l}}\Big)\alpha_{2l}+\sum_{l=1}^{L}\frac{\theta\alpha_{0l}}{\alpha_{2l}}\Big(\frac{n_{2l}}{n_2}-\alpha_{2l}\Big)+r, \quad (2.37)$$

where $r=\sum_{l=1}^{L}\Big(\frac{n_{0l}}{n_{0l}+n_{1l}}-\frac{\theta\alpha_{0l}}{\alpha_{2l}}\Big)\Big(\frac{n_{2l}}{n_2}-\alpha_{2l}\Big)$. Since $\sqrt{n}(\frac{n_{il}}{n_i}-\alpha_{il})$ is asymptotically normal and $\sqrt{n}(n_i/n-\rho_i)\to 0$, we have $\sqrt{n}(\frac{n_{0l}}{n_{0l}+n_{1l}}-\frac{\theta\alpha_{0l}}{\alpha_{2l}})$ is asymptotically normal. Furthermore,

$$r=O_p(n^{-1})=o_p(n^{-1/2}). \quad (2.38)$$

We can write the first term on the r.h.s. of (2.37) as

$$\begin{aligned}
&\sum_{l=1}^{L}\Big(\frac{n_{0l}}{n_{0l}+n_{1l}}-\frac{\theta\alpha_{0l}}{\alpha_{2l}}\Big)\alpha_{2l}\\
=&\sum_{l=1}^{L}\frac{n_{0l}-(n_{0l}+n_{ll})\theta\alpha_{0l}/\alpha_{2l}}{n_0+n_1}+\sum_{l=1}^{L}\frac{n_{0l}}{n_0+n_1}\cdot\frac{(\frac{n_{0l}+n_{1l}}{n_0+n_1}-\alpha_{2l})^2}{\frac{n_{0l}+n_{1l}}{n_0+n_1}\alpha_{2l}}\\
&\qquad-\sum_{l=1}^{L}\frac{1}{\alpha_{2l}}\Big(\frac{n_{0l}}{n_0+n_1}-\theta\alpha_{0l}\Big)\Big(\frac{n_{0l}+n_{1l}}{n_0+n_1}-\alpha_{2l}\Big)\\
=&\sum_{l=1}^{L}\frac{n_{0l}-(n_{0l}+n_{ll})\theta\alpha_{0l}/\alpha_{2l}}{n_0+n_1}+o_p(n^{-1/2}),
\end{aligned} \quad (2.39)$$

with the last equality follows from the fact that $\sqrt{n}\big(\frac{n_{0l}}{n_0+n_1}-\theta\alpha_{0l}\big)$ and $\sqrt{n}\big(\frac{n_{0l}+n_{1l}}{n_0+n_1}-\alpha_{2l}\big)$ are asymptotically normal. From (2.37), (2.38) and (2.39), we obtain

$$\begin{aligned}
&\sum_{l=1}^{L}\frac{n_{0l}}{n_{0l}+n_{1l}}\cdot\frac{n_{2l}}{n_2}-\theta\\
=&\sum_{l=1}^{L}\frac{n_{0l}-(n_{0l}+n_{ll})\theta\alpha_{0l}/\alpha_{2l}}{n_0+n_1}+\sum_{l=1}^{L}\frac{\theta\alpha_{0l}}{\alpha_{2l}}\Big(\frac{n_{2l}}{n_2}-\alpha_{2l}\Big)+o_p(n^{-1/2})\\
=&\sum_{l=1}^{L}\frac{(1-\theta)\alpha_{1l}}{\alpha_{2l}}\Big(\frac{n_{0l}}{n_0+n_1}-\theta\alpha_{0l}\Big)-\sum_{l=1}^{L}\frac{\theta\alpha_{0l}}{\alpha_{2l}}\Big(\frac{n_{1l}}{n_0+n_1}-(1-\theta)\alpha_{1l}\Big)\\
&\qquad+\sum_{l=1}^{L}\frac{\theta\alpha_{0l}}{\alpha_{2l}}\Big(\frac{n_{2l}}{n_2}-\alpha_{2l}\Big)+o_p(n^{-1/2}),
\end{aligned}$$

and thus by the CLT, $\sqrt{n}\big(\sum_{l=1}^{L}\frac{n_{0l}}{n_{0l}+n_{1l}}\cdot\frac{n_{2l}}{n_2}-\theta\big)$ is asymptotically normal with mean zero and variance given by

$$
\begin{aligned}
&\frac{\theta^2(1-\theta)^2}{\rho_0}\sum_{l,m=1}^{L}\frac{\alpha_{1l}\alpha_{1m}}{\alpha_{2l}\alpha_{2m}}\alpha_{0l}(\delta_{lm}-\alpha_{0m})+\\
&\quad\frac{\theta^2(1-\theta)^2}{\rho_1}\sum_{l,m=1}^{L}\frac{\alpha_{0l}\alpha_{0m}}{\alpha_{2l}\alpha_{2m}}\alpha_{1l}(\delta_{lm}-\alpha_{1m})+\frac{\theta^2}{\rho_2}\sum_{l,m=1}^{L}\frac{\alpha_{0l}\alpha_{0m}}{\alpha_{2l}\alpha_{2m}}\alpha_{2l}(\delta_{lm}-\alpha_{2m})\\
=\;&\frac{\theta^2(1-\theta)^2}{\rho_0}\Big[\sum_{l=1}^{L}\frac{\alpha_{1l}^2}{\alpha_{2l}^2}\alpha_{0l}-\Big(\sum_{l=1}^{L}\frac{\alpha_{1l}}{\alpha_{2l}}\alpha_{0l}\Big)^2\Big]+\\
&\quad\frac{\theta^2(1-\theta)^2}{\rho_1}\Big[\sum_{l=1}^{L}\frac{\alpha_{0l}^2}{\alpha_{2l}^2}\alpha_{1l}-\Big(\sum_{l=1}^{L}\frac{\alpha_{0l}}{\alpha_{2l}}\alpha_{1l}\Big)^2\Big]+\frac{\theta^2}{\rho_2}\Big[\sum_{l=1}^{L}\frac{\alpha_{0l}^2}{\alpha_{2l}}-1\Big],
\end{aligned}
\tag{2.40}
$$

where δ_{lm} denotes the Kronecker delta. Thus, by (2.36) and (2.40), $\sqrt{n}(\theta_{nL}-\theta)$ is asymptotically normal with mean 0 and variance Δ_L defined in (2.16). □

Proof of Theorem 2.6.

Suppose for some $f \neq g$ and $\theta \in (0,1)$, $\delta = \Delta(\theta, f, g, h_\theta) - V(\theta, f, g, h_\theta) > 0$. In view of the continuity of both V and Δ, we can choose step function densities $\bar{f}$, $\bar{g}$ and $\bar{M}_\theta = \theta\bar{f} + (1-\theta)\bar{g}$ such that $\Delta(\theta, \bar{f}, \bar{g}, \bar{M}_\theta) - V(\theta, \bar{f}, \bar{g}, \bar{M}_\theta) > \delta/2$. Decompose the real line R into regions $\{R_l\}$ in such a way that $\bar{f}$, $\bar{g}$ and $\bar{M}_\theta$ assume, respectively, a constant value on R_l for each l. When $(\bar{f}, \bar{g}, \bar{M}_\theta)$ is the true sequence of densities and θ_{nL} denotes the MLE of θ, by Theorem 2.5, $n^{1/2}(\theta_{nL}-\theta)$ is asymptotically normal with mean zero and variance $\Delta(\theta, \bar{f}, \bar{g}, \bar{M}_\theta)$. Since θ_{nL} is the MLE, $\Delta(\theta, \bar{f}, \bar{g}, \bar{M}_\theta) \leq \lim_{n\to\infty} nVar(\widetilde{\theta} - \theta) = V(\theta, \bar{f}, \bar{g}, \bar{M}_\theta)$. This contradicts $\Delta(\theta, \bar{f}, \bar{g}, \bar{M}_\theta) - V(\theta, \bar{f}, \bar{g}, \bar{M}_\theta) > \delta/2$ and so our assumption at the very beginning of the proof is incorrect and theorem must be true. □

Proof of Corollary 2.1.

Note that $\Delta_0 = Var[\frac{g(X_1)}{h_\theta(X_1)}] = \theta^2 Var[\frac{\partial \log h_\theta(X_1)}{\partial\theta}]$, $\Delta_1 = Var[\frac{f(Y_1)}{h_\theta(Y_1)}] = (1-\theta)^2 Var[\frac{\partial \log h_\theta(Y_1)}{\partial\theta}]$ and $\Delta_2 = Var[\frac{f(Z_1)}{h_\theta(Z_1)}] = (1-\theta)^2 Var[\frac{\partial \log h_\theta(Z_1)}{\partial\theta}]$, and therefore the asymptotic variance σ^2 derived in Theorem 2.4 of the proposed MHD estimator θ_n of (2.9) achieves the lower bound of (2.17) when $\rho_0/\rho_1 = \theta/(1-\theta)$. □

Proof of Remark 2.2.

As a result of Theorem 2.1.3 in Rao (1983), $\sup_x |\widehat{f}(x) - f(x)| \to 0$, $\sup_x |\widehat{g}(x) - g(x)| \to 0$, $\sup_x |\widehat{h}(x) - h_\theta(x)| \to 0$, and $\sup_{t\in[0,1]} \sup_x |\widetilde{h}_t(x) - h_t(x)| \leq \sup_x |\widehat{f}(x) - f(x)| + \sup_x |\widehat{g}(x) - g(x)| \to 0$ w.p.1 as $n \to \infty$.

By Devroye and Györfi (1995), $\int |\widehat{f}(x) - f(x)|dx \to 0$ w.p.1. Since $\int [\widehat{f}^{1/2}(x)$

$-f^{1/2}(x)\big]^2 dx \leq \int |\widehat{f}(x) - f(x)|dx$, we have $\| \widehat{f}^{1/2} - f^{1/2} \| \to 0$ w.p.1. Similarly, we have $\| \widehat{g}^{1/2} - g^{1/2} \| \to 0$ and $\| \widehat{h}^{1/2} - h_\theta^{1/2} \| \to 0$ w.p.1, and furthermore $\sup_{t\in[0,1]} \| \widetilde{h}_t^{1/2} - h_t^{1/2} \| \leq \sup_{t\in[0,1]} \big[\int |\widetilde{h}_t(x) - h_t(x)|dx \big]^{1/2} \leq \big[\int |\widehat{f}(x) - f(x)|dx + \int |\widehat{g}(x) - g(x)|dx \big]^{1/2} \to 0$ w.p.1. By Theorem 2.1, $\theta_n - \theta_0 = \widehat{T}(\widehat{h}) - T_0(M_{\theta_0}) \to 0$ w.p.1. □

CHAPTER THREE: MHD ESTIMATION IN A TWO-SAMPLE SEMIPARAMETRIC MODEL

3.1 Introduction

Semiparametric models have continued to receive increasing attention over the years from both practical and theoretical point of views due in large part to the fact that semiparametric models arise frequently in many areas, primarily in biostatistics and econometrics. The well-known semiparametric models include the Cox proportional hazard model in survival analysis, econometric index models, regression models and errors-in-variables models, among many others. More examples and theory on semiparametric models can be found in the monographs of Bickel et al. (1993), Van der Vaart (1998) and in the review articles of Bickel and Kwon (2001) and Forrester et al. (2003).

In this chapter, we consider the following two-sample semiparametric model: Let $X_1, \ldots, X_n$ be a random sample from a population with distribution function G and density function g. Independently of the X_i's, let $Z_1, \ldots, Z_m$ be another random sample from a population with distribution function H and density function h. The two unknown density functions g and h are linked by an "exponential tilt" $\exp[\alpha + r(x)\beta]$. Thus, we have

$$\begin{array}{ll} X_1, \ldots, X_n & \overset{\text{i.i.d.}}{\sim} \; g(x) \\ Z_1, \ldots, Z_m & \overset{\text{i.i.d.}}{\sim} \; g(x)\exp[\alpha + r(x)\beta], \end{array} \tag{3.1}$$

where $r(x) = (r_1(x), \ldots, r_p(x))$ is a $1 \times p$ vector of functions of x, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ parameter vector, and α is a normalizing parameter that makes $g(x)\exp[\alpha + r(x)\beta]$ integrate to 1. In most applications $r(x) = x$ or $r(x) = (x, x^2)$. We are concerned with estimation of parameters α and β.

For $r(x) = x$, model (3.1) encompasses many common distributions, including two exponential distributions with different means and two normal distributions with common variance but different means. Furthermore, model (3.1) with $r(x) = x$ or $r(x) = (x, x^2)$ has wide applications in the logistic discriminant analysis (Anderson, 1972, 1979) and in case-control studies (Prentice and Pyke, 1979; Breslow and Day, 1980). Suppose Y is a binary response variable and X is the associated covariate, then the (prospective) logistic regression model is of the form

$$P(Y=1|X=x)=\frac{\exp[\alpha^*+x\beta]}{1+\exp[\alpha^*+x\beta]}\,, \tag{3.2}$$

where α^* and β are parameters and the marginal distribution of X is not specified. In case-control studies, data are collected retrospectively in the sense that for samples of subjects having $Y=1$ ('case') and having $Y=0$ ('control'), the value x of X is observed. More specifically, suppose $X_1,\ldots,X_n$ is a random sample from $F(x|Y=0)$ and, independently of the X_i's, suppose $Z_1,\ldots,Z_m$ is a random sample from $F(x|Y=1)$. If $\pi=P(Y=1)=1-P(Y=0)$ and $f(x|Y=i)$ is the conditional density of X given $Y=i$, $i=0,1$, then it follows from (3.2) and Bayes rule that model (3.1) is satisfied with $g(x)=f(x|Y=0)$, $h(x)=f(x|Y=1)$, $\alpha=\alpha^*+\log[(1-\pi)/\pi]$ and $r(x)=x$.

Model (3.1) with $r(x)=(x,x^2)$ also coincides with exponential family of densities considered in Efron and Tibshirani (1996) in the case of two-sample problems. Moreover, model (3.1) can also be viewed as a biased sampling model with weight function $\exp[\alpha+r(x)\beta]$ depending on the unknown parameters α and β.

Vardi (1982, 1985), Gill et al. (1988) and Qin (1993) discussed estimating distribution functions in biased sampling models with known weight functions. Gilbert et al. (1998) have employed model (3.1) with $r(x)=(x,x^2)$ to analyze HIV vaccine trial data for assessing differential vaccine protection against human immunodeficiency virus types. Qin and Zhang (1997) considered a goodness-of-fit test for logistic regression model (3.2) based on case-control data by employing the maximum semiparametric likelihood estimator of G to test the validity of model (3.1) with $r(x)=x$. Zhang (2000) estimated quantiles of G under model (3.1). In this chapter, however, we are interested in the problem of estimating the parameters α and β when $g(x)$ is unknown. Note that since the form of $g(x)$ is not specified, statistical inference based on model (3.1) with unknown g would be more robust than those based on a full parametric model in which the form of $g(x)$ is known. Note that the test of equality of G and H can be regarded as a special case of model (3.1) with $\alpha=\beta=0$. The results of this chapter will help to solve this kind of problem.

In this chapter, we propose MHD estimation for the two-sample semiparametric model (3.1). This chapter is organized as follows. In Section 3.2, we investigate MHD estimators of the parameters $\theta=(\alpha,\beta)$ and study their existence and strong consistency. In Section 3.3, we derive the asymptotic distribution of the proposed estimators. Section 3.4 contains a simulation study where efficiency and robustness properties of the proposed MHD estimator are studied using a Monte Carlo study. A real data set is analyzed in Section 3.5. The detailed proof of asymptotic normality of the estimators (Theorem 3.4) is deferred to Section 3.6.

3.2 MHD Estimators of Regression Parameters

Define $\theta = (\alpha, \beta^T)^T$, where α and β are as in (3.1). Then the model (3.1) can be written as

$$\begin{array}{ll} X_1, \ldots, X_n & \overset{\text{i.i.d.}}{\sim} \; g(x) \\ Z_1, \ldots, Z_m & \overset{\text{i.i.d.}}{\sim} \; h_\theta(x), \end{array} \tag{3.3}$$

where $h_\theta(x) = g(x)\exp[(1, r(x))\theta]$, $r(x) = (r_1(x), \ldots, r_p(x))$ is a $1 \times p$ vector of continuous functions of x on $\mathbb{R}$, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ parameter vector and α is a normalizing parameter that makes $h_\theta(x)$ integrate to 1. We assume here and in what follows that $\theta \in \Theta$ and Θ is a compact subset of $\mathbb{R}^{p+1}$.

We first define following kernel density estimators of g and h_θ, respectively, based on data $X_1, \ldots, X_n$ and $Z_1, \ldots, Z_m$ of (3.3):

$$g_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K_0\Big(\frac{x - X_i}{b_n}\Big), \tag{3.4}$$

$$h_m(x) = \frac{1}{mb_m} \sum_{j=1}^{m} K_1\Big(\frac{x - Z_j}{b_m}\Big), \tag{3.5}$$

where K_0 and K_1 are symmetric density functions, bandwidths b_n and b_m are positive constants such that $b_n \to 0$ as $n \to \infty$ and $b_m \to 0$ as $m \to \infty$. We can also use adaptive kernel density estimators, which use $S_n b_n$ instead of b_n with S_n being a robust scale statistic. Here we use non-adaptive kernel density estimators (3.4) and (3.5) for simplicity. The results can be easily extended for adaptive kernel density estimators with some additional conditions on S_n.

Let $\mathcal{H}$ be the set of all densities w.r.t. Lebesgue measure on the real line. For $\phi \in \mathcal{H}$, the MHD functional $T_0(\phi)$ is defined as

$$T_0(\phi) = T\big(\{h_\theta\}_{\theta\in\Theta}, \phi\big) = \arg\min_{\theta\in\Theta} \parallel h_\theta^{1/2} - \phi^{1/2} \parallel . \tag{3.6}$$

If the family $\{h_\theta\}_{\theta\in\Theta}$ is identifiable, then the functional T_0 is Fisher consistent, i.e., $T_0(h_\theta) = \theta$ for any $\theta \in \Theta$. Since h_m defined by (3.5) is an estimator of h_θ, the MHD estimator of θ will be $T_0(h_m)$. However, this estimator is not available in reality since g and hence h_θ in (3.6) are unknown. Naturally, one can use the estimator g_n of g and then apply the plug-in rule to construct a parametric model, i.e., one can replace h_θ with

$$\widehat{h}_\theta(x) = \exp[(1, r(x))\,\theta] g_n(x). \tag{3.7}$$

Note that $\widehat{h}_\theta$ is a parametric density function with the unknown parameter being θ. Let $N = n + m$ be the total sample size here and in what follows. Now our proposed MHD estimator of θ is defined as

$$\theta_N = \widehat{T}(h_m) = T\big(\{\widehat{h}_\theta\}_{\theta\in\Theta}, h_m\big) = \arg\min_{\theta\in\Theta} \parallel \widehat{h}_\theta^{1/2} - h_m^{1/2} \parallel, \tag{3.8}$$

where h_m and $\widehat{h}_\theta$ are given by (3.5) and (3.7), respectively. That is, θ_N is the minimizer of the Hellinger distance between the parametric density $\widehat{h}_\theta$ and non-parametric density estimator h_m. This approach is in line with Beran's (1977) original mechanism of obtaining MHD estimators. Thus, we would expect θ_N to have good robustness and asymptotic efficiency properties. Since $\widehat{T}(h_m)$ may be multiple valued, we shall use the notation $\widehat{T}(h_m)$ to indicate any one of the possible values chosen arbitrarily. We are interested in both the asymptotic properties and the local properties of θ_N. So we let $n \to \infty$ and $m \to \infty$ as $N \to \infty$.

Note that in (3.8) we are not minimizing the Hellinger distance over a subset of Θ including those θ's which make $\widehat{h}_\theta$ densities, i.e., over $\{\theta \in \Theta : \int \widehat{h}_\theta(x)dx = 1\}$. The reason being that even for $\theta \in \Theta$ such that $\widehat{h}_\theta$ is not a density, it could make h_θ a density. The true parameter value θ may not make $\widehat{h}_\theta$ a density, but it is not reasonable to exclude θ as an estimate θ_N of itself defined by (3.8). Nevertheless, the definition of θ_N is equivalent to a minimization over a smaller parameter space, as shown in the following Lemma 3.1.

Lemma 3.1. *(i) Suppose that for any $\theta = (\alpha, \beta^T)^T \in \Theta$ there exists $\theta' = (\alpha', \beta^T)^T \in \Theta$ such that $\int \exp[\alpha' + r(x)\beta]g(x)dx = 1$. Let $\Theta_0 = \{\theta \in \Theta : \int \exp[(1, r(x))\theta]g(x)dx \le 1\}$. Then for any $\phi \in \mathcal{H}$,*

$$T_0(\phi) = \arg\min_{\theta\in\Theta} \parallel h_\theta^{1/2} - \phi^{1/2} \parallel = \arg\min_{\theta\in\Theta_0} \parallel h_\theta^{1/2} - \phi^{1/2} \parallel .$$

(ii) Suppose that for any $\theta = (\alpha, \beta^T)^T \in \Theta$ there exists $\theta' = (\alpha', \beta^T)^T \in \Theta$ such that $\int \exp[\alpha' + r(x)\beta]g_n(x)dx = 1$. Let $\Theta_n = \{\theta \in \Theta : \int \exp[(1, r(x))\theta]\cdot g_n(x)dx \le 1\}$. Then for any $\phi \in \mathcal{H}$,

$$\widehat{T}(\phi) = \arg\min_{\theta\in\Theta} \parallel \widehat{h}_\theta^{1/2} - \phi^{1/2} \parallel = \arg\min_{\theta\in\Theta_n} \parallel \widehat{h}_\theta^{1/2} - \phi^{1/2} \parallel,$$

where $\widehat{h}_\theta$ is defined by (3.7).

Proof. (i) For $\theta \in \Theta$, let $c = \int h_\theta(x)dx = \int \exp[\alpha + r(x)\beta]g(x)dx$ and suppose that $c > 1$. Obviously $\int \exp[(\alpha - \log c) + r(x)\beta]g(x)dx = 1$, and thus $\theta_1 = (\alpha - \log c, \beta^T)^T \in \Theta_0$. Note that

$$\begin{aligned} & \int \big(h_\theta^{1/2}(x) - \phi^{1/2}(x)\big)^2 dx - \int \big(h_{\theta_1}^{1/2}(x) - \phi^{1/2}(x)\big)^2 dx \\ = & \int \big(h_\theta(x) - h_{\theta_1}(x)\big) - 2\big[h_\theta^{1/2}(x) - h_{\theta_1}^{1/2}(x)\big]\phi^{1/2}(x)dx \end{aligned}$$

$$
\begin{aligned}
&= (c-1) - 2(\sqrt{c}-1)\int h_{\theta_1}^{1/2}(x)\phi^{1/2}(x)dx \\
&\geq (c-1) - 2(\sqrt{c}-1) \\
&= (\sqrt{c}-1)^2,
\end{aligned}
$$

i.e., $\int (h_\theta^{1/2}(x) - \phi^{1/2}(x))^2 dx > \int (h_{\theta_1}^{1/2}(x) - \phi^{1/2}(x))^2 dx$.

(ii) Proof is similar to that of (i). □

Remark 3.1. If $\int \exp[(1, r(x))\theta] g(x)dx < \infty$ for any $\theta \in \Theta$ and the parameter space Θ is of the form $\Theta = \mathbb{R} \times \Theta_p$ with $\mathbb{R}$ and Θ_p denote the parameter spaces for α and β, then the condition in Lemma 3.1 (i) holds. Furthermore, if g_n is defined by (3.4) with kernel K_0 compactly supported, then the condition in Lemma 3.1 (ii) also holds. Moreover, if $C < \sup_{\beta \in \Theta_p} \int \exp[r(x)\beta] g(x)dx < \infty$ (or $C < \sup_{\beta \in \Theta_p} \int \exp[r(x)\beta] g_n(x)dx < \infty$) for some constant $C > 0$, then the condition in Lemma 3.1 (i) (or (ii)) holds with $\Theta = [-M, M] \times \Theta_p$ for some finite positive value M.

We now discuss asymptotic properties of the proposed MHD estimator θ_N. First, we give some results on the functional $T(\cdot, \cdot)$ related to the existence, consistency and asymptotic uniqueness of the MHD estimator of θ. The next condition and lemma will be used to prove above properties.

(D1) There exists an ε-neighborhood $B(\theta, \varepsilon)$ of θ such that $h_t - h_\theta$ is bounded by an integrable function for any $t \in B(\theta, \varepsilon)$.

Lemma 3.2. *If (D1) holds for $\theta \in \Theta$, then $d(t) = \| h_t^{1/2} - \phi^{1/2} \|$ is continuous at point $t = \theta$ for any $\phi \in \mathcal{H}$.*

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

$$
| d(\theta_k) - d(\theta) | \leq \| h_{\theta_k}^{1/2} - h_\theta^{1/2} \| \leq \left[\int | h_{\theta_k}(x) - h_\theta(x) | dx \right]^{1/2}. \tag{3.9}
$$

By assumption (D1), $| h_{\theta_k} - h_\theta |$ is bounded by an integrable function, and therefore by the Dominated Convergence Theorem 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 $t = \theta$. □

Remark 3.2. Condition (D1) holds for many families including normal distributions. Suppose that $g(x)$ and $h(x)$ denotes density functions of the normal distribution $N(0, 1)$ and $N(\mu, 1)$, respectively. It is easy to see that $h(x) = h_\theta(x) = \exp[(1, r(x))\theta] g(x)$, where $r(x) = x$ and $\theta = (\alpha, \beta) = (-\frac{\mu^2}{2}, \mu)$. Obviously condition (D1) holds for this example.

Theorem 3.1. *Suppose that T_0 and $\widehat{T}$ are defined by (3.6) and (3.8), respectively, and (D1) holds for all $\theta \in \Theta$. Then we have following results.*

(i) For every $\phi \in \mathcal{H}$, there exists $\widehat{T}(\phi) \in \Theta$ satisfying (3.8) with $\widehat{h}_\theta$ and g_n defined by (3.7) and (3.4), respectively, and the kernel K_0 in (3.4) compactly supported. For every $\phi \in \mathcal{H}$, there exists $T_0(\phi) \in \Theta$ satisfying (3.6).

(ii) Suppose that $n \to \infty$ and $m \to \infty$ as $N \to \infty$ and $\theta_0 = T_0(\phi)$ is unique. Then $\theta_N = \widehat{T}(\phi_m) \to \theta_0$ as $N \to \infty$ for any density sequences $\{\phi_m\}_{m\in\mathbb{N}}$ and $\{\widehat{h}_\theta\}_{n\in\mathbb{N},\theta\in\Theta}$ such that $\| \phi_m^{1/2} - \phi^{1/2} \| \to 0$ and $\sup_{\theta\in\Theta} \| \widehat{h}_\theta^{1/2} - h_\theta^{1/2} \| \to 0$ as $N \to \infty$.

(iii) If $\{h_\theta\}_{\theta\in\Theta}$ is identifiable, then $T_0(h_{\theta_0}) = \theta_0$ uniquely for any $\theta_0 \in \Theta$.

Proof. (i) Let $d_n(t) = \| \widehat{h}_t^{1/2} - \phi^{1/2} \|$. 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 (3.9), we have

$$| d_n(t_k) - d_n(t) | \leq \Big[\int | \exp[(1, r(x))t_k] - \exp[(1, r(x))t] | \, g_n(x) \, dx \Big]^{1/2}.$$

Since g_n is compactly supported, we have by the Dominated Convergence Theorem that $d_n(t_k) \to d_n(t)$ as $k \to \infty$, i.e., $d_n(t)$ is continuous and achieves a minimum over $t \in \Theta$.

Let $d(t) = \| h_t^{1/2} - \phi^{1/2} \|$. By Lemma 3.2, $d(t)$ is continuous in t and therefore achieves a minimum over $t \in \Theta$.

(ii) Suppose $\| \phi_m^{1/2} - \phi^{1/2} \| \to 0$ and $\sup_{\theta\in\Theta} \| \widehat{h}_\theta^{1/2} - h_\theta^{1/2} \| \to 0$ as $N \to \infty$. Put $d_N(\theta) = \| \widehat{h}_\theta^{1/2}(x) - \phi_m^{1/2}(x) \|$ and $d(\theta) = \| h_\theta^{1/2}(x) - \phi^{1/2}(x) \|$. By Minkowski's inequality,

$$\begin{aligned} & | d_N(\theta) - d(\theta) | \\ \leq \; & \Big\{ \int [\widehat{h}_\theta^{1/2}(x) - \phi_m^{1/2}(x) - h_\theta^{1/2}(x) + \phi^{1/2}(x)]^2 dx \Big\}^{1/2} \\ \leq \; & \Big\{ 2 \int [\widehat{h}_\theta^{1/2}(x) - h_\theta^{1/2}(x)]^2 dx + 2 \int [\phi_m^{1/2}(x) - \phi^{1/2}(x)]^2 dx \Big\}^{1/2}, \end{aligned}$$

and consequently $\sup_{\theta\in\Theta} | d_N(\theta) - d(\theta) | \to 0$ as $N \to \infty$. Therefore, as $N \to \infty$, $d_N(\theta_0) \to d(\theta_0)$ and $d_N(\theta_N) - d(\theta_N) \to 0$. If $\theta_N \not\to \theta_0$, then there exists a subsequence $\{\theta_{N_i}\} \subseteq \{\theta_N\}$ such that $\theta_{N_i} \to \theta' \neq \theta_0$. Since Θ is compact, $\theta' \in \Theta$. Lemma 3.2 yields that $d(\theta_{N_i}) \to d(\theta')$. From above results we obtain $d_{N_i}(\theta_{N_i}) - d_{N_i}(\theta_0) \to d(\theta') - d(\theta_0)$. By the definition of θ_{N_i}, $d_{N_i}(\theta_{N_i}) - d_{N_i}(\theta_0) \leq 0$. Hence, $d(\theta') - d(\theta_0) \leq 0$. But by the definition and uniqueness of θ_0, $d(\theta') > d(\theta_0)$. This is a contradiction. Therefore, $\theta_N \to \theta_0$.

(iii) Since $\{h_\theta\}_{\theta\in\Theta}$ is identifiable, we now have $T_0(h_{\theta_0}) = \theta_0$ uniquely for any $\theta_0 \in \Theta$. □

Remark 3.3. If $\big(1, r(x)\big)$ are linearly independent, then $\{h_\theta\}_{\theta\in\Theta}$ is identifiable. To see this clearly, note that for $h_{\theta_1} = h_{\theta_2}$, we have $(1, r(x))(\theta_1 - \theta_2) = 0$, and then $\theta_1 = \theta_2$ when $(1, r(x))$ are linearly independent. Therefore, $\{h_\theta\}_{\theta\in\Theta}$ is identifiable for any continuous density function g.

With further assumptions on bandwidths and kernels in (3.4) and (3.5), the consistency of the MHD estimator of θ follows from the continuity of functional T in the Hellinger topology. This result is given next. First, we state conditions (D2), (D3) and (D4):

(D2) g and K_0 in (3.3) and (3.4), respectively, have compact supports.

(D3) $\sup_{\theta\in\Theta} \sup_x (1, r(x))\theta < +\infty$.

(D4) g in (3.3) has infinite support, K_0 in (3.4) is a bounded symmetric density with support $[-a_0, a_0]$, $0 < a_0 < \infty$, and there exists a sequence $\{\alpha_n\}$ of positive numbers such that as $n \to \infty$, $\alpha_n \to \infty$ and

$$\sup_{\theta\in\Theta} \int I_{\{|x|>\alpha_n\}} h_\theta(x)dx \to 0, \tag{3.10}$$

$$b_n^2 \sup_{\theta\in\Theta} \int I_{\{|x|>\alpha_n\}} h_\theta(x) \sup_{|t|\le a_0} \frac{|g^{(2)}(x + tb_n)|}{g(x)} dx \to 0, \tag{3.11}$$

$$n^{-1}b_n^{-1} \sup_{\theta\in\Theta} \int I_{\{|x|\le\alpha_n\}} h_\theta(x) \sup_{|t|\le a_0} \frac{g(x + tb_n)}{g^2(x)} dx \to 0, \tag{3.12}$$

$$b_n^4 \sup_{\theta\in\Theta} \int I_{\{|x|\le\alpha_n\}} h_\theta(x) \sup_{|t|\le a_0} \Big[\frac{g^{(2)}(x + tb_n)}{g(x)}\Big]^2 dx \to 0, \tag{3.13}$$

where $g^{(k)}$ denotes the k-th derivative of g and I_A denotes the indicator function of a set A.

Lemma 3.3. *If (D4) holds, then as $n \to \infty$,*

$$\sup_{\theta\in\Theta} \int \exp[(1, r(x))\theta][g_n^{1/2}(x) - g^{1/2}(x)]^2 dx \xrightarrow{P} 0.$$

Proof. By continuity of the function in θ and the compactness of Θ, there exists $\theta_n \in \Theta$ which maximizes $\int \exp[(1, r(x))\theta][g_n^{1/2}(x) - g^{1/2}(x)]^2 dx$. By (3.10), (3.11) and a Taylor expansion, one has

$$\begin{aligned} & E \Big| \int I_{\{|x|>\alpha_n\}} \exp[(1, r(x))\theta] g_n(x)dx \Big| \\ = & \int\int I_{\{|x|>\alpha_n\}} \exp[(1, r(x))\theta] \frac{1}{b_n} K_0\Big(\frac{y - x}{b_n}\Big) g(y)dy\, dx \end{aligned}$$

$$
\begin{aligned}
&= \int I_{\{|x|>\alpha_n\}} \exp[(1,r(x))\theta] \int K_0(t) g(x+tb_n) dt\, dx \\
&= \int I_{\{|x|>\alpha_n\}} \exp[(1,r(x))\theta] \int K_0(t) \big[g(x) + g^{(1)}(x) t b_n + \frac{1}{2} g^{(2)}(\xi) t^2 b_n^2\big] dt\, dx \\
&\le \int I_{\{|x|>\alpha_n\}} h_\theta(x) dx \\
&\qquad + \frac{1}{2} b_n^2 \int I_{\{|x|>\alpha_n\}} h_\theta(x) \sup_{|t|\le a_0} \frac{|g^{(2)}(x+tb_n)|}{g(x)} dx \int t^2 K_0(t) dt \\
&\le \sup_{\theta\in\Theta} \int I_{\{|x|>\alpha_n\}} h_\theta(x) dx \\
&\qquad + \frac{1}{2} b_n^2 \sup_{\theta\in\Theta} \int I_{\{|x|>\alpha_n\}} h_\theta(x) \sup_{|t|\le a_0} \frac{|g^{(2)}(x+tb_n)|}{g(x)} dx \int t^2 K_0(t) dt \\
&\to 0.
\end{aligned}
$$

Thus, as $n \to \infty$, $\int I_{\{|x|>\alpha_n\}} \exp[(1,r(x))\theta] g_n(x) dx \xrightarrow{P} 0$ and

$$
\begin{aligned}
&\int I_{\{|x|>\alpha_n\}} \exp[(1,r(x))\theta] \big(g_n^{1/2}(x) - g^{1/2}(x)\big)^2 dx \\
&\le 2 \int I_{\{|x|>\alpha_n\}} \exp[(1,r(x))\theta] g_n(x) dx + 2 \int I_{\{|x|>\alpha_n\}} h_\theta(x) dx \qquad (3.14) \\
&\xrightarrow{P} 0.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\Big| \int I_{\{|x|\le\alpha_n\}} \exp[(1,r(x))\theta] \big(g_n^{1/2}(x) - g^{1/2}(x)\big)^2 dx \Big| \\
&\le \int I_{\{|x|\le\alpha_n\}} \exp[(1,r(x))\theta] g^{-1}(x) \big(g_n(x) - g(x)\big)^2 dx \\
&\le 2\Big[\int I_{\{|x|\le\alpha_n\}} \exp[(1,r(x))\theta] g^{-1}(x) \big(g_n(x) - Eg_n(x)\big)^2 dx \\
&\qquad + \int I_{\{|x|\le\alpha_n\}} \exp[(1,r(x))\theta] g^{-1}(x) \big(Eg_n(x) - g(x)\big)^2 dx \Big] \\
&= 2(A_{1n} + A_{2n}), \quad \text{say.}
\end{aligned}
$$

By (3.12) as $n \to \infty$

$$
\begin{aligned}
E|A_{1n}| &= \int I_{\{|x|\le\alpha_n\}} \exp[(1,r(x))\theta] g^{-1}(x) E\big(g_n(x) - Eg_n(x)\big)^2 dx \\
&\le \int I_{\{|x|\le\alpha_n\}} \exp[(1,r(x))\theta] g^{-1}(x) \frac{1}{nb_n^2} \int K_0^2\Big(\frac{y-x}{b_n}\Big) g(y) dy\, dx \\
&= n^{-1} b_n^{-1} \int I_{\{|x|\le\alpha_n\}} \exp[(1,r(x))\theta] \int_{-a_0}^{a_0} K_0^2(t) g(x+tb_n) g^{-1}(x) dt\, dx \\
&\le n^{-1} b_n^{-1} \sup_{\theta\in\Theta} \int I_{\{|x|\le\alpha_n\}} h_\theta(x) \sup_{|t|\le a_0} \frac{g(x+tb_n)}{g^2(x)} dx \int_{-a_0}^{a_0} K_0^2(t) dt \\
&\to 0,
\end{aligned}
$$

i.e., $A_{1n} \xrightarrow{P} 0$ as $n \to \infty$. By a Taylor expansion and (3.13),

$$\begin{aligned}
|A_{2n}| &= \int I_{\{|x|\le\alpha_n\}} \exp[(1, r(x))\theta] g^{-1}(x) \Big[\int_{-a_0}^{a_0} K_0(t) \big(g(x+tb_n) - \\
&\qquad\qquad g(x)\big) dt \Big]^2 dx \\
&\le \frac{1}{4} b_n^4 \int I_{\{|x|\le\alpha_n\}} \exp[(1, r(x))\theta] g^{-1}(x) \cdot \\
&\qquad \Big[\sup_{|t|\le a_0} |g^{(2)}(x+tb_n)| \int_{-a_0}^{a_0} t^2 K_0(t) dt \Big]^2 dx \\
&\le \frac{1}{4} b_n^4 \sup_{\theta\in\Theta} \int I_{\{|x|\le\alpha_n\}} h_\theta(x) \sup_{|t|\le a_0} \Big[\frac{g^{(2)}(x+tb_n)}{g(x)} \Big]^2 dx \Big(\int_{-a_0}^{a_0} t^2 K_0(t) dt \Big)^2 \\
&\to 0.
\end{aligned}$$

Therefore $\int I_{\{|x|\le\alpha_n\}} \exp[(1, r(x))\theta] \big(g_n^{1/2}(x) - g^{1/2}(x)\big)^2 dx \xrightarrow{P} 0$ as $n \to \infty$. This combined with (3.14) gives $\int \exp[(1, r(x))\theta][g_n^{1/2}(x) - g^{1/2}(x)]^2 dx \xrightarrow{P} 0$ for any $\theta \in \Theta$. By the continuity of the function in θ and the compactness of Θ, hence the result. □

Remark 3.4. Condition (D3) is satisfied when g and h_θ are two normal density functions with different standard deviations. Assume that $g(x)$ and $h(x)$ denote density functions of $N(0,1)$ and $N(\mu, \sigma)$, respectively, where $\sigma < 1$. It is easy to see that $h(x) = h_\theta(x) = \exp[(1, r(x))\theta] g(x)$, where $r_1(x) = x$, $r_2(x) = x^2$ and $\theta = (\theta_0, \theta_1, \theta_2) = (-\frac{\mu^2}{2\sigma^2} - \log\sigma, \frac{\mu}{\sigma^2}, \frac{1}{2} - \frac{1}{2\sigma^2})$. If the parameter space Θ is such that its projection onto the third argument is to the left of zero, then obviously condition (D3) holds.

Remark 3.5. Condition (D4) holds for many families and one such example is stated in Remark 3.2, i.e., g and h are two normal density functions with the same standard deviation. Without loss of generality, we suppose the compact parameter space $\Theta = [\underline{\alpha}, \bar{\alpha}] \times [\underline{\beta}, \bar{\beta}]$ for some finite numbers $\bar{\alpha}$, $\underline{\alpha}$, $\bar{\beta}$ and $\underline{\beta}$. Then it is easy to show that (3.10)-(3.13) hold for some α_n, the log function of n, and any bandwidth b_n such that $b_n \to 0$ and $nb_n \to \infty$ as $n \to \infty$.

Theorem 3.2. *Let $n \to \infty$ and $m \to \infty$ as $N \to \infty$. Suppose that $\big(1, r(x)\big)$ are linearly independent, (D1) holds for any $\theta \in \Theta$, and bandwidths b_n and b_m in (3.4) and (3.5), respectively, satisfy $b_n, b_m \to 0$ and $nb_n, mb_m \to \infty$ as $N \to \infty$. Further, suppose that either (D2), (D3) or (D4) holds. Then $\| h_m^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$ and $\sup_{\theta\in\Theta} \| \widehat{h}_\theta^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$ as $N \to \infty$. Furthermore, $\theta_N \xrightarrow{P} \theta$ as $N \to \infty$, where θ_N is defined by (3.8) with g_n, h_m and $\widehat{h}_\theta$ given by (3.4), (3.5) and (3.7) respectively.*

Proof. Remark 3.3 yields that $\{h_\theta\}_{\theta\in\Theta}$ is identifiable. So if we can prove that

$\| h_m^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$ and $\sup_{\theta\in\Theta} \| \widehat{h}_\theta^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$ as $N \to \infty$, then $\theta_N \xrightarrow{P} \theta$ as $N \to \infty$ by Theorem 3.1.

It is known that $g_n \xrightarrow{P} g$ and $h_m \xrightarrow{P} h_\theta$ as $N \to \infty$ (see Rao, 1983). Since $\int h_\theta(x)dx = \int h_m(x)dx = 1$, $\int [h_\theta(x) - h_m(x)]^+ dx = \int [h_\theta(x) - h_m(x)]^- dx$ and $\| h_m^{1/2} - h_\theta^{1/2} \|^2 \leq \int \mid h_\theta(x) - h_m(x) \mid dx = 2\int [h_\theta(x) - h_m(x)]^+ dx$. Since $[h_\theta(x) - h_m(x)]^+ < h_\theta(x)$, by the Dominated Convergence Theorem, it follows that $\| h_m^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$ as $m \to \infty$.

Note that $\int [\widehat{h}_\theta^{1/2}(x) - h_\theta^{1/2}(x)]^2 dx = \int \exp[(1, r(x))\theta][g_n^{1/2}(x) - g^{1/2}(x)]^2 dx \leq \int \exp[(1, r(x))\theta] |g_n(x) - g(x)| dx$. If (D2) holds, then $g_n - g$ will have a compact support, on which $\exp[(1, r(x))\theta]$ is bounded. Therefore, $\int [\widehat{h}_\theta^{1/2}(x) - h_\theta^{1/2}(x)]^2 dx \leq C_1 \int |g_n(x) - g(x)| dx = 2C_1 \int [g(x) - g_n(x)]^+ dx$ for some positive number C_1. Since $g_n \xrightarrow{P} g$, by the Dominated Convergence Theorem we have $\sup_{\theta\in\Theta} \| \widehat{h}_\theta^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$. If (D3) holds, then $\exp[(1, r(x))\theta]$ is bounded and similarly $\sup_{\theta\in\Theta} \| \widehat{h}_\theta^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$. If (D4) holds, then Lemma 3.3 gives that $\sup_{\theta\in\Theta} \| \widehat{h}_\theta^{1/2} - h_\theta^{1/2} \| \xrightarrow{P} 0$. □

3.3 *Asymptotic Normality of MHD Estimator*

In this section, we develop the asymptotic distribution of the proposed MHD estimator θ_N. We first state following conditions (D5) and (D6):

(D5) There exists $B(\theta, \epsilon)$, an ϵ-neighborhood of θ for some $\epsilon > 0$, such that for $s = 1, 2$ and $i, j, k = 0, 1, \ldots, p$,

$$\sup_{t\in\Theta\cap B(\theta,\epsilon)} \sup_x \exp[\frac{1}{s}(1, r(x))t] |r_i(x) r_j(x) r_k(x)| < \infty,$$

where $r_0(x) = 1$.

(D6) There exists $B(\theta, \epsilon)$, an ϵ-neighborhood of θ for some $\epsilon > 0$, such that for $s = 1, 2$, $i, j, k = 0, 1, \ldots, p$, $r_0(x) = 1$, and $n \to \infty$

$$\int |r_i(x) r_j(x)|^2 \exp[(1, r(x))\theta] h_\theta(x) dx < \infty, \tag{3.15}$$

$$\int |r_i(x) r_j(x) r_k(x)|^s \sup_{t\in\Theta\cap B(\theta,\epsilon)} \exp[(1, r(x))t] \sup_{|t|\leq a_0} g(x + tb_n)\, dx = O(1), \tag{3.16}$$

$$\int |r_i(x) r_j(x)|^2 \exp[2(1, r(x))\theta] \sup_{|t|\leq a_0} g(x + tb_n)\, dx = O(1). \tag{3.17}$$

Under condition (D2), (D5) or (D6), we derive an expression for the bias term $\theta_N - \theta$, which is presented in the next theorem. We denote $I(\theta) = \int (1, r(x))^T (1, r(x)) h_\theta(x) dx$ and assume that $I(\theta)$ is finite and nonsingular.

Theorem 3.3. *Suppose that $\theta \in int(\Theta)$, K_0 in (3.4) has compact support, and assumptions in Theorem 3.2 hold. Further suppose that either (D2), (D5) or (D6) holds. Then, it follows that*

$$\begin{aligned} \theta_N - \theta \;=\; & [I^{-1}(\theta) + \mu_N] \times 2 \int \Big\{ \exp[\tfrac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) h_m^{1/2}(x) \\ & - \exp[(1, r(x))\theta] g_n(x) \Big\} (1, r(x))^T dx \end{aligned} \tag{3.18}$$

where θ_N is defined by (3.8) and μ_N is a $(p+1) \times (p+1)$ matrix with elements tending to zero in probability as $N \to \infty$.

Remark 3.6. An example in which condition (D5) holds is stated in Remark 3.4. In this example $\theta = (\theta_0, \theta_1, \theta_2)$ with $\theta_2 < 0$. Therefore, one can easily prove that condition (D5) is satisfied. It is also obvious that $I(\theta)$ is finite in this case.

Remark 3.7. Condition (D6) is satisfied for the example stated in Remark 3.2, i.e., two normal density functions with the same standard deviation.

Proof of Theorem 3.3. From Theorem 3.2 we have that $\theta_N \xrightarrow{P} \theta$ as $N \to \infty$. Since $t = \theta_N \in \Theta$ minimizes the Hellinger distance between $\widehat{h}_t$ and h_m, θ_N maximizes $\int \widehat{h}_t^{1/2}(x) h_m^{1/2}(x) dx - \frac{1}{2}\widehat{h}_t(x) dx$. Also since K_0 has compact support, we have $0 = \int \frac{\partial}{\partial t}[\widehat{h}_t^{1/2}(x) h_m^{1/2}(x) - \frac{1}{2}\widehat{h}_t(x)]|_{t=\theta_N} \, dx$, i.e.,

$$\begin{aligned} & \int \exp[\tfrac{1}{2}(1, r(x))\theta_N] g_n^{1/2}(x) h_m^{1/2}(x) (1, r(x))^T dx \\ & \qquad - \int \exp[(1, r(x))\theta_N] g_n(x) (1, r(x))^T dx \;=\; 0. \end{aligned} \tag{3.19}$$

We will prove in the following that under condition (D2), (D5) or (D6), (3.19) will reduce to

$$\begin{aligned} & \int \Big\{ \exp[\tfrac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) h_m^{1/2}(x) - \exp[(1, r(x))\theta] g_n(x) \Big\} (1, r(x))^T dx \\ & \qquad - [\tfrac{1}{2} \int h_\theta(x) (1, r(x))^T (1, r(x)) dx + c_N] (\theta_N - \theta) \;=\; 0, \end{aligned} \tag{3.20}$$

where c_N is a $(p+1) \times (p+1)$ matrix with elements tending to zero in probability as $N \to \infty$, i.e., (3.18) holds.

(i) Suppose that (D2) or (D5) holds. Then for any $t \in \Theta \cap B(\theta, \epsilon)$,

$$|\int r_i(x)r_j(x)r_k(x)\exp[(1,r(x))t]g_n(x)dx| \le C\int g_n(x)dx = C$$

$$|\int r_i(x)r_j(x)r_k(x)\exp[\frac{1}{2}(1,r(x))t]g_n^{1/2}(x)h_m^{1/2}(x)dx|$$
$$\le C\Big(\int g_n(x)dx\Big)^{1/2}\Big(\int h_m(x)dx\Big)^{1/2} = C$$

with some positive value C. Therefore, by a Taylor expansion of θ_N at θ, one obtains with $\theta_t = t\theta + (1-t)\theta_N$ for some $0 < t < 1$,

$$\begin{aligned}
&\int \exp[\frac{1}{2}(1,r(x))\theta_N]g_n^{1/2}(x)h_m^{1/2}(x)(1,r(x))^T dx \\
=\ &\int (1,r(x))^T\Big\{\exp[\frac{1}{2}(1,r(x))\theta] + \frac{1}{2}\exp[\frac{1}{2}(1,r(x))\theta](1,r(x))(\theta_N-\theta) \\
&+\frac{1}{8}\exp[\frac{1}{2}(1,r(x))\theta_t](\theta_N-\theta)^T(1,r(x))^T(1,r(x))(\theta_N-\theta)\Big\}g_n^{1/2}(x)h_m^{1/2}(x)dx \\
=\ &\int \exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)h_m^{1/2}(x)(1,r(x))^T dx \\
&+\frac{1}{2}\int \exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)h_m^{1/2}(x)(1,r(x))^T(1,r(x))dx(\theta_N-\theta) \\
&+a_N(\theta_N-\theta),
\end{aligned} \tag{3.21}$$

$$\begin{aligned}
&\int \exp[(1,r(x))\theta_N]g_n(x)(1,r(x))^T dx \\
=\ &\int \Big\{\exp[(1,r(x))\theta] + \exp[(1,r(x))\theta](1,r(x))(\theta_N-\theta) \\
&+\frac{1}{2}\exp[(1,r(x))\theta_t](\theta_N-\theta)^T(1,r(x))^T(1,r(x))(\theta_N-\theta)\Big\}g_n(x)(1,r(x))^T dx \\
=\ &\int \exp[(1,r(x))\theta]g_n(x)(1,r(x))^T dx \\
&+\int \exp[(1,r(x))\theta]g_n(x)(1,r(x))^T(1,r(x))dx(\theta_N-\theta) + b_N(\theta_N-\theta),
\end{aligned} \tag{3.22}$$

where a_N and b_N are $(p+1)\times(p+1)$ matrixes with elements tending to zero in probability as $N \to \infty$ by the fact that $\theta_N \to \theta$. From (3.19), (3.21) and (3.22), we obtain

$$\begin{aligned}
0 \ =\ &\int \Big\{\exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)h_m^{1/2}(x) - \exp[(1,r(x))\theta]g_n(x)\Big\}(1,r(x))^T dx \\
&+\Big\{\frac{1}{2}\int \exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)h_m^{1/2}(x)(1,r(x))^T(1,r(x))dx \\
&\qquad - \int \exp[(1,r(x))\theta]g_n(x)(1,r(x))^T(1,r(x))dx\Big\}(\theta_N-\theta) \\
&+[a_N - b_N](\theta_N-\theta).
\end{aligned} \tag{3.23}$$

Since either (D2) or (D5) holds,

$$\begin{aligned}
& \Big| \int \exp[\frac{1}{2}(1, r(x))\theta] \Big\{ g_n^{1/2}(x) h_m^{1/2}(x) - g^{1/2}(x) h_\theta^{1/2}(x) \Big\} (1, r(x))^T (1, r(x)) dx \Big| \\
\leq \; & C \Big\{ \int g_n^{1/2}(x) |h_m^{1/2}(x) - h_\theta^{1/2}(x)| dx + \int h_\theta^{1/2}(x) |g_n^{1/2}(x) - g^{1/2}(x)| dx \Big\} \\
\leq \; & C \Big\{ \Big[\int (h_m^{1/2}(x) - h_\theta^{1/2}(x))^2 dx \Big]^{1/2} + \Big[\int (g_n^{1/2}(x) - g^{1/2}(x))^2 dx \Big]^{1/2} \Big\}
\end{aligned}$$

with the r.h.s. of the preceding inequality goes to zero in probability using the results in Theorem 3.2. Thus,

$$\begin{aligned}
\int \exp[\frac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) h_m^{1/2}(x) (1, r(x))^T (1, r(x)) dx \xrightarrow{P} \\
\int h_\theta(x) (1, r(x))^T (1, r(x)) dx.
\end{aligned} \tag{3.24}$$

Similarly

$$\begin{aligned}
& \Big| \int \exp[(1, r(x))\theta] \big(g_n(x) - g(x) \big) (1, r(x))^T (1, r(x)) dx \Big| \\
\leq \; & C \int |(g_n^{1/2}(x) - g^{1/2}(x))(g_n^{1/2}(x) + g^{1/2}(x))| dx \\
\leq \; & C \Big[\int (g_n^{1/2}(x) - g^{1/2}(x))^2 dx \Big]^{1/2} \Big[\int (g_n^{1/2}(x) + g^{1/2}(x))^2 dx \Big]^{1/2} \\
\leq \; & 2C \Big[\int (g_n^{1/2}(x) - g^{1/2}(x))^2 dx \Big]^{1/2} \\
\xrightarrow{P} \; & 0,
\end{aligned}$$

i.e.,

$$\int \exp[(1, r(x))\theta] g_n(x) (1, r(x))^T (1, r(x)) dx \xrightarrow{P} \int h_\theta(x) (1, r(x))^T (1, r(x)) dx. \tag{3.25}$$

As a result, (3.23) reduces to (3.20).

(ii) Suppose (D6) holds. Then by (3.16),

$$\begin{aligned}
& E \Big| \int |r_i(x) r_j(x) r_k(x)| \sup_{t \in \Theta \cap B(\theta, \epsilon)} \exp[(1, r(x))t] \; g_n(x) dx \Big| \\
= \; & \int |r_i(x) r_j(x) r_k(x)| \sup_{t \in \Theta \cap B(\theta, \epsilon)} \exp[(1, r(x))t] \; E[g_n(x)] dx \\
= \; & \int |r_i(x) r_j(x) r_k(x)| \sup_{t \in \Theta \cap B(\theta, \epsilon)} \exp[(1, r(x))t] \int_{-a_0}^{a_0} K_0(t) g(x + t b_n) dt \, dx
\end{aligned}$$

$$
\begin{aligned}
&\leq \int |r_i(x)r_j(x)r_k(x)| \sup_{t\in\Theta\cap B(\theta,\epsilon)} \exp[(1,r(x))t] \sup_{|t|\leq a_0} g(x+tb_n)\, dx \\
&= O(1).
\end{aligned} \tag{3.26}
$$

Therefore, $\int |r_i(x)r_j(x)r_k(x)| \sup_{t\in\Theta\cap B(\theta,\epsilon)} \exp[(1,r(x)\theta_t]\, g_n(x)dx = O_P(1)$ and thus (3.22) holds. Similarly,

$$
\begin{aligned}
&E\Big[\int |r_i(x)r_j(x)r_k(x)| \sup_{t\in\Theta\cap B(\theta,\epsilon)} \exp[\frac{1}{2}(1,r(x))t]\, g_n^{1/2}(x)h_m^{1/2}(x)dx\Big]^2 \\
&\leq E\Big[\int |r_i(x)r_j(x)r_k(x)|^2 \sup_{t\in\Theta\cap B(\theta,\epsilon)} \exp[(1,r(x))t]\, g_n(x)dx \int h_m(x)dx\Big] \\
&= \int |r_i(x)r_j(x)r_k(x)|^2 \sup_{t\in\Theta\cap B(\theta,\epsilon)} \exp[(1,r(x))t]\, E[g_n(x)]dx \\
&\leq \int |r_i(x)r_j(x)r_k(x)|^2 \sup_{t\in\Theta\cap B(\theta,\epsilon)} \exp[(1,r(x))t] \sup_{|t|\leq a_0} g(x+tb_n)\, dx \\
&= O(1)
\end{aligned}
$$

and hence (3.21) holds. As a result (3.23) holds. By (3.15), (3.16) and a similar argument as in (3.26),

$$
\begin{aligned}
&\Big|\int r_i(x)r_j(x)\exp[\frac{1}{2}(1,r(x))\theta]\big\{g_n^{1/2}(x)h_m^{1/2}(x) - g^{1/2}(x)h_\theta^{1/2}(x)\big\}dx\Big| \\
\leq\ &\int |r_i(x)r_j(x)|\exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)|h_m^{1/2}(x) - h_\theta^{1/2}(x)|dx \\
&+\int |r_i(x)r_j(x)|\exp[\frac{1}{2}(1,r(x))\theta]h_\theta^{1/2}(x)|g_n^{1/2}(x) - g^{1/2}(x)|dx \\
\leq\ &\Big[\int |r_i(x)r_j(x)|^2\exp[(1,r(x))\theta]g_n(x)dx\Big]^{1/2}\Big[\int \big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big)^2dx\Big]^{1/2} \\
&+\Big[\int |r_i(x)r_j(x)|^2\exp[(1,r(x))\theta]h_\theta(x)dx\Big]^{1/2}\Big[\int \big(g_n^{1/2}(x) - g^{1/2}(x)\big)^2dx\Big]^{1/2} \\
=\ &O_P\big(\big[\int \big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big)^2dx\big]^{1/2}\big) + O\big(\big[\int \big(g_n^{1/2}(x) - g^{1/2}(x)\big)^2dx\big]^{1/2}\big)
\end{aligned}
$$

and thus (3.24) holds. By (3.15), (3.17) and using a similar argument as in (3.26),

$$
\begin{aligned}
&\Big|\int r_i(x)r_j(x)\exp[(1,r(x))\theta]\big(g_n(x) - g(x)\big)dx\Big|^2 \\
\leq\ &\Big[\int |r_i(x)r_j(x)|\exp[(1,r(x))\theta]\ \big|(g_n^{1/2}(x) - g^{1/2}(x))(g_n^{1/2}(x) + g^{1/2}(x))\big|dx\Big]^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \int |r_i(x)r_j(x)|^2 \exp[2(1, r(x))\theta]\big(g_n^{1/2}(x) + g^{1/2}(x)\big)^2 dx \times \\
&\qquad\qquad \int (g_n^{1/2}(x) - g^{1/2}(x))^2 dx \\
&\leq 2\Big[\int |r_i(x)r_j(x)|^2 \exp[2(1, r(x))\theta] g_n(x) + \\
&\qquad \int |r_i(x)r_j(x)|^2 \exp[(1, r(x))\theta] h_\theta(x) dx\Big] \times \int (g_n^{1/2}(x) - g^{1/2}(x))^2 dx \\
&= O\Big(\int (g_n^{1/2}(x) - g^{1/2}(x))^2 dx\Big) \\
&\xrightarrow{P} 0,
\end{aligned}
$$

i.e. (3.25) holds. As a result, (3.23) reduces to (3.20). □

We now state the asymptotic distribution of the proposed MHD estimator θ_N of θ. Following conditions are made in the next theorem:

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

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

(C1) The second derivatives of g and h_θ exist.

(C2) $n/N \to \rho \in (0, 1)$ as $N \to \infty$, and the bandwidths b_n and b_m in (3.4) and (3.5), respectively, converge to zero at the same rate as $N \to \infty$.

(C3) K_0 and K_1 in (3.4) and (3.5), respectively, are bounded symmetric densities with supports $[-a_0, a_0]$ and $[-a_1, a_1]$, $0 < a_0, a_1 < \infty$.

(C4) Both $I(\theta)$ and $J(\theta)$ are finite, where $I(\theta) = \int (1, r(x))^T (1, r(x)) h_\theta(x) dx$ and $J(\theta) = \int (1, r(x))^T (1, r(x)) \exp[(1, r(x))\theta] h_\theta(x) dx$.

(C5) The second derivative of g exists and satisfies for $i = 0, 1, \ldots, p$,

$$
b_n^2 \int \varepsilon_{Ni}^2(x) h_\theta(x) \sup_{|t| \leq a_0} \frac{|g^{(2)}(x + tb_n)|}{g(x)} dx = O(1) \quad \text{as } N \to \infty,
$$

where $\varepsilon_N(x) = (1, r(x))^T I_{\{|x| > \alpha_N\}} = (\varepsilon_{N0}(x), \varepsilon_{N1}(x), \ldots, \varepsilon_{Np}(x))^T$ and $g^{(k)}$ denotes the k-th derivative of g.

(C5') The second derivative of g exists and satisfies

$$
N^{1/2} b_n^2 \int |\varepsilon_N(x)| h_\theta(x) \sup_{|t| \leq a_0} \frac{|g^{(2)}(x + tb_n)|}{g(x)} dx = o(1) \quad \text{as } N \to \infty.
$$

(C6)

$$
\begin{aligned}
N \cdot P(|Z_1| > \alpha_N - a_1 b_m) &\to 0 \quad \text{as } N \to \infty, \\
N \cdot P(|X_1| > \alpha_N - a_0 b_n) &\to 0 \quad \text{as } N \to \infty.
\end{aligned}
$$

(C7) With $\delta_N(x) = (1, r(x))^T I_{\{|x| \le \alpha_N\}} = (\delta_{N0}(x), \delta_{N1}(x), \ldots, \delta_{Np}(x))^T$,

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

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

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

$$N^{1/2} b_n^4 \int |\delta_N(x)| h_\theta(x) \sup_{|t| \le a_0} \Big[\frac{g^{(2)}(x + t b_n)}{g(x)}\Big]^2 dx \to 0 \quad \text{as } N \to \infty.$$

(C8)

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

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

(C9)

$$\sup_{|x| \le \alpha_N} \sup_{|t| \le a_1} \frac{h_\theta(x + t b_m)}{h_\theta(x)} = O(1) \quad \text{as } N \to \infty,$$

$$\sup_{|x| \le \alpha_N} \sup_{|t| \le a_0} \frac{g(x + t b_n)}{g(x)} = O(1) \quad \text{as } N \to \infty.$$

(C10) $r(x)$ is differentiable and satisfies for $i = 0, 1, \ldots, p$,

$$b_m^2 \int I_{\{|x| \le \alpha_N\}} h_\theta(x) \sup_{|t| \le a_1} \big(r_i^{(1)}(x + t b_m)\big)^2 dx \to 0 \quad \text{as } N \to \infty,$$

$$b_n^2 \int I_{\{|x| \le \alpha_N\}} g(x) \sup_{|t| \le a_0} \Big[\frac{\partial r_i(y) \exp[(1, r(y))\theta]}{\partial y}\Big|_{y = x + t b_n}\Big]^2 dx \to 0 \quad \text{as } N \to \infty.$$

(C11)

$$N^{-1/2} b_m^{-1} \int |\delta_N(x)| \exp[\frac{1}{2}(1, r(x))\theta] dx \to 0 \quad \text{as } N \to \infty,$$

$$N^{1/2} b_m^4 \int |\delta_N(x)| \exp[\frac{1}{2}(1, r(x))\theta] dx \to 0 \quad \text{as } N \to \infty.$$

Theorem 3.4. *Suppose that θ_N defined by (3.8) satisfies (3.18). Further suppose that conditions (C0)-(C10) and (C5') hold. Then the asymptotic distribution of $N^{1/2}(\theta_N - \theta)$ is $N(0, \Sigma)$, where Σ is defined by*

$$\Sigma = I^{-1}(\theta)\big[\frac{1}{\rho}\Sigma_0 + \frac{1}{1-\rho}\Sigma_1\big]I^{-1}(\theta) \tag{3.27}$$

with

$$\begin{aligned}\Sigma_0 = &\int (1, r(x))^T (1, r(x)) \exp[(1, r(x))\theta] h_\theta(x) dx \\ &- \int (1, r(x))^T h_\theta(x) dx \int (1, r(x)) h_\theta(x) dx\end{aligned} \tag{3.28}$$

and

$$\Sigma_1 = \int (1, r(x))^T (1, r(x)) h_\theta(x) dx - \int (1, r(x))^T h_\theta(x) dx \int (1, r(x)) h_\theta(x) dx. \tag{3.29}$$

Proof. The sketch of the proof is as follows. Note that

$$\begin{aligned}&\exp[\frac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) h_m^{1/2}(x) (1, r(x))^T - \exp[(1, r(x))\theta] g_n(x) (1, r(x))^T \\ = \; &(1, r(x))^T \exp[\frac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) [h_m^{1/2}(x) - h_\theta^{1/2}(x)] \\ &-(1, r(x))^T \exp[(1, r(x))\theta] g_n^{1/2}(x) [g_n^{1/2}(x) - g^{1/2}(x)].\end{aligned}$$

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

$$N^{1/2} \int (1, r(x))^T \exp[\frac{1}{2}(1, r(x))\theta] [g_n^{1/2}(x) - g^{1/2}(x)][h_m^{1/2}(x) - h_\theta^{1/2}(x)] dx \xrightarrow{P} 0$$

and

$$N^{1/2} \int (1, r(x))^T \exp[(1, r(x))\theta] [g_n^{1/2}(x) - g^{1/2}(x)]^2 dx \xrightarrow{P} 0.$$

As a result we only need to give the asymptotic distribution of

$$N^{1/2} \int (1, r(x))^T h_\theta^{1/2}(x) [h_m^{1/2}(x) - h_\theta^{1/2}(x)] dx$$

and

$$N^{1/2} \int (1, r(x))^T \exp[(1, r(x))\theta] g(x) [g_n^{1/2}(x) - g^{1/2}(x)] dx.$$

For details see Section 3.6. □

Remark 3.8. Consider the example stated in Remark 3.2. It is easy to see that conditions (C0), (C1) and (C4) hold. We can easily choose bandwidths b_n and b_m, and kernels K_0 and K_1 satisfing conditions (C2) and (C3). Since for $k = 0, 1, 2,$

$$\int |x|^k h_\theta(x) \sup_{|t|\le a_0} \frac{|g^{(2)}(x + tb_n)|}{g(x)} dx = O(1) \quad \text{as } n \to \infty,$$

conditions (C5) and (C5') hold if $Nb_n^4 = O(1)$ as $N \to \infty$. Note that as $N \to \infty$,

$$N \int_{\alpha_N}^{\infty} \exp[-x^2/2] dx \le N \int_{\alpha_N}^{\infty} x \exp[-x^2/2] dx = N \exp[-\alpha_N^2/2].$$

Thus, if $N \exp[-\alpha_N^2/2] \to 0$ as $N \to \infty$, then condition (C6) holds. Since for $i = 0, 1$ and $j = 1, 2,$

$$\int |x|^i h_\theta(x) \sup_{|t|\le a_1} \Big| \frac{h_\theta^{(2)}(x + tb_m)}{h_\theta(x)} \Big|^j dx = O(1) \quad \text{as } N \to \infty$$

and

$$\int |x|^i h_\theta(x) \sup_{|t|\le a_0} \Big| \frac{g^{(2)}(x + tb_n)}{g(x)} \Big|^j dx = O(1) \quad \text{as } N \to \infty,$$

(C8) and the second and fourth expressions in (C7) hold if $Nb_n^4 \to 0$ as $N \to \infty$. If $b_n \alpha_N \to 0$ and $N^{-1/2} b_n^{-1} \alpha_N^2 \to 0$ as $N \to \infty$, then for $i = 0, 1$ and $N \to \infty$

$$\begin{aligned}
& N^{-1/2} b_m^{-1} \int_{-\alpha_N}^{\alpha_N} |x|^i h_\theta(x) \sup_{|t|\le a_1} \frac{h_\theta(x + tb_m)}{h_\theta^2(x)} dx \\
= \; & N^{-1/2} b_m^{-1} \int_{-\alpha_N}^{\alpha_N} |x|^i \sup_{|\epsilon| \le a_1 b_m} \exp[-\epsilon x + \epsilon\mu - \frac{\epsilon^2}{2}] dx \\
\le \; & 2 \exp[a_1 b_m |\mu|] \cdot N^{-1/2} b_m^{-1} \int_0^{\alpha_N} |x|^i \exp[a_1 b_m x] dx \\
\le \; & \frac{2}{a_1} \exp[a_1 b_m |\mu|] \cdot N^{-1/2} b_m^{-2} \alpha_N^i \big(\exp[a_1 b_m \alpha_N] - 1 \big) \\
= \; & O\big(N^{-1/2} b_m^{-1} \alpha_N^{i+1} \big) \\
\to \; & 0,
\end{aligned}$$

and therefore the first expression in (C7) holds. Similarly, for $i = 0, 1$ and $N \to \infty$, one has

$$\begin{aligned}
& N^{-1/2} b_n^{-1} \int_{-\alpha_N}^{\alpha_N} |x|^i h_\theta(x) \sup_{|t|\le a_0} \frac{g(x + tb_n)}{g^2(x)} dx \\
= \; & N^{-1/2} b_n^{-1} \int_{-\alpha_N}^{\alpha_N} |x|^i \exp[\mu x - \frac{\mu^2}{2}] \sup_{|\epsilon| \le a_0 b_n} \exp[-\epsilon x - \frac{\epsilon^2}{2}] dx
\end{aligned}$$

$$
\begin{aligned}
&\leq N^{-1/2}b_n^{-1}\alpha_N^i \int_0^{\alpha_N} \exp[(\mu + a_0 b_n)x]dx \\
&\qquad + N^{-1/2}b_n^{-1}\alpha_N^i \int_{-\alpha_N}^{0} \exp[(\mu - a_0 b_n)x]dx \\
&= N^{-1/2}b_n^{-1}\alpha_N^i(\mu + a_0 b_n)^{-1}\big(\exp[(\mu + a_0 b_n)\alpha_N] - 1\big) \\
&\qquad + N^{-1/2}b_n^{-1}\alpha_N^i(\mu - a_0 b_n)^{-1}\big(1 - \exp[-(\mu - a_0 b_n)\alpha_N]\big) \\
&= \begin{cases} O\big(N^{-1/2}b_n^{-1}\alpha_N^i \exp[|\mu|\alpha_N]\big) & \text{if } \mu \neq 0, \\ O\big(N^{-1/2}b_n^{-1}\alpha_N^{i+1}\big) & \text{if } \mu = 0. \end{cases}
\end{aligned}
$$

Therefore, if $N^{-1/2}b_n^{-1}\alpha_N \exp[|\mu|\alpha_N] \to 0$ as $N \to \infty$, then the third expression in (C7) holds. If $b_n\alpha_N = O(1)$ as $N \to \infty$, then (C9) holds. It is easy to check that (C10) is satisfied. Note that as $N \to \infty$,

$$
\int_{-\alpha_N}^{\alpha_N} \exp[\frac{1}{2}(1, r(x))\theta]dx = \begin{cases} O\big(\exp[|\mu|\alpha_N/2]\big) & \text{if } \mu \neq 0, \\ O(\alpha_N) & \text{if } \mu = 0, \end{cases}
$$

and

$$
\int_{-\alpha_N}^{\alpha_N} |x| \exp[\frac{1}{2}(1, r(x))\theta]dx = \begin{cases} O\big(\alpha_N \exp[|\mu|\alpha_N/2]\big) & \text{if } \mu \neq 0, \\ O\big(\alpha_N^2\big) & \text{if } \mu = 0. \end{cases}
$$

So if $N^{-1}b_m^{-2}\alpha_N^2 \exp[|\mu|\alpha_N] \to 0$ and $Nb_m^4\alpha_N^2 \exp[|\mu|\alpha_N] \to 0$ as $N \to \infty$, then (C11) hold. In summary, if we choose

$$
b_n = O\big(N^{-r}\big), \quad 1/4 < r < 1/2
$$

and

$$
\alpha_N = O\big((\log N)^q\big), \quad 1/2 < q < 1,
$$

then conditions (C0)-(C10) and (C5') are satisfied. Also by Remarks 3.2, 3.5 and 3.7, (3.18) holds. As a result, (3.27) holds by Theorem 3.4.

Remark 3.9. Again consider the example investigated in Remark 3.8. Simple calculation yields that the asymptotic variance for our proposed estimator θ_N of θ is

$$
\begin{aligned}
\Sigma &= \frac{1}{\rho}\begin{bmatrix} \mu^4 \exp[\mu^2] - \mu^2 \exp[\mu^2] + \exp[\mu^2] - 1 & -\mu^3 \exp[\mu^2] \\ -\mu^3 \exp[\mu^2] & \mu^2 \exp[\mu^2] + \exp[\mu^2] \end{bmatrix} \\
&\quad + \frac{1}{1-\rho}\begin{bmatrix} \mu^2 & -\mu \\ -\mu & 1 \end{bmatrix}.
\end{aligned}
$$

Zhang (2000) estimated $\theta = (\alpha, \beta)$ by using semiparametric likelihood under model (3.1). He derived the asymptotic variance, say $\overline{\Sigma}$, of his proposed estima-

tor of θ. It is hard to give an explicit expression for the asymptotic variance $\overline{\Sigma}$ in this example. So, here we compare asymptotic variances in the simplest case when $\mu = 0$. If $\mu = 0$ then the asymptotic variance of our proposed estimator θ_N is

$$\Sigma = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\rho(1-\rho)} \end{bmatrix},$$

which is exactly the same as that of $\overline{\Sigma}$. More detailed comparison of Σ with $\overline{\Sigma}$ are shown in Section 3.4.

3.4 Simulation Studies

In this section, we report the results of simulation studies. We use Monte Carlo methods to demonstrate that the proposed MHD estimator θ_N defined in (3.8) has good robustness and efficiency properties.

In this simulation study, we considered the example stated in Remark 3.2. We assumed $g(x)$ and $h(x)$ as density functions of the normal distributions $N(0,1)$ and $N(\mu,1)$, respectively. Thus $h(x) = h_\theta(x) = \exp[(1, r(x))\theta]g(x)$, where $r(x) = x$ and $\theta = (\alpha, \beta) = (-\frac{\mu^2}{2}, \mu)$. For different μ and ρ values, Table 3.1 compares Σ defined in (3.27) with the asymptotic variance matrix $\overline{\Sigma}$ of the maximum semiparametric likelihood estimator $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta})$ of Zhang (2000). From Table 3.1, we can see that smaller μ values give smaller values for the variance of estimator $\theta_N = (\widehat{\alpha}, \widehat{\beta})$. The correlations are all negative since $\alpha = -\frac{\beta^2}{2}$. When $\mu = 0$, the asymptotic variance of θ_N is exactly the same as that of $\widetilde{\theta}$ as shown in Remark 3.9. When $\mu = 0.1$, the asymptotic variance of θ_N is almost the same as that of $\widetilde{\theta}$ for all different ρ values. But for large μ values, θ_N has much larger asymptotic variance compared with those of $\widetilde{\theta}$. In fact, we can expect this behavior from the expression of asymptotic variance derived in Remark 3.9. However, we have shown below in our simulation that θ_N could have smaller bias and mean squared error (MSE) than those of $\widetilde{\theta}$, and at the same time θ_N is much more robust to outliers than $\widetilde{\theta}$.

Our aim of this simulation is to compare the performance of our proposed estimator θ_N defined at (3.8) with that of Zhang's maximum semiparametric likelihood estimator $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta})$, by examining their biases, MSEs and α-IFs. In our simulations, we let $\mu = 0.5$ be fixed and therefore $\theta = (\alpha, \beta) = (-0.125, 0.5)$. For each pair (n, m), we generated ten independent sets of combined random samples of size $N = n + m = 60$ from the $N(0,1)$ and $N(\mu,1)$ distributions. Here the pair (n, m) takes varying values $(10, 50)$, $(20, 40)$, $(30, 30)$, $(40, 20)$ and $(50, 10)$. For each pair (n, m) considered, we obtained estimates of the bias and MSE as follows:

$$\widehat{\text{Bias}} = \frac{1}{N_s}\sum_{i=1}^{N_s}(\widehat{\gamma}_i - \gamma)$$

Tab. 3.1: The asymptotic variance matrixes Σ and $\overline{\Sigma}$ of θ_N and $\widetilde{\theta}$ defined in (3.8) and Zhang (2000), respectively, when g and h are the densities of $N(0,1)$ and $N(\mu,1)$, respectively.

	$\mu = 0.1$		$\mu = 0.5$		$\mu = 1$	
ρ	Σ	$\overline{\Sigma}$	Σ	$\overline{\Sigma}$	Σ	$\overline{\Sigma}$
1/6	$\begin{bmatrix} 0.01 & -0.13 \\ -0.13 & 7.32 \end{bmatrix}$	$\begin{bmatrix} 0.01 & -0.12 \\ -0.12 & 7.22 \end{bmatrix}$	$\begin{bmatrix} 0.56 & -1.56 \\ -1.56 & 10.83 \end{bmatrix}$	$\begin{bmatrix} 0.33 & -0.73 \\ -0.73 & 7.74 \end{bmatrix}$	$\begin{bmatrix} 11.51 & -17.51 \\ -17.51 & 33.82 \end{bmatrix}$	$\begin{bmatrix} 1.74 & -2.33 \\ -2.33 & 9.70 \end{bmatrix}$
2/6	$\begin{bmatrix} 0.02 & -0.15 \\ -0.15 & 4.56 \end{bmatrix}$	$\begin{bmatrix} 0.02 & -0.15 \\ -0.15 & 4.52 \end{bmatrix}$	$\begin{bmatrix} 0.50 & -1.23 \\ -1.23 & 6.32 \end{bmatrix}$	$\begin{bmatrix} 0.41 & -0.88 \\ -0.88 & 5.01 \end{bmatrix}$	$\begin{bmatrix} 6.65 & -9.65 \\ -9.65 & 17.81 \end{bmatrix}$	$\begin{bmatrix} 2.02 & -2.54 \\ -2.54 & 6.67 \end{bmatrix}$
3/6	$\begin{bmatrix} 0.02 & -0.20 \\ -0.20 & 4.04 \end{bmatrix}$	$\begin{bmatrix} 0.02 & -0.20 \\ -0.20 & 4.02 \end{bmatrix}$	$\begin{bmatrix} 0.59 & -1.32 \\ -1.32 & 5.21 \end{bmatrix}$	$\begin{bmatrix} 0.53 & -1.13 \\ -1.13 & 4.50 \end{bmatrix}$	$\begin{bmatrix} 5.44 & -7.44 \\ -7.44 & 12.87 \end{bmatrix}$	$\begin{bmatrix} 2.55 & -3.05 \\ -3.05 & 6.09 \end{bmatrix}$
4/6	$\begin{bmatrix} 0.03 & -0.30 \\ -0.30 & 4.53 \end{bmatrix}$	$\begin{bmatrix} 0.03 & -0.30 \\ -0.30 & 4.52 \end{bmatrix}$	$\begin{bmatrix} 0.81 & -1.74 \\ -1.74 & 5.41 \end{bmatrix}$	$\begin{bmatrix} 0.78 & -1.63 \\ -1.63 & 5.01 \end{bmatrix}$	$\begin{bmatrix} 5.58 & -7.08 \\ -7.08 & 11.15 \end{bmatrix}$	$\begin{bmatrix} 3.62 & -4.13 \\ -4.13 & 6.67 \end{bmatrix}$
5/6	$\begin{bmatrix} 0.06 & -0.60 \\ -0.60 & 7.22 \end{bmatrix}$	$\begin{bmatrix} 0.06 & -0.60 \\ -0.60 & 7.22 \end{bmatrix}$	$\begin{bmatrix} 1.55 & -3.19 \\ -3.19 & 7.93 \end{bmatrix}$	$\begin{bmatrix} 1.54 & -3.14 \\ -3.14 & 7.74 \end{bmatrix}$	$\begin{bmatrix} 8.06 & -9.26 \\ -9.26 & 12.52 \end{bmatrix}$	$\begin{bmatrix} 6.78 & -7.37 \\ -7.37 & 9.70 \end{bmatrix}$

and

$$\widehat{\mathrm{MSE}} = \frac{1}{N_s}\sum_{i=1}^{N_s}(\widehat{\gamma}_i - \gamma)^2,$$

where N_s is the number of replications ($N_s = 10$ in our case), and $\widehat{\gamma}_i$ denotes an estimate of γ for the ith replication. Here $\gamma = \alpha$ or β, and $\widehat{\gamma}$ denotes either the proposed MHD estimators $\widehat{\alpha}$ and $\widehat{\beta}$ in (3.8), or the maximum semiparametric likelihood estimators $\widetilde{\alpha}$ and $\widetilde{\beta}$ of Zhang (2000). The bandwidths b_n and b_m in (3.4) and (3.5), respectively, were taken to be $h_n = n^{-2/5}$ and $h_m = m^{-2/5}$. We used Epanechnikov kernel function given by

$$K(x) = \frac{3}{4}\left(1 - x^2\right) I_{[-1,1]}(x), \tag{3.30}$$

for both K_0 and K_1. According to the discussion in Remark 3.8, our choice of kernel functions and bandwidths satisfy conditions (C0)-(C10) and (C5'), and therefore Theorem 3.4 holds. The simulation results are summarized in Table 3.2. From Table 3.2, we can see that for each pair (n, m) considered, $\widetilde{\alpha}$ is better than $\widehat{\alpha}$ considering the estimated bias and MSE. However, the MHD estimator $\widehat{\beta}$ is uniformly better than $\widetilde{\beta}$ in the sense of having smaller estimated bias and MSE. Note that β is the coefficient of $r(x) = x$ while α is only a normalizing parameter that makes $g(x)\exp[\alpha + r(x)\beta]$ integrate to one. We believe that β plays a more important role than α in most applications. For instance, in the Cox model, the value $\exp[\beta]$ can be interpreted as the ratio of the hazards of two individuals whose covariates are $Z = 1$ and $Z = 0$, respectively, but who are identical otherwise.

Tab. 3.2: Estimates of the biases and MSEs of $\theta_N = (\widehat{\alpha}, \widehat{\beta})$ and $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta})$ defined in (3.8) and Zhang (2000), respectively, when g and h are the densities of $N(0,1)$ and $N(0.5,1)$, respectively.

(n,m)	$\widehat{\mathrm{Bias}}(\widehat{\alpha})$	$\widehat{\mathrm{MSE}}(\widehat{\alpha})$	$\widehat{\mathrm{Bias}}(\widehat{\beta})$	$\widehat{\mathrm{MSE}}(\widehat{\beta})$	$\widehat{\mathrm{Bias}}(\widetilde{\alpha})$	$\widehat{\mathrm{MSE}}(\widetilde{\alpha})$	$\widehat{\mathrm{Bias}}(\widetilde{\beta})$	$\widehat{\mathrm{MSE}}(\widetilde{\beta})$
(10,50)	-0.77	0.63	0.41	0.21	-0.41	0.19	0.58	0.37
(20,40)	-0.67	0.52	0.58	0.83	-0.51	0.35	0.86	1.39
(30,30)	-0.65	0.50	0.51	0.38	-0.42	0.21	0.56	0.40
(40,20)	-0.67	0.53	0.47	0.37	-0.42	0.22	0.68	0.55
(50,10)	-0.74	0.58	0.39	0.25	-0.48	0.26	0.59	0.42

For the ten simulated replications, we examined at the same time the resistance of our MHD estimator θ_N to a single outlying observation, and compared it with that of $\widetilde{\theta}$. For this purpose, the α-IF given in Beran (1977) is a suitable measure of the change in the estimator. Here we have used the adapted version of the α-IF employed by Lu et al. (2003), among many others. Note that the

outlying observation could come from either density $g(x)$ or density $h(x)$. Here we only considered the case that the outlying observation comes from $h(x)$ and similar result applies to the other case. After drawing two data sets of the specified sizes n and m, we replaced the last observation from density $h(x)$ by an integer from -9 to 11. The contamination rate is then $1/60$ and the α-IFs are calculated by averaging the following value over ten replications

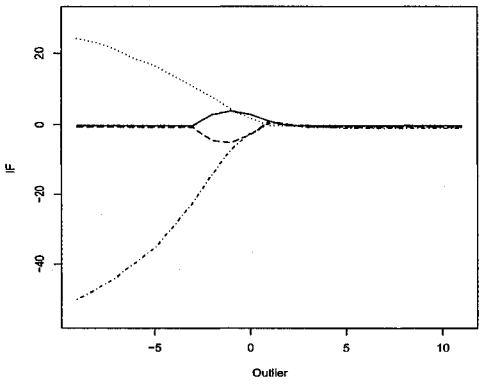
$$IF(x) = \frac{W\big((X_i)_{i=1}^n, (x, Z_i)_{i=1}^{m-1}\big) - W\big((X_i)_{i=1}^n, (Z_i)_{i=1}^m\big)}{1/60},$$

where W could be any functional (estimator of θ) based on data sets from $g(x)$ and $h(x)$, respectively. In our case W is either θ_N or $\widetilde{\theta}$. For the average of the ten replications, the α-IFs for different pairs (n, m) are displayed in Figure 3.1, which shows that θ_N is more robust than $\widetilde{\theta}$ in the sense of resistance to a single outlying observation.

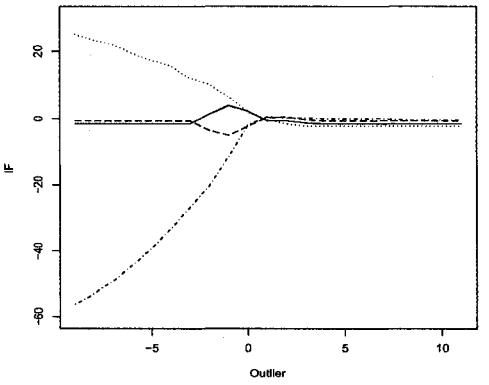
We can see from Figure 3.1 that as the outlier increases in its absolute value, the α-IFs of θ_N (solid and dashed lines) appear to converge to constants. In fact, the absolute values of the α-IFs of θ_N reach their peaks when outlying observation is around -1 and then slide down to the 0 baseline on both directions with a constant outside the interval $[-4, 4]$. For $\widetilde{\theta}$, however, when the outlying observation moves to the left from -1 , its α-IF increases dramatically in absolute value. When the outlier is bigger than -1, θ_N and $\widetilde{\theta}$ are competitive. The behavior of the α-IF of $\widetilde{\theta}$ could be expected from the fact that the semiparametric likelihood is proportional in some sense to the quantity $\prod_{i=1}^{m} \frac{\exp[\alpha+\beta Z_i]}{n+m\exp[\alpha+\beta Z_i]}$. Without an outlying observation, $\widetilde{\beta}$ should be a value around $\beta = 0.5$. When the outlying observation x is a positive large value, $\frac{\exp[\widetilde{\alpha}+\widetilde{\beta}x]}{n+m\exp[\widetilde{\alpha}+\widetilde{\beta}x]}$ is not an extremely small value and therefore $\widetilde{\beta}$ is not much affected. If x is a negative value with $|x|$ large enough, then $\frac{\exp[\widetilde{\alpha}+\widetilde{\beta}x]}{n+m\exp[\widetilde{\alpha}+\widetilde{\beta}x]}$ will be extremely small and hence the maximizing process will tend to assign $\widetilde{\beta}$ a negative value with a large absolute value. Therefore, when x is negative with $|x|$ large enough, the α-IF will be negative with large absolute values as shown in Figure 3.1.

3.5 An Example

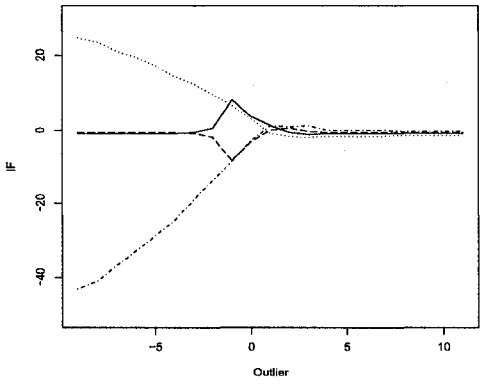
Hosmer and Lemeshow (1989) analyzed the relationship between age and coronary disease status. Table 1.1 in Hosmer and Lemeshow (1989) lists age in years (AGE), and presence or absence of evidence of significant coronary heart disease (CHD) for 100 subjects selected to participate in a study. The ourcome variable is CHD, which is coded with a value of 0 to indicate CHD is absent, or 1 to indicate that it is present in the individual. A summary of the data is also



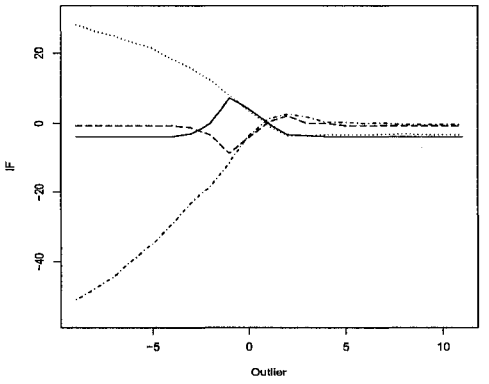
(a) $(n, m) = (10, 50)$



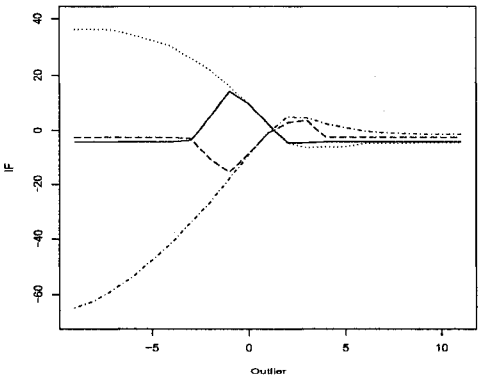
(b) $(n, m) = (20, 40)$



(c) $(n, m) = (30, 30)$



(d) $(n, m) = (40, 20)$



(e) $(n, m) = (50, 10)$

Fig. 3.1: The α-influence functions for $\widehat{\alpha}$ (solid), $\widehat{\beta}$ (dashed), $\widetilde{\alpha}$ (dotted) and $\widetilde{\beta}$ (dot-dashed) with respect to single outlier, where $\theta_N = (\widehat{\alpha}, \widehat{\beta})$ and $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta})$ are defined in (3.8) and Zhang (2000), respectively.

given below in Table 3.3.

Tab. 3.3: Age and coronary heart disease status (CHD) of 100 subjects.

AGE	CHD	AGE	CHD	AGE	CHD	AGE	CHD	AGE	CHD
20	0	34	0	41	0	48	1	57	0
23	0	34	0	42	0	48	1	57	1
24	0	34	1	42	0	49	0	57	1
25	0	34	0	42	0	49	0	57	1
25	1	34	0	42	1	49	1	57	1
26	0	35	0	43	0	50	0	58	0
26	0	35	0	43	0	50	1	58	1
28	0	36	0	43	1	51	0	58	1
28	0	36	1	44	0	52	0	59	1
29	0	36	0	44	0	52	1	59	1
30	0	37	0	44	1	53	1	60	0
30	0	37	1	44	1	53	1	60	1
30	0	37	0	45	0	54	1	61	1
30	0	38	0	45	1	55	0	62	1
30	0	38	0	46	0	55	1	62	1
30	1	39	0	46	1	55	1	63	1
32	0	39	1	47	0	56	1	64	0
32	0	40	0	47	0	56	1	64	1
33	0	40	1	47	1	56	1	65	1
33	0	41	0	48	0	57	0	69	1

They analyzed the relationship between AGE and CHD based on those 100 subjects by employing the logistic regression model (3.2). Let X denote the age and $Y = 1$ or 0 represent the presence or absence of coronary heart disease. Then the sample data (X_i, Y_i), $i = 1, \ldots, 100$, can be thought of as being drawn independently and identically from the joint distribution of (X, Y). The proposed MHD estimate can be applied to this data set with $n = 57$ and $m = 43$. We again take the bandwidths $h_n = n^{-2/5}$ and $h_m = m^{-2/5}$ and use Epanechnikov kernel function defined in (3.30) for the two kernels K_0 and K_1 in (3.4) and (3.5), respectively. By fitting model (3.1), we obtained estimates $\theta_N = (\widehat{\alpha}, \widehat{\beta}) = (-4.64, 0.09)$. When compared with Zhang's (2000) estimates, $(\widetilde{\alpha}, \widetilde{\beta}) = (-5.03, 0.11)$, our estimates seem more conservative; in other words, our estimates are smaller in absolute values than Zhang's (2000) estimates.

3.6 Proof of Asymptotic Normality

To prove Theorem 3.4, we first state a series of lemmas that are employed in the proof.

Lemma 3.4. *Suppose that (C3)-(C6) hold. Then as $N \to \infty$,*

$$N^{1/2} \int \varepsilon_N(x) \exp[\frac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) h_m^{1/2}(x) dx \xrightarrow{P} 0, \tag{3.31}$$

$$N^{1/2} \int \varepsilon_N(x) \exp[\frac{1}{2}(1, r(x))\theta] h_\theta^{1/2}(x) g_n^{1/2}(x) dx \xrightarrow{P} 0. \tag{3.32}$$

Proof. By Cauchy-Schwarz Inequality,

$$\begin{aligned} & N \cdot E\Big[\int \varepsilon_{Ni}(x) \exp[\frac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) h_m^{1/2}(x) dx\Big]^2 \\ \leq\ & N \cdot E\Big[\int \varepsilon_{Ni}^2(x) \exp[(1, r(x))\theta] g_n(x) dx\Big] \cdot E\Big[\int I_{\{|x|>\alpha_N\}} h_m(x) dx\Big] \\ =\ & N \cdot \Delta_1 \cdot \Delta_2, \quad \text{say.} \end{aligned}$$

Note that by a Taylor expansion and using assumptions (C4) and (C5)

$$\begin{aligned} |\Delta_1| &= \int\int \varepsilon_{Ni}^2(x) \exp[(1, r(x))\theta] \frac{1}{b_n} K_0(\frac{y-x}{b_n}) g(y) dy\, dx \\ &= \int \varepsilon_{Ni}^2(x) \exp[(1, r(x))\theta] \int_{-a_0}^{a_0} K_0(t) g(x + tb_n) dt\, dx \\ &= \int \varepsilon_{Ni}^2(x) \exp[(1, r(x))\theta] \int_{-a_0}^{a_0} K_0(t) \big(g(x) + g^{(1)}(x) t b_n + \\ &\qquad \frac{1}{2} g^{(2)}(\xi) t^2 b_n^2\big) dt\, dx \\ &\leq \int r_i^2(x) h_\theta(x) dx \\ &\quad + \frac{1}{2} b_n^2 \int \varepsilon_{Ni}^2(x) h_\theta(x) \sup_{|t| \leq a_0} \frac{|g^{(2)}(x + tb_n)|}{g(x)} dx \int_{-a_0}^{a_0} t^2 K_0(t) dt \\ &= O(1), \end{aligned}$$

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

$$\begin{aligned} |\Delta_2| &= \int\int I_{\{|x|>\alpha_N\}} \frac{1}{b_m} K_1(\frac{y-x}{b_m}) h_\theta(y) dy\, dx \\ &= \int\int I_{\{|x|>\alpha_N\}} K_1(t) h_\theta(x + tb_m) dt\, dx \\ &= \int_{-a_1}^{a_1} K_1(t) \int_{|z - tb_m| > \alpha_N} h_\theta(z) dz\, dt \\ &\leq \int_{-a_1}^{a_1} K_1(t) dt \int_{|z| > \alpha_N - a_1 b_m} h_\theta(z) dz \\ &= P(|Z_1| > \alpha_N - a_1 b_m). \end{aligned} \tag{3.33}$$

By assumption (C6) we have that

$$N \cdot E\Big[\int \varepsilon_{Ni}(x)\exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)h_m^{1/2}(x)dx\Big]^2 \to 0,$$

i.e., (3.31) holds.

By Cauchy-Schwarz Inequality and using a similar argument as in (3.33),

$$\begin{aligned}
& N \cdot E\Big[\int \varepsilon_{Ni}(x)\exp[\frac{1}{2}(1,r(x))\theta]h_\theta^{1/2}(x)g_n^{1/2}(x)dx\Big]^2 \\
\le\ & N \cdot \int r_i^2(x)\exp[(1,r(x))\theta]h_\theta(x)dx \cdot E\Big[\int I_{\{|x|>\alpha_N\}}g_n(x)dx\Big] \\
=\ & N \cdot \int r_i^2(x)\exp[(1,r(x))\theta]h_\theta(x)dx \cdot \int\int I_{\{|x|>\alpha_N\}}\frac{1}{b_n}K_0(\frac{y-x}{b_n})g(y)dy\,dx \\
\le\ & N \cdot \int r_i^2(x)\exp[(1,r(x))\theta]h_\theta(x)dx \cdot P(|X_1| > \alpha_N - a_0b_n),
\end{aligned}$$

and by assumptions (C4) and (C6) we have that (3.32) holds. □

Lemma 3.5. *Suppose that (C0)-(C3) and (C7) hold. Then as $N \to \infty$,*

$$N^{1/2}\int |\delta_N(x)|\big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big)^2dx \xrightarrow{P} 0, \tag{3.34}$$

$$N^{1/2}\int |\delta_N(x)|\exp[(1,r(x))\theta]\big(g_n^{1/2}(x) - g^{1/2}(x)\big)^2dx \xrightarrow{P} 0. \tag{3.35}$$

Proof. Note that

$$\begin{aligned}
& N^{1/2}\int |\delta_N(x)|\big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big)^2dx \\
\le\ & N^{1/2}\int |\delta_N(x)|h_\theta^{-1}(x)\big(h_m(x) - h_\theta(x)\big)^2dx \\
\le\ & 2\Big[N^{1/2}\int |\delta_N(x)|h_\theta^{-1}(x)\big(h_m(x) - Eh_m(x)\big)^2dx \\
& \qquad +N^{1/2}\int |\delta_N(x)|h_\theta^{-1}(x)\big(Eh_m(x) - h_\theta(x)\big)^2dx\Big] \\
=\ & 2(A_{1N} + A_{2N}), \quad \text{say.}
\end{aligned}$$

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

$$\begin{aligned}
E|A_{1N}| &= N^{1/2}\int |\delta_N(x)|h_\theta^{-1}(x)E\big(h_m(x) - Eh_m(x)\big)^2dx \\
&\le N^{1/2}\int |\delta_N(x)|h_\theta^{-1}(x)\,\frac{1}{mb_m^2}\int K_1^2(\frac{y-x}{b_m})h_\theta(y)dy\,dx \\
&= N^{1/2}m^{-1}b_m^{-1}\int |\delta_N(x)|\int_{-a_1}^{a_1} K_1^2(t)h_\theta(x+tb_m)h_\theta^{-1}(x)dt\,dx
\end{aligned}$$

$$\begin{aligned} &\leq N^{1/2}m^{-1}b_m^{-1}\int |\delta_N(x)| \sup_{|t|\leq a_1} \frac{h_\theta(x+tb_m)}{h_\theta(x)}dx \int_{-a_1}^{a_1} K_1^2(t)dt \\ &\to 0, \end{aligned}$$

i.e., $A_{1N} \xrightarrow{P} 0$ as $N \to \infty$. By a Taylor expansion and using conditions (C1) and (C7),

$$\begin{aligned} |A_{2N}| &= N^{1/2}\int |\delta_N(x)|h_\theta^{-1}(x)\Big[\int_{-a_1}^{a_1} K_1(t)\big(h_\theta(x+tb_m)-h_\theta(x)\big)dt\Big]^2 dx \\ &\leq \frac{1}{4}N^{1/2}b_m^4 \int |\delta_N(x)|h_\theta^{-1}(x)\Big[\sup_{|t|\leq a_1}|h_\theta^{(2)}(x+tb_m)|\int_{-a_1}^{a_1} t^2K_1(t)dt\Big]^2 dx \\ &\leq \frac{1}{4}N^{1/2}b_m^4 \int |\delta_N(x)|h_\theta(x) \sup_{|t|\leq a_1}\Big[\frac{h_\theta^{(2)}(x+tb_m)}{h_\theta(x)}\Big]^2 dx \Big(\int_{-a_1}^{a_1} t^2K_1(t)dt\Big)^2 \\ &\to 0. \end{aligned}$$

Hence (3.34) holds. Proof of (3.35) is similar to that of (3.34). □

Lemma 3.6. *Suppose that (C0)-(C7) hold. Then the asymptotic distribution of*

$$N^{1/2}\int (1,r(x))^T \exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)dx \tag{3.36}$$

is the same as that of

$$N^{1/2}\int \delta_N(x)h_\theta^{1/2}(x)\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)dx.$$

Proof. From Lemma 3.4,

$$N^{1/2}\int \varepsilon_N(x)\exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)dx \xrightarrow{P} 0,$$

and as a result the asymptotic distribution of (3.36) is the same as that of

$$N^{1/2}\int \delta_N(x)\exp[\frac{1}{2}(1,r(x))\theta]g_n^{1/2}(x)\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)dx.$$

By Cauchy-Schwarz Inequality

$$\begin{aligned} &\Big\{N^{1/2}\int \delta_{Ni}(x)\exp[\frac{1}{2}(1,r(x))\theta]\big(g_n^{1/2}(x)-g^{1/2}(x)\big)\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)dx\Big\}^2 \\ \leq\ & N^{1/2}\int |\delta_{Ni}(x)|\exp[(1,r(x))\theta]\big(g_n^{1/2}(x)-g^{1/2}(x)\big)^2dx \\ &\qquad \times N^{1/2}\int |\delta_{Ni}(x)|\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)^2dx, \end{aligned}$$

which is $o_P(1)$ by Lemma 3.5. Hence the result. □

Remark 3.10. In fact, the asymptotic distribution of (3.36) is the same as that of

$$N^{1/2}\int (1, r(x))^T h_\theta^{1/2}(x)\big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big)dx.$$

The reason being that as $N \to \infty$,

$$N^{1/2}\int \varepsilon_N(x) h_\theta^{1/2}(x)\big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big)dx \xrightarrow{P} 0$$

under conditions (C3), (C4) and (C6). The proof is similar to that of Lemma 3.4 and therefore be omitted.

Remark 3.11. Instead of condition (C7), if h_θ and g have bounded second derivatives and conditions (C9) and (C11) hold, then Lemma 3.6 still holds. Since

$$\begin{aligned}
&\Big\{N^{1/2}\int \delta_{Ni}(x)\exp[\frac{1}{2}(1, r(x))\theta]\big(g_n^{1/2}(x) - g^{1/2}(x)\big)\big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big)dx\Big\}^2 \\
\le\ & N^{1/2}\int |\delta_{Ni}(x)|\exp[\frac{1}{2}(1, r(x))\theta]\big(g_n^{1/2}(x) - g^{1/2}(x)\big)^2 dx \\
&\times N^{1/2}\int |\delta_{Ni}(x)|\exp[\frac{1}{2}(1, r(x))\theta]\big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big)^2 dx,
\end{aligned}$$

similar arguments as in the proof of Lemmas 3.5 and 3.6 give above conclusion.

Lemma 3.7. *Suppose that (C4) and (C6) hold. Then as $N \to \infty$,*

$$N^{1/2}\int |\varepsilon_N(x)| h_\theta(x)dx \to 0,$$

$$N^{1/2}\cdot\frac{1}{m}\sum_{i=1}^{m}\varepsilon_N(Z_i) \xrightarrow{P} 0,$$

$$N^{1/2}\cdot\frac{1}{n}\sum_{i=1}^{n}\varepsilon_N(X_i)\exp[(1, r(X_i))\theta] \xrightarrow{P} 0.$$

Proof. By Cauchy-Schwarz Inequality,

$$\begin{aligned}
N^{1/2}\int |\varepsilon_{Ni}(x)| h_\theta(x)dx &\le \big[N\int I_{\{|x|>\alpha_N\}} h_\theta(x)dx\big]^{1/2}\big[\int r_i^2(x) h_\theta(x)dx\big]^{1/2} \\
&= \big[NP(|Z_1| > \alpha_N)\big]^{1/2}\big[\int r_i^2(x) h_\theta(x)dx\big]^{1/2} \\
&\to 0.
\end{aligned}$$

As a result,

$$\begin{aligned} E\big|N^{1/2}\cdot\frac{1}{m}\sum_{i=1}^{m}\varepsilon_N(Z_i)\big| &\le E\big[N^{1/2}\cdot\frac{1}{m}\sum_{i=1}^{m}|\varepsilon_N(Z_i)|\big] \\ &= N^{1/2}\int|\varepsilon_N(x)|h_\theta(x)dx \\ &\to 0, \end{aligned}$$

$$\begin{aligned} & E\big|N^{1/2}\cdot\frac{1}{n}\sum_{i=1}^{n}\varepsilon_N(X_i)\exp[(1,r(X_i))\theta]\big| \\ \le\ & E\big[N^{1/2}\cdot\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_N(X_i)|\exp[(1,r(X_i))\theta]\big] \\ =\ & N^{1/2}\int|\varepsilon_N(x)|h_\theta(x)dx \\ \to\ & 0, \end{aligned}$$

and hence the results. □

Lemma 3.8. *Suppose that (C0)-(C4) and (C8)-(C10) hold. Then as* $N\to\infty$,

$$N^{1/2}\int\delta_N(x)h_m(x)dx - N^{1/2}\frac{1}{m}\sum_{i=1}^{m}\delta_N(Z_i)\xrightarrow{P}0,$$

$$N^{1/2}\int\delta_N(x)\exp[(1,r(x))\theta]g_n(x)dx - N^{1/2}\frac{1}{n}\sum_{i=1}^{n}\delta_N(X_i)\exp[(1,r(X_i))\theta]\xrightarrow{P}0.$$

Proof. We give only the proof for the second convergence, and the proof for the first convergence is similar. For $i=0,1,\ldots,p$, let

$$D_{Ni}=N^{1/2}\int\delta_{Ni}(x)\exp[(1,r(x))\theta]g_n(x)dx-N^{1/2}\frac{1}{n}\sum_{i=1}^{n}\delta_{Ni}(X_i)\exp[(1,r(X_i))\theta].$$

Then by (C8)

$$\begin{aligned} \big|E[D_{Ni}]\big| &= N^{1/2}\Big|\int\delta_{Ni}(x)\exp[(1,r(x))\theta]E[g_n(x)]dx-\int\delta_{Ni}(x)h_\theta(x)dx\Big| \\ &= N^{1/2}\Big|\int\delta_{Ni}(x)\exp[(1,r(x))\theta]\int_{-a_0}^{a_0}K_0(t)\big(g(x+tb_n)-g(x)\big)dt\,dx\Big| \\ &\le N^{1/2}b_n^2\int|\delta_{Ni}(x)|h_\theta(x)\sup_{|t|\le a_0}\frac{|g^{(2)}(x+tb_n)|}{g(x)}dx\int_{-a_0}^{a_0}t^2K_0(t)dt \\ &\to 0. \end{aligned}$$

Note that

$$
\begin{aligned}
& Var[D_{Ni}] \\
\leq\ & \frac{N}{n}E\Big[\int \delta_{Ni}(x)\exp[(1,r(x))\theta]\frac{1}{b_n}K_0\big(\frac{x-X_1}{b_n}\big)dx - \delta_{Ni}(X_1)\exp[(1,r(X_1))\theta]\Big]^2 \\
=\ & \frac{N}{n}E\Big[\int_{-a_0}^{a_0} K_0(t)\Big(\delta_{Ni}(X_1+tb_n)\exp[(1,r(X_1+tb_n))\theta] \\
& \qquad -\delta_{Ni}(X_1)\exp[(1,r(X_1))\theta]\Big)dt\Big]^2 \\
=\ & \frac{N}{n}E\Big[\int_{-a_0}^{a_0} K_0(t)r_i(X_1+tb_n)\exp[(1,r(X_1+tb_n))\theta]\Big(I_{\{|X_1+tb_n|\leq\alpha_N\}} \\
& -I_{\{|X_1|\leq\alpha_N\}}\Big)dt + \int_{-a_0}^{a_0} K_0(t)I_{\{|X_1|\leq\alpha_N\}}\Big(r_i(X_1+tb_n)\exp[(1,r(X_1+tb_n))\theta] \\
& \qquad -r_i(X_1)\exp[(1,r(X_1))\theta]\Big)dt\Big]^2 \\
\leq\ & \frac{2N}{n}\Big\{E\Big[\int_{-a_0}^{a_0} K_0(t)r_i(X_1+tb_n)\exp[(1,r(X_1+tb_n))\theta]\Big(I_{\{|X_1+tb_n|\leq\alpha_N\}} \\
& -I_{\{|X_1|\leq\alpha_N\}}\Big)dt\Big]^2 + E\Big[\int_{-a_0}^{a_0} K_0(t)I_{\{|X_1|\leq\alpha_N\}}\Big(r_i(X_1+tb_n)\cdot \\
& \qquad \exp[(1,r(X_1+tb_n))\theta] - r_i(X_1)\exp[(1,r(X_1))\theta]\Big)dt\Big]^2\Big\} \\
=\ & \frac{2N}{n}(B_{Ni}+C_{Ni}), \quad \text{say.}
\end{aligned}
$$

By Cauchy-Schwarz Inequality,

$$
\begin{aligned}
B_{Ni} \leq\ & E\int_{-a_0}^{a_0} K_0(t)r_i^2(X_1+tb_n)\exp[2(1,r(X_1+tb_n))\theta]\big(I_{\{|X_1+tb_n|\leq\alpha_N\}} \\
& \qquad -I_{\{|X_1|\leq\alpha_N\}}\big)^2dt \\
=\ & \int_0^{a_0} K_0(t)\Big[\int_{-\alpha_N-tb_n}^{-\alpha_N} r_i^2(y+tb_n)\exp[2(1,r(y+tb_n))\theta]g(y)dy \\
& \qquad + \int_{\alpha_N-tb_n}^{\alpha_N} r_i^2(y+tb_n)\exp[2(1,r(y+tb_n))\theta]g(y)dy\Big]dt \\
& + \int_{-a_0}^{0} K_0(t)\Big[\int_{-\alpha_N}^{-\alpha_N-tb_n} r_i^2(y+tb_n)\exp[2(1,r(y+tb_n))\theta]g(y)dy \\
& \qquad + \int_{\alpha_N}^{\alpha_N-tb_n} r_i^2(y+tb_n)\exp[2(1,r(y+tb_n))\theta]g(y)dy\Big]dt.
\end{aligned} \tag{3.37}
$$

Note that $r_i^2(x)\exp[(1,r(x))\theta]h_\theta(x)$ is bounded by (C4) and therefore by (C9)

$$
\begin{aligned}
& \int_0^{a_0} K_0(t)\int_{-\alpha_N-tb_n}^{-\alpha_N} r_i^2(y+tb_n)\exp[2(1,r(y+tb_n))\theta]g(y)dy \\
=\ & \int_0^{a_0} K_0(t)\int_{-\alpha_N}^{-\alpha_N+tb_n} r_i^2(y)\exp[2(1,r(y))\theta]g(y-tb_n)dy\,dt
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sup_{|x|\leq\alpha_N}\sup_{|t|\leq a_0}\frac{g(x+tb_n)}{g(x)}\int_0^{a_0}K_0(t)\int_{-\alpha_N}^{-\alpha_N+tb_n} r_i^2(y)\exp[(1,r(y))\theta]h_\theta(y)dy\,dt\\
&= O\big(b_n\int_0^{a_0} tK_0(t)dt\big)\\
&\to 0,
\end{aligned}
$$

as $N\to\infty$, and other three terms on the r.h.s. of (3.37) go to zero using similar arguments. Thus $B_{Ni}\to 0$ as $N\to\infty$. For C_{Ni}, by Cauchy-Schwarz inequality and (C10) we have

$$
\begin{aligned}
C_{Ni} \;\leq\; & E\Big[\int_{-a_0}^{a_0}K_0(t)I_{\{|X_1|\leq\alpha_N\}}\Big(r_i(X_1+tb_n)\exp[(1,r(X_1+tb_n))\theta]\\
&\qquad -r_i(X_1)\exp[(1,r(X_1))\theta]\Big)^2dt\Big]\\
=\;& \int_{-a_0}^{a_0}K_0(t)\int I_{\{|x|\leq\alpha_N\}}\Big(r_i(x+tb_n)\exp[(1,r(x+tb_n))\theta]\\
&\qquad -r_i(x)\exp[(1,r(x))\theta]\Big)^2 g(x)dx\,dt\\
\leq\;& b_n^2\int I_{\{|x|\leq\alpha_N\}}g(x)\sup_{|t|\leq a_0}\Big[\frac{\partial r_i(y)\exp[(1,r(y))\theta]}{\partial y}|_{y=x+tb_n}\Big]^2dx\int_{-a_0}^{a_0}t^2K_0(t)dt\\
\to\;& 0.
\end{aligned}
$$

Thus $Var[D_{Ni}]\to 0$ as $N\to\infty$. This yields that $E[D_{Ni}^2]=Var[D_{Ni}]+(E[D_{Ni}])^2\to 0$, and therefore $D_{Ni}\xrightarrow{P}0$ as $N\to\infty$. □

Corollary 3.1. *Suppose that (C0)-(C10) hold. Then the asymptotic distribution of (3.36) is* $N\big(0,\frac{1}{4(1-\rho)}\Sigma_1\big)$ *with* Σ_1 *defined by (3.29).*

Proof. In view of Lemma 3.6, we only need to give the asymptotic distribution of
$N^{1/2}\int\delta_N(x)h_\theta^{1/2}(x)\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)dx$. Applying the following algebraic expression, with $b\geq 0$, $a>0$,

$$
b^{1/2}-a^{1/2}=\frac{b-a}{2a^{1/2}}-\frac{\big(b^{1/2}-a^{1/2}\big)^2}{2a^{1/2}}, \tag{3.38}
$$

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

$$
\begin{aligned}
&N^{1/2}\int\delta_N(x)h_\theta^{1/2}(x)\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)dx\\
=\;&\frac{1}{2}N^{1/2}\int\delta_N(x)\big(h_m(x)-h_\theta(x)\big)dx+\frac{1}{2}N^{1/2}\int\delta_N(x)\big(h_m^{1/2}(x)-h_\theta^{1/2}(x)\big)^2dx\\
=\;&\frac{1}{2}N^{1/2}\int\delta_N(x)\big(h_m(x)-h_\theta(x)\big)dx+o_P(1)\;\;\text{(by Lemma 3.5)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}N^{1/2}\Big[\frac{1}{m}\sum_{i=1}^{m}\delta_N(Z_i) - \int \delta_N(x)h_\theta(x)dx\Big] \\
&\qquad + \frac{1}{2}N^{1/2}\Big[\int \delta_N(x)h_m(x)dx - \frac{1}{m}\sum_{i=1}^{m}\delta_N(Z_i)\Big] + o_P(1) \\
&= \frac{1}{2}N^{1/2}\Big[\frac{1}{m}\sum_{i=1}^{m}\delta_N(Z_i) - \int \delta_N(x)h_\theta(x)dx\Big] + o_P(1) \quad \text{(by Lemma 3.8)} \\
&= \frac{1}{2}N^{1/2}\Big[\frac{1}{m}\sum_{i=1}^{m}(1, r(Z_i))^T - \int (1, r(x))^T h_\theta(x)dx\Big] + o_P(1) \quad \text{(by Lemma 3.7).}
\end{aligned}
$$

Obviously the asymptotic distribution of $m^{1/2}\big[\frac{1}{m}\sum_{i=1}^{m}(1, r(Z_i))^T - \int (1, r(x))^T \cdot h_\theta(x)dx\big]$ is $N\big(0, \Sigma_1\big)$. Hence the result. □

Lemma 3.9. *Suppose that (C0)-(C7) and (C5') hold. Then the asymptotic distribution of*

$$
N^{1/2}\int (1, r(x))^T \exp[(1, r(x))\theta] g_n^{1/2}(x)\big(g_n^{1/2}(x) - g^{1/2}(x)\big)dx \tag{3.39}
$$

is the same as that of

$$
N^{1/2}\int \delta_N(x)\exp[(1, r(x))\theta] g^{1/2}(x)\big(g_n^{1/2}(x) - g^{1/2}(x)\big)dx.
$$

Proof. Note that by Cauchy-Schwarz Inequality, a Taylor expansion, (C5') and Lemma 3.7,

$$
\begin{aligned}
& E\big|N^{1/2}\int \varepsilon_{Ni}(x)\exp[(1, r(x))\theta] g_n(x)dx\big| \\
\le\ & N^{1/2}\int |\varepsilon_{Ni}(x)|\exp[(1, r(x))\theta]\int_{-a_0}^{a_0} K_0(t)g(x + tb_n)dt\,dx \\
\le\ & N^{1/2}\int |\varepsilon_{Ni}(x)|\exp[(1, r(x))\theta]\int_{-a_0}^{a_0} K_0(t)\big(g(x) + g^{(1)}(x)tb_n \\
& \qquad + \frac{1}{2}t^2 b_n^2 \sup_{|t|\le a_0}|g^{(2)}(x + tb_n)|\big)dt\,dx \\
\le\ & N^{1/2}\int |\varepsilon_{Ni}(x)|h_\theta(x)dx \\
& \qquad + \frac{1}{2}N^{1/2}b_n^2\int |\varepsilon_{Ni}(x)|h_\theta(x)\sup_{|t|\le a_0}\frac{|g^{(2)}(x + tb_n)|}{g(x)}\int_{-a_0}^{a_0} t^2 K_0(t)dt \\
\to\ & 0.
\end{aligned}
$$

Thus $N^{1/2}\int \varepsilon_N(x)\exp[(1, r(x))\theta] g_n(x)dx \xrightarrow{P} 0$. Combined with the result in

Lemma 3.4, we therefore have

$$N^{1/2}\int \varepsilon_N(x)\exp[(1,r(x))\theta]g_n^{1/2}(x)\big(g_n^{1/2}(x)-g^{1/2}(x)\big)dx \xrightarrow{P} 0,$$

and so the asymptotic distribution of (3.39) is the same as that of

$$N^{1/2}\int \delta_N(x)\exp[(1,r(x))\theta]g_n^{1/2}(x)\big(g_n^{1/2}(x)-g^{1/2}(x)\big)dx \xrightarrow{P} 0.$$

The result now follows from Lemma 3.5. □

Corollary 3.2. *Suppose that (C0)-(C10) and (C5') hold. Then the asymptotic distribution of (3.39) is* $N\big(0,\frac{1}{4\rho}\Sigma_0\big)$ *with* Σ_0 *defined by (3.28).*

Proof. Similar to that of Corollary 3.1.

Again in view of Lemma 3.9, we only need to give the asymptotic distribution of $N^{1/2}\int \delta_N(x)\exp[(1,r(x))\theta]g^{1/2}(x)\big(g_n^{1/2}(x)-g^{1/2}(x)\big)dx$. Applying the algebraic expression (3.38) we have that as $N\to\infty$,

$$\begin{aligned}
& N^{1/2}\int \delta_N(x)\exp[(1,r(x))\theta]g^{1/2}(x)\big(g_n^{1/2}(x)-g^{1/2}(x)\big)dx \\
= \;& \frac{1}{2}N^{1/2}\int \delta_N(x)\exp[(1,r(x))\theta]\big(g_n(x)-g(x)\big)dx \\
& \qquad +\frac{1}{2}N^{1/2}\int \delta_N(x)\exp[(1,r(x))\theta]\big(g_n^{1/2}(x)-g^{1/2}(x)\big)^2dx \\
= \;& \frac{1}{2}N^{1/2}\int \delta_N(x)\exp[(1,r(x))\theta]\big(g_n(x)-g(x)\big)dx+o_P(1) \quad \text{(by Lemma 3.5)} \\
= \;& \frac{1}{2}N^{1/2}\Big\{\frac{1}{n}\sum_{i=1}^n \delta_N(X_i)\exp[(1,r(X_i))\theta]-\int \delta_N(x)h_\theta(x)dx\Big\} \\
& +\frac{1}{2}N^{1/2}\Big\{\int \delta_N(x)\exp[(1,r(x))\theta]g_n(x)dx-\frac{1}{n}\sum_{i=1}^n \delta_N(X_i)\exp[(1,r(X_i))\theta]\Big\} \\
& +o_P(1) \\
= \;& \frac{1}{2}N^{1/2}\Big\{\frac{1}{n}\sum_{i=1}^n \delta_N(X_i)\exp[(1,r(X_i))\theta]-\int \delta_N(x)h_\theta(x)dx\Big\}+o_P(1) \\
& \qquad \text{(by Lemma 3.8)} \\
= \;& \frac{1}{2}N^{1/2}\Big\{\frac{1}{n}\sum_{i=1}^n (1,r(X_i))^T\exp[(1,r(X_i))\theta]-\int (1,r(x))^T h_\theta(x)dx\Big\}+o_P(1) \\
& \qquad \text{(by Lemma 3.7).}
\end{aligned}$$

Obviously the asymptotic distribution of $n^{1/2}\big[\frac{1}{n}\sum_{i=1}^n (1,r(X_i))^T\exp[(1,r(X_i))\theta]-\int(1,r(x))^T h_\theta(x)dx\big]$ is $N\big(0,\Sigma_0\big)$. Hence the result. □

Proof of Theorem 3.4. Note that by Lemmas 3.6 and 3.9

$$
\begin{aligned}
& N^{1/2} \int \Big\{ \exp[\frac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) h_m^{1/2}(x) - \exp[(1, r(x))\theta] g_n(x) \Big\} (1, r(x))^T dx \\
= \; & N^{1/2} \int (1, r(x))^T \exp[\frac{1}{2}(1, r(x))\theta] g_n^{1/2}(x) \big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big) dx \\
& \qquad - N^{1/2} \int (1, r(x))^T \exp[(1, r(x))\theta] g_n^{1/2}(x) \big(g_n^{1/2}(x) - g^{1/2}(x)\big) dx \\
= \; & N^{1/2} \int \delta_N(x) h_\theta^{1/2}(x) \big(h_m^{1/2}(x) - h_\theta^{1/2}(x)\big) dx \\
& \qquad - N^{1/2} \int \delta_N(x) \exp[(1, r(x))\theta] g^{1/2}(x) \big(g_n^{1/2}(x) - g^{1/2}(x)\big) dx + o_P(1)
\end{aligned}
$$

and the first two terms on the r.h.s. of the preceding expression are independent. By Corollaries 3.1 and 3.2 and Slutsky's theorem, the result follows. □

CHAPTER FOUR: MHD ESTIMATION IN SEMIPARAMETRIC MODELS OF GENERAL FORM

4.1 Introduction

Consider the situation where we observe a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$ with continuous density function f. Assume that f belongs to a class of general semiparametric models of the form

$$\{f_{\theta,\eta} : \theta \in \Theta \subseteq \mathbb{R}^p, \eta \in \mathcal{H}\}, \tag{4.1}$$

where Θ is a compact subset of $\mathbb{R}^p$ and $\mathcal{H}$ is an arbitrary set of infinite dimension. The problem is to estimate the parameter θ assuming that η as a nuisance parameter. The support of density functions may be finite or infinite in the Euclidean space, unless otherwise specified.

Numerous examples fall into the class (4.1), well-known examples include semiparametric mixture models (Van der Vaart, 1996), errors-in-variables models (Bickel and Ritov, 1987 and Murphy and Van der Vaart, 1996), regression models (Van der Vaart, 1998) and Cox model for survival analysis (Cox, 1972). More examples and theory can be found in the monographs of Pfzangel (1990), Bickel et al. (1993) and Van der Vaart (1998) and in the articles of Murphy and Van der Vaart (2000), Bickel and Kwon (2001), and Forrester et al. (2003) and in the references therein. The two-component mixture model and the two-sample model considered in Chapters 2 and 3, respectively, are two special cases of general semiparametric models (4.1).

If η is known, then θ can be easily estimated using the maximum likelihood approach. If η is unknown, then replacing η by an appropriate estimator the maximum likelihood approach still may be implemented; see, e.g., Van der Vaart (1998, Section 25.8). These estimators are usually asymptotically efficient, but may perform poorly if the parametric assumption is slightly violated. Applications of MHD estimators in the two semiparametric models considered in Chapters 2 and 3 suggest that MHD estimators have good efficiency and robustness properties in semiparametric models. In this chapter, we investigate the efficiency and robustness of MHD estimators in semiparametric models (4.1) of general form.

In a parametric class of density functions of the form

$$\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}, \tag{4.2}$$

a MHD estimator of θ is defined as a functional $T_0(g) = T(\{f_t\}_{t\in\Theta}, g)$ at f_n (Beran, 1977) such that

$$T_0(f_n) = T(\{f_t\}_{t\in\Theta}, f_n) = \arg\min_{t\in\Theta} \|f_t^{1/2} - f_n^{1/2}\|, \tag{4.3}$$

where f_n is a nonparametric density estimator of f based on the observations $X_1, X_2, \ldots, X_n$. Various asymptotic and robustness properties of $T_0(f_n)$ have been studied under some regularity conditions in Beran (1977), Stather (1981) and Tamura and Boos (1986), among others. MHD estimators in semiparametric models have not been yet obtained in the literature.

In this chapter, we extend the Hellinger distance approach to general semiparametric models (4.1). Roughly speaking, a MHD estimator of θ in semiparametric models (4.1) can be defined as

$$\theta_n = T_n(f_n) = T(\{f_{t,\eta_n}\}_{t\in\Theta}, f_n) = \arg\min_{t\in\Theta} \|f_{t,\eta_n}^{1/2} - f_n^{1/2}\|, \tag{4.4}$$

where η_n is a suitable estimator of η. Alternatively, one could also construct an estimator of θ as

$$T_1(f_n) = T(\{f_{t,h}\}_{t\in\Theta,h\in\mathcal{H}}, f_n) = \arg\min_{t} \min_{t\in\Theta,h\in\mathcal{H}} \|f_{t,h}^{1/2} - f_n^{1/2}\|,$$

which we will call a minimum *profile* Hellinger distance (MPHD) estimator. Both types of these estimators will be investigated in this chapter. The main question is whether or not the proposed estimators retained any of the desirable properties of MHD in fully parametric models. In particular, we wish to examine the following important questions. Are the proposed estimators consistent and asymptotically normal? Do they possess similar efficiency properties as in the parametric case? Are the proposed estimators still robust? What about other properties such as adaptivity? How does the presence of nuisance parameter affect the overall process of construction and efficiency? Clearly, it is of theoretical and practical interest to investigate above issues. The main purpose of this chapter is to attempt to answer these questions systematically. This chapter is organized as follows. Sections 4.2 and 4.3 discuss the efficiency of the MHD estimator (4.4) in parametric and semiparametric senses, respectively. Minimum profile Hellinger distance (MPHD) estimator is constructed in Section 4.4. Section 4.5 studies robustness properties of the estimator (4.4). Simulation studies, examples and concluding remarks are given in Sections 4.6, 4.7 and 4.8, respectively.

4.2 *Efficiency in the Parametric Sense*

In this section, we first give a general result on the asymptotic efficiency of MHD estimators in the parametric family (4.2). Beran (1977) has shown that the MHD estimator defined by (4.3) is efficient and robust, at least for continuous distributions with compact support. Stather (1981) extended Beran's results to the case of discrete distributions and continuous distributions with infinite support. Brown and Hwang (1993) examined a MHD estimator using a histogram type estimator for f_n. Tamura and Boos (1986) considered MHD estimators for multivariate location and scale models. In the next theorem we obtain the efficiency of the MHD estimator defined by (4.3) in a more general sense; i.e., without assuming any specific form of f_n. Let $s_\theta = f_\theta^{1/2}$ and suppose for $\theta \in \Theta$, there exist a $p \times 1$ vector $\dot{s}_\theta(x)$ with components in L_2 and a $p \times p$ matrix $\ddot{s}_\theta(x)$ with components in L_2 such that for every $p \times 1$ real vector e of unit Euclidean length and for every scalar α in a neighborhood of zero,

$$s_{\theta+\alpha e}(x) = s_\theta(x) + \alpha e^T \dot{s}_\theta(x) + \alpha e^T u_\alpha(x) \tag{4.5}$$

$$\dot{s}_{\theta+\alpha e}(x) = \dot{s}_\theta(x) + \alpha \ddot{s}_\theta(x)e + \alpha v_\alpha(x)e, \tag{4.6}$$

where $u_\alpha(x)$ is $p \times 1$, $v_\alpha(x)$ is $p \times p$, and the components of u_α and v_α tend to zero in L_2 as $\alpha \to 0$. The family $\{f_\theta : \theta \in \Theta\}$ is called identifiable if $\theta_1 \neq \theta_2$ implies $f_{\theta_1} \neq f_{\theta_2}$ on a set of positive Lebesgue measure. For notational simplicity, we write $|(a_1, \ldots, a_p)| = \sum_{i=1}^{p} |a_i|$.

Theorem 4.1. *Suppose that the family $\{f_t : t \in \Theta\}$ is identifiable with Θ being a compact subset of $\mathbb{R}^p$. Further suppose that $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_\theta(x)$ with $\theta \in int(\Theta)$, $t \mapsto s_t = f_\theta^{1/2}$ is continuous in L_2, (4.5) and (4.6) hold for every $\theta \in int(\Theta)$, and $I_\theta = 4 \int \dot{s}_\theta(x)\dot{s}_\theta^T(x)dx$ is nonsingular. If a sequence of density functions $\{f_n\}$ satisfies, as $n \to \infty$, that*

$$\int \left(f_n^{1/2}(x) - s_\theta(x)\right)^2 dx \xrightarrow{P} 0,$$

$$n^{1/2}\Big[\int \frac{\dot{s}_\theta(x)}{s_\theta(x)} f_n(x)dx - \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{s}_\theta(X_i)}{s_\theta(X_i)}\Big] \xrightarrow{P} 0,$$

$$n^{1/2} \int \frac{|\, \dot{s}_\theta(x)\,|}{s_\theta^3(x)} \left(f_n(x) - f_\theta(x)\right)^2 dx \xrightarrow{P} 0,$$

then the MHD estimator defined in (4.3) is asymptotically efficient; i.e.,

$$n^{1/2}(T_0(f_n) - \theta) \xrightarrow{\mathcal{L}} N\left(0, I_\theta^{-1}\right).$$

Proof. Similar to the proof of Theorem 4 of Beran (1977). □

Remark 4.1. Suppose f_n is the kernel density estimator given by $f_n(x) = \frac{1}{nb_n s_n}\sum_{i=1}^{n} K(\frac{x-X_i}{b_n s_n})$, where $\{b_n\}$ being a sequence of bandwidths such that $\lim_{n\to\infty} n^{1/2} b_n = \infty$ and $\lim_{n\to\infty} n^{1/2} b_n^2 = 0$, K being a symmetric smooth density with compact support, and $s_n = s_n(X_1, X_2, ..., X_n)$ being a robust scale estimator such that $n^{1/2}(s_n - s) = O_P(1)$ for some positive constant s depending on f_θ. If the underlying model f_θ has compact support and satisfies certain smoothness properties, then Theorem 4.1 holds. This can be seen from Theorem 4 of Beran (1977).

The asymptotic variance of $T_0(f_n)$ attains the Fisher information I_θ in parametric models (4.2), and therefore $T_0(f_n)$ is an efficient estimator. For semiparametric models (4.1), the lower bound of the asymptotic variance I_θ^{-1} is attained only when a sequence of very good estimators η_n of η is available. This result is given next, and it is an extension over previous results given for parametric models. Let us denote $I_\theta(\eta) = \int \big(\frac{\partial \log f_{\theta,\eta}}{\partial \theta}\big)\big(\frac{\partial \log f_{\theta,\eta}}{\partial \theta}\big)^T f_{\theta,\eta}\, dx$.

Theorem 4.2. *Suppose that*

(i) $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_{\theta,\eta} \in \{f_{t,h} : t \in \Theta, h \in \mathcal{H}\}$ *with* $\theta \in int(\Theta)$, *where* Θ *is a compact subset of* $\mathbb{R}^p$ *and* $\mathcal{H}$ *is an infinite dimensional set.*

(ii) *For every* $\eta \in \mathcal{H}$, *the family* $\{f_{t,\eta} : t \in \Theta\}$ *is identifiable,* $t \mapsto s_t = f_{t,\eta}^{1/2}$ *is continuous in* L_2, *and (4.5) and (4.6) hold for* s_t *and for every* $t \in int(\Theta)$.

(iii) $\{f_n\}_{n\in\mathbb{N}}$ *is a sequence of estimators of* $f_{\theta,\eta}$ *based on* $(X_1, \ldots, X_n)$ *such that for some* $r > 1/2$,

$$\int \big(f_n^{1/2}(x) - s_\theta(x)\big)^2 dx = O_P(n^{-r}), \tag{4.7}$$

$$n^{1/2} \int \frac{|\dot{s}_\theta(x)|}{s_\theta^3(x)} \big(f_n(x) - f_{\theta,\eta}(x)\big)^2 dx = o_P(1), \tag{4.8}$$

$$n^{1/2}\Big(\int \frac{\dot{s}_\theta(x)}{s_\theta(x)} f_n(x) dx - \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{s}_\theta(X_i)}{s_\theta(X_i)}\Big) = o_P(1). \tag{4.9}$$

(iv) $\{\eta_n\}$ *is a sequence of estimators of* η *such that with* $\widehat{s}_t = f_{t,\eta_n}^{1/2}$ *and* $\dot{\widehat{s}}_t = \frac{\partial}{\partial t}\widehat{s}_t$

$$\sup_{t\in\Theta} \int (\widehat{s}_t(x) - s_t(x))^2 dx = O_P(n^{-r}), \tag{4.10}$$

$$\int (\dot{\hat{s}}_{t_n}(x) - \dot{s}_{t_n}(x))^2 dx = o_P(n^{-(1-r)}), \tag{4.11}$$

$$\int \dot{\hat{s}}_{t_n}(x) s_{t_n}(x) dx = o_P(n^{-1/2}) \tag{4.12}$$

for any sequence of random variables $\{t_n\}$ *such that* $t_n = \theta + O_P(n^{-r/2})$.

Then the MHD estimator defined by (4.4) satisfies

$$\theta_n - \theta = I_\theta^{-1}(\eta) \, \frac{1}{n} \sum_{j=1}^{n} \frac{2\dot{s}_\theta}{s_\theta}(X_j) + o_P(n^{-1/2}). \tag{4.13}$$

Consequently,

$$n^{1/2}(\theta_n - \theta) \xrightarrow{\mathcal{L}} N\big(0, I_\theta^{-1}(\eta)\big).$$

Proof. Note that θ_n is a minimizer of the function d_n and θ is the unique minimizer of the function d, where

$$d_n(t) = \|f_{t,\eta_n}^{1/2} - f_n^{1/2}\| \quad \text{and} \quad d(t) = \|f_{t,\eta}^{1/2} - f_{\theta,\eta}^{1/2}\|, \quad t \in \Theta.$$

Observe that

$$d_n^2(t) = 2 - 2\int f_{t,\eta_n}^{1/2}(x) f_n^{1/2}(x)\, dx \quad \text{and} \quad d^2(t) = 2 - 2\int f_{t,\eta}^{1/2}(x) f_{\theta,\eta}^{1/2}(x)\, dx.$$

Since $f_{t,\eta}^{1/2}$ is continuous in t in L_2 by assumption (ii), d_n and d are continuous and θ_n is well defined. By Minkowski inequality

$$|d_n(t) - d(t)| \le \|f_{t,\eta_n}^{1/2} - f_n^{1/2} - f_{t,\eta}^{1/2} + f_{\theta,\eta}^{1/2}\| \le \|f_{t,\eta_n}^{1/2} - f_{t,\eta}^{1/2}\| + \|f_n^{1/2} - f_{\theta,\eta}^{1/2}\|.$$

Thus, by (4.7) and (4.10), we obtain

$$\Delta_n := \sup_{t\in\Theta} |d_n(t) - d(t)| = O_P(n^{-r/2}). \tag{4.14}$$

We have from (4.5) that

$$d^2(t) = \|s_t - s_\theta\|^2 = \frac{1}{4}(t-\theta)^T I_\theta(\eta)(t-\theta) + o(\|t-\theta\|^2)$$

and therefore $d(t) \ge c|t-\theta|$ for some positive constant c and for all t close to θ. The preceding result and the continuity of d show that

$$\phi(s) \ge cs, \qquad 0 < s < \delta, \tag{4.15}$$

for some $\delta > 0$, where ϕ is given by $\phi(s) = \inf_{t\in\Theta,|t-\theta|\geq s} d(t)$, $s > 0$. Next we can show that the events $\{|\theta_n - \theta| \geq s\}$ and $\{\Delta_n < \phi(s)/2\}$ are disjoint for $0 < s < \delta$. Indeed on their intersection we can conclude that $d_n(\theta) < d(\theta) + \phi(s)/2 = \phi(s)/2$ and $d_n(\theta_n) > d(\theta_n) - \phi(s)/2 \geq \phi(s) - \phi(s)/2 = \phi(s)/2$, and therefore $d_n(\theta) < d_n(\theta_n)$, which yields a contradiction to the definition of θ_n. Thus by (4.15) we have for all $\epsilon > 0$,

$$P(|\theta_n - \theta| \geq \epsilon n^{-r/2}) \leq P(\Delta_n \geq \phi(\epsilon n^{-r/2})/2) \leq P(\Delta_n \geq c\epsilon n^{-r/2}/2).$$

This and (4.14) establish that

$$\theta_n = \theta + O_P(n^{-r/2}). \tag{4.16}$$

As a consequence of (4.5), (4.7) and (4.16) we obtain

$$\begin{aligned} \|f_n^{1/2} - f_{\theta_n,\eta}^{1/2}\| &\leq \|f_n^{1/2} - f_{\theta,\eta}^{1/2}\| + \|f_{\theta_n,\eta}^{1/2} - f_{\theta,\eta}^{1/2}\| \\ &= O_P(n^{-r/2}) + O_P(\|\theta_n - \theta\|) \\ &= O_P(n^{-r/2}). \end{aligned} \tag{4.17}$$

It follows from (ii) that $\int \dot{s}_t f_{t,\eta}^{1/2}(x)\,dx = 0$ for all $t \in \text{int}(\Theta)$ and that the map $t \mapsto \int \widehat{s}_t f_n^{1/2}(x)\,dx$ is differentiable at each $t \in \text{int}(\Theta)$ with derivative $\int \dot{\widehat{s}}_t(x) f_n^{1/2}(x)\,dx$. Since θ_n maximizes this map, we see that $\int \dot{\widehat{s}}_{\theta_n}(x) f_n^{1/2}(x)dx = 0$ on the event that θ_n is an interior point of Θ. This event has probability tending to one since θ_n is a consistent estimator of $\theta \in \text{int}(\Theta)$ as shown in (4.16). On this event we also have $\int \dot{s}_{\theta_n} f_{\theta_n,\eta}^{1/2}(x)\,dx = 0$ and thus

$$-\int \dot{s}_{\theta_n}(x) f_n^{1/2}(x)\,dx = \int [\dot{\widehat{s}}_{\theta_n}(x) - \dot{s}_{\theta_n}(x)] f_n^{1/2}(x)\,dx = \int \dot{\widehat{s}}_{\theta_n}(x) f_{\theta_n,\eta}^{1/2}(x)\,dx + R_n,$$

where

$$R_n = \int \left[\dot{\widehat{s}}_{\theta_n}(x) - \dot{s}_{\theta_n}(x)\right]\left[f_n^{1/2}(x) - f_{\theta_n,\eta}^{1/2}(x)\right] dx.$$

It follows from (4.11), (4.17) and Cauchy–Schwarz inequality that

$$|R_n| \leq \|\dot{\widehat{s}}_{\theta_n} - \dot{s}_{\theta_n}\| \cdot \|f_n^{1/2} - f_{\theta_n,\eta}^{1/2}\| = o_p(n^{-(1-r)/2}) O_p(n^{-r/2}) = o_P(n^{-1/2}).$$

The preceding result and (4.12) yield that

$$\int \dot{s}_{\theta_n}(x) f_n^{1/2}(x) dx = o_P(n^{-1/2}). \tag{4.18}$$

Note that

$$\begin{aligned} & \int \dot{s}_{\theta_n}(x) f_n^{1/2}(x) dx \\ = & \int \dot{s}_{\theta_n}(x)(f_n^{1/2}(x) - s_\theta(x)) dx - \int \dot{s}_{\theta_n}(x)(s_{\theta_n}(x) - s_\theta(x)) dx \\ = & I_1 - I_2, \quad \text{say.} \end{aligned} \tag{4.19}$$

Then from (4.5), (4.6), (4.7) and (4.16) we obtain

$$\begin{aligned} I_1 &= \int \dot{s}_\theta(x)\big(f_n^{1/2}(x) - s_\theta(x)\big) dx + \int \big(\dot{s}_{\theta_n}(x) - \dot{s}_\theta(x)\big)\big(f_n^{1/2}(x) - s_\theta(x)\big) dx \\ &= \int \dot{s}_\theta(x) f_n^{1/2}(x) dx + O(\|\dot{s}_{\theta_n} - \dot{s}_\theta\| \cdot \|f_n^{1/2} - s_\theta\|) \\ &= \int \dot{s}_\theta(x) f_n^{1/2}(x) dx + O_P(n^{-r}) \\ &= \int \dot{s}_\theta(x) f_n^{1/2}(x) dx + o_P(n^{-1/2}) \end{aligned} \tag{4.20}$$

$$\begin{aligned} I_2 &= \int \dot{s}_\theta(x)(s_{\theta_n}(x) - s_\theta(x)) dx + \int (\dot{s}_{\theta_n}(x) - \dot{s}_\theta(x))(s_{\theta_n}(x) - s_\theta(x)) dx \\ &= [\frac{1}{4} I_\theta(\eta)(\theta_n - \theta) + o_P(|\theta_n - \theta|)] + O_P(\|\theta_n - \theta\|^2) \\ &= \frac{1}{4} I_\theta(\eta)(\theta_n - \theta) + o_P(|\theta_n - \theta|) \end{aligned} \tag{4.21}$$

Equations (4.18)-(4.21) give

$$\theta_n - \theta = 4I_\theta^{-1}(\eta) \int \dot{s}_\theta(x) f_n^{1/2}(x) dx + a_n \int \dot{s}_\theta(x) f_n^{1/2}(x) dx + o_P(n^{-1/2}) \tag{4.22}$$

with $a_n \xrightarrow{P} 0$ as $n \to \infty$. Applying the algebraic identity

$$b^{1/2} - a^{1/2} = (b-a)/(2a^{1/2}) - (b-a)^2/[2a^{1/2}(b^{1/2} + a^{1/2})^2]$$

for $b \geq 0$ and $a > 0$, we have by assumption (iii) that

$$\begin{aligned} n^{1/2} \int \dot{s}_\theta(x) f_n^{1/2}(x) dx &= n^{1/2} \int \dot{s}_\theta(x)[f_n^{1/2}(x) - s_\theta(x)] dx \\ &= n^{1/2} \int \frac{\dot{s}_\theta(x)}{2s_\theta(x)}[f_n(x) - s_\theta^2(x)] dx + R_n \\ &= n^{1/2} \int \frac{\dot{s}_\theta(x)}{2s_\theta(x)} f_n(x) dx + R_n \\ &= n^{1/2} \cdot \frac{1}{2n} \sum_{i=1}^{n} \frac{\dot{s}_\theta}{s_\theta}(X_i) + o_P(1) + R_n \end{aligned} \tag{4.23}$$

with $|R_n| \le n^{1/2} \int \frac{|\dot{s}_\theta(x)|}{2s_\theta^3(x)}[f_n(x) - s_\theta^2(x)]^2 dx \xrightarrow{P} 0$. By the CLT, the asymptotic distribution of $n^{1/2}\big(\frac{1}{n}\sum_{i=1}^{n} \frac{\dot{s}_\theta}{s_\theta}(X_i)\big)$ is $N(0, \frac{1}{4}I_\theta(\eta))$. Therefore, (4.22) and (4.23) give the desired result (4.13). This also shows that the asymptotic distribution of $n^{1/2}(\theta_n - \theta)$ is $N(0, I_\theta^{-1}(\eta))$. □

Remark 4.2. When kernel density estimators f_n and η_n are used to estimate f_θ and η, respectively, Theorem 4.2 holds for semiparametric models $f_{\theta,\eta}$ of certain form. The symmetric location model is one such particular family and it is shown that conditions of Theorem 4.2 are satisfied for the preceding family, see Section 4.7.

Corollary 4.1. *Suppose that the conditions (i) and (ii) in Theorem 4.2 hold, $\{f_n\}_{n\in\mathbb{N}}$ is a sequence of estimators of $f_{\theta,\eta}$ based on $(X_1, \ldots, X_n)$ such that $\int \big(f_n^{1/2}(x) - f_{\theta,\eta}^{1/2}(x)\big)^2 dx = O_P(n^{-1})$, and $\{\eta_n\}$ is a sequence of estimators of η such that $\sup_{t\in\Theta} \int (f_{t,\eta_n}^{1/2}(x) - f_{t,\eta}^{1/2}(x))^2 dx = O_P(n^{-1})$. Then the MHD estimator defined by (4.4) is $n^{1/2}$-consistent; i.e., $n^{1/2}(\theta_n - \theta) = O_P(1)$.*

Remark 4.3. The conditions stated in Theorem 4.2 are typical assumptions made in this context (see, e.g., Beran, 1977) and are easily satisfied by many families, except the conditions (4.11) and (4.12) in assumption (iv). The condition (4.12) is analogous to but stronger than condition (2.3) in Schick (1986) if $r < 1$. If it is known that θ_n is an $n^{1/2}$-consistent estimator of θ, then (4.12) can be weakened to hold only for sequences $\{t_n\}$ such that $n^{1/2}(t_n - \theta) = O_P(1)$. For the mixture model $\theta f(x) + (1-\theta)g(x)$, in Chapter 2 we constructed a MHD estimator θ_n and proved that $n^{1/2}(\theta_n - \theta)$ is asymptotically normal. But to weaken the condition (4.12) further, a general result about the boundedness of $n^{1/2}(\theta_n - \theta)$ for the MHD estimator θ_n may be needed.

Remark 4.4. The condition (4.11) in some sense requires the rate of convergence of η_n to η to be of order $o_P(n^{-(1-r)/2})$. This could be satisfied by certain nonparametric estimators. The above convergence requirement of η_n, for example, is fulfilled by most kernel density estimators of η in the mixture model $\theta f(x) + (1-\theta)g(x)$ considered in Chapter 2 with $\eta = (f, g)$. In fact, in Chapter 2 we have shown that $\int (\hat{s}_{t_n}(x) - \dot{s}_{t_n}(x))^2 dx = O_P(n^{-1/2})$ (see (2.24) and the argument given just below (2.29)). But (4.12) was not satisfied by the MHD estimator θ_n constructed in Chapter 2. However, if we change the setup of the model somewhat (in other words, we regard the data from the mixture as the whole sample, and the estimators of the two components are based on other resources with sample sizes converge to infinity faster than that of the size of the sample from the mixture) then faster convergence rate of the estimators of the two components can be obtained and the lower bound $I_\theta^{-1}(\eta)$ can be achieved;

see Theorem 2.4 in Chapter 2. However, the unknown feature of η will usually bring an extra variance to the estimator. In fact, one cannot expect in most cases that the lower bound of the asymptotic variance for semiparametric models (4.1) to be the same as that for the parametric models (4.2): the former is always larger than the latter.

Remark 4.5. The property that one can estimate θ as well asymptotically not knowing η as knowing η is so called *adaptivity*. A sequence of estimators $\{\widehat{\theta}_n\}$ is *adaptive* if and only if, under $f_{\theta_n,\eta}$,

$$n^{1/2}(\widehat{\theta}_n - \theta_n) \xrightarrow{\mathcal{L}} N\big(0, I_\theta^{-1}(\eta)\big)$$

whenever $n^{1/2}(\theta_n - \theta) = O_P(1)$. The preceding expression is equivalent to

$$n^{1/2}\big(\widehat{\theta}_n - \theta - \frac{1}{n}\sum_{j=1}^{n} I_\theta^{-1}(\eta)\dot{l}_{\theta,\eta}(X_j)\big) = o_P(1),$$

where $l_{\theta,\eta} = \log f_{\theta,\eta}$ and $\dot{l}_{\theta,\eta} = \frac{\partial}{\partial\theta} l_{\theta,\eta}$. This follows from Theorem 6.3 of Fabian and Hannan (1982) and Theorem 6.1 of Bickel (1982) and the note thereafter.

Remark 4.6. Given any $n^{1/2}$-consistent estimator, Bickel (1982) used sample splitting techniques to give a general procedure for constructing adaptive estimators in semiparametric models (4.1). Schick (1987) gave sufficient conditions for the construction of efficient estimators without sample splitting, which are stronger and more cumbersome to verify than the necessary and sufficient conditions for the existence of efficient estimators which suffice for the construction based on sample splitting. Forrester et al. (2003) used a conditioning argument to weaken those conditions of Schick (1987) and showed that the resulting weaker conditions reduce to minimal conditions for the construction with sample splitting in a large class of semiparametric models and for properly chosen estimators of the score function. Theorem 4.2 in fact gives sufficient conditions for the estimator θ_n of θ defined in (4.4) to be adaptive. If the MHD estimator has been proved to be $n^{1/2}$-consistent (as the cases in Chapters 2 and 3), we can use one of the procedures given above to construct adaptive estimators based on the MHD estimator.

4.3 Efficiency in the Semiparametric Sense

The requirement of adaptivity is much stronger than efficiency. Also it is more reasonable to use the efficiency in the semiparametric sense, instead of the usual parametric sense. Next we construct non-adaptive but efficient estimators in the semiparametric sense (for the definition see (4.30) below).

In order to investigate the efficiency for semiparametric models (4.1), we first need to introduce a lower bound of the asymptotic variance under these models. For simplicity, suppose the parameter space is a compact interval $\Theta = [a, b] \subseteq \mathbb{R}^p$. The results could be easily extended to a more general space.

Recall that the root-density $f_{\theta,\eta}^{1/2}$ is said to be Hellinger-differentiable at $(\theta, \eta) \in \Theta \times \mathcal{H}$ if there exists $\rho_\theta \in L_2$ and a bounded linear operator $A : L_2 \to L_2$ such that

$$\frac{\|f_{\theta_n,\eta_n}^{1/2} - f_{\theta,\eta}^{1/2} - [\rho_\theta(\theta_n - \theta) + A(\eta_n^{1/2} - \eta^{1/2})]\|}{|\theta_n - \theta| + \|\eta_n^{1/2} - \eta^{1/2}\|} \longrightarrow 0 \quad \text{as } n \to \infty \tag{4.24}$$

for all sequences $\{\theta_n\} \subseteq \Theta$ and $\{\eta_n\} \subseteq \mathcal{H}$ such that $\theta_n \to \theta$ and $\|\eta_n^{1/2} - \eta^{1/2}\| \to 0$ as $n \to \infty$. If η is known, then ρ_θ is typically just the usual parametric score function $\dot{l}_{\theta,\eta}$ for θ times $\frac{1}{2} f_{\theta,\eta}^{1/2}$. The operator A can be regarded as yielding a "score for η". Here we use the Hellinger perturbations to define the differentiability. The rationale for choosing Hellinger differentiability here because it is consistent with previous sections and it nicely ties in with local asymptotic normality (LAN). Define classes

$$\begin{aligned} \mathcal{B} \;=\; \{\beta \in L_2 : \|n^{1/2}(\eta_n^{1/2} - \eta^{1/2}) - \beta\| \to 0 \;\text{ as }\; n \to \infty \\ \text{for some sequence } \{\eta_n\} \subseteq \mathcal{H}\}, \end{aligned} \tag{4.25}$$

$$\mathcal{A} \;=\; \{\alpha \in L_2 : \alpha = h\rho_\theta + A\beta \;\text{ for some }\; h \in \mathbb{R},\; \beta \in \mathcal{B}\}, \tag{4.26}$$

and make the following assumption:

ASSUMPTION S. The set $\mathcal{B}$ defined in (4.25) is a subspace of L_2 and $\{A\beta : \beta \in \mathcal{B}\}$ is closed.

It is known that finding the "information" for estimation of θ in the presence of nuisance parameters requires orthogonal projection of the score for the parameter of interest onto the space of nuisance parameter scores $\{A\beta : \beta \in \mathcal{B}\}$, thereby yielding the "effective" component of ρ_θ orthogonal to the nuisance parameter scores. Under ASSUMPTION S, there exists a $\beta^* \in \mathcal{B}$ minimizing $\|\rho_\theta - A\beta\|$, i.e.,

$$\beta^* = \arg\min_{\beta \in \mathcal{B}} \|\rho_\theta - A\beta\|. \tag{4.27}$$

Here β^* represents a "least favorable" or worst possible direction of approach to η for the problem of estimating θ. Let

$$S^*(x, \theta, \eta) = \rho_\theta(x) - A\beta^*(x) \tag{4.28}$$

and

$$I_* = 4\|S^*(\cdot, \theta, \eta)\|^2. \tag{4.29}$$

Assume that $I_* \neq 0$. Obviously, $I_* \leq I_\theta$ (defined just above Theorem 4.2), and $S^* \perp A\beta$ for any $\beta \in \mathcal{B}$, where $\alpha \perp \beta$ denotes $\int \alpha(x)\beta(x)dx = 0$. Under some regularity conditions, Begun et al. (1983) proved that I_*^{-1} is the achievable lower bound of the asymptotic variance. Informally, an estimator θ_n of θ is said to be asymptotically efficient in the *semiparametric sense* if

$$n^{1/2}(\theta_n - \theta) \xrightarrow{\mathcal{L}} N\big(0, I_*^{-1}\big). \tag{4.30}$$

This definition can be made precise in the sense of a convolution and local asymptotic minimax (LAM) theorem, as is explained in Begun et al. (1983). We now construct an estimator of θ based on the Hellinger distance, which achieves the semiparametric efficiency bound in the sense of (4.30).

When η is known, the maximum likelihood method can usually be reduced to solving the score equation $\sum_{i=1}^{n} \dot{l}_\theta(X_i) = 0$. A natural generalization of estimating the parameter θ in semiparametric models (4.1) is to solve θ from the efficient score equations $\sum_{i=1}^{n} \widetilde{l}_{\theta,\eta}(X_i) = 0$, where $\widetilde{l}_{\theta,\eta}$ is the efficient score function for θ under the semiparametric sense, i.e., the projection of $\dot{l}_{\theta,\eta}$ onto the orthogonal complement of $\{A\beta : \beta \in \mathcal{B}\}$. We can substitute an estimator η_n for the unknown nuisance parameter η, which results in solving the equation for θ from the equation $\sum_{i=1}^{n} \widetilde{l}_{\theta,\eta_n}(X_i) = 0$. Van der Vaart (1998) proved that such an estimator of θ is asymptotically efficient under certain assumptions. Intuitively, we could make the definition of MHD estimator accommodates to semiparametric models similarly. From (4.4) we have that $\theta_n = \arg\max_{\theta\in\Theta} \int f_{\theta,\eta_n}^{1/2}(x) f_n^{1/2}(x)dx$, or equivalently (in most situations) θ_n solves $\int \rho_\theta(x)|_{\eta=\eta_n} f_n^{1/2}(x)dx = 0$, where ρ_θ is given by (4.24). We now propose a MHD estimator of θ as the solution of

$$\int S^*(x, t, \eta_n) f_n^{1/2}(x)dx = 0, \tag{4.31}$$

where S^* is given by (4.28). Suppose the solution exists and we denote it as $\widehat{\theta}_n$. A similar estimator was investigated by Huang (1982) in a different context. He proved that his estimator is efficient under certain conditions including the consistency of the estimator. Schick (1986) pointed out that proving consistency of the estimator may pose difficult mathematical problems and therefore limit the use of Huang's estimator. Next we prove the consistency of the estimator $\widehat{\theta}_n$ under some reasonable conditions.

Lemma 4.1. *For ρ_θ, A, $\beta \in \mathcal{B}$ and $\alpha \in \mathcal{A}$ defined in (4.24), (4.25) and (4.26), we have $\rho_\theta \perp f_{\theta,\eta}^{1/2}$, $A\beta \perp f_{\theta,\eta}^{1/2}$ and $\alpha \perp f_{\theta,\eta}^{1/2}$.*

Proof. Since $\frac{\|f_{\theta,\eta_n}^{1/2} - f_{\theta,\eta}^{1/2} - A(\eta_n^{1/2} - \eta^{1/2})\|}{\|\eta_n^{1/2} - \eta^{1/2}\|} \to 0$ and the definition of β imply that $n^{1/2}\|\eta_n^{1/2} - \eta^{1/2}\| \to \|\beta\|$, we have $\|n^{1/2}(f_{\theta,\eta_n}^{1/2} - f_{\theta,\eta}^{1/2}) - n^{1/2}A(\eta_n^{1/2} - \eta^{1/2})\| \to 0$. Further,

$$\begin{aligned} & \|n^{1/2}(f_{\theta,\eta_n}^{1/2} - f_{\theta,\eta}^{1/2}) - A\beta\| \\ \leq\ & \|n^{1/2}(f_{\theta,\eta_n}^{1/2} - f_{\theta,\eta}^{1/2}) - n^{1/2}A(\eta_n^{1/2} - \eta^{1/2})\| + \|A\| \cdot \|n^{1/2}(\eta_n^{1/2} - \eta^{1/2}) - \beta\| \\ \to\ & 0, \end{aligned}$$

and thus $n\|f_{\theta,\eta_n}^{1/2} - f_{\theta,\eta}^{1/2}\|^2 = O(1)$ and $\|n^{1/2}f_{\theta,\eta_n}^{1/2} - A\beta\| - \|n^{1/2}f_{\theta,\eta}^{1/2}\| \to 0$. This gives

$$\begin{aligned} & \|n^{1/2}(f_{\theta,\eta_n}^{1/2} - f_{\theta,\eta}^{1/2}) - A\beta\|^2 \\ =\ & \|n^{1/2}f_{\theta,\eta_n}^{1/2} - A\beta\|^2 + \|n^{1/2}f_{\theta,\eta}^{1/2}\|^2 - 2n < f_{\theta,\eta}^{1/2}, f_{\theta,\eta_n}^{1/2} > +2n^{1/2} < f_{\theta,\eta}^{1/2}, A\beta > \\ =\ & 2n < f_{\theta,\eta}^{1/2},\ f_{\theta,\eta}^{1/2} - f_{\theta,\eta_n}^{1/2} > +2n^{1/2} < f_{\theta,\eta}^{1/2},\ A\beta > +o(1) \\ \to\ & 0. \end{aligned}$$

Hence,

$$\begin{aligned} & n^{1/2} < f_{\theta,\eta}^{1/2},\ f_{\theta,\eta}^{1/2} - f_{\theta,\eta_n}^{1/2} > + < f_{\theta,\eta}^{1/2},\ A\beta > \\ =\ & \tfrac{1}{2}n^{1/2} < f_{\theta,\eta}^{1/2} - f_{\theta,\eta_n}^{1/2},\ f_{\theta,\eta}^{1/2} - f_{\theta,\eta_n}^{1/2} > + < f_{\theta,\eta}^{1/2}, A\beta > \\ \to\ & 0 \end{aligned}$$

and thus $A\beta \perp f_{\theta,\eta}^{1/2}$. Similarly, one can prove that $\rho_\theta \perp f_{\theta,\eta}^{1/2}$ by the definition of ρ_θ. Furthermore, $\alpha = (h\rho_\theta + A\beta) \perp f_{\theta,\eta}^{1/2}$. □

Theorem 4.3. *Suppose that $(t, \eta^{1/2}) \mapsto S^*(\cdot, t, \eta)$ is continuous in L_2 at $(t, \eta^{1/2})$ for any $t \in int(\Theta)$, $\|f_n^{1/2} - f_{\theta,\eta}^{1/2}\| \xrightarrow{P} 0$ and $\|\eta_n^{1/2} - \eta^{1/2}\| \xrightarrow{P} 0$ as $n \to \infty$. Further suppose that equation $\int S^*(x, t, \eta) f_{\theta,\eta}^{1/2}(x)dx = 0$ has unique solution in t. Then the MHD estimator defined in (4.31) satisfies $\widehat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$.*

Proof. First suppose that $\|f_n^{1/2} - f_{\theta,\eta}^{1/2}\| \to 0$ w.p.1 and $\|\eta_n^{1/2} - \eta^{1/2}\| \to 0$ w.p.1, as $n \to \infty$. Lemma 4.1 gives $\int S^*(x, \theta, \eta) f_{\theta,\eta}^{1/2}(x)dx = 0$; i.e., $t = \theta$ is the unique solution to the equation $\int S^*(x, t, \eta) f_{\theta,\eta}^{1/2}(x)dx = 0$. Note that θ and $\widehat{\theta}_n$ satisfy

$$0 \ = \ \int \left[S^*(x, \widehat{\theta}_n, \eta_n) f_n^{1/2}(x) - S^*(x, \theta, \eta) f_{\theta,\eta}^{1/2}(x)\right] dx$$

$$
\begin{aligned}
= & \int [S^*(x,\widehat{\theta}_n,\eta_n) - S^*(x,\widehat{\theta}_n,\eta)] f_n^{1/2}(x)dx \\
& + \int S^*(x,\widehat{\theta}_n,\eta)[f_n^{1/2}(x) - f_{\theta,\eta}^{1/2}(x)]dx \qquad (4.32) \\
& + \int [S^*(x,\widehat{\theta}_n,\eta) - S^*(x,\theta,\eta)] f_{\theta,\eta}^{1/2}(x)dx.
\end{aligned}
$$

In view of the compactness of Θ, the continuity of $(t,\eta^{1/2}) \mapsto S^*(\cdot,t,\eta)$ in L_2 implies that $\|S^*(\cdot,\widehat{\theta}_n,\eta_n) - S^*(\cdot,\widehat{\theta}_n,\eta)\| \to 0$ as $n \to \infty$ and that $\sup_{t\in\Theta}\|S^*(\cdot,t,\eta)\|$ is bounded. As a result, as $n \to \infty$,

$$
\begin{aligned}
& \Big| \int [S^*(x,\widehat{\theta}_n,\eta_n) - S^*(x,\widehat{\theta}_n,\eta)] f_n^{1/2}(x)dx \Big| \\
\leq\ & \|S^*(\cdot,\widehat{\theta}_n,\eta_n) - S^*(\cdot,\widehat{\theta}_n,\eta)\| \cdot \|f_n^{1/2}\| \\
\leq\ & \|S^*(\cdot,\widehat{\theta}_n,\eta_n) - S^*(\cdot,\widehat{\theta}_n,\eta)\| \cdot (\|f_n^{1/2} - f_{\theta,\eta}^{1/2}\| + \|f_{\theta,\eta}^{1/2}\|) \\
\to\ & 0
\end{aligned}
$$

and

$$
\Big| \int S^*(x,\widehat{\theta}_n,\eta)[f_n^{1/2}(x) - f_{\theta,\eta}^{1/2}(x)]dx \Big| \ \leq\ \|S^*(\cdot,\widehat{\theta}_n,\eta)\| \cdot \|f_n^{1/2} - f_{\theta,\eta}^{1/2}\| \ \to\ 0.
$$

Thus (4.32) gives

$$
\int [S^*(x,\widehat{\theta}_n,\eta) - S^*(x,\theta,\eta)] f_{\theta,\eta}^{1/2}(x)dx \to 0. \qquad (4.33)
$$

Suppose $\widehat{\theta}_n \nrightarrow \theta$ as $n \to \infty$. By the compactness of Θ, there exists a subsequence $\{\widehat{\theta}_m\} \subset \{\widehat{\theta}_n\}$ such that $\widehat{\theta}_m \to \theta' \neq \theta$ for some $\theta' \in \Theta$ as $m \to \infty$. Then (4.33) gives that $\int [S^*(x,\theta',\eta) - S^*(x,\theta,\eta)] f_{\theta,\eta}^{1/2}(x)dx = 0$, i.e., $\int S^*(x,\theta',\eta) f_{\theta,\eta}^{1/2}(x)dx = 0$ and thus $t = \theta'$ is a solution to $\int S^*(x,t,\eta) f_{\theta,\eta}^{1/2}(x)dx = 0$. This contradicts to the uniqueness of the solution, and thus $\widehat{\theta}_n \to \theta$ as $n \to \infty$. Therefore, the solution to $\int S^*(x,t,h) f^{1/2}(x)dx = 0$ as a functional of (f,h) is continuous at $(f_{\theta,\eta},\eta)$ in the Hellinger metric. As a result, $\widehat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$ for any sequences $\{f_n\}$ and $\{\eta_n\}$ such that $\|f_n^{1/2} - f_{\theta,\eta}^{1/2}\| \xrightarrow{P} 0$ and $\|\eta_n^{1/2} - \eta^{1/2}\| \xrightarrow{P} 0$ as $n \to \infty$. □

We now summarize all the conditions needed for the efficiency of the MHD estimator $\widehat{\theta}_n$ defined by (4.31) as follows:

S1. $f_{\theta,\eta}(x)$ has compact support, twice absolutely continuous and the second derivative $f_{\theta,\eta}^{(2)}(x)$ is bounded. Further, $f_{\theta,\eta}^{1/2}$ is Hellinger-differentiable as defined in (4.24).

S2. $(t,\eta^{1/2}) \mapsto S^*(\cdot,t,\eta)$ is continuous in L_2 at $(t,\eta^{1/2})$ for any $t \in \text{int}(\Theta)$; equation $\int S^*(x,t,\eta) f_{\theta,\eta}^{1/2}(x)dx = 0$ has a unique solution in t; $S^*(\cdot,\theta,\eta)$ is

Hellinger-differentiable at (θ, η) and $\int \frac{\partial}{\partial\theta} S^*(x, \theta, \eta) f_{\theta,\eta}^{1/2}(x)dx$ is finite and nonzero.

S3. $\|f_n^{1/2} - f_{\theta,\eta}^{1/2}\| \xrightarrow{P} 0$ and $n^{1/2} \int \sigma(x)\big(f_n^{1/2}(x) - f_{\theta,\eta}^{1/2}(x)\big)dx \xrightarrow{\mathcal{L}} N(0, \frac{1}{4}\|\sigma\|^2)$ as $n \to \infty$ for all $\sigma \in L_2$ and $\int \sigma(x) f_{\theta,\eta}^{1/2}(x)dx = 0$.

S4. $\|\eta_n^{1/2} - \eta^{1/2}\| \xrightarrow{P} 0$, and $S^*(\cdot, t, \eta_n)$ is well-defined for large n and all $t \in \Theta$.

Theorem 4.4. *Under conditions S1-S4, any solution θ_n of (4.31) is an asymptotically efficient estimator of θ; i.e., (4.30) holds for $\widehat{\theta}_n$.*

Proof. Similar to the proof of Theorem 5.2.1 of Huang (1982). □

Remark 4.7. This remark is parallel to Remark 4.6, and we consider the case that we only have a $n^{1/2}$-consistency of the estimator θ_n of θ defined in (4.4). In this case, we can use one of the procedures mentioned in Remark 4.6 to construct asymptotically efficient estimators in the sense of (4.30). The only difference from the construction of an adaptive estimator is now we are using S^* defined in (4.28) instead of ρ_θ.

Remark 4.8. Consider the estimator θ_n defined by (4.4). Suppose that η_n is a consistent estimator of η in the Hellinger metric, and $f_{t,\eta}^{1/2}$ is Hellinger-differentiable for each $t \in \Theta$ with $A = A_t$ in (4.24) satisfying $\sup_{t\in\Theta}\|A_t\| \le M$ for some $M > 0$. Then the condition (4.10) in Theorem 4.2 could be reduced to

$$\|\eta_n^{1/2} - \eta^{1/2}\|^2 = O_P(n^{-r}).$$

Suppose $\dot{s}_t$ is Hellinger-differentiable for each $t \in B(\theta, \varepsilon)$ with some $\varepsilon > 0$ and $B(\theta, \varepsilon)$ is an ε-neighborhood of θ, then there exists a bounded linear operator $B_t : L_2 \to L_2$ such that

$$\frac{\|\dot{\widehat{s}}_t - \dot{s}_t - B_t(\eta_n^{1/2} - \eta^{1/2})\|}{\|\eta_n^{1/2} - \eta^{1/2}\|} \longrightarrow 0 \quad \text{as } n \to \infty.$$

The condition (4.11) is now equivalent to, for any $t_n = \theta + O_P(n^{-r/2})$,

$$\begin{aligned} o_P(n^{-(1-r)/2}) = \|\dot{\widehat{s}}_{t_n} - \dot{s}_{t_n}\| &= \|B_{t_n}(\eta_n^{1/2} - \eta^{1/2})\| + o(\|\eta_n^{1/2} - \eta^{1/2}\|) \\ &= \|B_{t_n}(\eta_n^{1/2} - \eta^{1/2})\| + o_P(n^{-r/2}), \end{aligned}$$

and since $r > 1/2$, equivalently

$$\|B_{t_n}(\eta_n^{1/2} - \eta^{1/2})\| = o_P(n^{-(1-r)/2}).$$

Therefore, if

$$\|B_{t_n}\| = o_P(n^{r-1/2})$$

for any $t_n = \theta + O_P(n^{-r/2})$, then (4.11) holds.

4.4 *Minimum Profile Hellinger Distance Estimation*

The MHD estimator defined by (4.4) in semiparametric models (4.1) is based on minimizing the Hellinger distance between a density estimator f_n and the parametric family f_{θ,η_n}, i.e., the nuisance parameter η in $f_{\theta,\eta}$ is replaced by an estimator η_n. This approach is in line with Beran's (1977) original mechanism of deriving MHD estimators. Intuitively, one could also define a MHD estimator of θ in semiparametric families (4.1) via profiles.

For any density function g, define a functional $\eta(t, g)$ by

$$\eta(t, g) = \arg\min_{h\in\mathcal{H}} \|f_{t,h}^{1/2} - g^{1/2}\|. \tag{4.34}$$

Set

$$s_{t,g} = f_{t,\eta(t,g)}^{1/2} \tag{4.35}$$

and define the MHD functional $T_1(g)$ as

$$T_1(g) = \arg\min_{t\in\Theta} \|s_{t,g} - g^{1/2}\| = \arg\max_{t\in\Theta} < s_{t,g},\ g^{1/2} > . \tag{4.36}$$

Here we don't require that $f_{t,h}$ is a density function for any $h \in \mathcal{H}$, but we do require that the second equality in (4.36) holds for convenience. In case that the second equality does not hold, we can use the r.h.s. of (4.36) as the definition and the results of this section still hold. We call $\|s_{t,g} - g^{1/2}\|$ the "profile" Hellinger distance between $f_{t,h}$ and g. Now the minimum profile Hellinger distance (MPHD) estimator is defined as $T_1(g_n)$, where g_n is an nonparametric estimator of g based on observed data $X_1, ..., X_n$. Clearly, $T_1(h_{\theta,\eta}) = \theta$ uniquely if $\{f_{t,h}\}_{t\in\Theta,h\in\mathcal{H}}$ is identifiable. Assume that $T_1(g) \in \text{int}(\Theta)$ is uniquely defined and Hellinger continuous at g in the sense that $T_1(g_n) \to T_1(g)$ for any sequence $\{g_n\}_{n\in\mathbb{N}}$ such that $\|g_n^{1/2} - g^{1/2}\| \to 0$ as $n \to \infty$. Assume further that for any t in a small neighborhood of $T_1(g)$ and any h in a small Hellinger neighborhood of g, the map $t \mapsto s_{t,h}$ satisfies (4.5) and (4.6) with continuous gradient vector $\dot{s}_{t,h}$ and continuous Hessian matrix $\ddot{s}_{t,h}$. Let

$$H(t, g) =< \dot{s}_{t,g},\ g^{1/2} > . \tag{4.37}$$

Then $H(T_1(g), g) = 0$. Assume that $\{g_n\}_{n\in\mathbb{N}}$ is a sequence of estimators of g such that $\|g_n^{1/2} - g^{1/2}\| \xrightarrow{P} 0$ and $M(g_n, g) \xrightarrow{P} 0$ as $n \to \infty$, where M is some metric. Thus, there exists a version of $\{g_n\}$, defined on a suitable probability

space, such that $T_1(g_n) \to T_1(g)$ and $T_1(g_n) \in \text{int}(\Theta)$ w.p.1 and

$$\begin{aligned} 0 &= H(T_1(g_n), g_n) - H(T_1(g), g) \\ &= \big[H(T_1(g_n), g_n) - H(T_1(g), g_n)\big] + \big[H(T_1(g), g_n) - H(T_1(g), g)\big]. \end{aligned}$$

Since the map $t \mapsto s_{t,h}$ satisfies (4.5) and (4.6), $H(t,h)$ is differentiable in t with derivative

$$\dot{H}(t,h) = < \ddot{s}_{t,h},\ h^{1/2} >$$

that is continuous in t. Suppose that for any t in a small neighborhood of $T_1(g)$, $\dot{H}(t,h)$ is continuous at $h = g$ w.r.t. metric M. Then

$$\begin{aligned} &H(T_1(g_n), g_n) - H(T_1(g), g_n) \\ &\qquad = \big(T_1(g_n) - T_1(g)\big) \int_0^1 \dot{H}\big(T_1(g) + u(T_1(g_n) - T_1(g)), g_n\big) du. \end{aligned}$$

Suppose further that there is a ψ_g such that $< \psi_g,\ g^{1/2} >= 0$ and

$$H(T_1(g), g_n) - H(T_1(g), g) = < \psi_g,\ g_n^{1/2} - g^{1/2} > + o(\|g_n^{1/2} - g^{1/2}\|) \quad (4.38)$$

for any sequence $\{g_n\}_{n \in \mathbb{N}}$ such that $\|g_n^{1/2} - g^{1/2}\| \to 0$ as $n \to \infty$. If $< \ddot{s}_{T_1(g),g},\ g^{1/2} >$ is invertible, then we have

$$\begin{aligned} T_1(g_n) - T_1(g) &= -\big(< \ddot{s}_{T_1(g),g},\ g^{1/2} >^{-1} + o(1)\big) < \psi_g,\ g_n^{1/2} - g^{1/2} > \\ &\quad + o(\|g_n^{1/2} - g^{1/2}\|), \end{aligned}$$

i.e., the MHD functional T_1 is Hellinger differentiable provided that $\|\psi_g\| < \infty$. If g_n satisfies

$$H(T_1(g), g_n) - H(T_1(g), g) = \frac{1}{n} \sum_{i=1}^{n} \frac{\psi_g(X_i)}{2g^{1/2}(X_i)} + o_P(n^{-1/2}), \quad (4.39)$$

then

$$T_1(g_n) - T_1(g) = -\big(< \ddot{s}_{T_1(g),g},\ g^{1/2} >^{-1} + o(1)\big) \frac{1}{n} \sum_{i=1}^{n} \frac{\psi_g(X_i)}{2g^{1/2}(X_i)} + o_P(n^{-1/2}),$$

and therefore the asymptotic distribution of $n^{1/2}(T_1(g_n) - T_1(g))$ is normal with mean zero and variance Σ defined by

$$\begin{aligned} \Sigma &= \tfrac{1}{4}\, \dot{H}^{-1}(T_1(g), g) < \psi_g,\ \psi_g^T > \dot{H}^{-1}(T_1(g), g) \\ &= \tfrac{1}{4}\ < \ddot{s}_{T_1(g),g},\ g^{1/2} >^{-1} < \psi_g,\ \psi_g^T > < \ddot{s}_{T_1(g),g},\ g^{1/2} >^{-1}. \end{aligned} \quad (4.40)$$

With $\theta := T_1(g)$, note that

$$
\begin{aligned}
& H(T_1(g), g_n) - H(T_1(g), g) \\
= \; & < \dot{s}_{\theta,g_n},\ g_n^{1/2} > - < \dot{s}_{\theta,g}\ g^{1/2} > \\
= \; & 2 < \dot{s}_{\theta,g},\ g_n^{1/2} - g^{1/2} > + < \dot{s}_{\theta,g_n} - \dot{s}_{\theta,g},\ g_n^{1/2} - g^{1/2} > \\
& \qquad + < \dot{s}_{\theta,g_n},\ g^{1/2} > - < g_n^{1/2},\ \dot{s}_{\theta,g} > \\
= \; & 2 < \dot{s}_{\theta,g},\ g_n^{1/2} - g^{1/2} > + \big[< \dot{s}_{\theta,g_n},\ g^{1/2} > - < g_n^{1/2},\ \dot{s}_{\theta,g} > \big] \\
& \qquad + O\big(\|\dot{s}_{\theta,g_n} - \dot{s}_{\theta,g}\| \cdot \|g_n^{1/2} - g^{1/2}\|\big).
\end{aligned}
$$

So if $\|\dot{s}_{\theta,g_n} - \dot{s}_{\theta,g}\| \xrightarrow{P} 0$ and

$$
< \dot{s}_{\theta,g_n},\ g^{1/2} > - < g_n^{1/2},\ \dot{s}_{\theta,g} >= o_P(\|g_n^{1/2} - g^{1/2}\|), \tag{4.41}
$$

then $\psi_g = 2\dot{s}_{T_1(g),g}$. These results are summarized in the next theorem.

Theorem 4.5. *Suppose that*

(i) $T_1(g) \in int(\Theta)$ is uniquely defined and Hellinger continuous at g.

(ii) For any t in a small neighborhood of $T_1(g)$ and any h in a small Hellinger neighborhood of g, the map $t \mapsto s_{t,h}$ defined in (4.35) satisfies (4.5) and (4.6) with continuous gradient vector $\dot{s}_{t,h}$ and continuous Hessian matrix $\ddot{s}_{t,h}$; $< \ddot{s}_{T_1(g),g},\ g^{1/2} >$ is invertible.

(iii) For any t in a small neighborhood of $T_1(g)$, $H(t,g)$ defined in (4.37) satisfies (4.38) with $< \psi_g,\ g^{1/2} >= 0$ and $\|\psi_g\| < \infty$, and the derivative $\dot{H}(t,h)$ is continuous at $h = g$ w.r.t. some metric M.

(iv) $\{g_n\}_{n\in\mathbb{N}}$ is a sequence of estimators of g such that $\|g_n^{1/2} - g^{1/2}\| \xrightarrow{P} 0$ and $M(g_n, g) \xrightarrow{P} 0$ as $n \to \infty$, and satisfies (4.39).

Then T_1 is Hellinger differentiable and the asymptotic distribution of $n^{1/2}(T_1(g_n) - T_1(g))$ is $N(0, \Sigma)$ with variance matrix Σ defined by (4.40). Furthermore, if g_n satisfies (4.41) and $\|\dot{s}_{T_1(g),g_n} - \dot{s}_{\theta,g}\| \xrightarrow{P} 0$, then the above result holds with $\psi_g = 2\dot{s}_{T_1(g),g}$.

Remark 4.9. Condition (4.38) requires in some sense that H defined in (4.37) is Hellinger differentiable. In most cases, $< \psi_g,\ g_n^{1/2} - g^{1/2} >= \frac{1}{n}\sum_{i=1}^{n} \frac{\psi_g(X_i)}{2g^{1/2}(X_i)} + o_P(n^{-1/2})$ (as shown in (4.23)). Therefore, it is reasonable to assume that both (4.38) and (4.39) hold. The example on symmetric location models given in Section 4.7 satisfies the conditions of Theorem 4.5.

In what follows, we consider the case that $g = f_{\theta,\eta}$. We suppose that $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_{\theta,\eta}$, and f_n is a nonparametric density estimator of $f_{\theta,\eta}$ based on observed data. Then a MPHD estimator is $T_1(f_n)$. Theorem 4.5 investigates the asymptotic normality of MPHD estimator $T_1(f_n)$. In order to see the effi-

cient of $T_1(f_n)$ in the semiparametric sense, we need to examine the achievable lower bound of the asymptotic variance, i.e., the explicit form of I_* defined by (4.29). The next theorem achieves this goal. For notational convenience, we let for $t \in \Theta$,

$$\eta_t = \eta(t, f_{\theta,\eta}) = \arg\min_{h\in\mathcal{H}} \|f_{t,h}^{1/2} - f_{\theta,\eta}^{1/2}\| = \arg\max_{h\in\mathcal{H}} < f_{t,h}^{1/2},\ f_{\theta,\eta}^{1/2} > . \tag{4.42}$$

Obviously, $\eta_\theta = \eta$ if the models $\{f_{t,h}\}_{t\in\Theta,h\in\mathcal{H}}$ is identifiable. Define

$$\mathcal{H}_1 = \{\eta_t \in \mathcal{H} : \eta_\theta = \eta, t \mapsto f_{t,\eta_t}^{1/2} \text{ is differentiable in } L_2 \text{ at point } t = \theta\}.$$

Theorem 4.6. *Suppose that $\{f_{t,h}\}_{t\in\Theta,h\in\mathcal{H}}$ is identifiable and $s_t = f_{t,\eta_t}^{1/2}$ is differentiable in L_2 at point $t = \theta$ with gradient $\dot{s}_\theta$, where η_t is defined by (4.42). Then η_t is a least favorable curve among $\mathcal{H}_1$ in the sense of (4.27). Furthermore, $I_* = 4 < \dot{s}_\theta,\ \dot{s}_\theta^T >$ with I_* defined by (4.29).*

Proof. Clearly, $\eta_t \in \mathcal{H}_1$. For any other $\eta_{1t} \in \mathcal{H}_1$, let $s_{1t} = f_{t,\eta_{1t}}^{1/2}$ and $\dot{s}_{1\theta}$ be the gradient of s_{1t} at point $t = \theta$. By the definition of η_t in (4.42),

$$< s_t - s_{1t},\ s_\theta > \geq 0 \quad \text{for all } t \in \Theta. \tag{4.43}$$

Note that

$$\begin{aligned} < s_{\theta+t} - s_\theta,\ s_\theta > &= < s_{\theta+t},\ s_\theta > - < s_\theta,\ s_\theta > \\ &= < s_{\theta+t}\ , s_\theta > -1 \\ &= < s_{\theta+t},\ s_\theta > -\tfrac{1}{2}(< s_{\theta+t},\ s_{\theta+t} > + < s_\theta,\ s_\theta >) \\ &= -\tfrac{1}{2} < s_{\theta+t} - s_\theta,\ s_{\theta+t} - s_\theta > \\ &= -\tfrac{1}{2}\, t^T < \dot{s}_\theta,\ \dot{s}_\theta^T > t + o(\|t\|^2). \end{aligned}$$

Similarly, $< s_{1(\theta+t)} - s_{1\theta},\ s_{1\theta} >= -\frac{1}{2}\, t^T < \dot{s}_{1\theta},\ \dot{s}_{1\theta}^T > t + o(\|t\|^2)$, and thus we obtain

$$\begin{aligned} & < s_{\theta+t} - s_{1(\theta+t)},\ s_\theta > \\ = \ & < s_{\theta+t} - s_\theta,\ s_\theta > - < s_{1(\theta+t)} - s_{1\theta},\ s_{1\theta} > \\ = \ & -\tfrac{1}{2}\, t^T\big(< \dot{s}_\theta,\ \dot{s}_\theta^T > - < \dot{s}_{1\theta},\ \dot{s}_{1\theta}^T > \big)t + o(\|t\|^2). \end{aligned} \tag{4.44}$$

From (4.43) and (4.44), we have that

$$< \dot{s}_\theta,\ \dot{s}_\theta^T > \leq < \dot{s}_{1\theta},\ \dot{s}_{1\theta}^T > .$$

Since η_{1t} is arbitrary, this implies that η_t is a least favorable curve and by definition (4.29) $I_* = 4 < \dot{s}_\theta,\ \dot{s}_\theta^T >$. □

Remark 4.10. With $g = f_{\theta,\eta}$ and $\psi_g = 2\dot{s}_{T_1(g),g}$, the asymptotic variance defined by (4.40) is reduced to

$$\Sigma = < \ddot{s}_{\theta,f_{\theta,\eta}},\ f_{\theta,\eta}^{1/2} >^{-1} < \dot{s}_{\theta,f_{\theta,\eta}},\ \dot{s}_{\theta,f_{\theta,\eta}} > < \ddot{s}_{\theta,f_{\theta,\eta}},\ f_{\theta,\eta}^{1/2} >^{-1}.$$

It follows that $< \ddot{s}_{\theta,f_{\theta,\eta}},\ f_{\theta,\eta}^{1/2} >= -2 < \dot{s}_{\theta,f_{\theta,\eta}},\ \dot{s}_{\theta,f_{\theta,\eta}} >$ (see, e.g., the symmetric location models discuss in Section 4.7). Then Σ is further reduced to $\Sigma = \left[4 < \dot{s}_{\theta,f_{\theta,\eta}},\ \dot{s}_{\theta,f_{\theta,\eta}} > \right]^{-1} = I_*^{-1}$. Therefore, Theorem 4.5 shows that the MPHD estimator is efficient in the semiparametric sense. Theorem 4.5 in a certain sense shows the best possible MHD type estimator and gives a set of sufficient conditions to achieve this best estimator. Theorem 4.5 also demonstrates when an adaptive MHD type estimator exists. If $\dot{s}_{\theta,f_{\theta,\eta}} = \frac{\partial}{\partial\theta} f_{\theta,\eta}/(2f_{\theta,\eta}^{1/2})$, then there exists an adaptive estimator.

4.5 *Robustness*

In this section, we examine some robustness properties of the MHD estimator θ_n defined by (4.4). As many authors have pointed out, the robustness of an estimator would be ideally be studied by considering what happens to the distribution of the estimator as the distribution of the data is varied.

From Theorem 4.2 it follows that the estimator θ_n defined in (4.4) is continuous as a functional of f_n and η_n. A small Hellinger-metric change in f_n and η_n induced by data recording errors or other mechanisms will typically induce correspondingly a small change in the value of θ_n by virtue of the continuity of this estimator.

To this end, we suppose that the true density of data is not strictly from the class defined in (4.1). Instead, we suppose that $X_1, \ldots, X_n \overset{iid}{\sim} g$ with g in a small Hellinger neighborhood of $f_{\theta,\eta}$, i.e., $\|g^{1/2}(x) - f_{\theta,\eta}^{1/2}(x)\| \leq \varepsilon$ for some positive small ε. Then the actual parameter estimated is

$$\theta = T_0(g) = \arg\min_{t\in\Theta} \int \left(f_{t,\eta}^{1/2}(x) - g^{1/2}(x)\right)^2 dx, \tag{4.45}$$

where T_0 is in fact defined in (4.3). Suppose that $\{g_n\}$ is a sequence of estimators of g based on $(X_1, \ldots, X_n)$, and η_n is a sequence of estimators of η that may based on the same data or from other resources. Define a MHD estimator θ_n of θ as

$$\theta_n = T_n(g_n) = \arg\min_{t\in\Theta} \|f_{t,\eta_n}^{1/2} - g_n^{1/2}\| \tag{4.46}$$

where T_n is defined in (4.4). Clearly, definition (4.46) is a generalization of (4.4). The next theorem shows that the estimator θ_n defined in (4.46) is still $\sqrt{n}$−consistent even when the actual density is not from the class defined in (4.1), exhibiting a desirable robustness property of θ_n; i.e., θ_n is not affected by a small Hellinger perturbation of the density of data.

Theorem 4.7. *Suppose that*

(i) *θ defined in (4.45) is unique and $\theta \in int(\Theta)$, where Θ is a compact subset of $\mathbb{R}^p$.*

(ii) *For every $\eta \in \mathcal{H}$, the family $\{f_{t,\eta} : t \in \Theta\}$ is identifiable, $t \mapsto s_t = f_{t,\eta}^{1/2}$ is continuous in L_2, and (4.5) and (4.6) hold for s_t and for every $t \in int(\Theta)$ with $\int \ddot{s}_\theta(x) g^{1/2}(x) dx$ nonsingular.*

(iii) *$\{g_n\}_{n\in\mathbb{N}}$ is a sequence of estimators of g based on $(X_1, \ldots, X_n)$ such that for some $r > 1/2$,*

$$\int \big(g_n^{1/2}(x) - g^{1/2}(x)\big)^2 dx = O_P(n^{-r}), \tag{4.47}$$

$$n^{1/2} \int \frac{|\dot{s}_\theta(x)|}{g^{3/2}(x)} \big(g_n(x) - g(x)\big)^2 dx = o_P(1), \tag{4.48}$$

$$n^{1/2}\Big(\int \frac{\dot{s}_\theta(x)}{g^{1/2}(x)} g_n(x) dx - \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{s}_\theta(X_i)}{g^{1/2}(X_i)} \Big) = o_P(1). \tag{4.49}$$

(iv) *$\{\eta_n\}$ is a sequence of estimators of η such that with $\widehat{s}_t = f_{t,\eta_n}^{1/2}$ and $\dot{\widehat{s}}_t = \frac{\partial}{\partial t}\widehat{s}_t$*

$$\sup_{t\in\Theta} \int (\widehat{s}_t(x) - s_t(x))^2 dx = O_P(n^{-r}), \tag{4.50}$$

$$\int (\dot{\widehat{s}}_{t_n}(x) - \dot{s}_{t_n}(x))^2 dx = o_P(n^{-(1-r)}) \tag{4.51}$$

$$\int [\dot{\widehat{s}}_{t_n}(x) - \dot{s}_{t_n}(x)] g^{1/2}(x)\, dx = o_P(n^{-1/2}) \tag{4.52}$$

for any sequence of random variables $\{t_n\}$ such that $t_n = \theta + O_P(n^{-r/4})$.

Then the MHD estimator defined by (4.46) satisfies

$$\theta_n - \theta = -\Big[\int \ddot{s}_\theta(x) g^{1/2}(x) dx \Big]^{-1} \frac{1}{2n} \sum_{j=1}^{n} \frac{\dot{s}_\theta}{g^{1/2}}(X_j) + o_P(n^{-1/2}). \tag{4.53}$$

Consequently,

$$n^{1/2}(\theta_n - \theta) \xrightarrow{\mathcal{L}} N\big(0, 4^{-1} \int \rho_g(x) \rho_g^T(x) dx\big),$$

where

$$\rho_g(x) = -\Big[\int \ddot{s}_\theta(x) g^{1/2}(x) dx \Big]^{-1} \dot{s}_\theta(x). \tag{4.54}$$

Remark 4.11. Theorem 4.7 is parallel to Theorem 4.5. Both are for a general underlying density function g (may not be exactly the semiparametric model $f_{\theta,\eta}$). Theorem 4.5 discusses the asymptotic efficiency of the MPHD estimator of θ in semiparametric sense, while Theorem 4.8 examines the asymptotic efficiency of the MHD estimator of θ in parametric sense (adaptivity). As discussed in Remark 4.10, the MPHD estimator will be reduced to an adaptive estimator if it exists.

Proof of Theorem 4.7. The proof follows along the same line as the proof of Theorem 4.2.

Note that θ_n defined by (4.46) is a minimizer of the function d_n and θ is the unique minimizer of the function d, where

$$d_n(t) = \|\widehat{s}_t - g_n^{1/2}\| \text{ and } d(t) = \|s_t - g^{1/2}\|, \quad t \in \Theta.$$

Observe that

$$d_n^2(t) = 2 - 2 < \widehat{s}_t,\ g_n^{1/2} > \text{ and } d^2(t) = 2 - 2 < s_t,\ g^{1/2} > .$$

Since $t \mapsto s_t$ is continuous in L_2 by assumption (ii), d_n and d are continuous and θ_n is well defined. By Minkowski inequality

$$\begin{aligned} \left|d_n^2(t) - d^2(t)\right| &= \left(\|\widehat{s}_t - g_n^{1/2}\| + \|s_t - g^{1/2}\|\right) \cdot \left|\|\widehat{s}_t - g_n^{1/2}\| - \|s_t - g^{1/2}\|\right| \\ &\leq \left(\|\widehat{s}_t\| + \|g_n^{1/2}\| + \|s_t\| + \|g^{1/2}\|\right) \cdot \|\widehat{s}_t - g_n^{1/2} - s_t + g^{1/2}\| \\ &\leq 4\left(\|\widehat{s}_t - s_t\| + \|g_n^{1/2} - g^{1/2}\|\right). \end{aligned}$$

Thus, from (4.47) and (4.50), we obtain

$$\Delta_n := \sup_{t\in\Theta} |d_n^2(t) - d^2(t)| = O_P(n^{-r/2}). \tag{4.55}$$

Now define

$$\phi(s) = \inf_{t\in\Theta, |t-\theta|\geq s} d^2(t) - d^2(\theta), \qquad s > 0.$$

If g is a member of models (4.1), then $d(\theta) = 0$ and we can follow the same line as in the proof of Theorem 4.2 to prove that for some $\delta > 0$,

$$\phi(s) \geq cs^2, \qquad 0 < s < \delta. \tag{4.56}$$

If g is not from the semiparametric models defined in (4.1), then $d(\theta) > 0$. Since $t = \theta \in \text{int}(\Theta)$ is the unique maximizer of $< s_t,\ g^{1/2} >$, we have $< \dot{s}_\theta,\ g^{1/2} >= 0$ and $< \ddot{s}_\theta,\ g^{1/2} >$ is negative definite. Then by (4.5) and (4.6),

$$
\begin{aligned}
& d^2(t) - d^2(\theta) \\
= \; & -2 < s_t - s_\theta,\ g^{1/2} > \\
= \; & -2(t-\theta)^T < \dot{s}_\theta,\ g^{1/2} > -(t-\theta)^T < \ddot{s}_\theta,\ g^{1/2} > (t-\theta) + o(\|t-\theta\|^2) \\
= \; & (t-\theta)^T < -\ddot{s}_\theta,\ g^{1/2} > (t-\theta) + o(\|t-\theta\|^2),
\end{aligned}
$$

and therefore $d^2(t) - d^2(\theta) \geq c|t-\theta|^2$ for some positive constant c and all t close to θ. The preceding result and the continuity of d show that (4.56) holds. Next we can show that the events $\{|\theta_n - \theta| \geq s\}$ and $\{\Delta_n < \phi(s)/2\}$ are disjoint for $0 < s < \delta$. Indeed, on their intersection we can conclude that $d_n^2(\theta) - d^2(\theta) < \phi(s)/2$ and $d_n^2(\theta_n) - d^2(\theta) > \big(d^2(\theta_n) - d^2(\theta)\big) - \phi(s)/2 \geq \phi(s) - \phi(s)/2 = \phi(s)/2$, and therefore $d_n(\theta) < d_n(\theta_n)$, which yields a contradiction to the definition of θ_n. Thus, by (4.55) and (4.56) we have

$$
P(|\theta_n - \theta| \geq \epsilon n^{-r/4}) \leq P(\Delta_n \geq \phi(\epsilon n^{-r/4})/2) \leq P(\Delta_n \geq c\epsilon^2 n^{-r/2}/2) \to 0
$$

for all $\epsilon > 0$. This establishes that

$$
\theta_n = \theta + O_P(n^{-r/4}).
$$

It follows from assumption (ii) that $< \dot{s}_t,\ s_t >= 0$ for every $t \in \text{int}(\Theta)$ and that $< \dot{\hat{s}}_{\theta_n},\ g_n^{1/2} >= 0$ on the event that θ_n is an interior point of Θ. This event has probability tending to one since θ_n is a consistent estimator of $\theta \in \text{int}(\Theta)$. On this event we also have $< \dot{s}_{\theta_n},\ s_{\theta_n} >= 0$ and thus

$$
- < \dot{s}_{\theta_n},\ g_n^{1/2} > = < \dot{\hat{s}}_{\theta_n} - \dot{s}_{\theta_n},\ g_n^{1/2} > = < \dot{\hat{s}}_{\theta_n} - \dot{s}_{\theta_n},\ g^{1/2} > + R_n, \qquad (4.57)
$$

where

$$
R_n \; = < \dot{\hat{s}}_{\theta_n} - \dot{s}_{\theta_n},\ g_n^{1/2} - g^{1/2} > .
$$

From (4.47), (4.51) and the Cauchy–Schwarz inequality , we obtain

$$
|R_n| \; \leq \; \|\dot{\hat{s}}_{\theta_n} - \dot{s}_{\theta_n}\| \cdot \|g_n^{1/2} - g^{1/2}\| \; = \; o_p(n^{-(1-r)/2}) O_p(n^{-r/2}) \; = \; o_P(n^{-1/2}).
$$

The above result together with (4.57) and (4.52) yield that

$$
< \dot{s}_{\theta_n},\ g_n^{1/2} > \; = \; o_P(n^{-1/2}).
$$

Now from (4.6), we have

$$
\begin{aligned}
o_P(n^{-1/2}) \; &= \; < \dot{s}_{\theta_n},\ g_n^{1/2} > \\
&= \; < \dot{s}_\theta + \ddot{s}_\theta(\theta_n - \theta) + v_n(\theta_n - \theta),\ g_n^{1/2} >,
\end{aligned}
$$

where the components of $p \times p$ matrix $v_n(x)$ converge in L_2 to zero as $n \to \infty$.

Thus, for n sufficiently large, one obtains

$$\begin{aligned}\theta_n - \theta &= -<\ddot{s}_\theta + v_n,\ g_n^{1/2}>^{-1}<\dot{s}_\theta,\ g_n^{1/2}> + o_P(n^{-1/2}) \\ &= -<\ddot{s}_\theta + v_n,\ g_n^{1/2}>^{-1}<\dot{s}_\theta,\ g_n^{1/2} - g^{1/2}> + o_P(n^{-1/2}) \\ &= -<\ddot{s}_\theta,\ g^{1/2}>^{-1}<\dot{s}_\theta,\ g_n^{1/2} - g^{1/2}> \\ &\qquad + a_n <\dot{s}_\theta,\ g_n^{1/2} - g^{1/2}> + o_P(n^{-1/2}),\end{aligned} \tag{4.58}$$

where $a_n \xrightarrow{P} 0$ as $n \to \infty$. Applying the algebraic identity

$$b^{1/2} - a^{1/2} = (b-a)/(2a^{1/2}) - (b-a)^2/[2a^{1/2}(b^{1/2} + a^{1/2})^2]$$

for $b \geq 0$ and $a > 0$, we have by assumption (iii) that

$$\begin{aligned}n^{1/2} <\dot{s}_\theta,\ g_n^{1/2} - g^{1/2}> &= n^{1/2} \int \frac{\dot{s}_\theta(x)}{2g^{1/2}(x)}[g_n(x) - g(x)]dx + R_n \\ &= n^{1/2} \int \frac{\dot{s}_\theta(x)}{2g^{1/2}(x)} g_n(x)dx + R_n \\ &= n^{1/2} \cdot \frac{1}{2n} \sum_{i=1}^{n} \frac{\dot{s}_\theta}{g^{1/2}}(X_i) + o_P(1) + R_n\end{aligned} \tag{4.59}$$

with $|R_n| \leq n^{1/2} \int \frac{|\dot{s}_\theta(x)|}{2g^{3/2}(x)}[g_n(x) - g(x)]^2 dx \xrightarrow{P} 0$. By the CLT, the asymptotic distribution of $n^{1/2}\big(\frac{1}{n}\sum_{i=1}^{n} \frac{\dot{s}_\theta}{g^{1/2}}(X_i)\big)$ is $N(0, \int \dot{s}_\theta(x)\dot{s}_\theta^T(x)dx)$. Therefore, (4.58) and (4.59) give the desired result (4.53). This result also shows that the asymptotic distribution of $n^{1/2}(\theta_n - \theta)$ is $N(0, 4^{-1} \int \rho_g(x)\rho_g^T(x)dx)$ as well. □

We now consider a special form of contamination. Let the true density function be g and the contamination model be $g_{\alpha,y} = (1-\alpha)g + \alpha u_y$ with u_y denotes the uniform density on $(y - \varepsilon, y + \varepsilon)$ for small $\varepsilon \geq 0$. Here $g_{\alpha,y}$ models the situation where a proportion α $(0 \leq \alpha \leq 1)$ of outliers located at (or near) y occurs in a sample from the density g. Note that the Hellinger distance between $g_{\alpha,y}$ and g is no more than $(2\alpha)^{1/2}$, since $\|g_{\alpha,y}^{1/2} - g^{1/2}\|^2 \leq \int |g_{\alpha,y}(x) - g(x)|dx = \int \alpha|u_y(x) - g(x)|dx \leq 2\alpha$. Define

$$\theta_{\alpha,y} = T_0(g_{\alpha,y}) = \arg\min_{t\in\Theta} \int \big(f_{t,\eta}^{1/2}(x) - g_{\alpha,y}^{1/2}(x)\big)^2 dx. \tag{4.60}$$

The next theorem compares $\theta_{\alpha,y}$ with θ defined in (4.45), which is a generalization of Theorem 7 in Beran (1977) to any density function g.

Theorem 4.8. *Suppose that Θ is a compact subset of $\mathbb{R}^p$. Further suppose that the family $\{f_{t,\eta} : t \in \Theta\}$ is identifiable, $t \mapsto s_t = f_{t,\eta}^{1/2}$ is continuous in L_2, and θ defined in (4.45) is unique. Then*

(i) $\lim_{\alpha \to 0} \theta_{\alpha,y} = \theta$ *for any* y.

(ii) *If* $\theta_{\alpha,y}$ *defined in (4.60) is unique for every* y, *then* $\theta_{\alpha,y}$ *is a continuous bounded function of* y *such that* $\lim_{|y| \to \infty} \theta_{\alpha,y} = \theta$.

(iii) *If* $\theta \in int(\Theta)$, *(4.5) and (4.6) hold for* $s_t = f_{t,\eta}^{1/2}$ *and for every* $t \in int(\Theta)$, *and* $\int \ddot{s}_\theta(x) g^{1/2}(x) dx$ *is nonsingular, then for every* y

$$\lim_{\alpha \to 0} \alpha^{-1}(\theta_{\alpha,y} - \theta) = \frac{1}{2} \int g^{-1/2}(x) \rho_g(x) u_y(x) dx, \tag{4.61}$$

where $\rho_g(x)$ *is defined in (4.54) and* $\dot{s}_\theta$ *and* $\ddot{s}_\theta$ *are defined in (4.5) and (4.6).*

Since Theorem 4.8 holds for any semiparametric model $f_{t,\eta}$, we can replace $f_{t,\eta}$ throughout with f_{t,η_n}, where η_n is an estimator of η. If further we replace g with its estimator g_n, then Theorem 4.8 holds with $s_t = f_{t,\eta_n}^{1/2}$ and corresponding $g = g_n$ and $\theta = \theta_n$ defined in (4.46).

Theorem 4.8 (i) is a special case of the consistency of MHD estimators. A more general result than Theorem 4.8 (i) is that the MHD estimator θ_n defined in (4.46) is robust in the sense that small Hellinger-metric perturbation in the underlying density g can only induce small changes in the density estimates g_n, and this in turn will only lead to small changes in the MHD estimator θ_n.

Theorem 4.8 (ii) represents the effect on MHD estimator (4.46) of adding some outliers with large values around y. It shows that for any fixed contamination rate $\alpha \in (0, 1)$ (even close to 1), MHD estimators based on the contaminated data set are close to those based on data sets without contamination for large enough y. This behavior is exhibited in the figures in Chapters 2 and 3, see particular Figure 3.1. Simulation studies in Section 4.6 further demonstrates this fact.

The limit defined in (4.61) gives the IF (a function of y) of the functional $\theta = T_0(g)$ defined in (4.45) at g, with modifications to Hampel's (1968) definition to suit functionals on a space of densities. As discussed above, (4.61) with $s_t = f_{t,\eta_n}^{1/2}$ and corresponding $g = g_n$ gives the IF of $T_n(g_n) = \theta_n$ defined in (4.46). These IFs are generally unbounded, but this does not rule out the robustness of MHD estimators, as in the parametric case (Beran, 1977) and in semiparametric cases considered in Chapters 2 and 3. In other words, a statistic does not need to have a bounded IF in order to be robust, as noted by Beran (1977) and many others. As shown in Theorem 4.8 (ii), the so called α-IF $\alpha^{-1}(\theta_{\alpha,y} - \theta)$ is a bounded continuous function of y such that $\lim_{|y| \to \infty} \alpha^{-1}(\theta_{\alpha,y} - \theta) = 0$. Hence the MHD estimator (4.46) is robust at g_n against $100\alpha\%$ contamination by gross errors at arbitrary real y. Thus the usage of the α-IF might be better than IF

to assess the robustness of statistics in the present context. See Beran (1977, pp 456-7) for further discussion on this issue.

Proof of Theorem 4.8. (i) Denote

$$d(t) = \|f_{t,\eta}^{1/2} - g^{1/2}\| \quad \text{and} \quad d_\alpha(t) = \|f_{t,\eta}^{1/2} - g_{\alpha,y}^{1/2}\|.$$

By Minkowski inequality,

$$\sup_{t\in\Theta}|d_\alpha(t) - d(t)| \le \|g_{\alpha,y}^{1/2} - g^{1/2}\| \le (2\alpha)^{1/2} \to 0 \quad \text{as } \alpha \to 0.$$

Therefore we have

$$d_\alpha(\theta) - d(\theta) \to 0 \quad \text{and} \quad d_\alpha(\theta_{\alpha,y}) - d(\theta_{\alpha,y}) \to 0 \quad \text{as } \alpha \to 0.$$

If $\theta_{\alpha,y} \nrightarrow \theta$ as $\alpha \to 0$, then there exists a sequence $\{\alpha_n\}$ such that $\alpha_n \to 0$ and $\theta_{\alpha_n,y} \to \theta' \neq \theta$ as $n \to \infty$. It is easy to prove that $\|f_{t,\eta}^{1/2} - \psi\|$ is continuous in t for any function $\psi \in L_2$, and thus $d(\theta_{\alpha_n,y}) \to d(\theta')$ as $n \to \infty$. From above results, we have $d_{\alpha_n}(\theta_{\alpha_n,y}) - d_{\alpha_n}(\theta) \to d(\theta') - d(\theta)$ as $n \to \infty$. Furthermore, we have $d_{\alpha_n}(\theta_{\alpha_n,y}) - d_{\alpha_n}(\theta) \le 0$ by the definition of $\theta_{\alpha,y}$, and hence $d(\theta') - d(\theta) \le 0$. But by the definition of θ and the uniqueness of it, $d(\theta') - d(\theta) > 0$. This is a contradiction. Therefore, $\theta_{\alpha,y} \to \theta$.

(ii) Let $\bar{g}_{\alpha,y} = [(1-\alpha)^{1/2} g^{1/2} + \alpha^{1/2} u_y^{1/2}]^2$. Since as $|y| \to \infty$,

$$\begin{aligned}
& \sup_{t\in\Theta}\big|\|f_{t,\eta}^{1/2} - g_{\alpha,y}^{1/2}\| - \|f_{t,\eta}^{1/2} - \bar{g}_{\alpha,y}^{1/2}\|\big| \\
\le\ & \|g_{\alpha,y}^{1/2} - \bar{g}_{\alpha,y}^{1/2}\| \\
\le\ & \Big[\int |g_{\alpha,y}(x) - \bar{g}_{\alpha,y}(x)|dx\Big]^{1/2} \\
=\ & (4\alpha(1-\alpha))^{1/4}\Big[\int g^{1/2}(x) u_y^{1/2}(x)dx\Big]^{1/2} \\
\to\ & 0
\end{aligned}$$

and

$$\begin{aligned}
& \|f_{t,\eta}^{1/2} - \bar{g}_{\alpha,y}^{1/2}\|^2 \\
=\ & 2 - 2(1-\alpha)^{1/2}\int f_{t,\eta}^{1/2}(x) g^{1/2}(x)dx - 2\alpha^{1/2}\int f_{t,\eta}^{1/2}(x) u_y^{1/2}(x)dx \\
& + 2\alpha^{1/2}(1-\alpha)^{1/2}\int g^{1/2}(x) u_y^{1/2}(x)dx \\
\to\ & 2 - 2(1-\alpha)^{1/2} + (1-\alpha)^{1/2}\|f_{t,\eta}^{1/2} - g^{1/2}\|^2,
\end{aligned}$$

we have for any $t \in \Theta$

$$\|f_{t,\eta}^{1/2} - g_{\alpha,y}^{1/2}\|^2 \to 2 - 2(1-\alpha)^{1/2} + (1-\alpha)^{1/2}\|f_{t,\eta}^{1/2} - g^{1/2}\|^2 \quad \text{as } |y| \to \infty. \quad (4.62)$$

If $\theta_{\alpha,y} \nrightarrow \theta$ as $|y| \to \infty$, then there exists a sequence $\{y_n\}$ such that $y_n \to \infty$ and $\theta_{\alpha,y_n} \to \theta' \neq \theta$ as $n \to \infty$. From (4.62), we have as $n \to \infty$,

$$\|f_{\theta',\eta}^{1/2} - g_{\alpha,y_n}^{1/2}\|^2 \to 2 - 2(1-\alpha)^{1/2} + (1-\alpha)^{1/2}\|f_{\theta',\eta}^{1/2} - g^{1/2}\|^2,$$

$$\|f_{\theta,\eta}^{1/2} - g_{\alpha,y_n}^{1/2}\|^2 \to 2 - 2(1-\alpha)^{1/2} + (1-\alpha)^{1/2}\|f_{\theta,\eta}^{1/2} - g^{1/2}\|^2. \tag{4.63}$$

Since $\left|\|f_{\theta_{\alpha,y_n},\eta}^{1/2} - g_{\alpha,y_n}^{1/2}\| - \|f_{\theta',\eta}^{1/2} - g_{\alpha,y_n}^{1/2}\|\right| \leq \|f_{\theta_{\alpha,y_n},\eta}^{1/2} - f_{\theta',\eta}^{1/2}\| \to 0$ by the continuity of $s_t = f_{t,\eta}^{1/2}$ in L_2, we have

$$\|f_{\theta_{\alpha,y_n},\eta}^{1/2} - g_{\alpha,y_n}^{1/2}\|^2 \to 2 - 2(1-\alpha)^{1/2} + (1-\alpha)^{1/2}\|f_{\theta',\eta}^{1/2} - g^{1/2}\|^2. \tag{4.64}$$

By definition, $\|f_{\theta_{\alpha,y_n},\eta}^{1/2} - g_{\alpha,y_n}^{1/2}\| \leq \|f_{\theta,\eta}^{1/2} - g_{\alpha,y_n}^{1/2}\|$. This together with (4.63) and (4.64) yield $\|f_{\theta',\eta}^{1/2} - g^{1/2}\| \leq \|f_{\theta,\eta}^{1/2} - g^{1/2}\|$. But by the uniqueness of θ, one has $\|f_{\theta',\eta}^{1/2} - g^{1/2}\| > \|f_{\theta,\eta}^{1/2} - g^{1/2}\|$. This is a contradiction. Therefore, $\theta_{\alpha,y} \to \theta$ as $|y| \to \infty$.

Note that $\|g_{\alpha,y+\delta}^{1/2} - g_{\alpha,y}^{1/2}\|^2 \leq \int |g_{\alpha,y+\delta}(x) - g_{\alpha,y}(x)|dx = \alpha \int |u_{y+\delta}(x) - u_y(x)|dx = \alpha\delta/\ \varepsilon \to 0$ as $\delta \to 0$. Hence $y \mapsto g_{\alpha,y}^{1/2}$ is continuous in L_2. Since the functional T_0 in (4.60) is continuous at $g_{\alpha,y}$ in the Hellinger topology (see Theorem 1 of Beran (1977)), one has that $\theta_{\alpha,y+\delta} \to \theta_{\alpha,y}$ as $\delta \to 0$, i.e. $\theta_{\alpha,y}$ is a continuous function of y. The boundedness of $\theta_{\alpha,y}$ follows immediately.

(iii) Obviously Theorems 1 and 2 in Beran (1977) hold. As a result, The proof follows along the same line as the proof of Theorem 7 in Beran (1977). □

4.6 *Simulation Studies*

In this section, we report the results of a Monte Carlo study designed to demonstrate the efficiency and robustness of the proposed MHD estimator defined in (4.4). We considered MHD estimation in mixture models. Specifically, we considered the semiparametric models

$$\{f_{\theta,\eta} : f_{\theta,\eta} = \theta\phi(0,1) + (1-\theta)\eta, 0 \leq \theta \leq 1, \eta \text{ is a density function}\},$$

where $\phi(\mu,\sigma)$ denotes the normal density function with mean μ and standard deviation σ. We examined the situation where $\eta = \phi(a,b)$, i.e., normal mixture models. Let $\Phi(\mu,\sigma)$ denote the distribution function of $\phi(\mu,\sigma)$. For different values of θ, a and b, we considered ten normal mixture models displayed in Table 4.1. The value of a was chosen to provide the desired overlap between components, as defined by Woodward et al. (1995).

Tab. 4.1: Summary of mixture models under study.

θ	Scale parameter b	Overlap	Mixture model
0.25	1	0.03	$0.25\Phi(0,1)+0.75\Phi(3.6,1)$ (I)
		0.1	$0.25\Phi(0,1)+0.75\Phi(2.32,1)$ (II)
0.5	1	0.03	$0.5\Phi(0,1)+0.5\Phi(3.76,1)$ (III)
		0.1	$0.5\Phi(0,1)+0.5\Phi(2.56,1)$ (IV)
0.25	$\sqrt{2}$	0.03	$0.25\Phi(0,1)+0.75\Phi(4.46,2)$ (V)
		0.1	$0.25\Phi(0,1)+0.75\Phi(2.96,2)$ (VI)
0.5	$\sqrt{2}$	0.03	$0.5\Phi(0,1)+0.5\Phi(4.52,2)$ (VII)
		0.1	$0.5\Phi(0,1)+0.5\Phi(3.07,2)$ (VIII)
0.75	$\sqrt{2}$	0.03	$0.75\Phi(0,1)+0.25\Phi(4.20,2)$ (IX)
		0.1	$0.75\Phi(0,1)+0.25\Phi(2.57,2)$ (X)

1. Robustness

This subsection analyzes the robustness of the proposed MHD estimator defined by (4.4) for the normal mixture models labeled I to X in Table 4.1. We examined the resistance of the MHD estimator to a single outlying observation. For this purpose, the α-IF given in Beran (1977) is a suitable measure of the change in the estimator. Here we have used the adapted version of the α-IF employed by Lu et al. (2003).

For the ten models in Table 4.1, we chose a sample of size $n = 100$ from the mixture model $f_{\theta,\eta}$. To construct an estimator η_n of η, we chose another sample of size $n_0 = 40$ from the distribution η, i.e., the second component in the mixture model. So our data structure is

$$\begin{aligned} X_1,\ldots,X_n &\overset{iid}{\sim} \theta\Phi(0,1)+(1-\theta)\Phi(a,b) \\ Y_1,\ldots,Y_{n_0} &\overset{iid}{\sim} \Phi(a,b). \end{aligned} \tag{4.65}$$

Note that the outlying observation could come from either the X_i's or the Y_i's. Thus, after drawing data sets of the specified sizes, 98 alternate versions of the data were created by replacing the last observation in the sample X_i's, or the last observation in the sample Y_i's by an integer x from -24 to 24. We have done ten replications and calculated the average of the ten replications. The contamination rate α is then $1/140$ and the two α-IFs are given by

$$IF(x) = \frac{W\big((x,X_i)_{i=1}^{n-1},(Y_i)_{i=1}^{n_0}\big) - W\big((X_i)_{i=1}^{n},(Y_i)_{i=1}^{n_0}\big)}{1/140} \tag{4.66}$$

and

$$IF_0(x) = \frac{W\big((X_i)_{i=1}^n, (x, Y_i)_{i=1}^{n_0-1}\big) - W\big((X_i)_{i=1}^n, (Y_i)_{i=1}^{n_0}\big)}{1/140}, \tag{4.67}$$

where W could be any functional (estimator of θ) based on two data sets from $f_{\theta,\eta}$ and η, respectively. In our case, W is functional T_n defined in (4.4). Next we define following adaptive kernel density estimators (see, e.g., Silverman, 1986) of $f_{\theta,\eta}$ and η, respectively, based on data $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_{n_0}$ of (4.65):

$$f_n(x) = \frac{1}{nS_n b_n} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{S_n b_n}\Big), \tag{4.68}$$

$$\eta_n(x) = \frac{1}{n_0 S_{n_0} b_{n_0}} \sum_{j=1}^{n_0} K_0\Big(\frac{x - Y_j}{S_{n_0} b_{n_0}}\Big), \tag{4.69}$$

where K and K_0 are two smooth density functions, bandwidths b_n and b_{n_0} are positive constants such that $b_n \to 0$ as $n \to \infty$ and $b_{n_0} \to 0$ as $n_0 \to \infty$, and $S_n = S_n(X_1, \ldots, X_n)$ and $S_{n_0} = S_{n_0}(Y_1, \ldots, Y_{n_0})$ are robust scale statistics (these statistics generally estimate the scale parameters of respective distributions). We used the compact-supported Epanechnikov kernel function

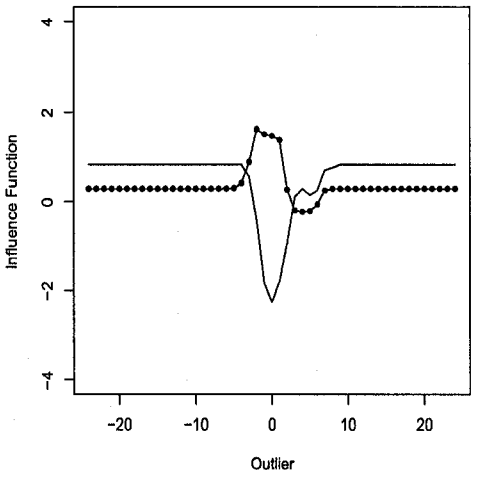
$$K(x) = \frac{3}{4}\left(1 - x^2\right) I_{[-1,1]}(x) \tag{4.70}$$

for kernels K and K_0 in (4.68) and (4.69), respectively. The bandwidths b_n and b_{n_0} in (4.68) and (4.69), respectively, were taken to be $b_n = n^{-1/3}$ and $b_{n_0} = n_0^{-1/3}$. For scale statistics S_n and S_{n_0} in (4.68) and (4.69), respectively, we used the following robust scale estimator proposed by Rousseeuw and Croux (1993),

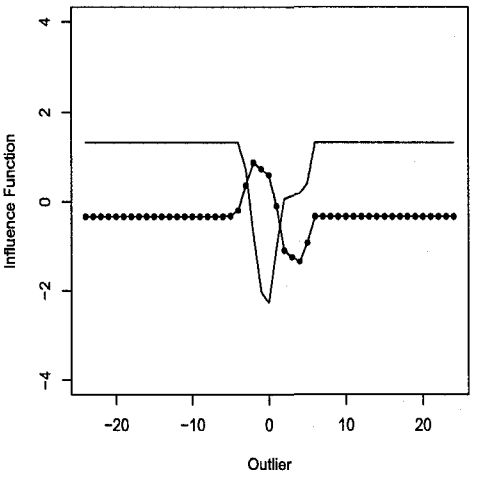
$$S_n = 1.1926 \ \mathrm{med}_i\big(\mathrm{med}_j(|X_i - X_j|)\big).$$

For the average of the ten replications, the α-IFs (4.66) and (4.67) under the ten models in Table 4.1 are calculated, of which four are graphically displayed in Figure 4.1. The α-IFs under other models are similar. From Figure 4.1, we can see that as the outlier approaches $\pm\infty$, the α-IF appears to converge to a constant, i.e., $\lim_{x\to\infty} IF(x) = \lim_{x\to-\infty} IF(x)$ and $\lim_{x\to\infty} IF_0(x) = \lim_{x\to-\infty} IF_0(x)$. This phenomenon is partially explained by Theorem 4.8 (ii). In fact, the α-IFs outside the interval $[-7, 10]$ seem to be constant, while they take varying values inside the interval $[-7, 10]$. Specifically, IF_0 has a lower value inside the interval $[-7, 10]$ than outside the interval.

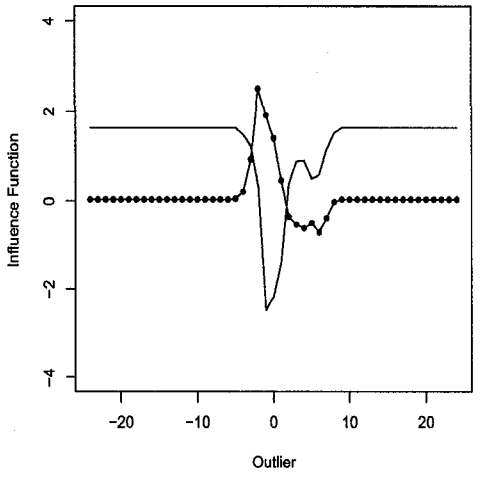
We also compared our MHD estimator with two MLEs. We examined the two MLEs based on following likelihood functions combined with the data $(X_1, \ldots, X_n)$:



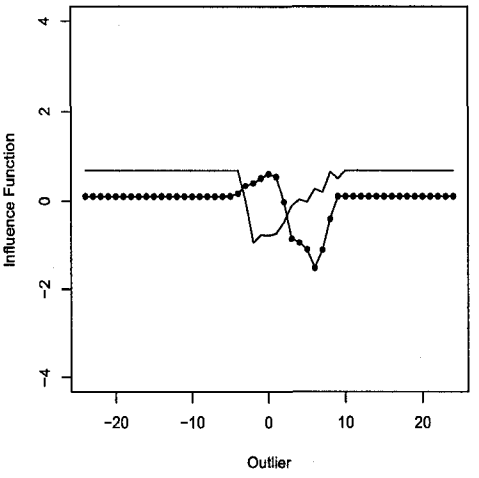
(a) Model I



(b) Model IV



(c) Model VI



(d) Model IX

Fig. 4.1: The α-influence function of MHD estimator θ_n with respect to single outlier, with $\bullet$ - IF and $-$ - IF_0.

$$L = \prod_{i=1}^{n} [\theta f(X_i) + (1-\theta)\eta(X_i)]$$

and

$$L_n = \prod_{i=1}^{n} [\theta f(X_i) + (1-\theta)\eta_n(X_i)],$$

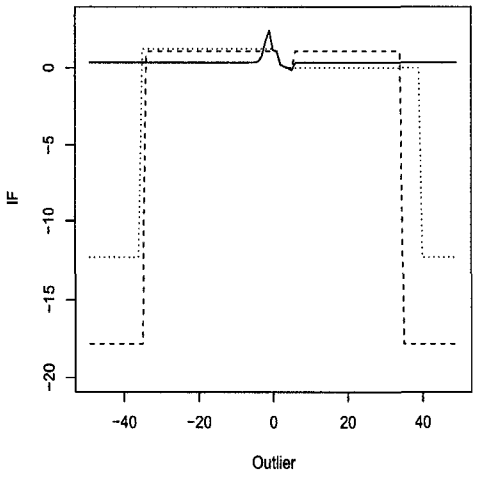
where $f = \phi(0,1)$ and η_n is the kernel density estimator of η defined by (4.69). In other words, the likelihood L is constructed assuming that density functions f and g are completely known, whereas L_n is obtained by replacing η by its estimator η_n. Thus, L and L_n are rather naturally constructed for simulation purposes. We define

$$\theta_{\text{MLE}} = \arg\max_{\theta \in [0,1]} L \tag{4.71}$$

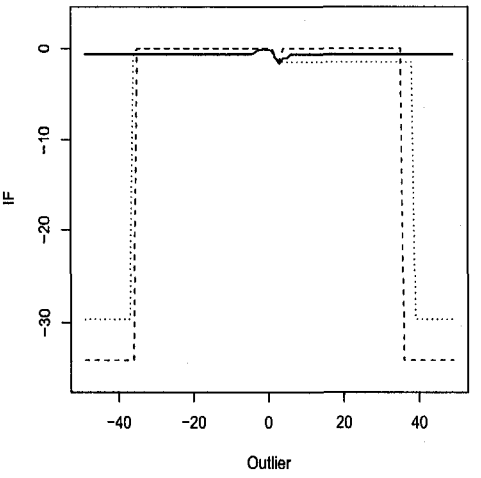
and

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta \in [0,1]} L_n \tag{4.72}$$

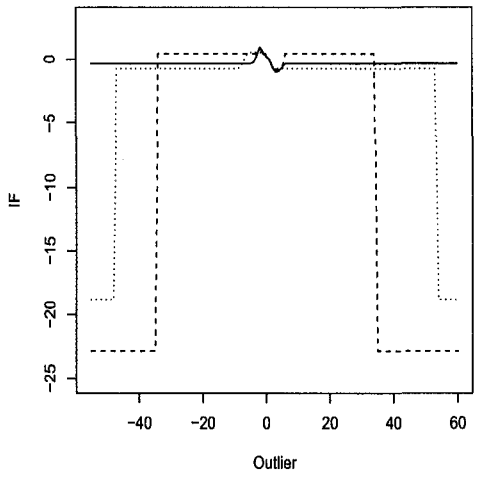
as the MLEs of θ based on L and L_n, respectively. In our simulation, the data were again generated from the models defined in Table 4.1. For each model, samples of sizes $n = 50$ and $n_0 = 20$ were obtained from the corresponding distributions. For instance, for Model I, samples of size $n = 50$ were obtained from the mixture distribution $0.25\Phi(0,1) + 0.75\Phi(3.6,1)$, while a sample of size $n_0 = 20$ was obtained from the distribution $\Phi(3.6,1)$. We used (4.66) to calculate α-IFs for θ_n, θ_{MLE} and $\widehat{\theta}_{\text{MLE}}$ defined in (4.4), (4.71) and (4.72), respectively. For the sake of consistency, we used the contamination rate $\alpha = 1/50 = 0.02$ in (4.66). For a single sample, the α-IFs of the three estimators for Model I, IV, VI and IX are displayed in Figure 4.2. Influence functions under other models are similar. From Figure 4.2, we can see that all the α-IFs of θ_n, θ_{MLE} and $\widehat{\theta}_{\text{MLE}}$ are approximately symmetric about zero. When the outlier is between -30 and 30, the three estimators are competitive and the α-IFs take values between -3 and 3. As mentioned in the Figure 4.1, the α-IF of θ_n outside the interval $[-7,7]$ seems to be constant, while the α-IFs of θ_{MLE} and $\widehat{\theta}_{\text{MLE}}$ have explored at some point around ± 40 and they take values as high as 41.27. Nevertheless, θ_{MLE} works better than $\widehat{\theta}_{\text{MLE}}$ in the sense that the 'exploration' point of θ_{MLE} is higher than that of $\widehat{\theta}_{\text{MLE}}$ and the α-IF of θ_{MLE} after the exploration point has smaller absolute value than that of $\widehat{\theta}_{\text{MLE}}$. This behavior can be expected since θ_{MLE} employs more information (i.e., knowing η, or in other words $n_0 = \infty$) than either θ_n or $\widehat{\theta}_{\text{MLE}}$. Note that θ_{MLE} is not available in practice and the sole purpose of analyzing it here is to examine the amount of loss in performance when η is unknown. Figure 4.2 shows that θ_n is more robust than either θ_{MLE} or $\widehat{\theta}_{\text{MLE}}$ in the sense of resistance to a single outlying observation.



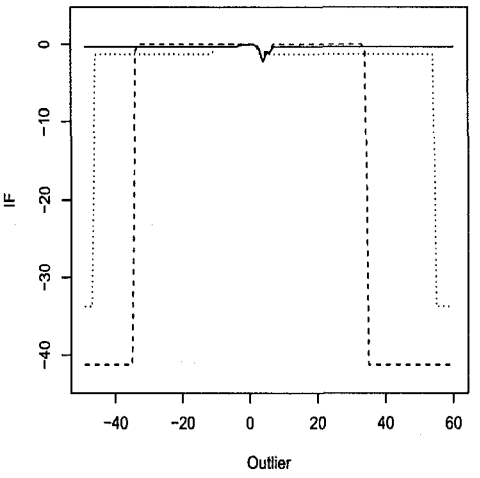
(a) Model I



(b) Model IV



(c) Model VI



(d) Model IX

Fig. 4.2: The α-influence functions for θ_n (solid), $\widehat{\theta}_{\mathrm{MLE}}$ (dashed) and θ_{MLE} (dotted) with respect to single outlier.

The breakdown point is the smallest fraction of data that, when strategically placed, can cause an estimator to give an arbitrarily bad answer. Tamura and Boos (1986) gave breakdown results for MHD estimators of multivariate location and covariance. Cutler and Cordero-Braña (1996) investigated the breakdown point of MHD estimators for mixture models. The models considered in both of these papers are parametric models, while ours is a semiparametric model (η unknown). We again considered the normal mixture model $f_{\theta,\eta} = \theta\phi(0,1)+(1-\theta)\phi(\mu,b)$ with $b = 1, \sqrt{2}$, $\theta = 0.25, 0.5, 0.75$, and varying μ values. Define the contamination model

$$(1-\alpha)(\theta\phi(0,1) + (1-\theta)\phi(\mu,b)) + \alpha I_{\{10\}}$$

with contamination of the point mass function $I_{\{10\}}$ and contamination rate α. Here we numerically compared the behavior of θ_n and $\widehat{\theta}_{\text{MLE}}$ defined in (4.4) and (4.72), respectively, as we vary the value of μ. For given values of θ, μ and b, consider increasing α until θ_n jumps to fit the contamination, and similarly for $\widehat{\theta}_{\text{MLE}}$. We used sample sizes $n = 50$ and $n_0 = 20$ for one single sampling. To increase α, we replaced the last observation X_{50} from the mixture model with a value 10, and then the second last, and so on. The values of μ are $\mu = 0.5k$, $k = 1, 2, \ldots, 14$. If the estimator jumps to and stays at value 1 as α increases, then the estimator is fitting the contamination. The reason for this is that we are using a compact-supported kernel function (4.70) for density estimation. The results for the models $(\theta, b) = (0.25, 1)$ and $(0.5, \sqrt{2})$ are shown in Figure 4.3. The breakdown points under other normal mixture models are similar. From Figure 4.3 we can see that the breakdown point α for θ_n seems to be constant 0.5 for any μ value between 0.5 and 7.0, while for $\widehat{\theta}_{\text{MLE}}$ it is around 0.25 for μ values between 0.5 and 7.0. So the breakdown point for θ_n is about twice of that for $\widehat{\theta}_{\text{MLE}}$. In other words, MHD estimator θ_n shows more robust behavior than the MLE estimator $\widehat{\theta}_{\text{MLE}}$ in our simulation.

2. Efficiency

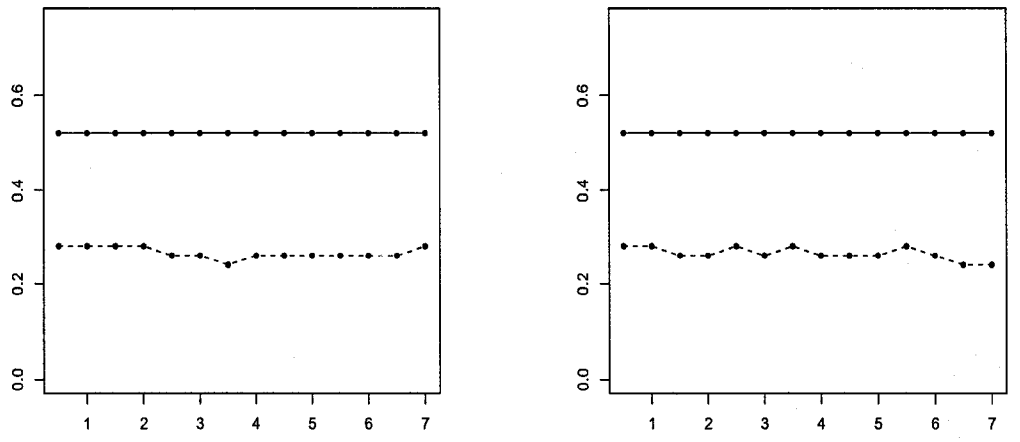
In each of the distributional situations considered in Table 4.1, we obtained estimates of the bias and mean squared error (MSE) as follows:

$$\widehat{\text{Bias}} = \frac{1}{N_s}\sum_{i=1}^{N_s}(\widehat{\mu}_i - \mu)$$

and

$$\widehat{\text{MSE}} = \frac{1}{N_s}\sum_{i=1}^{N_s}(\widehat{\mu}_i - \mu)^2,$$

where N_s is the number of replications, and $\widehat{\mu}_i$ denotes an estimate of μ for the



(a) $\theta = 0.25$ and $b = 1$ (b) $\theta = 0.5$ and $b = \sqrt{2}$

Fig. 4.3: The smallest proportion α of contamination at which θ_n (solid) and $\widehat{\theta}_{\text{MLE}}$ (dashed) fit the contamination, as a function of μ, with the contamination model $(1-\alpha)(\theta\phi(0,1) + (1-\theta)\phi(\mu, b)) + \alpha I_{\{10\}}$.

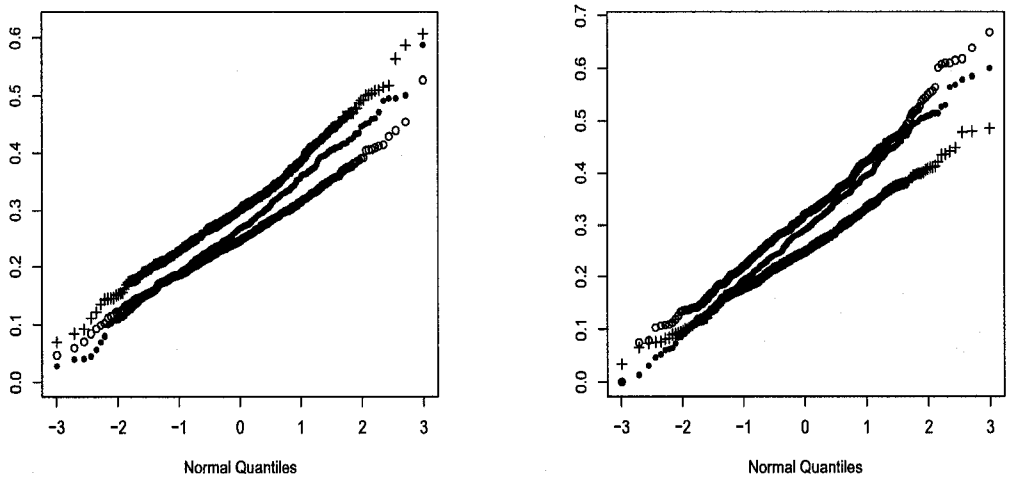
ith replication. Here $\mu = \theta$ and $\widehat{\mu}$ denotes either the proposed MHD estimator θ_n or the MLEs θ_{MLE} and $\widehat{\theta}_{\text{MLE}}$. We chose $N_s = 500$, $n = 50$ and $n_0 = 20$ in our simulation. Kernel estimators f_n and η_n are the same as those employed in the robustness study above. Simulation results are summarized in Table 4.2.

We found that the MHD estimator θ_n performed competitively with the MLE $\widehat{\theta}_{\text{MLE}}$ for all ten models. Thus, it is not surprising that in many circumstances the MHD estimator achieves about the same efficiency as that of the MLE under semiparametric models. On the other hand, the MLE θ_{MLE}, which is based on assuming η is known, showed the best performance among the three estimators for all ten models. This behavior can be expected for the reason mentioned in the robustness study and the fact that the lower bound of the asymptotic variance is higher when η is unknown than when it is known. In Figure 4.4, we have given the normal probability plots of the three estimators for Models I and VI. Figure 4.4 demonstrates that the sampling distribution of θ_n closely approximates a normal curve for each model considered. We have observed very similar plots for other models considered as well.

We also investigated the relative biases and relative MSEs of θ_n to $\widehat{\theta}_{\text{MLE}}$ for the contamination model $(1-\alpha)f_{\theta,\eta} + \alpha I_{\{10\}}$ with $f_{\theta,\eta}$ being one of the models defined in Table 4.1. We again chose $N_s = 500$, $n = 50$ and $n_0 = 20$ in our simulation. We considered four contamination rates, 2%, 4%, 10% and 20%. For the contamination rate 2%, we replaced the last observation X_{50} with a

Tab. 4.2: Estimates of the biases and mean squared errors of θ_n, $\widehat{\theta}_{\mathrm{MLE}}$ and θ_{MLE} with no contamination.

Model	$\widehat{\mathrm{Bias}}(\theta_n)$	$\widehat{\mathrm{MSE}}(\theta_n)$	$\widehat{\mathrm{Bias}}(\widehat{\theta}_{\mathrm{MLE}})$	$\widehat{\mathrm{MSE}}(\widehat{\theta}_{\mathrm{MLE}})$	$\widehat{\mathrm{Bias}}(\theta_{\mathrm{MLE}})$	$\widehat{\mathrm{MSE}}(\theta_{\mathrm{MLE}})$
I	0.0210	0.0075	0.0564	0.0099	0.0002	0.0044
II	0.0375	0.0138	0.0809	0.0189	-0.0021	0.0059
III	0.0392	0.0088	0.0359	0.0084	-0.0011	0.0060
IV	0.0511	0.0119	0.0533	0.0115	-0.0013	0.0069
V	0.0308	0.0084	0.0587	0.0115	0.0017	0.0046
VI	0.0430	0.0127	0.0705	0.0154	0.0026	0.0060
VII	0.0439	0.0087	0.0378	0.0081	0.0022	0.0054
VIII	0.0483	0.0117	0.0404	0.0098	-0.0009	0.0069
IX	0.0501	0.0078	0.0166	0.0044	-0.0001	0.0037
X	0.0556	0.0112	0.0182	0.0070	-0.0041	0.0060



(a) Model I

(b) Model VI

Fig. 4.4: Normal probability plots of estimates θ_n (•), $\widehat{\theta}_{\mathrm{MLE}}$ (∘) and θ_{MLE} (+).

value 10, for the contamination rate 4% we replaced the last two observations X_{49} and X_{50} with a value 10, and so on. Simulation results are summarized in Table 4.3. From Table 4.3 one can see that most of the relative values are less than one with exceptions on models with $\theta = 0.75$. The relative biases and relative MSEs are especially small for models with $\theta = 0.25$. An interesting observation is that the relative biases and relative MSEs are uniformly smaller for higher contamination rate α than for lower α. In particular, the relative MSEs for models VII and VIII are bigger than one when $\alpha = 2\%$, while those are less than one when $\alpha = 4\%$. All the relative biases and relative MSEs decrease when the contamination rate α increases. One could probably expect that all the relative bias and relative MSE values would be close to or less than one when the contamination rate increases. This is another indication that θ_n seems to show more robust behavior than $\widehat{\theta}_{\text{MLE}}$ in our simulation.

4.7 An Example

In this section, we consider a specific semiparametric model, the symmetric location model. Here we construct and investigate the MHD estimator (4.4) and MPHD estimator (4.36) for the parameter of interest. We will show that the MPHD estimator of the location turns out to be an adaptive estimator, and the MHD estimator of the location is also efficient in the parametric sense.

Symmetric Location Model. Assume that the data $X_1, \ldots, X_n \in \mathbb{R}$ are i.i.d. and satisfy the model

$$X = \theta + \varepsilon,$$

where the center θ is the parameter to be estimated and the error ε has a continuous density $\eta(\cdot)$ that is symmetric about the origin.

Therefore, the semiparametric model under consideration is

$$\{f_{\theta,\eta}(x) = \eta(x-\theta) : \theta \in \mathbb{R}, \eta \in \mathcal{H}\}, \tag{4.73}$$

where

$$\mathcal{H} = \{h \in L_1 : h \geq 0,\ h \neq 0,\ h(-x) = h(x), h \text{ is continuous}\}.$$

Although the parameter space for θ is the real line in this case, it is reasonable to set $\Theta = [-C, C]$ with C being a large positive number such that the true parameter $\theta \in \text{int}(\Theta)$. Such a C could be decided based on the observations X_i's, e.g., one could let $C = \max_{i=1,2\ldots,n} \{|X_i|\}$. With this assumption, we will not lose any information about θ and at the same time we can guarantee the consistency of the MHD estimator in most cases.

Tab. 4.3: Relative bias (RB) and relative MSEs (RM) of θ_n to $\widehat{\theta}_{\mathrm{MLE}}$ for the contamination model $(1-\alpha)f_{\theta,\eta} + \alpha I_{\{10\}}$ with $f_{\theta,\eta}$ being one of the models defined in Table 4.1.

Model $f_{\theta,\eta}$	RB $\alpha = 2\%$	RM $\alpha = 2\%$	RB $\alpha = 4\%$	RM $\alpha = 4\%$	RB $\alpha = 10\%$	RM $\alpha = 10\%$	RB $\alpha = 20\%$	RM $\alpha = 20\%$
I	0.2518	0.6760	0.2398	0.5946	0.1851	0.3406	0.1093	0.2061
II	0.3571	0.6850	0.3225	0.5461	0.2472	0.3653	0.1688	0.1817
III	0.8386	0.9650	0.7015	0.9032	0.4799	0.6644	0.2831	0.4195
IV	0.7775	0.9648	0.6603	0.9268	0.4139	0.6526	0.2502	0.4168
V	0.4128	0.6153	0.3072	0.5932	0.2420	0.3428	0.1668	0.1906
VI	0.4490	0.7187	0.4189	0.6031	0.2541	0.3967	0.1883	0.2002
VII	0.9381	1.0391	0.7870	0.9795	0.4819	0.7032	0.3347	0.3843
VIII	0.8748	1.0759	0.6664	0.9696	0.4811	0.7388	0.3014	0.4460
IX	2.3012	1.7285	1.9038	1.6732	1.0932	1.5204	0.6657	1.0013
X	1.9267	1.5780	1.4644	1.4653	0.9483	1.3286	0.5442	0.9537

1. MPHD estimator of θ

To construct an efficient MPHD estimator for the location parameter θ in models (4.73), we first look at the MHD functional T_1 defined in (4.36). Note that for any density function g,

$$g^{1/2}(x+t) = \frac{1}{2}\big(g^{1/2}(t+x) + g^{1/2}(t-x)\big) + \frac{1}{2}\big(g^{1/2}(t+x) - g^{1/2}(t-x)\big)$$

and that the first term on the r.h.s. in the above expression is an even function of x while the second is odd. Then

$$\|f_{t,\eta}^{1/2} - g^{1/2}\|^2 = \int \big(\eta^{1/2}(x) - g^{1/2}(x+t)\big)^2 dx \geq \frac{1}{4}\int \big(g^{1/2}(t+x) - g^{1/2}(t-x)\big)^2 dx$$

with equality if $\eta^{1/2}(x) = \frac{1}{2}\big(g^{1/2}(t+x) + g^{1/2}(t-x)\big)$. Thus we have

$$\eta(t,g) = \frac{1}{4}\big(g^{1/2}(t-x) + g^{1/2}(t+x)\big)^2$$

and

$$s_{t,g} = \frac{1}{2}\big(g^{1/2}(2t-x) + g^{1/2}(x)\big).$$

With $\phi_g = g^{(1)}/(2g^{1/2})$, we have $\dot{s}_{t,g}(x) = \phi_g(2t-x)$. The function H defined in (4.37) becomes

$$H(t,g) = \int \phi_g(2t-x)g^{1/2}(x)dx = \int \phi_g(x)g^{1/2}(2t-x)dx,$$

and thus

$$\dot{H}(t,g) = 2\int \phi_g(x)\phi_g(2t-x)dx.$$

The fact that $\int g_n^{1/2}(2t-x)g^{1/2}(x)dx = \int g_n^{1/2}(x)g^{1/2}(2t-x)dx$ gives

$$\int \phi_{g_n}(2t-x)g^{1/2}(x)dx = \int g_n^{1/2}(x)\phi_g(2t-x)dx,$$

i.e., (4.41) holds. Hence we have

$$\begin{aligned} & H(t,g_n) - H(t,g) \\ = & \int \phi_{g_n}(2t-x)\big(g_n^{1/2}(x) - g^{1/2}(x)\big)dx \\ & \qquad + \int \big(\phi_{g_n}(2t-x) - \phi_g(2t-x)\big)g^{1/2}(x)dx \\ = & \int \phi_{g_n}(2t-x)\big(g_n^{1/2}(x) - g^{1/2}(x)\big)dx + \int \phi_g(2t-x)\big(g_n^{1/2}(x) - g^{1/2}(x)\big)dx \end{aligned}$$

$$= 2\int \phi_g(2t-x)\big(g_n^{1/2}(x) - g^{1/2}(x)\big)dx + O\big(\|g_n^{1/2} - g^{1/2}\| \cdot \|\phi_{g_n} - \phi_g\|\big).$$

Thus if $\|\phi_{g_n} - \phi_g\| \to 0$, then (4.38) holds with $\psi_g(x) = 2\phi_g(2T_1(g) - x)$. For models (4.73), $g(x) = f_{\theta,\eta}(x) = \eta(x-\theta)$, and then

$$\psi_{f_{\theta,\eta}} = 2\phi_{f_{\theta,\eta}}(2\theta - x) = \frac{\eta^{(1)}(\theta - x)}{\eta^{1/2}(\theta - x)} = -\frac{\eta^{(1)}(x-\theta)}{\eta^{1/2}(x-\theta)},$$

$$< \psi_{f_{\theta,\eta}},\ f_{\theta,\eta}^{1/2} >= 0,$$

$$s_t := s_{t,f_{\theta,\eta}} = \frac{1}{2}\big(\eta^{1/2}(2t - x - \theta) + \eta^{1/2}(x-\theta)\big),$$

$$\dot{s}_\theta := \dot{s}_{\theta,f_{\theta,\eta}} = -\frac{\eta^{(1)}(x-\theta)}{2\eta^{1/2}(x-\theta)},$$

$$\ddot{s}_\theta := \ddot{s}_{\theta,f_{\theta,\eta}} = \frac{\eta^{(2)}(x-\theta)}{\eta^{1/2}(x-\theta)} - \frac{(\eta^{(1)}(x-\theta))^2}{2\eta^{3/2}(x-\theta)}.$$

Define $M(f_1, f_2) = \|\phi_{f_1} - \phi_{f_2}\|$ for any density functions f_1 and f_2. Then it is easy to see that M is a metric and $\dot{H}$ is continuous in the sense that $\dot{H}(t_n, g_n) \to \dot{H}(t, g)$ whenever $t_n \to t$ and $M(g_n, g) \to 0$ as $n \to \infty$. Suppose that the Fisher information of θ,

$$I_\theta = \int (\eta^{(1)}(x))^2/\eta(x)dx,$$

is finite and nonzero. Then $\|\psi_{f_{\theta,\eta}}\|^2 = I_\theta < \infty$. Therefore, condition (iii) in Theorem 4.5 holds. Further assume that $\{g_n\}_{n\in\mathbb{N}}$ is a sequence of estimators of $f_{\theta,\eta}$ such that $\|g_n^{1/2} - f_{\theta,\eta}^{1/2}\| \xrightarrow{P} 0$ and $M(g_n, f_{\theta,\eta}) \xrightarrow{P} 0$ as $n \to \infty$ and satisfies (4.39), i.e., condition (iv) in Theorem 4.5 holds. Then the MPHD estimator, as defined in (4.36), is

$$\begin{aligned} T_1(g_n) &= \arg\min_{t\in\Theta} \|s_{t,g_n} - g_n^{1/2}\| = \arg\min_{t\in\Theta} \|g_n^{1/2}(2t-x) - g_n^{1/2}(x)\| \\ &= \arg\max_{t\in\Theta} \int g_n^{1/2}(2t-x)g_n^{1/2}(x)dx = \arg\max_{t\in\Theta} \int s_{t,g_n}(x)g_n^{1/2}(x)dx. \end{aligned}$$

The preceding estimator is identical to the estimator proposed in Beran (1978). In other words, Beran's estimator is a special case of the MPHD estimator.

Since any function in $\mathcal{H}$ can have only one symmetric point, the models defined in (4.73) is identifiable and thus $T_1(f_{\theta,\eta}) = \theta$ is well defined and unique. This fact is also shown in Lemma 1 of Beran (1978). Lemma 2 in Beran (1978) proves that $T_1(g)$ is Hellinger continuous at $g = f_{\theta,\eta}$. Thus condition (i) in Theorem 4.5 holds. Note that

$$- < \ddot{s}_\theta,\ f_{\theta,\eta}^{1/2} >= 2 < \dot{s}_\theta,\ \dot{s}_\theta >= I_\theta/2.$$

If $\frac{\eta^{(1)}}{\eta^{1/2}}$, $\frac{(\eta^{(1)})^2}{\eta^{3/2}}$ and $\frac{\eta^{(2)}}{\eta^{1/2}}$ are all in L_2 and continuous, then condition (ii) in Theorem 4.5 is easily satisfied. Combined with all above discussion, Theorem 4.5 holds. As a result, $n^{1/2}(T_n(g_n) - \theta)$ is asymptotically normally distributed with mean zero and variance I_θ^{-1}. Note that I_θ is the regular Fisher information for θ when η is known. This means that the MPHD estimator $T_1(g_n)$ is an adaptive estimator for the location parameter θ, provided one can construct an estimator g_n of $f_{\theta,\eta}$ that satisfies condition (iv) in Theorem 4.5. Here we choose g_n as the smoothly truncated kernel density estimator proposed in Beran (1978). Under certain conditions, Beran (1978) proved that $\|g_n^{1/2} - f_{\theta,\eta}^{1/2}\| \xrightarrow{P} 0$ (Theorem 1) and $M(g_n, f_{\theta,\eta}) \to 0$ w.p.1 (Lemma 4). The proof of Theorem 2 in Beran (1978) also shows that (4.39) holds. Since our MPHD estimator for location is the same as that in Beran (1978), a detailed construction and proofs are omitted here. A detailed construction of a MHD estimator is given in the next subsection.

2. MHD estimator of θ

In this subsection, we construct and investigate a MHD estimator of the location parameter θ. To avoid technical difficulties, here we only consider the case that η has a finite support. Clearly, for every $\eta \in \mathcal{H}$, the model $\{f_{t,\eta} : t \in \Theta\}$ is identifiable, $t \mapsto s_t = f_{t,\eta}^{1/2}$ is continuous in L_2, and (4.5) and (4.6) hold for s_t and for every $t \in \text{int}(\Theta)$; i.e., condition (ii) of Theorem 4.2 holds.

We define following kernel density estimator of $f_{\theta,\eta}$ based on data $X_1, \ldots, X_n$:

$$f_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K_1\left(\frac{x - X_i}{b_n}\right), \tag{4.74}$$

where K_1 is a differentiable density function and bandwidths $\{b_n\}$ is a sequence of positive numbers such that $b_n \to 0$ as $n \to \infty$. We can also use an adaptive kernel density estimator (see, e.g., Silverman, 1986), which uses $S_n b_n$ instead of b_n with S_n a robust scale statistic. Here we employed a non-adaptive kernel density estimator for simplicity.

An estimator η_n of η can be constructed based on the same data set $X_1, \ldots, X_n$ or it could be based on other resources, such as another data set from the density η. When no other resources than X_i's are available, we can split the X_i's into two groups $\{X_1, \ldots, X_m\}$ and $\{X_{m+1}, \ldots, X_n\}$ with $m = [n/2]$, the integer part of $n/2$. Based on the second group, one can construct an initial estimator of θ (for example the mean or median) and denote the corresponding estimator by $\overline{X}_{n-m}$. Then, based on the transformed values $Z_i = X_i - \overline{X}_{n-m}$, $i = 1, \ldots, m$, one can construct an estimator η_m of η by using kernel or by any other suitable

nonparametric density estimation technique. For simplicity, we suppose there is another data set $Y_1, \ldots, Y_m$ from the density function η. Another important reason why we chose this situation here is that for classical estimators of location (mean or median) it is not easy to utilize the information contained in the Y_i's and the second data set will likely be ignored. While the sample mean is an efficient but non-robust estimator and the sample median is a robust but non-efficient estimator, we will use the MHD method to construct an efficient and robust estimator based on the information contained in both the X_i's and the Y_i's.

To construct a symmetric estimator of η, one can generate pseudo data by reflecting all the Y_i's around the origin. Based on these $2m$ values, $Y_1, \ldots, Y_{2m}$, one can define following kernel density estimator of η,

$$\eta_m(x) = \frac{1}{2mb_m} \sum_{i=1}^{2m} K_2\left(\frac{x - Y_i}{b_m}\right), \tag{4.75}$$

where K_2 is a differentiable density function symmetric about origin and bandwidths $\{b_m\}$ is a sequence of positive numbers such that $b_m \to 0$ as $m \to \infty$. Obviously $\eta_m \in \mathcal{H}$ and $f_{t,\eta_m}(x) = \eta_m(x-t)$ is an estimator of $f_{t,\eta}$ for any $t \in \Theta$. Denote $s_t = \eta^{1/2}(x-t)$, $\widehat{s}_t = \eta_m^{1/2}(x-t)$ and $\dot{\widehat{s}}_t = \frac{\partial}{\partial t}\widehat{s}_t$. Now we can define the MHD estimator θ_n of the location θ as in (4.4). The next theorem establishes the efficiency of θ_n in the parametric sense, i.e. the adaptivity.

Theorem 4.9. *Suppose that $\eta > 0$ on its compact support W_η, K_1 and K_2 in (4.74) and (4.75), respectively, are differentiable and symmetric about origin, $K_1 > 0$ and $K_2 > 0$ on their compact supports. Further suppose that $m = O(n^\alpha)$ with $\alpha > 0$, b_n and b_m in (4.74) and (4.75), respectively, satisfy $b_n = O(n^{-w})$ and $b_m = O(m^{-u})$ with $1/4 < w < 1/2$, $u < 1/4$, $\alpha u > 1/7$, $\alpha(1-2u) > 1/2$, $\alpha(1+u) > 1$, $3\alpha u - w > 0$ and $\alpha(1-3u) - w > 0$. Then*

$$n^{1/2}(\theta_n - \theta) \xrightarrow{\mathcal{L}} N(0, I_\theta^{-1}),$$

where $I_\theta = \int (\eta^{(1)}(x))^2 \eta^{-1}(x) dx$.

Remark 4.12. If we take $u = 1/5$, then $\alpha > 5/6$. This shows that m could converge to infinity at a lower rate compared to n. This means that one can use a comparatively smaller sample of Y_i's to estimate the nonparametric component η.

To prove Theorem 4.9, we need following two lemmas.

Lemma 4.2. *Suppose that $\eta > 0$ on its compact support W_η, K_1 in (4.74) is differentiable and symmetric about origin and has compact support W_{K_1} on*

which $K_1 > 0$, *and* b_n *in (4.74) satisfies* $b_n = O(n^{-w})$ *with* $w > 0$. *Then,*

$$\int \big(f_n^{1/2}(x) - f_{\theta,\eta}^{1/2}(x)\big)^2 dx \;=\; O_P\big(n^{-(1-w)} + n^{-4w}\big),$$

$$n^{1/2} \int \frac{|\dot{s}_\theta(x)|}{s_\theta^3(x)} \big(f_n(x) - f_{\theta,\eta}(x)\big)^2 dx \;=\; O_P\big(n^{-(1/2-w)} + n^{-(4w-1/2)}\big),$$

$$n^{1/2}\Big(\int \frac{\dot{s}_\theta(x)}{s_\theta(x)} f_n(x) dx - \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{s}_\theta(X_i)}{s_\theta(X_i)}\Big) \;=\; O_P\big(n^{-(2w-1/2)}\big).$$

Proof. Note that, with $\delta = \min_{x \in W_\eta} \eta(x)$ and $\delta' = \min_{x \in W_{K_1}} K_1(x)$,

$$\begin{aligned}
& \int \big(f_n^{1/2}(x) - f_{\theta,\eta}^{1/2}(x)\big)^2 dx \\
\leq\; & \int \frac{\big(f_n(x) - f_{\theta,\eta}(x)\big)^2}{f_n(x) + f_{\theta,\eta}(x)} dx \\
\leq\; & \int \frac{\big(f_n(x) - f_{\theta,\eta}(x)\big)^2}{f_n(x)} dx + \int \frac{\big(f_n(x) - f_{\theta,\eta}(x)\big)^2}{f_{\theta,\eta}(x)} dx \\
\leq\; & \frac{b_n}{\delta'} \int \big(f_n(x) - f_{\theta,\eta}(x)\big)^2 dx + \frac{1}{\delta} \int \big(f_n(x) - f_{\theta,\eta}(x)\big)^2 dx \\
=\; & O\Big(\int \big(f_n(x) - f_{\theta,\eta}(x)\big)^2 dx\Big).
\end{aligned}$$

For kernel density estimator f_n, it is known that $f_n(x) - f_{\theta,\eta}(x) = O_P\big((nb_n)^{-1/2} + b_n^2\big)$, and as a result

$$\int \big(f_n^{1/2}(x) - f_{\theta,\eta}^{1/2}(x)\big)^2 dx \;=\; O_P\big((nb_n)^{-1} + b_n^4\big) \;=\; O_P\big(n^{-(1-w)} + n^{-4w}\big),$$

$$\begin{aligned}
& n^{1/2} \int \frac{|\dot{s}_\theta(x)|}{s_\theta^3(x)} \big(f_n(x) - f_{\theta,\eta}(x)\big)^2 dx \\
=\; & n^{1/2} \int \frac{|\eta^{(1)}(x-\theta)|}{2 f_{\theta,\eta}^2(x)} \big(f_n(x) - f_{\theta,\eta}(x)\big)^2 dx \\
=\; & O\big(n^{1/2} \int \big(f_n(x) - f_{\theta,\eta}(x)\big)^2 dx\big) \\
=\; & O_P\big(n^{-1/2} b_n^{-1} + n^{1/2} b_n^4\big), \\
=\; & O_P\big(n^{-(1/2-w)} + n^{-(4w-1/2)}\big),
\end{aligned}$$

and

$$
\begin{aligned}
& n^{1/2}\Big[\int \frac{\dot{s}_\theta(x)}{s_\theta(x)} f_n(x)dx - \frac{1}{n}\sum_{i=1}^{n}\frac{\dot{s}_\theta(X_i)}{s_\theta(X_i)}\Big] \\
= \;& (2^{-1}n^{1/2})\cdot\frac{1}{n}\sum_{i=1}^{n}\Big[\int \frac{1}{b_n}\frac{\eta^{(1)}(x-\theta)}{\eta(x-\theta)}K_1\big(\frac{x-X_i}{b_n}\big)dx - \frac{\eta^{(1)}(X_i-\theta)}{\eta(X_i-\theta)}\Big] \\
= \;& (2^{-1}n^{1/2})\cdot\frac{1}{n}\sum_{i=1}^{n}\Big[\int K_1(t)\frac{\eta^{(1)}(X_i-\theta+b_nt)}{\eta(X_i-\theta+b_nt)}dt - \int K_1(t)\frac{\eta^{(1)}(X_i-\theta)}{\eta(X_i-\theta)}dt\Big] \\
= \;& (2^{-1}n^{1/2})\cdot\frac{1}{n}\sum_{i=1}^{n}\int K_1(t)\big[(\log\eta)^{(2)}(X_i-\theta)b_nt \\
& \qquad\qquad +(\log\eta)^{(3)}(X_i-\theta+\xi_i)b_n^2t^2/2\big]dt \\
= \;& O_P\big(n^{1/2}b_n^2\big) \\
= \;& O_P\big(n^{-(2w-1/2)}\big).
\end{aligned}
$$

□

Lemma 4.3. *Suppose that $\eta > 0$ on its compact support W_η, K_2 in (4.75) is differentiable and symmetric about origin and has compact support W_{K_2} on which $K_2 > 0$, and b_m in (4.75) satisfies $b_m = O(m^{-u})$ with $u > 0$. Then*

$$\sup_{t\in\Theta}\int \big(\widehat{s}_t(x) - s_t(x)\big)^2dx = O_P(m^{-(1-u)} + m^{-4u}),$$

$$\int \big(\dot{\widehat{s}}_t(x) - \dot{s}_t(x)\big)^2dx = o_P(m^{-(u+2v)})$$

and

$$\int \dot{\widehat{s}}_t(x)s_t(x)dx = 0$$

for any $t \in \Theta$ and v such that $0 < v < \min\{u, \frac{1}{2} - 2u\}$.

Proof. Using a similar proof as of Lemma 4.2, we have

$$
\begin{aligned}
\sup_{t\in\Theta}\int (\widehat{s}_t(x) - s_t(x))^2dx &= \int (\eta_m^{1/2}(x) - \eta^{1/2}(x))^2dx \\
&= O\big(\int(\eta_m(x) - \eta(x))^2dx\big) \\
&= O_P\big((mb_m)^{-1} + b_m^4\big) \\
&= O_P\big(m^{-(1-u)} + m^{-4u}\big).
\end{aligned}
$$

Since $\sup_x m^v|\eta_m^{(1)}(x) - \eta^{(1)}(x)| \xrightarrow{a.e.} 0$ as $m \to \infty$ and for any $0 < v < \min\{u, \frac{1}{2} - 2u\}$ (Schuster, 1969),

$$
\begin{aligned}
& \int (\dot{\widehat{s}}_t(x) - \dot{s}_t(x))^2 dx \\
= \; & \frac{1}{4}\int (\eta_m^{-1/2}(x)\eta_m^{(1)}(x) - \eta^{-1/2}(x)\eta^{(1)}(x))^2 dx \\
= \; & \frac{1}{4}\int [\eta_m^{-1/2}(x)(\eta_m^{(1)}(x) - \eta^{(1)}(x)) + (\eta_m^{-1/2}(x) - \eta^{-1/2}(x))\eta^{(1)}(x)]^2 dx \\
\leq \; & \frac{1}{2}\int \eta_m^{-1}(x)(\eta_m^{(1)}(x) - \eta^{(1)}(x))^2 dx + \frac{1}{2}\int (\eta_m^{-1/2}(x) - \eta^{-1/2}(x))^2(\eta^{(1)}(x))^2 dx \\
= \; & o_P\big(m^{-(u+2v)}\big) + O\big(m^{-u}\int (\eta_m^{1/2}(x) - \eta^{1/2}(x))^2 dx\big) \\
= \; & o_P\big(m^{-(u+2v)}\big) + O_P\big(m^{-1} + m^{-5u}\big) \\
= \; & o_P\big(m^{-(u+2v)}\big).
\end{aligned}
$$

Since η_m and η are symmetric about origin, one has

$$
\int \dot{\widehat{s}}_t(x) s_t(x) dx = \frac{1}{2}\int \eta_m^{-1/2}(x)\eta_m^{(1)}(x)\eta^{1/2}(x) dx = 0.
$$

□

Proof of Theorem 4.9. Since $1/4 < w < 1/2$, Lemma 4.2 yields that (4.7), (4.8) and (4.9) hold for some $r \leq 1 - w$. By assumptions in the theorem, $\alpha(1-u) > \alpha(1-2u) > 1/2$, $4\alpha u > 4/7 > 1/2$. Also, $\alpha(1-u)+\alpha(u+2u) = \alpha(1+2u) > 1$, $\alpha(1-u) + \alpha[u + (1-4u)] = \alpha(2-4u) > 1$, $4\alpha u + \alpha(u+2u) = 7\alpha u > 1$ and $4\alpha u+\alpha[u+(1-4u)] = \alpha(1+u) > 1$, and thus there exists some $r' > 1/2$ such that (4.10), (4.11) and (4.12) hold. Furthermore, $1-w+\alpha(u+2u) = 3\alpha u-w+1 > 1$ and $1-w+\alpha[u+(1-4u)] = \alpha(1-3u)-w+1 > 1$, and thus there exists some common $r > 1/2$ such that (iii) and (iv) of Theorem 4.2 hold. Now the result follows from Theorem 4.2. □

4.8 Concluding Remarks

The Hellinger distance approach has been applied to variety of parametric models in statistical inference. This approach yields statistics that have good efficiency and robustness properties. In this chapter, we have shown that the Hellinger distance approach can be extended successfully to semiparametric models of general form as well. As in the parametric case, the resulting MHD estimators are robust and have good asymptotic efficiency properties - in many cases our estimators are fully efficient in the semiparametric sense. We have supported our theoretical findings with extensive finite sample simulation studies. We have also introduced a new distance measure; namely, profile Hellinger distance, and have constructed the corresponding optimal estimator. The preceding approach is in some sense analogous to the profile likelihood approach.

The success of Hellinger distance and profile Hellinger distance approaches in semiparametric models considered in this chapter should encourage its application to other models and problems as well. We consider following problems, among others, to be worthy candidates for application: hypothesis testing, regression models and perhaps to quantal assay models. To best of our knowledge, minimum distance procedures have not been studied in general semiparametric models (4.1) in the literature.

A few words comparing the MHD estimator $T_n(f_n)$ and the MPHD estimator $T_1(f_n)$ defined by (4.4) and (4.36), respectively, would be appropriate here. In practice, the exact determination of $T_1(f_n)$ may not be easily possible due to computational difficulties in calculating Hellinger profiles, and one may only be able to come up with some numerical approximations. This is the rationale behind the establishment of Theorem 4.2, which to some degree eases off some computational difficulties. In the definition (4.4), a single η_n value is used to replace $\eta(t, f_n)$ defined in (4.34) for all $t \in \Theta$. The above discussion thus appears to suggest that the estimator at (4.4) may have a smaller asymptotic variance than that of the estimator defined by (4.36). Indeed, from Theorems 4.2 and 4.5 and Remark 4.10, it follows that $T_n(f_n)$ is efficient in the parametric sense with asymptotic variance I_θ^{-1}, while $T_1(f_n)$ is efficient in the semiparametric sense with a generally larger asymptotic variance I_*^{-1}. However, this does not imply that $T_n(f_n)$ is a better estimator than $T_1(f_n)$, since the theorems are proved under different conditions. From a practical point of view, $T_n(f_n)$ may be preferred over $T_1(f_n)$ when a good estimator of η is available, while $T_1(f_n)$ may be preferred over $T_n(f_n)$ when one can easily calculate the profiles.

An heuristic argument of describing robustness of $T_n(f_n) = T(\{f_{t,\eta_n}\}_{t\in\Theta}, f_n)$ defined by (4.4) is as follows. From Theorem 4.2 it follows that the estimator $T_n(f_n)$ is a Hellinger continuous functional of f_n and η_n. Thus, small Hellinger distance perturbations in the underlying density will only result in small changes in the MHD estimator $T_n(f_n)$. In fact, the MHD functional is optimally insensitive (in a certain sense) to small changes in the density (Beran, 1977). Theorems 4.7 and 4.8 have confirmed above arguments theoretically. Furthermore, the numerical results presented in Section 4.6 again displayed the behavior suggested in our theoretical findings.

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. Begun, J., Hall, W., Huang, W. and Wellner, J. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432-452.

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

5. Beran, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6**, 292-313.

6. Bickel, P.J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-671.

7. Bickel, P.J., Klaassen, C. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

8. Bickel, P.J. and Kwon, J. (2001). Inference for semiparametric models: some questions and an answer. *Statistica Sinica* **11**, 863-960.

9. Bickel, P.J. and Ritov, Y. (1987). Efficient estimation in the errors in variables models. *Ann. Statist.* **15**, 513-540.

10. Breslow, N. and Day, N.E. (1980). *Statistical Methods in Cancer Research, Vol. 1: The Analysis of Case-Control Studies.* Lyon: IARC.

11. Brown, L.D. and Hwang, J.T.G. (1993). How to approximate a histogram by a normal density. *Amer. Statist.* 47,251-255

12. Chen, J. (1995). Optimal rate of convergence in finite mixture models. *Ann. Statist.* **23**, 221-233.

13. Chen, J. (1998). Penalized likelihood-ratio test for finite mixture models with multinomial observations. *Canad. J. Statist.* **26**, 583-599.

14. Communications in Statistics (1976). Special Issue on Remote Sensing. **Vol. A5**, No. 12.

15. Cordero-Braña, O.I. (1994). *Minimum Hellinger distance estimation for finite mixture models.* Unpublished Ph.D. dissertation, Utah State University.

16. Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Royal Statist. Society Series B* **34**, 187-202.

17. Cutler, A. and Cordero-Braña, O.I. (1996). Minimum Hellinger distance estimation for finite mixture models. *J. Amer. Statist. Assoc.* **91**, 1716-1723.

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

19. Devroye, L.P. and Wagner, T.J. (1979). The L_1 convergence of kernel density estimates. *Ann. Statist.* **7**, 1136-1139.

20. Donoho, D.L. and Liu, R.C. (1988). The "automatic" robustness of minimum distance functionals. *Ann. Statist.* **16**, 552-586.

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

22. Fabian, V. and Hannan, J. (1982). On estimation and adaptive estimation for locally asymptotically normal families. *Z. Wahrsch. verw. Gebiete* **59**, 459-487.

23. Fernholz, L. (1983). Von Mises calculus for statistical functionals. *Lecture Notes in Statistics* Vol. **19**, Springer Verlag, New York.

24. Forrester, J., Hooper, W., Peng, H. and Schick, A. (2003). On the construction of efficient estimators in semiparametric models. *Statist. Decisions* **21**, 109-137.

25. Gilbert, P.B., Self, S.G. and Ashby, M.A. (1998). Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics* **54**, 799-814.

26. Gill, R.D., Vardi, Y. and Wellner, J.A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069-1112.

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

28. Hall, P. (1983). Orthogonal series distribution function estimation, with applications. *J. Roy. Statist. Soc. B* **45**, 81-88.

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

30. Hampel, F.R. (1968). *Contributions to the Theory of Robust Estimation.* Ph.D. dissertation, University of California, Berkeley.

31. Hampel, F. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42**, 1887-1896.

32. Hosmer, D.W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three types of samples. *Biometrics* **29**, 761-770.

33. Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression.* New York: Wiley.

34. Huang, W.M. (1982). *Parameter Estimation When There Are Nuisance Functions.* Ph.D. dissertation, University of Rochester.

35. Huber, P.J. (1980). *Robust Statistics.* Wiley & Sons, New York.

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

37. Karlis, D. and Xekalaki, E. (2001). Robust inference for finite Poisson mixtures. *J. Statist. Plan. Inf.* **93**, 93–115.

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

39. Karlis, D. and Xekalaki, E. (2001). Robust inference for finite poisson mixtures. *J. Statist. Planning Inference* **93**, 93-115.

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

41. Lindsay, B.G. (1995). *Mixture models: theory, geometry and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, IMS, Hayward, California, USA.

42. 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.

43. Maronna, R.A., Martin, D.R. and Yohai, V. (2007). *Robust Statistics: Theory and Methods.* John Wiley, New York.

44. Mclachlan, G.J. and Peel, D. (2000). *Finite mixture models.* Wiley, New York.

45. Murphy, S.A. and Van der Vaart, A.W. (1996). Likelihood inference in the errors-in-variables model. *J. Multi. Anal.* **59**, 81-108.

46. Murphy, S.A. and Van der Vaart, A.W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95**, 449-485.

47. Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation.* Academic Press, Orlando.

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

49. Qin, J. (1993). Empirical likelihood in biased sample problems. *Ann. Statist.*, **21**, 1182-1196.

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

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

52. Rousseeuw, P.J. and Croux, C. (1993). Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.* **88**, 1273-1283.

53. Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14**, 1139-1151.

54. Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Planning Inference* **16**, 89-105. Correction, **22**, 269-270, 1989.

55. Schuster, E.F. (1969). Estimation of a probability density function and its derivatives. *Ann. Math. Statist.* **40**, 1187-1195.

56. Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* John Wiley, New York.

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

58. Silverman, B.W. (1986). *Density Estimation.* Chapman and Hall, New York.

59. Simpson, D.G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.* **82**, 802-807.

60. Skorokhod, A.V. (1956). Limit theorems for stochastic processes. *Theor. Probability Appl.* **1**, 261-290.

61. Sriram, T.N. and Vidyashankar, A.N. (2000). Minimum Hellinger distance estimation for supercritical Galton-Watson processes. *Statist. Probab. Lett.* **50**, 331-342.

62. Stather, C.R. (1981). *Robust Statistical Inference Using Hellinger Distance Methods.* Ph.D. dissertation, LaTrobe University, Australia.

63. Tamura, R.N. and Boos, D.D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer. Statist. Assoc.* **81**, 223-229.

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

65. Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* Wiley, New York.

66. Van der Vaart, A.W. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.* **24**, 862-878.

67. Van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press.

68. Vardi, Y. (1982). Nonparametric estimation in presence of length bias. *Ann. Statist.* **10**, 616-620.

69. Vardi, Y. (1985). Empirical distribution in selection bias models. *Ann. Statist.* **13**, 178-203.

70. Woo, Mi-Ja and Sriram, T.N. (2006). Robust estimation of mixture complexity. *J. Amer. Statist. Assoc.* **101**, 1475-1486.

71. Woo, Mi-Ja and Sriram, T.N. (2007). Robust estimation of mixture complexity for count data. *Comput. Statis. Data Anal.* **51**, 4379-4392.

72. Woodward, W.A., Whitney, P. and Eslinger, P.W. (1995). Minimum Hellinger distance estimation of mixture proportions. *J. Statist. Plan. Inf.* **48**, 303-319.

73. Yang, S. (1991). Minimum Hellinger distance estimation of parameter in the random censorship model. *Ann. Statist.* **19**, 579-602.

74. Ying, Z. (1992). Minimum Hellinger-type distance estimation for censored data. *Ann. Statist.* **20**, 1361-1390.

75. Zhang, B. (2000). Quantile estimation under a two-sample semi-parametric model. *Bernoulli* **6**, 491-511.

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