Journal of Data Science 465-484 , DOI: 10.6339/JDS.201807_16(3).0002

MINIMUM PROFILE HELLINGER DISTANCE ESTIMATION FOR A TWO-SAMPLE LOCATION-SHIFTED MODEL

Haocheng Li[*1], Jingjing Wu[*1], Jian Yang[1]

[1]*Department of Mathematics and Statistics, University of Calgary*

[*] *These two authors contribute equally to the work*

Abstract: Minimum Hellinger distance estimation (MHDE) for parametric model is obtained by minimizing the Hellinger distance between an assumed parametric model and a nonparametric estimation of the model. MHDE receives increasing attention for its efficiency and robustness. Recently, it has been extended from parametric models to semiparametric models. This manuscript considers a two-sample semiparametric location-shifted model where two independent samples are generated from two identical symmetric distributions with different location parameters. We propose to use profiling technique in order to utilize the information from both samples to estimate unknown symmetric function. With the profiled estimation of the function, we propose a minimum profile Hellinger distance estimation (MPHDE) for the two unknown location parameters. This MPHDE is similar to but dif- ferent from the one introduced in Wu and Karunamuni (2015), and thus the results presented in this work is not a trivial application of their method. The difference is due to the two-sample nature of the model and thus we use different approaches to study its asymptotic properties such as consistency and asymptotic normality. The efficiency and robustness properties of the proposed MPHDE are evaluated empirically though simulation studies. A real data from a breast cancer study is analyzed to illustrate the use of the proposed method.

Key words: Minimum Hellinger distance estimation, profiling, two-sample location-shifted model, efficiency, robustness.

1 Introduction

Minimum distance estimation of unknown parameters in a parametric model is obtained by minimizing the distance between a nonparametric distribution esti- mation (such as empirical, kernel, etc) and an assumed parametric model. Some well-known examples of minimum distance estimation include least-squares esti- mation and minimum Chi-square estimation. Among different minimum distance estimations, minimum Hellinger distance estimation (MHDE) receives increasing attention for its superior properties in efficiency and robustness. The idea of the estimation using Hellinger distance was firstly introduced by Beran (1977) for parametric models. Simpson (1987) examined the MHDE for discrete data. Yang (1991) and Ying (1992) studied censored data in survival analysis by using the MHDE. Woo and Sriram (2006) and Woo and Sriram (2007) employed the MHDE method to investigate mixture complexity in finite mixture models. The MHDEs for mixture models were also studied by many literatures such as Lu et al. (2003) and Xiang et al. (2008). Other applications of the MHDE method can be referred to Takada (2009), N'drin and Hili (2013) and Prause et al. (2016).

Recently, the MHDE method was extended to semiparametric model of gen- eral form. Wu and Karunamuni (2009) and Wu and Karunamuni (2012) proved that MHDE retains good efficiency and robustness properties for semiparametric model of general form under certain conditions. However, the MHDE usually re- quires an estimate of infinite-dimensional nuisance parameter in semiparametric models, which leads to computational difficulty. To solve this problem, Wu and Karunamuni (2015) proposed a minimum profile Hellinger distance estimation (MPHDE). The MPHDE method profiles out the infinite-dimensional nuisance parameter and thus circumvents the computational obstacle. Wu and Karuna- muni (2015) derived the MPHDE for one-sample symmetric location model, while in real applications two-sample location-shifted symmetric model is often encoun- tered. For example, as we will show in data application section, the comparison of biomarkers across different patient groups requires two-sample models. There- fore, in this manuscript we extend the MPHDE approach to two-sample location- shifted symmetric model.

The idea of using profiling approach is quite intuitive but its theoretical frame- work is often complicated to study. For one-sample case, Wu and Karunamuni (2015) used the Hellinger distance between the location model and its nonpara- metric estimation. However, this method does not work for our two-sample case because it cannot utilize the information for the nuisance parameter contained in both samples. To handle the two-sample estimation, we propose a new Hellinger distance between the location-shifted model and its estimation that involves the nuisance density estimation and the location parameters of our interest. Conse- quently, a novel approach is proposed to explore the asymptotic properties of the resulted estimator obtained from minimizing the new Hellinger distance.

The remainder of this paper is organized as follows. In Section 2 we propose a MPHDE for the two-sample semiparametric location-shifted model. Section 3 presents the asymptotic properties of the proposed MPHDE with all proofs deferred to appendix. In Section 4, we evaluate the performance of the MPHDE by simulation studies compared with commonly used least-squares and maximum

likelihood estimation methods. A data from a real breast cancer study is analyzed in Section 5 to demonstrate the use of the proposed method. Concluding remarks are given in Section 6.

2 MPHDE for Two-Sample Location Model

Suppose we have two samples with n_0 and n_1 independent and identically dis- tributed (i.i.d.) random variables (r.v.s), respectively. Denote the two samples as $X_i, i = 1,\cdots,n_0$, and $Y_j, j = 1,\cdots,n_1$. We assume that the two samples are independent and they follow

$$\begin{aligned} X_1,\cdots,X_{n0} &\overset{i.i.d.}{\sim} f_0(\cdot) = f(\cdot - \theta_0) \\ Y_1,\cdots,Y_{n1} &\overset{i.i.d.}{\sim} f_1(\cdot) = f(\cdot - \theta_1) \end{aligned} \tag{1}$$

where $\theta = (\theta_0, \theta_1)^\top$ is the location parameter vector of our interest and the unknown $f \in H$ is treated as the nuisance parameter. Here H is the collection of all continuous even densities. We focus on model (1) in this paper and work on the inference for the location parameter θ. Model (1) could be represented in a regression form. Let total sample size be $n = n_0 + n_1$. The r.v.s in model (1) could be written as $\{(Z_i, Di): i = 1,\cdots,n\}$, where $(Z_1, \ldots, Z_n)^\top = \left(X_1,\cdots,X_{n_0}, Y_1,\cdots,\cdot Y_{n_1}\right)^\top$ and Di is an indicator function taking $D_i = 1$ if Z_i is from f_1 and 0 otherwise.Model (1) can be equivalently represented as

$$Z_i = \theta_0 + (\theta_1 - \theta_0)D_i + \epsilon_i$$

where the i.i.d. error terms ϵ_i's are from f.

For any given θ, since $X_1 - \theta 0, \ldots, X_{n0} - \theta_0, Y_1 - \theta_1, \ldots, Y_{n1} - \theta 1$ are i.i.d. r.v.s from f, we can estimate the unknown f using the following kernel density estimator based on the pooled sample:

$$\begin{aligned} \hat{f}(\theta ; x) &= \frac{1}{nb_n}\left\{\sum_{i=1}^{n0} K\left[\frac{x-(X_i-\theta_0)}{b_n}\right] + \sum_{i=1}^{n_1} K\left[\frac{x-(Y_i-\theta_1)}{b_n}\right]\right\} \\ &= \frac{n0}{n}\left\{\frac{1}{n_0 b_n}\sum_{i=1}^{n0} K\left[\frac{x-(X_i-\theta_0)}{b_n}\right] + \frac{n_1}{n}\sum_{i=1}^{n_1} K\left[\frac{x-(Y_i-\theta_1)}{b_n}\right]\right\} \\ &= \rho_0\hat{f}_0(x+\theta_0) + \rho_1\hat{f}_1(x+\theta_1) \end{aligned}$$

where $\rho_0 = n0/n, \rho_1 = 1 - \rho_0 = n1/n$, kernel function K is a symmetric density function, the bandwidth bn is a sequence of positive constants such that $b_n \to 0$ as $n \to \infty$, and $\hat{f}_0$ and $\hat{f}_1$ are kernel density estimators of f_0 and f_1, respectively. To be specific, f_0 and f_1 have

$$\hat{f}_0(x) = \frac{1}{n_0 b_n}\sum_{i=1}^{n_0} K\left(\frac{x-X_i}{b_n}\right) \tag{3}$$

and

$$\hat{f}_1(x) = \frac{1}{n_1 b_n}\sum_{i=1}^{n_1} K\left(\frac{x-Y_i}{b_n}\right) \tag{4}$$

Even though ρ_0 and ρ_1 depend on n, we depress their dependence for notation simplicity. We generally require that $n_i/n \to \rho_i$ as $n \to \infty$ with $\rho_i \in (0, 1)$, $i = 0, 1$. Based on (2), f_0 and f_1 can also be estimated respectively by

$$\tilde{f}_0(x) = \hat{f}(\theta ; x-\theta_0) = \rho_0\hat{f}_0(x) + \rho_1\hat{f}_1(x-\theta_0+\theta_1)$$

and

$$\widetilde{f_1}(x) = \hat{f}(\theta\ ; x - \theta_1) = \rho_0 \hat{f}_0(x + \theta_0 - \theta_1) + \rho_1 \hat{f}_1(x)$$

To obtain the MPHDE of θ, we firstly profile the unknown nuisance parameter f out by minimizing the sum of the squared Hellinger distance for the two samples, i.e.

$$\min_{f\in H}\left\{\left\|\widetilde{f_0}^{\frac{1}{2}}(x) - f_0^{\frac{1}{2}}(x)\right\|^2 + \left\|\widetilde{f_1}^{\frac{1}{2}}(x) - f_1^{\frac{1}{2}}(x)\right\|^2\right\}$$

$$= \min_{f\in H}\left\{\left\|[\rho_0\hat{f}_0(x) + \rho_1\hat{f}_1(x - \theta_0 + \theta_1)]^{\frac{1}{2}} - f^{\frac{1}{2}}(x - \theta_0)\right\|^2\right.$$

$$\left. + \left\|[\rho_0\hat{f}_0(x + \theta_0 - \theta_1) + \rho_1\hat{f}_1(x)]^{\frac{1}{2}} - f^{\frac{1}{2}}(x - \theta_1)\right\|^2\right\}$$

$$= \min_{f\in H}\left\{2\left\|[\rho_0\hat{f}_0(x + \theta_0) + \rho_1\hat{f}_1(x + \theta_1)]^{\frac{1}{2}} - f^{\frac{1}{2}}(x)\right\|^2\right\}$$

$$= 4\left\{1 - \max_{f\in H}\int [\rho_0\hat{f}_0(x + \theta_0) + \rho_1\hat{f}_1(x + \theta_1)]^{\frac{1}{2}} f^{\frac{1}{2}}(x)\, dx\right\}$$

$$= 4\left\{1 - \max_{f\in H}\int \hat{f}^{\frac{1}{2}}(\theta\ ; x)\, f^{\frac{1}{2}}(x)\, dx\right\}$$

Note that $\hat{f}^{\frac{1}{2}}(\theta;\cdot)$ can be represented as

$$\hat{f}^{\frac{1}{2}}(\theta; x) = \frac{1}{2}\left[\hat{f}^{\frac{1}{2}}(\theta; x) + \hat{f}^{\frac{1}{2}}(\theta; -x)\right] + \frac{1}{2}\left[\hat{f}^{\frac{1}{2}}(\theta; x) - \hat{f}^{\frac{1}{2}}(\theta; -x)\right] = \widehat{\eta_1}(\theta; x) + \widehat{\eta_2}(\theta; x),$$

say,

where $\widehat{\eta_1}(\theta;\ \cdot)$ is an even function while $\widehat{\eta_2}(\theta;\ \cdot)$ is an odd function. As a result,

$$\int \hat{f}^{\frac{1}{2}}(\theta\ ; x)\, f^{\frac{1}{2}}(x)\, dx = \int [\widehat{\eta_1}(\theta; x) + \widehat{\eta_2}(\theta; x)] f^{\frac{1}{2}}(x)\, dx = \int \widehat{\eta_1}(\theta; x) f^{\frac{1}{2}}(x)\, dx$$

By the Cauchy-Schwarz inequality, $\int \widehat{\eta_1}(\theta; x) f^{\frac{1}{2}}(x)\, dx$ is uniquely maximized by

$f^{\frac{1}{2}}(x) = \widehat{\eta_1}(\theta; x)/\|\widehat{\eta_1}(\theta; x)\|$ and thus $f^{\frac{1}{2}}(x) = \widehat{\eta_1}^2(\theta; x)/\|\widehat{\eta_1}(\theta; x)\|^2$ is the profiled f.

Therefore, after replacing $f^{\frac{1}{2}}(x)$ with $\widehat{\eta_1}(\theta; x)/\|\widehat{\eta_1}(\theta; x)\|$, the MPHDE $\hat{\theta}$ of θ is given by

$$\hat{\theta} = \underset{t=(t_0,t_1)^\top\in\mathbb{R}^2}{arg\,max} \int \widehat{\eta_1}(t; x)\frac{\widehat{\eta_1}(t;x)}{\|\widehat{\eta_1}(t;x)\|}\, dx$$

$$= \underset{t\in\mathbb{R}^2}{arg\,max}\, \|\widehat{\eta_1}(t; x)\|$$

$$= \underset{t\in\mathbb{R}^2}{arg\,max}\left\|\hat{f}^{\frac{1}{2}}(t; x) + \hat{f}^{\frac{1}{2}}(t; -x)\right\| \tag{5}$$

$$= \underset{t\in\mathbb{R}^2}{arg\,max}\int [\rho_0\hat{f}_0(x + t_0) + \rho_1\hat{f}_1(x + t_1)]^{\frac{1}{2}}[\rho_0\hat{f}_0(-x + t_0) + \rho_1\hat{f}_1(-x + t_1)]^{\frac{1}{2}}\, dx$$

$$= \underset{t\in\mathbb{R}^2}{arg\,min}\left\|\hat{f}^{\frac{1}{2}}(t; x) - \hat{f}^{\frac{1}{2}}(t; -x)\right\|$$

$$=: T(\hat{f}_0, \hat{f}_1)$$

where in the last equality we represent $\hat{\theta}$ as a functional T which only depends on $\hat{f}_0$ and $\hat{f}_1$. As there is no explicit expression of the solution to the above optimization in (5), $\hat{\theta}$ has to be calculated numerically. In this manuscript, the computation was implemented by the R function "nlm" with the medians of X_i and Y_i to be the initial values of θ_0 and θ_1, respectively. The numerical optimization leads to satisfactory results in our simulation and data application studies. All of them successfully achieve convergence.

Remark 1. Even though θ can take any value in $\mathbb{R}^2$, we can use a large enough compact subset of $\mathbb{R}^2$, say $\Theta = [-A, A]^2$ with A to be a large positive number, so that θ is an interior point of Θ, i.e. $\theta \in int(\Theta)$. Thus in what follows we will optimize θ over Θ instead of $\mathbb{R}^2$ simply for technical convenience.

Remark 2. The proposed MPHDE involves the mixture model $\rho_0 f(\cdot - \theta_0) + \rho_1 f(\cdot - \theta_1)$ which has been studied by many literatures such as ecently Xi- ang et al. (2014), Erisoglu and Erisoglu (2013), and Ngatchou-Wandji and Bulla(2013). For the identifiability of this model, we can assume $\theta_0 < \theta_1$ without any loss of generality. By Theorem 2 of Hunter et al. (2007), this mixture model is identifiable if $\rho_0 \in (0, 0.5) \cup (0.5, 1)$. If f is unimodal, then this mixture model is identifiable even when $\rho_0 = 0.5$. Therefore the identifiability is not a problem for the MPHDE and we will assume from now on that the mixture model is identifiable for the sake of simplicity.

Remark 3. For one-sample location model $f(\cdot - \theta)$, the Hellinger distance is between the location model, involving both f and θ together, and its nonparametric estimation. For this two-sample model, in order to use the information about the nuisance parameter f contained in both the first and second samples, the Hellinger distance is between f and its estimation that involves the nuisance density estimation and the location parameters of our interest.

3 Asymptotic Properties

In this section, we discuss the asymptotic distribution of the MPHDE $\hat{\theta}$ given in (5) for the two-sample semiparametric location-shifted model (1). Note that $\hat{\theta}$ given in (5) is a bit different than the MPHDE defined in Wu and Karunamuni (2015) for general semiparametric models in the sense that the former incorpo- rates the model assumption in the nonparametric estimation of f while the later uses a completely nonparametric estimation of f not depending on the model at all. In this sense, we can not apply the asymptotics obtained in Wu and Karuna- muni (2015) to our model (1). Instead we will directly derive below the existence,consistency and asymptotic normality of $\hat{\theta}$. Let F be the set of all densities with respect to (w.r.t.) Lebesgue measure on the real line. We first give in the next theorem the existence and uniqueness of the MPHDE $\hat{\theta}$.

Theorem 1. Suppose that T is defined by (5). Then,

(i) For any $g_0, g_1 \in F$, there exists $T(g_0, g_1)$ satisfying (5).

(ii) If $T(g_0, g_1)$ is unique, then T is continuous at $(g_0, g_1)^\top$ in the Hellinger metric. In another word, $T(g_{n0}, g_{n1}) \rightarrow T(g_0, g_1)$ for any sequences g_{n0} and g_{n1} such that $\|g_{ni}{}^{1/2} - g_i{}^{1/2}\| \rightarrow 0$ as $n \rightarrow \infty$, i = 0, 1.

(iii) The MPHDE is Fisher-consistent, i.e. $T(f_0, f_1) = \theta$ uniquely for every $(f_0, f_1)^\top$ satisfying model (1).

The following theorem is a consequence of Theorem 1 which gives the consis- tency of the MPHDE $\hat{\theta}$ defined in (5).

Theorem 2. Suppose that the kernel K in (3) and (4) are absolutely continuous, has compact support and bounded first derivative, and the bandwidth b_n satisfies $b_n \rightarrow 0$ and $n^{1/2} b_n \rightarrow \infty$ as $n \rightarrow \infty$. If f in model (1) is uniformly continuous on its support, then $\left\|\hat{f}_i{}^{1/2} - f_i{}^{1/2}\right\| \xrightarrow{p} 0$, i = 0, 1, and furthermore the MPHDE $\hat{\theta} \xrightarrow{p} \theta$.

The next Theorem 3 gives the expression of the different $\hat{\theta} - \theta$ which will be used to establish the asymptotic normality of θˆ in Theorem 4.

Theorem 3. Assume that the conditions in Theorem 2 are satisfied. Further suppose f has uniformly continuous first derivative. Then

$$\Sigma_1(\hat{\theta} - \theta) = -\binom{\rho_0}{\rho_1}\left[\rho_0 \int \frac{f'}{f}(x)\, \hat{f}_0(x+\theta_0)\, dx + \rho_1 \int \frac{f'}{f}(x)\, \hat{f}_1(x+\theta_1)\, dx\right] \cdot \left(1 + o_p(1)\right), \quad (6)$$

where

$$\sum\nolimits_1 = \begin{pmatrix} \rho_0{}^2 & \rho_0\rho_1 \\ \rho_0\rho_1 & \rho_1{}^2 \end{pmatrix} \int \frac{(f')^2}{f}(x)\, dx - \begin{pmatrix} \rho_0 & 0 \\ 0 & \rho_0 \end{pmatrix} \int f''(x)\, dx$$

With (6) and some regularity condition we can immediately derive the asymptotic distribution of $\hat{\theta} - \theta$ given in the next theorem.

Theorem 4. Assume that the conditions in Theorem 3 are satisfied and in addition f has continuous third derivative, $\int f''(x)\, dx \neq \int \frac{(f')^2}{f}(x)\, dx$ and the bandwidth satisfies $n b^6{}_n \rightarrow 0$ as $n \rightarrow \infty$. Then the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is $N(0, \Sigma)$ with covariance matrix Σ defined by

$$\sum = \begin{cases} \int \frac{(f')^2}{f} dx\, \Sigma_1^{-1} \begin{pmatrix} \rho_0{}^2 & \rho_0\rho_1 \\ \rho_0\rho_1 & \rho_1{}^2 \end{pmatrix} \Sigma_1^{-1}, & if \int f''(x)\, dx \neq 0 \\ [\int \frac{(f')^2}{f} dx]^{-1} \begin{pmatrix} 1/\rho_0 & 0 \\ 0 & 1/\rho_0 \end{pmatrix}, & if \int f''(x)\, dx = 0 \end{cases}$$

Remark 3. Distributions satisfying $\int f''(x)\, dx = 0$ include those with support on the whole real line, such as normal and t distributions. The distributions satisfying $\int f''(x)\, dx \neq 0$ include those with finite support and its first derivative evaluated at boundary of support is non-zero, such as $f(x) = \frac{3}{4}(1 - x^2) for |x| \leq 1$.

Remark 4. If the two samples in (1) are actually a single sample from the mixture $\rho_0 f(\cdot - \theta_0) + \rho_1 f(\cdot - \theta_1)$ with known classification for each data point, then by comparing the lower bound of asymptotic variance described in Wu and Karunamuni (2015) with the results in our Theorem 4, we can conclude that the proposed MPHDE $\hat{\theta}$ defined in (5) is efficient, in the semiparametric sense, for any f. In addition, if $\int f''(x)\,dx = 0$, then this semiparametric model is an adaptive model and the proposed MPHDE $\hat{\theta}$ is an adaptive estimator.

4 Simulation Studies

We assess the empirical performance of the proposed MPHDE in Section 2 for the two-sample location-shifted model. Five hundred simulations are run for each parameter configuration. We consider a parameter setting of $(\theta_0, \theta_1)^\top = (0,1)^\top$ and simulate four different distributions for $f(x)$: normal, Student's t, triangular and Laplace. We set the standard deviation to be 1 for normal distribution, the degrees of freedom to be 4 for t distribution. The triangular distribution has density function

$$f(x) = \frac{c - |x|}{c^2}, |x| \leq c,$$

and we set $c = 1$. The Laplace distribution has density function

$$f(x) = \frac{1}{2b} exp\left(-\frac{|x|}{b}\right),$$

and we set $b = 1$. The bandwidth bn is chosen to be $b_n = n^{-1/5}$according to the bandwidth requirement in Theorem 4. The biweight kernel $K(t) = \frac{15}{16}(1 - t^2)^2$for $|t| \leq 1$ is employed in the simulation studies. We consider both smaller sample sizes $n_0 = n_1 = 20$ and larger sample sizes $n_0 = n_1 = 50$.

As a comparison, we also give both least-squares estimation (LSE) and max- imum likelihood estimation (MLE). For the two-sample location-shifted model (1) under our consideration, simple calculation shows that the LSEs of θ_0 and θ_1 are essentially the sample means $\bar{X}$ and $\bar{Y}$ respectively. With f assumed known, straight calculation says that the MLEs of θ_0 and θ_1 are sample means for normal case and sample medians for Laplace case, while there is no explicit expression of the MLEs for Student's t and Triangular populations. Tables 1 and 2 display the simulation results of MPHDE, LSE and MLE methods for sample sizes $n_0 = n_1 = 20$ and $n_0 = n_1 = 50$, respectively. In the tables, the term Bias represents the average of biases over the 500 repetitions; the terms RMSE and SE are the average of root mean squared errors and empirical standard errors, respectively; and the term CR represents the empirical coverage rate for 95% confidence intervals. From Tables 1 and 2 we can see that all the three estimation approaches have fairly small bias. In terms of standard errors, the MPHDE has worse performance than the LSE and the MLE regardless of sample size.

To investigate the robustness properties of the proposed MPHDE and make comparison, we examine the performance of the three methods under data con- tamination. In this simulation, the data from model (1) is intentionally contami- nated by a single outlying observation. This is implemented, say for $n_0 = n_1 = 20$,by replacing the last observation X_{20} with an integer number z varying from −20 and 20. To quantify the robustness, the α-influence function (α-IF) discussed by Lu et al. (2003) is used. The α-IF for parameter θi, i = 0, 1, is defined as

$$IF(z) = n_i(\hat{\theta}_i^z - \hat{\theta}_i),$$

where $\hat{\theta}_i^z$ represents the estimate based on the contaminated data with outlying observation $X_{20} = z$ and $\hat{\theta}_i$ denotes the estimate based on the uncontaminated

Table 1: Simulation results for $n_0 = n_1 = 20$.

MPHDE	Bias(θ_0)	RMSE(θ_0)	SE(θ_0)	CR(θ_0)	Bias(θ_1)	RSME(θ_1)	SE(θ_1)	CR(θ_1)
Normal	-0.025	0.307	0.331	0.924	-0.012	0.325	0.341	0.948
Student's t	-0.008	0.350	0.380	0.936	-0.025	0.340	0.389	0.946
Triangular	0.002	0.115	0.115	0.918	-0.005	0.114	0.115	0.918
Laplace	0.006	0.310	0.343	0.952	-0.024	0.297	0.344	0.964
LSE	Bias(θ_0)	RMSE(θ_0)	SE(θ_0)	CR(θ_0)	Bias(θ_1)	RSME(θ_1)	SE(θ_1)	CR(θ_1)
Normal	-0.001	0.234	0.213	0.944	-0.005	0.229	0.214	0.926
Student's t	-0.014	0.312	0.298	0.944	-0.030	0.319	0.299	0.930
Triangular	-0.0001	0.095	0.088	0.914	-0.002	0.095	0.088	0.910
Laplace	-0.004	0.332	0.298	0.916	-0.002	0.327	0.298	0.916
MLE	Bias(θ_0)	RMSE(θ_0)	SE(θ_0)	CR(θ_0)	Bias(θ_1)	RSME(θ_1)	SE(θ_1)	CR(θ_1)
Normal	-0.001	0.234	0.213	0.944	-0.005	0.229	0.214	0.926
Student's t	-0.011	0.266	0.292	0.950	-0.030	0.277	0.293	0.958
Triangular	-0.001	0.107	0.106	0.928	-0.008	0.102	0.105	0.916
Laplace	-0.0009	0.274	0.287	0.932	-0.023	0.257	0.287	0.950

Table 2: Simulation results for $n_0 = n_1 = 50$.

MPHDE	Bias(θ0)	RMSE(θ0)	SE(θ0)	CR(θ0)	Bias(θ1)	RSME(θ1)	SE(θ1)	CR(θ1)
Normal	-0.009	0.200	0.220	0.942	0.007	0.205	0.218	0.946
Student's t	0.008	0.221	0.244	0.962	0.012	0.228	0.243	0.942
Triangular	-0.004	0.070	0.074	0.950	-0.004	0.070	0.074	0.950
Laplace	-0.010	0.187	0.207	0.968	-0.007	0.187	0.205	0.956
LSE	Bias(θ0)	RMSE(θ0)	SE(θ0)	CR(θ0)	Bias(θ1)	RSME(θ1)	SE(θ1)	CR(θ1)
Normal	-0.007	0.139	0.139	0.958	0.003	0.143	0.138	0.944
Student's t	-0.003	0.197	0.193	0.954	-0.001	0.194	0.192	0.958
Triangular	-0.003	0.057	0.057	0.940	-0.002	0.056	0.057	0.942
Laplace	-0.006	0.199	0.194	0.944	-0.006	0.189	0.194	0.952
MLE	Bias(θ0)	RMSE(θ0)	SE(θ0)	CR(θ0)	Bias(θ1)	RSME(θ1)	SE(θ1)	CR(θ1)
Normal	-0.007	0.139	0.139	0.958	0.003	0.143	0.138	0.944
Student's t	-0.001	0.169	0.174	0.954	0.003	0.173	0.173	0.936
Triangular	-0.003	0.060	0.066	0.946	-0.003	0.062	0.066	0.952
Laplace	-0.012	0.154	0.167	0.940	-0.011	0.157	0.164	0.954

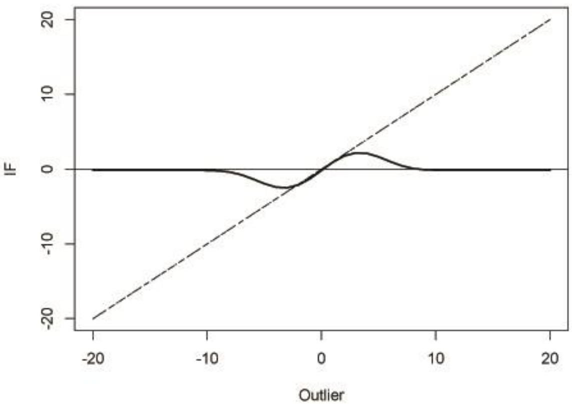
data. The α-IF is calculated by using the change in the estimate before and after contamination divided by the contamination rate, i.e. $1/n_i$. We can similarly calculate the α-IF when outlying observations contaminate the second sample. The simulation results in Figure 1 are for θ_0, $n_0 = n_1 = 20$ and the case that the first sample is contaminated. The results for θ_1, $n_0 = n_1 = 50$ or the case that the second sample is contaminated are very similar to those in Figure 1 and thus omitted to save space.

Figure 1 presents the average α-IFs over 500 simulation runs for the MPHDE, MLE and LSE of θ_0 under normal, t, triangular and Laplace distributions. Regardless of the population distribution, the α-IF of the MPHDE are bounded and converge to the same small constant when the value of the outlying observation gets larger and larger on either side, while the α-IFs of the MLE and LSE are unbounded in general. Therefore, compared to the MLE and LSE methods, the MPHDE has a little lower efficiency but this limitation is compensated by its excellent robustness. In summary, the MPHDE method always results in reasonable estimates no matter data is contaminated or not, whereas the MLE and LSE methods under contaminated data lead to significantly biased estimates.

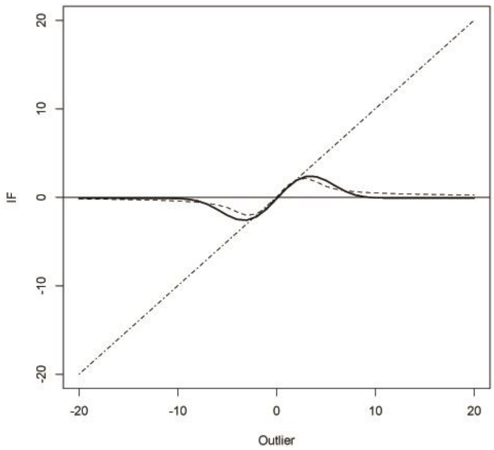
5 Data Applications

In this section, we demonstrate the use of the proposed MPHDE method through analyzing a breast cancer data collected in Calgary, Canada (Feng et al., 2016). Breast cancer is regarded as the most common cancer and the second leading cause of cancer death for females in North America. Existing studies suggest that it would be more informative to use some protein expression levels as indicators of biological behavior (Feng et al., 2015). These biomarkers could reflect genetic properties in cancer formation and cancer aggressiveness. Our dataset has 316 patients diagnosed with breast cancer between years 1985 and 2000. Two interested biomarkers measured on these patients are Ataxia telangiectasia mutated (ATM) and Ki67. ATM is a protein to support maintaining genomic stability. Comparing with normal breast tissue, ATM could be significantly reduced in the tissue with breast cancer. Ki67 is a protein expressed exclusively in proliferating cells. It is often used as a prognostic marker in breast cancer.

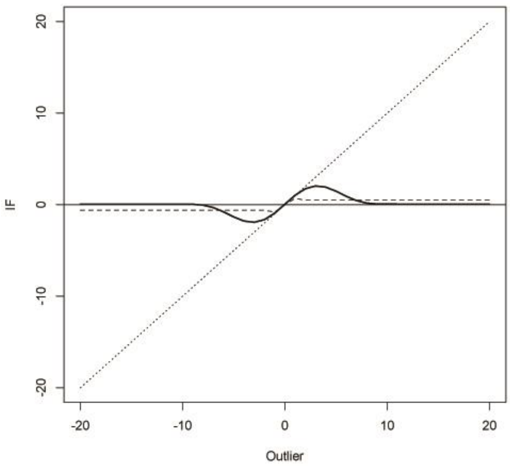
Let $\theta^{(1)}$and $\theta^{(2)}$ denote the location parameters in the distributions of the protein expression level of ATM and Ki67 biomarkers, respectively. Our research focuses on the comparison of the protein expression levels across both cancer stages (Stage) and lymph node (LN). As for cancer stage, ${\theta_0}^{(k)}$ and ${\theta_1}^{(k)}$ $(k = 1, 2)$ denote the location parameters in the distributions of protein expression level for Stage I and Stage II/III patients, respectively. Regarding LN status, ${\theta_0}^{(k)}$ and ${\theta_1}^{(k)}$ $(k = 1, 2)$ denote the location parameters in the distributions of protein expression level for negative LN (LN-) and positive LN (LN+) patients, respectively. Figure 2 displays the boxplots for ATM and Ki67 expression levels across both cancer stages and LN statuses, respectively. From this figure we do see the difference in location of both ATM and Ki67 variables across both cancer stages and LN statuses, especially for Ki67 considering the smaller variation in expression level.



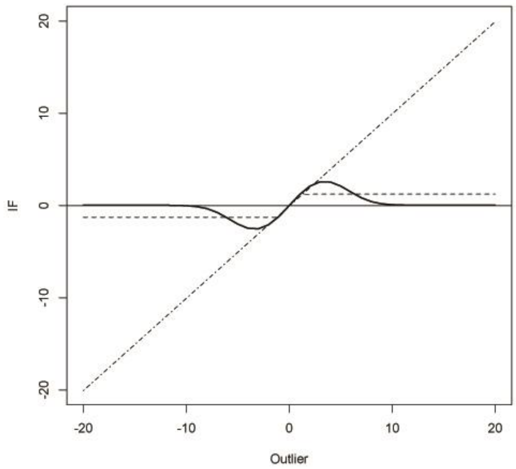
(a) Normal



(b) Student’s t



(c) Triangular



(d) Laplace

Figure 1: The average α-IFs under (a) normal distribution, (b) Student’s t dis- tribution, (c) triangular distribution and (d) Laplace distribution. Thin-solid line represents the zero horizontal baseline, and the thick-solid, dot-dashed and dashed lines represent respectively the MPHDE, LSE and MLE approaches.

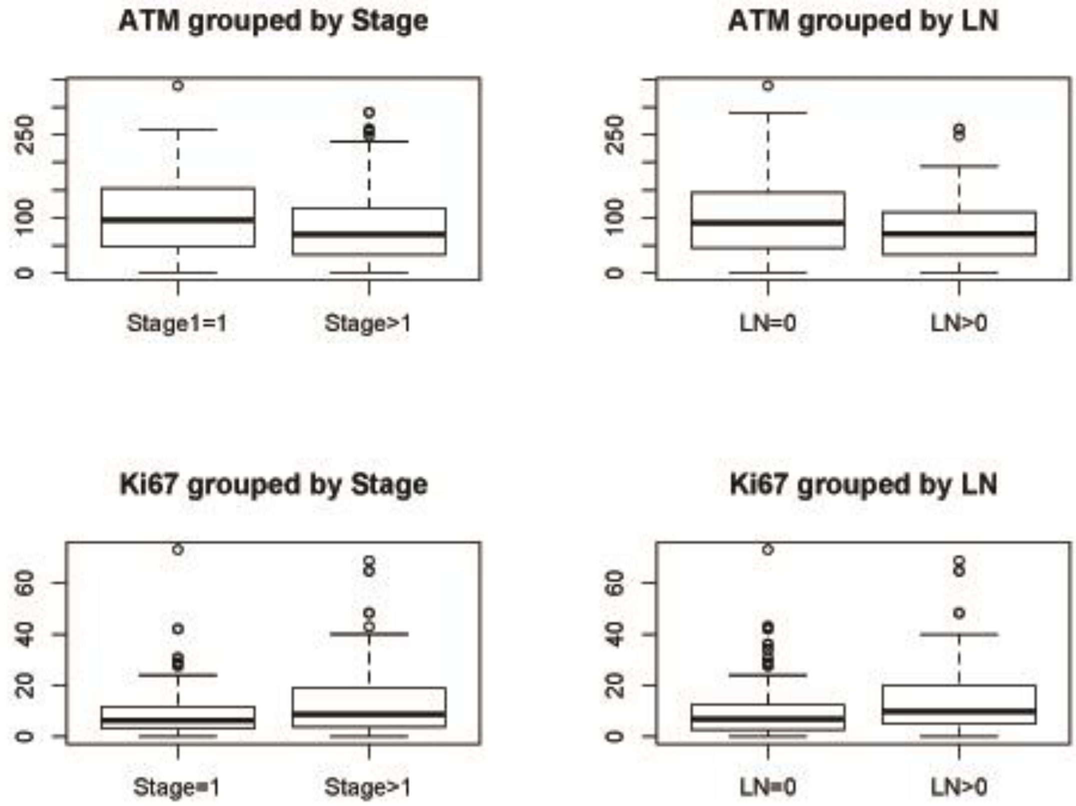


Figure 2: Boxplots for ATM and Ki67 expression levels across cancer stages and LN statuses (0: negative; > 0: positive).

To compare the two biomarkers ATM and Ki67, we calculate the MPHDEs $\theta_0{}^{(k)}$ and $\theta_1{}^{(k)}$ for both k = 1 and k = 2. The parameter estimates (Est.), estimated standard errors (SE), 95% confidence intervals (CI) and p-values are reported in Table 3. Based on the results in this table, both of the two biomarkers have significant difference across cancer stages and LN statuses. For cancer stage, ATM has higher expression level in stage I group than in stage II/III group ($p = 0.046$). On the other hand, Ki67 has lower expression level in stage I group than in stage II/III group ($p < 0.001$). For LN status, ATM has higher expression

Table 3: Breast cancer data analysis results based on MPHDE.

Cancer Stage					
Biomarker	Group	Est.	SE	95%CI	p-value
ATM	I	$\hat{\theta}_0^{(1)}$	90.35	(73.33,107.37)	0.046
ATM	II/III	$\hat{\theta}_1^{(1)}$	68.11	(54.39,81.84)	
Ki67	I	$\hat{\theta}_0^{(2)}$	6.29	(5.05,7.53)	<0.001
Ki67	II/III	$\hat{\theta}_1^{(2)}$	8.63	(5.55,11.70)	
Lymph node(LN)Status					
Biomarker	Group	Est.	SE	95%CI	p-value
ATM	LN−	$\hat{\theta}_0^{(1)}$	95.89	(76.92, 114.87)	0.019
ATM	LN+	$\hat{\theta}_1^{(1)}$	70.17	(60.10,80.25)	
Ki67	LN−	$\hat{\theta}_0^{(2)}$	6.82	(6.13,7.49)	<0.001
Ki67	LN+	$\hat{\theta}_1^{(2)}$	10.02	(5.20,14.81)	

level in negative LN group than in positive LN group ($p = 0.019$), while Ki67 has lower expression level in negative LN group than in positive LN group ($p < 0.001$).

6 Concluding Remarks

In this paper, we propose to use MPHDE for the inferences of the two-sample semiparametric location-shifted model. Compared with commonly used least- squares and maximum likelihood approaches, the proposed method leads to ro- bust inferences. Simulation results demonstrate satisfactory performance and the analysis for the breast cancer data exemplifies its utility in real practice.

Acknowledgments

The authors thank Dr. Xiaolan Feng for providing the breast cancer data. This research was supported by the Natural Sciences and Engineering Research Coun- cil of Canada (Haocheng Li, RGPIN-2015-04409; Jingjing Wu, RGPIN-355970- 2013).

Appendix

The proofs of Theorems 1, 2, 3 and 4 are presented in this section. The techniques used in the proofs are similar to those in Karunamuni and Wu (2009).

A1. Proof of Theorem 1. (i) With $t = (t_0, t_1)^\top$, define $\mathrm{d}(\mathrm{t}) = \| [\rho_0 g_0(x+t_0) + \rho_1 g_1(x+t_1)]^{1/2} + [\rho_0 g_0(-x+t_0) + \rho_1 g_1(-x+t_1)]^{1/2} \|$. For any sequence $t_n = (t_{n0}, t_{n1})^\top$ such that $t_n \to t$ *as* $n \to \infty$,

$$|d(t_n) - d(t)|$$

$$\leq \left\| [\rho_0 g_0(x+t_{n0}) + \rho_1 g_1(x+t_{n1})]^{\frac{1}{2}} + [\rho_0 g_0(-x+t_{n0}) + \rho_1 g_1(-x+t_{n1})]^{\frac{1}{2}} - [\rho_0 g_0(x+t_0) + \rho_1 g_1(x+t_1)]^{\frac{1}{2}} - [\rho_0 g_0(-x+t_0) + \rho_1 g_1(-x+t_1)]^{\frac{1}{2}} \right\|$$

$$\leq \left\| [\rho_0 g_0(x+t_{n0}) + \rho_1 g_1(x+t_{n1})]^{\frac{1}{2}} - [\rho_0 g_0(x+t_0) + \rho_1 g_1(x+t_1)]^{\frac{1}{2}} \right\| + \left\| [\rho_0 g_0(-x+t_{n0}) + \rho_1 g_1(-x+t_{n1})]^{\frac{1}{2}} - [\rho_0 g_0(-x+t_0) + \rho_1 g_1(-x+t_1)]^{\frac{1}{2}} \right\|$$

$$\leq 2\left\{\rho_0 \int |g_0(x+t_{n0}) - g_0(x+t_0)|\, dx + \rho_1 \int |g_1(x+t_{n1}) - g_1(x+t_1)|\, dx\right\}^{\frac{1}{2}}$$

$$2\left\{\rho_0 \int |g_0(x+t_{n0}) - g_0(x+t_{n0}-t_0)|\, dx + \rho_1 \int |g_1(x+t_{n1}) - g_1(x+t_{n1}-t_1)|\, dx\right\}^{\frac{1}{2}}.$$

Since $\int g_0(x)\,dx = \int g_0(x+t_{n0}-t_0)\,dx = 1$, $\int [g_0(x) - g_0(x+t_{n0}-t_0)]^+\,dx = \int [g_0(x) - g_0(x+t_{n0}-t_0)]^-\,dx$. Thus $\int |g_0(x) - g_0(x+t_{n0}-t_0)|\,dx = 2\int [g_0(x) - g_0(x+t_{n0}-t_0)]^+\,dx$. We also have $|g_0(x) - g_0(x+t_{n0}-t_0)|^+ \leq g_0(x)$, thus by the Dominated Convergence Theorem $\int |g_0(x) - g_0(x+t_{n0}-t_0)|\,dx \to 0$ as $n \to \infty$. Similarly $\int |g_1(x+t_{n1}) - g_1(x+t_{n1}-t_1)|\,dx \to 0$. Therefore $d(t_n) \to d(t)$ *as* $n \to \infty$, i.e. *d(t)* is continuous in *t* and then the maximum can be achieved over Θ.

(ii) Define $d_n(t) = \| [\rho_0 g_{n0}(x+t_0) + \rho_1 g_{n1}(x+t_1)]^{1/2} + [\rho_0 g_{n0}(-x+t_0) + \rho_1 g_{n1}(-x+t_1)]^{1/2} \|$ and write $\theta_n = T(g_{n0}, g_1)$ and $\theta = T(g_{n0}, g_1)$. Then by similar argument as in (i),

$$|d_n(t) - d(t)| \leq 2\left\{\rho_0 \int |g_{n0}(x) - g_0(x)|\,dx + \rho_1 \int |g_{n1}(x) - g_1(x)|\,dx\right\}^{\frac{1}{2}}.$$

By Hölder's inequality, $\int |g_{n0}(x) - g_0(x)|\,dx \leq \|{g_{n0}}^{1/2} + {g_0}^{1/2}\| \cdot \|{g_{n0}}^{1/2} - {g_0}^{1/2}\| \leq 4\|{g_{n0}}^{1/2} - {g_0}^{1/2}\|$. Similarly $\int |g_{n1}(x) - g_1(x)|\,dx \to 0$, and thus $\sup_t |d_n(t) - d(t)| \to 0$. This implies $d_n(\theta) - d(\theta) \to 0$ and $d_n(\theta_n) - d(\theta_n) \to 0$. If $\theta_n \neq \theta$, then the compactness of Θ ensures that there exists a subsequence $\{\theta_m\} \subseteq \{\theta_n\}$ such that $\theta_m \to \theta' \neq \theta$, implying $\theta' \in \Theta$ and $d(\theta_m) \to d(\theta')$ by continuity of d from (i). From above results, we have $d_m(\theta_m) - d_m(\theta) \to d(\theta') - d(\theta)$. By the definition of θ_m, $d_m(\theta_m) - d_m(\theta) \geq 0$. Hence, $d(\theta') - d(\theta) \geq 0$. But by the definition of θ and its uniqueness, $d(\theta) > d(\theta')$. This is a contradiction. Therefore, $\theta_n \to \theta$.

(iii) For $(f_0, f_1)^\top$ satisfying model (1),

$$T(f_0, f_1) = \underset{t\in\Theta}{arg\ min}\, ||[\rho_0 f(x-\theta_0+t_0)+\rho_1 f(x-\theta_1+t_1)]^{\frac{1}{2}}$$

$$-[\rho_0 f(-x-\theta_0+t_0)+\rho_1 f(-x-\theta_1+t_1)]^{\frac{1}{2}}||.$$

By the symmetry of f and the identifiability of $\rho_0 f(\cdot - \theta_0) + \rho_1 f(\cdot - \theta_1)$, the minimum is achieved when $\theta_i - t_i = -\theta_i + t_i$, i.e. $t_i = \theta_i$ for i = 0, 1. Thus $T(f_0, f_1) = \theta$.

A2. Proof of Theorem 2. If we can prove that $\left\|\hat{f}_i^{\frac{1}{2}} - f_i^{\frac{1}{2}}\right\| \xrightarrow{p} 0\ as\ n \to \infty$, $i = 0, 1$, then by (ii) and (iii) of Theorem 1, $T(\hat{f}_0, \hat{f}_1) \xrightarrow{p} T(f_0, f_1)$, i.e.$\hat{\theta} \xrightarrow{p} \theta$. It is easy to show that $sup_x|\hat{f}_0(x) - f_0(x)| \xrightarrow{p} 0$. Note that $\left\|\hat{f}_0^{\frac{1}{2}} - f_0^{\frac{1}{2}}\right\|^2 \leq \int |f_0(x) - \hat{f}_0(x)| dx$ by the same technique used in the proof of Theorem 1 (i) and Dominated Convergence Theorem we have $\int |f_0(x) - \hat{f}_0(x)| dx \xrightarrow{p} 0$ and thus $\left\|\hat{f}_0^{\frac{1}{2}} - f_0^{\frac{1}{2}}\right\|^2 \xrightarrow{p} 0$. Similarly $\left\|\hat{f}_1^{\frac{1}{2}} - f_1^{\frac{1}{2}}\right\|^2 \xrightarrow{p} 0$.

A3. Proof of Theorem 3. By Theorem 2, $\hat{\theta} \to \theta\ as\ n \to \infty$. Thus for large n, $\hat{\theta} \in int(\Theta)$ since $\theta \in int(\Theta)$. Denote $h_t(x) = \rho_o f_0(x + t_0) + \rho_1 f_1(x + t_1)$, $\hat{h}_t(x) = \rho_o \hat{f}_0(x + t_0) + \rho_1 \hat{f}_1(x + t_1)$, $s_t = {h_t}^{1/2}$ and $\hat{s}_t = {\hat{h}_t}^{1/2}$. Note that $h_\theta = f$. We claim that for any $t \in int(\Theta)$, any 2×1 real vector e of unit euclidean length and any scalar ϵ close to zero,

$$s_{t+\epsilon e}(x) = s_t(x) + \epsilon e^\top \dot{s}_t(x) + \epsilon e^\top u_\epsilon(x) \tag{A.1}$$

and

$$\dot{s}_{t+\epsilon e}(x) = \dot{s}_t(x) + \epsilon \ddot{s}_t(x) e + \epsilon v_\epsilon(x) e \tag{A.2}$$

hold for both s_t and $\hat{s}_t$, where $\dot{s}_t = \frac{\partial s_t}{\partial t}$ and $\ddot{s}_t = \frac{\partial^2 s_t}{\partial t^2}$ with components in $L2$, and the components of u_ϵ and v_ϵ individually tend to zero in $L2$ as $\epsilon \to 0$. The proof of this statement is shown at the end of this proof. (A.1) yields that

$$\lim_{\epsilon\to 0} \epsilon^{-1} \int \left[\hat{h}_{t+\epsilon e}^{\frac{1}{2}}(x) \hat{h}_{t+\epsilon e}^{\frac{1}{2}}(-x) - \hat{h}_t^{\frac{1}{2}}(x) \hat{h}_t^{\frac{1}{2}}(-x) \right] dx$$

$$= \int \left[e^\top \dot{\hat{s}}_t(x) \hat{s}_t(-x) + e^\top \dot{\hat{s}}_t(-x) \hat{s}_t(x) \right] dx$$

$$= 2e^\top \int \dot{\hat{s}}_t(x) \hat{s}_t(-x)\, dx$$

Since $\hat{\theta}$ is the optimizer defined in (5), we have $\int \dot{\hat{s}}_{\hat{\theta}}(x) \hat{s}_{\hat{\theta}}(-x)\, dx = 0$. Similarly $\int \dot{s}_\theta(x) s_\theta(-x)\, dx = 0$, $\int \frac{\partial \hat{h}_t}{\partial t}(x)\, dx = \frac{\partial}{\partial t} \int \hat{h}_t(x)\, dx = 0$ and $\int \frac{\partial h_t}{\partial t}(x)\, dx = \frac{\partial}{\partial t} \int h_t(x)\, dx = 0$ for any $t \in int(\Theta)$. Thus (A.1) and (A.2) give

$$0 = 2 \int \dot{\hat{s}}_{\hat{\theta}}(x) \hat{s}_{\hat{\theta}}(-x)\, dx$$

$$= 2\int \left[\dot{\hat{s}}_\theta(x)\hat{s}_\theta(-x) + \left(\hat{\theta}-\theta\right)^\top \dot{\hat{s}}_\theta(-x)\hat{s}_\theta(x) + \ddot{\hat{s}}_\theta(x)\left(\hat{\theta}-\theta\right)\hat{s}_\theta(-x)\right] dx + o_p\left(\hat{\theta}-\theta\right)$$

$$= 2\int \dot{\hat{s}}_\theta(x)\hat{s}_\theta(-x)\, dx + 2\int [\dot{\hat{s}}_\theta(-x)\dot{\hat{s}}_\theta{}^\top(x) + \hat{s}_\theta(-x)\ddot{\hat{s}}_\theta(x)]\, dx \left(\hat{\theta}-\theta\right) + o_p\left(\hat{\theta}-\theta\right) \tag{A.3}$$

Since $\hat{f}_0 \to f_0$ and$\hat{f}_1 \to f_1$ uniformly by the proof of Theorem 2,

$$-2\int \left[\dot{\hat{s}}_\theta(-x)\dot{\hat{s}}_\theta{}^\top(x) + \hat{s}_\theta(-x)\ddot{\hat{s}}_\theta(x)\right] dx$$

$$= -2\int \left[\dot{s}_\theta(-x)\dot{s}_\theta{}^\top(x) + s_\theta(-x)\ddot{s}_\theta(x)\right] dx + o_p(1)$$

$$= -2\int \left\{\frac{1}{4f(x)}\begin{pmatrix}\rho_0 f'(x)\\ \rho_1 f'(x)\end{pmatrix}\begin{pmatrix}\rho_0 f'(-x)\\ \rho_1 f'(-x)\end{pmatrix}^\top + \frac{1}{2}\begin{pmatrix}\rho_0 f''(x) & 0\\ 0 & \rho_1 f''(x)\end{pmatrix}\right.$$
$$\left. - \frac{1}{4f(x)}\begin{pmatrix}\rho_0 f'(x)\\ \rho_1 f'(x)\end{pmatrix}\begin{pmatrix}\rho_0 f'(-x)\\ \rho_1 f'(-x)\end{pmatrix}^\top\right\} + o_p(1)$$

$$= \begin{pmatrix}\rho_0{}^2 & \rho_0\rho_1\\ \rho_0\rho_1 & \rho_1{}^2\end{pmatrix}\int \frac{(f')^2}{f}(x)\, dx - \begin{pmatrix}\rho_0 & 0\\ 0 & \rho_0\end{pmatrix}\int f''(x)\, dx + o_p(1)$$

$$= \Sigma_1 + o_p(1)$$

Direct calculation gives

$$\int 2\dot{\hat{s}}_\theta(x)\hat{s}_\theta(-x)\, dx = \int 2\dot{\hat{s}}_\theta(x)[\hat{s}_\theta(-x) - \hat{s}_\theta(x)]\, dx$$

$$= \int \hat{h}_\theta^{-\frac{1}{2}}(x)\frac{\partial \hat{h}_\theta}{\partial\theta}(x)\left[\hat{h}_\theta^{\frac{1}{2}}(-x) - \hat{h}_\theta^{\frac{1}{2}}(x)\right] dx$$

$$= \int \hat{h}_\theta^{-\frac{1}{2}}(x)\frac{\partial \hat{h}_\theta}{\partial\theta}(x)\left[\hat{h}_\theta^{\frac{1}{2}}(-x) - \hat{h}_\theta^{\frac{1}{2}}(x)\right]^{-1}\left[\hat{h}_\theta(-x) - \hat{h}_\theta(x)\right] dx$$

$$= \int U_n(x)\{\rho_0[\hat{f}_0(-x+\theta_0) - f(-x)] - \rho_0[\hat{f}_0(x+\theta_0) - f(x)] + \rho_1[\hat{f}_0(-x+\theta_1) - f(-x)]$$
$$- \rho_1[\hat{f}_1(x+\theta_1) - f(x)]\}\, dx$$

where $U_n(\mathrm{x}) = \hat{h}_\theta^{-\frac{1}{2}}(x)\frac{\partial \hat{h}_\theta}{\partial\theta}(x)\left[\hat{h}_\theta^{\frac{1}{2}}(-x) - \hat{h}_\theta^{\frac{1}{2}}(x)\right]^{-1}$. With $U(x) = h_\theta^{-\frac{1}{2}}(x)\frac{\partial h_\theta}{\partial\theta}(x)\left[h_\theta^{\frac{1}{2}}(-x) - h_\theta^{\frac{1}{2}}(x)\right]^{-1} = \frac{1}{2}\begin{pmatrix}\rho_0\\ \rho_1\end{pmatrix}\frac{f'}{f}(x)$,

we have

$$\int U_n(x)[\hat{f}_0(-x+\theta_0) - f(-x)]\, dx$$

$$= \int U_n(-x)[\hat{f}_0(x+\theta_0) - f(x)]\, dx$$

$$= \int U(-x)[\hat{f}_0(x+\theta_0) - f(x)]\, dx + \int [U_n(-x) - U(-x)][\hat{f}_0(x+\theta_0) - f(x)]\, dx$$

$$= -\frac{1}{2}\begin{pmatrix}\rho_0\\ \rho_1\end{pmatrix}\int \frac{f'}{f}(x)\,\hat{f}_0(x+\theta_0)\, dx \cdot \left(1 + o_p(1)\right).$$

Similarly,

$$\int U_n(x)\big[\hat{f}_0(x+\theta_0)-f(x)\big]\,dx=\int U(x)\big[\hat{f}_0(x+\theta_0)-f(x)\big]\,dx\cdot\big(1+o_p(1)\big)$$

$$=\frac{1}{2}\binom{\rho_0}{\rho_1}\int\frac{f'}{f}(x)\,\hat{f}_0(x+\theta_0)\,dx\cdot\big(1+o_p(1)\big).$$

$$\int U_n(x)\big[\hat{f}_1(-x+\theta_1)-f(-x)\big]\,dx=\int U(-x)\big[\hat{f}_1(x+\theta_1)-f(x)\big]\,dx$$

$$=-\frac{1}{2}\binom{\rho_0}{\rho_1}\int\frac{f'}{f}(x)\,\hat{f}_1(x+\theta_1)\,dx\cdot\big(1+o_p(1)\big).$$

$$\int U_n(x)\big[\hat{f}_1(x+\theta_1)-f(x)\big]\,dx=\int U(x)\big[\hat{f}_1(x+\theta_1)-f(x)\big]\,dx$$

$$=\frac{1}{2}\binom{\rho_0}{\rho_1}\int\frac{f'}{f}(x)\,\hat{f}_1(x+\theta_1)\,dx\cdot\big(1+o_p(1)\big).$$

Thus

$$\int 2\dot{\hat{s}}_\theta(x)\hat{s}_\theta(-x)\,dx$$

$$=-\binom{\rho_0}{\rho_1}[\rho_0\int\frac{f'}{f}(x)\,\hat{f}_0(x+\theta_0)\,dx+\rho_1\int\frac{f'}{f}(x)\,\hat{f}_1(x+\theta_1)\,dx]\cdot\big(1+o_p(1)\big)$$

and (A.3) is reduced to (6).

A4. Proof of Theorem 4.

If $f''(x)dx\neq 0$, then $det(\Sigma_1)\neq 0$ and thus Σ_1^{-1} exists. The expression of $\hat{\theta}-\theta$ from (6) indicates that $\hat{\theta}_0-\theta_0$ and $\hat{\theta}_1-\theta_1$ have asymptotic correlation 1. Since

$$\int\frac{f'}{f}(x)\,\hat{f}_0(x+\theta_0)\,dx=\frac{1}{n_0}\sum_{i=1}^{n_0}\int\frac{f'}{f}(x)\frac{1}{b_n}K\Big(\frac{x-(X_i-\theta_0)}{b_n}\Big)\,dx,$$

by CLT and $nb^6{}_n\to 0,\sqrt{n_0}\int\frac{f'}{f}(x)\,\hat{f}_0(x+\theta_0)\,dx$ has asymptotic normal distribution with mean 0 and variance $\int\frac{(f')^2}{f}(x)\,dx$. Similarly, $\sqrt{n_1}\int\frac{f'}{f}(x)\,\hat{f}_1(x+\theta_1)\,dx$ has asymptotic normal distribution with mean 0 and variance $\int\frac{(f')^2}{f}(x)\,dx$. By the independence of the two samples X_i's and Y_j's, hence the result.

If $\int f''(x)dx=0$, then $\Sigma_1=(\rho_0,\rho_1)^\top(\rho_0,\rho_1)\int\frac{(f')^2}{f}dx$ is not invertible. Expression (6) says

$$\int\frac{(f')^2}{f}dx\binom{\rho_0}{\rho_1}(\rho_0,\rho_1)\big(\hat{\theta}-\theta\big)=-\binom{\rho_0}{\rho_1}(\rho_0,\rho_1)\begin{pmatrix}\int\frac{f'}{f}(x)\,\hat{f}_0(x+\theta_0)\,dx\\ \int\frac{f'}{f}(x)\,\hat{f}_1(x+\theta_1)\,dx\end{pmatrix}\big(1+o_p(1)\big)$$

holds for any $\rho_0,\rho_1\in(0,0.5)\cup(0.5,1)$, thus

$$\hat{\theta} - \theta = -[\int \frac{(f')^2}{f} dx]^{-1} \begin{pmatrix} \int \frac{f'}{f}(x)\, \hat{f}_0(x+\theta_0)\, dx \\ \int \frac{f'}{f}(x)\, \hat{f}_1(x+\theta_1)\, dx \end{pmatrix} \left(1 + o_p(1)\right)$$

and thus the result.

References

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

[2] Erisoglu, U. and Erisoglu, M. (2013). L-moments estimations for mixture of weibull distributions. Journal of Data Science, 12(1):87–106.

[3] Feng, X., Li, H., Dean, M., Wilson, H. E., Kornaga, E., Enwere, E. K., Tang,P., Paterson, A., Lees-Miller, S. P., Magliocco, A. M., et al. (2015). Low atm protein expression in malignant tumor as well as cancer-associated stroma are independent prognostic factors in a retrospective study of early-stage hormone- negative breast cancer. Breast Cancer Research, 17(1):65.

[4] Feng, X., Li, H., Kornaga, E. N., Dean, M., Lees-Miller, S. P., Riabowol, K.,

[5] Magliocco, A. M., Morris, D., Watson, P. H., Enwere, E. K., et al. (2016). Low ki67/high atm protein expression in malignant tumors predicts favorable prognosis in a retrospective study of early stage hormone receptor positive breast cancer. Oncotarget, 7(52):85798–85812.

[6] Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007). Inference for mix- tures of symmetric distributions. The Annals of Statistics, 35(1):224–251.

[7] Karunamuni, R. and Wu, J. (2009). Minimum hellinger distance estimation in a nonparametric mixture model. Journal of Statistical Planning and Inference, 139(3):1118–1133.

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

[9] N'drin, J. A. and Hili, O. (2013). Parameter estimation of one-dimensional dif- fusion process by minimum hellinger distance method. Random Operators and Stochastic Equations, 21(4):403–424.

[10] Ngatchou-Wandji, J. and Bulla, J. (2013). On choosing a mixture model for clustering. Journal of Data Science, 11(1):157–179.

[11] Prause, A., Steland, A., and Abujarad, M. (2016). Minimum hellinger distance estimation for bivariate samples and time series with applications to nonlinear regression and copula-based models. Metrika, 79(4):425–455.

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

[13] Takada, T. (2009). Simulated minimum hellinger distance estimation of stochastic volatility models. Computational Statistics & Data Analysis, 53(6):2390–2403.

[14] Woo, M.-J. and Sriram, T. (2006). Robust estimation of mixture complexity.Journal of the American Statistical Association, 101(476):1475–1486.

[15] Woo, M.-J. and Sriram, T. (2007). Robust estimation of mixture complexity for count data. Computational Statistics & Data Analysis, 51(9):4379–4392.

[16] Wu, J. and Karunamuni, R. J. (2009). On minimum hellinger distance estimation.Canadian Journal of Statistics, 37(4):514–533.

[17] Wu, J. and Karunamuni, R. J. (2012). Efficient hellinger distance estimates for semiparametric models. Journal of Multivariate Analysis, 107:1–23.

[18] Wu, J. and Karunamuni, R. J. (2015). Profile hellinger distance estimation.Statistics, 49(4):711–740.

[19] Xiang, L., Yau, K. K., Van Hui, Y., and Lee, A. H. (2008). Minimum hellinger distance estimation for k-component poisson mixture with random effects. Bio- metrics, 64(2):508–518.

[20] Xiang, S., Yao, W., and Wu, J. (2014). Minimum profile hellinger distance estimation for a semiparametric mixture model. Canadian Journal of Statistics, 42(2):246–267.

[21] Yang, S. (1991). Minimum hellinger distance estimation of parameter in the random censorship model. The Annals of Statistics, 19(2):579–602.

[22] Ying, Z. (1992). Minimum hellinger-type distance estimation for censored data.The Annals of Statistics, 20(3):1361–1390.

Haocheng Li
Department of Mathematics and Statistics
University of Calgary
Calgary, Alberta, Canada T2N 1N4

Jingjing Wu
Department of Mathematics and Statistics University of Calgary
Calgary, Alberta, Canada T2N 1N4

Jian Yang
Department of Mathematics and Statistics University of Calgary
Calgary, Alberta, Canada T2N 1N4