The Canadian Journal of Statistics
Vol. 42, No. 2, 2014, Pages 246–267
La revue canadienne de statistique

Minimum profile Hellinger distance estimation for a semiparametric mixture model

Sijia XIANG[1], Weixin YAO[2]* and Jingjing WU[3]

[1] *School of Mathematics and Statistics, Zhejiang University of Finance and Economics, Hangzhou, Zhejiang 310018, P.R. China*
[2] *Department of Statistics, Kansas State University, Manhattan, KS 66506, USA*
[3] *Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4*

Key words and phrases: Minimum profile Hellinger distance; semiparametric EM algorithm; semiparametric mixture models.

MSC 2010: Primary 62G05; secondary 62G35

Abstract: In this paper, we propose a new effective estimator for a class of semiparametric mixture models where one component has known distribution with possibly unknown parameters while the other component density and the mixing proportion are unknown. Such semiparametric mixture models have been often used in multiple hypothesis testing and the sequential clustering algorithm. The proposed estimator is based on the minimum profile Hellinger distance (MPHD), and its theoretical properties are investigated. In addition, we use simulation studies to illustrate the finite sample performance of the MPHD estimator and compare it with some other existing approaches. The empirical studies demonstrate that the new method outperforms existing estimators when data are generated under contamination and works comparably to existing estimators when data are not contaminated. Applications to two real data sets are also provided to illustrate the effectiveness of the new methodology. *The Canadian Journal of Statistics* 42: 246–267; 2014 © 2014 Statistical Society of Canada

Résumé: Les auteurs proposent un nouvel estimateur efficace pour une classe de modèles de mélange semi-paramétriques où l'une des composantes provient d'une distribution connue dont les paramètres peuvent être inconnus, mais où la distribution des autres composantes et les poids sont inconnus. De tels modèles de mélange semi-paramétriques sont souvent utilisés pour les tests d'hypothèse multiples et pour l'algorithme séquentiel de mise en grappe. L'estimateur proposé est basé sur le profil de distance de Hellinger minimal. Les auteurs étudient les propriétés théoriques de l'estimateur proposé et illustrent sa performance sur des échantillons de taille finie à l'aide de simulations en le comparant aux approches existantes. Cette étude empirique montre que la nouvelle méthode offre des performances supérieures aux méthodes existantes lorsque les données sont générées avec de la contamination, et des performances semblables aux méthodes classiques en absence de contamination. Les auteurs illustrent l'efficacité de la nouvelle méthode en l'appliquant à deux jeux de données réelles. *La revue canadienne de statistique* 42: 246–267; 2014 © 2014 Société statistique du Canada

1. INTRODUCTION

The two-component mixture model considered in this paper is defined by

$$h(x) = \pi f_0(x;\xi) + (1-\pi) f(x-\mu), \quad x \in \mathbb{R}, \tag{1}$$

where $f_0(x;\xi)$ is a known probability density function (pdf) with possibly unknown parameter ξ, f is an unknown pdf with non-null location parameter $\mu \in \mathbb{R}$ and π is the unknown mixing proportion.

* *Author to whom correspondence may be addressed.*
E-mail: wxyao@ksu.edu

Bordes, Delmas, & Vandekerkhove (2006) studied a special case when ξ is assumed to be known, that is, the first component density is completely known and model (1) becomes

$$h(x) = \pi f_0(x) + (1 - \pi) f(x - \mu), \quad x \in \mathbb{R}. \tag{2}$$

Model (2) is motivated by multiple hypothesis testing to detect differentially expressed genes under two or more conditions in microarray data. For this purpose, we build a test statistic for each gene. The test statistics can be considered as coming from a mixture of two distributions: the known distribution f_0 under null hypothesis, and the other distribution $f(\cdot - \mu)$, the unknown distribution of the test statistics under the alternative hypothesis. Please see Section 4 for such an application on multiple hypothesis testing.

Song, Nicolae, & Song (2010) studied another special case of model (1),

$$h(x) = \pi \phi_\sigma(x) + (1 - \pi) f(x), \quad x \in \mathbb{R}, \tag{3}$$

where ϕ_σ is a normal density with mean 0 and unknown standard deviation σ and $f(x)$ is an unknown density. Model (3) was motivated by a sequential clustering algorithm (Song & Nicolae, 2009), which works by finding a local centre of a cluster first, and then identifying whether an object belongs to that cluster or not. If we assume that the objects belonging to the cluster come from a normal distribution with known mean (such as zero) and unknown variance σ^2 and that the objects not belonging to the cluster come from an unknown distribution f, then identifying the points in the cluster is equivalent to estimating the mixing proportion in model (3).

Bordes, Delmas, & Vandekerkhove (2006) proposed to estimate model (2) based on symmetrization of the unknown distribution f and proved the consistency of their estimator. However, the asymptotic distribution of their estimator has not been provided. Song, Nicolae, & Song (2010) also proposed an EM-type estimator and a maximizing π-type estimator (inspired by the constraints imposed to achieve identifiability of the parameters and Swanepoel's approach (Swanepoel, 1999)) to estimate model (3) without providing any asymptotic properties.

In this article, we propose a new estimation procedure for the unified model (1) based on minimum profile Hellinger distance (MPHD) (Wu, Schick, & Karunamuni, 2011). We will investigate the theoretical properties of the proposed MPHD estimator for the semiparametric mixture model, such as existence, consistency and asymptotic normality. A simple and effective algorithm is also given to compute the proposed estimator. Using simulation studies, we illustrate the effectiveness of the MPHD estimator and compare it with the estimators suggested by Bordes, Delmas, & Vandekerkhove (2006) and Song, Nicolae, & Song (2010). Compared to the existing methods (Bordes, Delmas, & Vandekerkhove, 2006; Song, Nicolae, & Song, 2010), the new method can be applied to the more general model (1). In addition, the MPHD estimator works competitively under semiparametric model assumptions, while it is more robust than the existing methods when data are contaminated.

Donoho & Liu (1988) have shown that the class of minimum distance estimators has automatic robustness properties over neighbourhoods of the true model-based on the distance functional defining the estimator. However, minimum distance estimators typically obtain this robustness at the expense of not being optimal at the true model. Beran (1977) has suggested the use of the minimum Hellinger distance (MHD) estimator that has certain robustness properties and is asymptotically efficient at the true model. For a comparison between MHD estimators, MLEs and other minimum distance type estimators, and the balance between robustness and efficiency of estimators, see Lindsay (1994).

There are other well-known robust approaches within the mixture model-based clustering literature. García-Escudero, Gordaliza, & Matrán (2003) proposed exploratory graphical tools based on trimming for detecting main clusters in a given dataset, where the trimming is obtained by resorting to trimmed k-means methodology. García-Escudero et al. (2008) introduced a new method for performing clustering with the aim of fitting clusters with different scatters and weights. García-Escudero et al. (2010) reviewed different robust clustering approaches in the literature, emphasizing on methods based on trimming which try to discard most outlying data when carrying out the clustering process. A more recent work by Punzo & McNicholas (2013) introduced a family of 14 parsimonious mixtures of contaminated Gaussian distributions models within the general model-based classification framework.

The rest of the article is organized as follows. In Section 2, we introduce the proposed MPHD estimator and discuss its asymptotic properties. Section 3 presents simulation results for comparing the new estimation with some existing methods. Applications to two real data sets are also provided in Section 4 to illustrate the effectiveness of the proposed methodology. A discussion section ends the paper.

2. MPHD ESTIMATION

2.1. Introduction of MPHD Estimator

In this section, we develop a MPHD estimator for model (1). Let

$$\mathscr{H} = \{h_{\boldsymbol{\theta},f}(x) = \pi f_0(x;\xi) + (1-\pi)f(x-\mu) : \boldsymbol{\theta} \in \Theta, f \in \mathscr{F}\},$$

where

$$\Theta = \{\boldsymbol{\theta} = (\pi,\xi,\mu) : \pi \in (0,1), \xi \in \mathbb{R}, \mu \in \mathbb{R}\},$$
$$\mathscr{F} = \{f : f \geq 0, \int f(x)\mathrm{d}x = 1\}$$

be the functional space for the semiparametric model (1). In practice, the parameter space of ξ depends on its interpretation. For example, if ξ is the standard deviation of f_0, then the parameter space of ξ will be $\mathbb{R}^+$. For model (2), ξ is known and thus the parameter space of ξ is a singleton and, as a result, $\boldsymbol{\theta} = (\pi, \mu)$.

Let $\|\cdot\|$ denote the $L_2(v)$-norm. For any $g_1, g_2 \in L_2(v)$, the Hellinger distance between them is defined as

$$d_{\mathrm{H}}(g_1, g_2) = \left\| g_1^{1/2} - g_2^{1/2} \right\|.$$

Suppose a sample $X_1, X_2, ..., X_n$ is from a population with density function $h_{\boldsymbol{\theta},f} \in \mathscr{H}$. We propose to estimate $\boldsymbol{\theta}$ and f by minimizing the Hellinger distance

$$\left\| h_{t,l}^{1/2} - \hat{h}_n^{1/2} \right\| \tag{4}$$

over all $t \in \Theta$ and $l \in \mathscr{F}$, where $\hat{h}_n$ is an appropriate nonparametric density estimator of $h_{\boldsymbol{\theta},f}$. Note that the above objective function (4) contains both the parametric component t and the nonparametric component l. Here, we propose to use the profile idea to implement the calculation.

For any density function g and t, define functional $f(t,g)$ as

$$f(t,g) = \arg\min_{l \in \mathscr{F}} \left\| h_{t,l}^{1/2} - g^{1/2} \right\|$$

and then define the profile Hellinger distance as

$$d_{\mathrm{PH}}(t,g) = \| h_{t,f(t,g)}^{1/2} - g^{1/2} \|.$$

Now the MPHD functional $T(g)$ is defined as

$$T(g) = \arg\min_{t\in\Theta} d_{\mathrm{PH}}(t, g) = \arg\min_{t\in\Theta} \left\| h_{t,f(t,g)}^{1/2} - g^{1/2} \right\|. \tag{5}$$

Given the sample $X_1, X_2, ..., X_n$, one can construct an appropriate nonparametric density estimator of $h_{\boldsymbol{\theta},f}$, say $\hat{h}_n$, and then the proposed MPHD estimator of $\boldsymbol{\theta}$ is given by $T(\hat{h}_n)$. In the examples of Sections 3 and 4, we use the kernel density estimator for $\hat{h}_n$ and the bandwidth h is chosen based on Botev, Grotowski, & Kroese (2010).

2.2. Algorithm

In this section, we propose the following two-step algorithm to calculate the MPHD estimator. Suppose the initial estimates of $\boldsymbol{\theta} = (\pi, \xi, \mu)$ and f are $\boldsymbol{\theta}^{(0)} = (\pi^{(0)}, \xi^{(0)}, \mu^{(0)})$ and $f^{(0)}$.

Step 1: Given $\pi^{(k)}$, $\xi^{(k)}$ and $\mu^{(k)}$, find $f^{(k+1)}$ which minimizes

$$\left\| [\pi^{(k)} f_0(\cdot; \xi^{(k)}) + (1-\pi^{(k)}) f^{(k+1)}(\cdot - \mu^{(k)})]^{1/2} - \hat{h}_n^{1/2}(\cdot) \right\|.$$

Similar to Wu, Schick, & Karunamuni (2011), we obtain that

$$f^{(k+1)}(x-\mu^{(k)}) = \begin{cases} \dfrac{\alpha}{1-\pi^{(k)}} \hat{h}_n(x) - \dfrac{\pi^{(k)}}{1-\pi^{(k)}} f_0(x; \xi^{(k)}), & \text{if } x \in M, \\ 0, & \text{if } x \in M^C, \end{cases}$$

where $M = \{x : \alpha \hat{h}_n(x) \geq \pi^{(k)} f_0(x; \xi^{(k)})\}$ and $\alpha = \sup\limits_{0<\alpha\leq 1} \{\pi^{(k)} \int_M f_0(x; \xi^{(k)}) \mathrm{d}x + (1 - \pi^{(k)}) \geq \alpha \int_M \hat{h}_n(x) \mathrm{d}x\}$.

Step 2: Given fixed $f^{(k+1)}$, find $\pi^{(k+1)}$, $\xi^{(k+1)}$ and $\mu^{(k+1)}$ which minimize

$$\left\| [\pi^{(k+1)} f_0(\cdot; \xi^{(k+1)}) + (1-\pi^{(k+1)}) f^{(k+1)}(\cdot - \mu^{(k+1)})]^{1/2} - \hat{h}_n^{1/2}(\cdot) \right\|. \tag{6}$$

Then go back to **Step 1**.

Each of the above two steps monotonically decreases the objective function (4) until convergence. In Step 1, if $f(\cdot)$ is assumed to be symmetric, then we can further symmetrize $f^{(k+1)}(\cdot)$ as

$$\tilde{f}^{(k+1)}(x) = \frac{f^{(k+1)}(x) + f^{(k+1)}(-x)}{2}.$$

Note that there is no closed form for (6) in Step 2 and thus some numerical algorithms, such as the Newton–Raphson algorithm, are needed to minimize (6). In our examples, we used the "fminsearch" function in Matlab to find the minimizer numerically. "fminsearch" function uses the Nelder–Mead simplex algorithm as described in Lagarias et al. (1998).

2.3. Asymptotic Results

Note that $\boldsymbol{\theta}$ and f in the semiparametric mixture model (1) are not generally identifiable without any assumptions for f. Bordes, Delmas, & Vandekerkhove (2006) showed that model (2) is not generally identifiable if we do not put any restrictions on the unknown density f, but identifiability can be achieved under some sufficient conditions. One of these conditions is that $f(\cdot)$ is symmetric about 0. Under these conditions, Bordes, Delmas, & Vandekerkhove (2006) proposed an elegant estimation procedure based on the symmetry of f. Song, Nicolae, & Song (2010) also addressed the non-identifiability problem and noticed that model (3) is not generally identifiable. However, due to the additional unknown parameter σ in the first component, Song, Nicolae, & Song (2010) mentioned that it is hard to find the conditions to avoid unidentifiability of model (3) and proposed using simulation studies to check the performance of the proposed estimators. Please refer to Bordes, Delmas, & Vandekerkhove (2006) and Song, Nicolae, & Song (2010) for detailed discussions on the identifiability of model (1).

Next, we discuss some asymptotic properties of the proposed MPHD estimator. Here, for simplicity of explanation, we will only consider model (2) for which Bordes, Delmas, & Vandekerkhove (2006) has proved identifiability. However, we conjecture that all the results presented in this section also apply to the unified model (1) when it is identifiable. But this is beyond the scope of the article and requires more research to find the identifiable conditions for the general model (1).

The next theorem gives results on the existence and uniqueness of the proposed estimator, and the continuity of the functional defined in (5), which is in line with Theorem 1 of Beran (1977).

Theorem 1. *With T defined by (5), if model (2) is identifiable, then we have*

1. *For every $h_{\boldsymbol{\theta},f} \in \mathscr{H}$, there exists $T(h_{\boldsymbol{\theta},f}) \in \Theta$ satisfying (5);*
2. *$T(h_{\boldsymbol{\theta},f}) = \boldsymbol{\theta}$ uniquely for any $\boldsymbol{\theta} \in \Theta$;*
3. *$T(h_n) \to T(h_{\boldsymbol{\theta},f})$ for any sequences $\{h_n\}_{n\in\mathbb{N}}$ such that $\left\|h_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2}\right\| \to 0$ and $\sup_{t\in\Theta}$ $\left\|h_{t,f(t,h_n)} - h_{t,f(t,h_{\boldsymbol{\theta},f})}\right\| \to 0$
as $n \to \infty$.*

Remark 1. *Without the global identifiability of model (2), the local identifiability of model (2) proved by Bordes, Delmas, & Vandekerkhove (2006) tells that there exists one solution that has the asymptotic properties presented in Theorem 1.*

Define a kernel density estimator based on $X_1, X_2, ..., X_n$ as

$$\hat{h}_n(x) = \frac{1}{nc_ns_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{c_ns_n}\right), \tag{7}$$

where $\{c_n\}$ is a sequence of constants (bandwidths) converging to zero at an appropriate rate and s_n is a robust scale statistic. Under further conditions on the kernel density estimator defined in (7), the consistency of the MPHD estimator is established in the next theorem.

Theorem 2. *Suppose that*

1. *The kernel function $K(\cdot)$ is absolutely continuous and bounded with compact support.*
2. $\lim_{n\to\infty} c_n = 0$, $\lim_{n\to\infty} n^{1/2}c_n = \infty$.
3. *The model (2) is identifiable and $h_{\boldsymbol{\theta},f}$ is uniformly continuous.*

Then $\|\hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2}\| \xrightarrow{p} 0$ as $n \to \infty$, and therefore $T(\hat{h}_n) \xrightarrow{p} T(h_{\boldsymbol{\theta},f})$ as $n \to \infty$.

Define the map $\boldsymbol{\theta} \mapsto s_{\boldsymbol{\theta},g}$ as $s_{\boldsymbol{\theta},g} = h_{\boldsymbol{\theta},f(\boldsymbol{\theta},g)}^{1/2}$, and suppose that for $\boldsymbol{\theta} \in \Theta$ there exists a 2×1 vector $\dot{s}_{\boldsymbol{\theta},g}$ with components in L_2 and a 2×2 matrix $\ddot{s}_{\boldsymbol{\theta},g}$ with components in L_2 such that for every 2×1 real vector e of unit Euclidean length and for every scalar α in a neighborhood of zero,

$$s_{\boldsymbol{\theta}+\alpha e,g}(x) = s_{\boldsymbol{\theta},g}(x) + \alpha e^T \dot{s}_{\boldsymbol{\theta},g}(x) + \alpha e^T u_{\alpha,g}(x), \tag{8}$$

$$\dot{s}_{\boldsymbol{\theta}+\alpha e,g}(x) = \dot{s}_{\boldsymbol{\theta},g}(x) + \alpha \ddot{s}_{\boldsymbol{\theta},g}(x) e + \alpha v_{\alpha,g}(x) e, \tag{9}$$

where $u_{\alpha,g}(x)$ is 2×1, $v_{\alpha,g}(x)$ is 2×2, and the components of $u_{\alpha,g}$ and $v_{\alpha,g}$ tend to zero in L_2 as $\alpha \to 0$.

The next theorem shows that the MPHD estimator has an asymptotic normal distribution.

Theorem 3. *Suppose that*

1. *Model (2) is identifiable.*
2. *The conditions in Theorem 2 hold.*
3. *The map $\boldsymbol{\theta} \mapsto s_{\boldsymbol{\theta},g}$ satisfies (8) and (9) with continuous gradient vector $\dot{s}_{\boldsymbol{\theta},g}$ and continuous Hessian matrix $\ddot{s}_{\boldsymbol{\theta},g}$ in the sense that $\|\dot{s}_{\boldsymbol{\theta}_n,g_n} - \dot{s}_{\boldsymbol{\theta},g}\| \to 0$ and $\|\ddot{s}_{\boldsymbol{\theta}_n,g_n} - \ddot{s}_{\boldsymbol{\theta},g}\| \to 0$ whenever $\boldsymbol{\theta}_n \to \boldsymbol{\theta}$ and $\|g_n^{1/2} - g^{1/2}\| \to 0$ as $n \to \infty$.*
4. *$< \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h^{1/2}_{\boldsymbol{\theta},f} >$ is invertible.*

Then, with T defined in (5) for model (2), the asymptotic distribution of $n^{1/2}(T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f}))$ is $N(0, \Sigma)$ with variance matrix Σ defined by

$$\Sigma = < \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h^{1/2}_{\boldsymbol{\theta},f} >^{-1} < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \dot{s}^T_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}} > < \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h^{1/2}_{\boldsymbol{\theta},f} >^{-1} .$$

3. SIMULATION STUDIES

In this section, we investigate the finite sample performance of the proposed MPHD estimator and compare it to Maximizing-π type estimator (Song, Nicolae, & Song, 2010), EM-type estimator (Song, Nicolae, & Song, 2010) and Symmetrization estimator (Bordes, Delmas, & Vandekerkhove, 2006) under both models (2) and (3).

Model (3) that Song, Nicolae, & Song (2010) considered does not have a location parameter in the second component. However, we can equivalently replace $f(x)$ with $f(x - \mu)$, where $\mu \in \mathbb{R}$ is a location parameter. Throughout this section, we will consider this equivalent form of (3). Under this model, after we have $\hat{\pi}$ and $\hat{\sigma}$, we can simply estimate μ by

$$\hat{\mu} = \frac{\sum_{i=1}^{n} (1 - \hat{Z}_i) X_i}{\sum_{i=1}^{n} (1 - \hat{Z}_i)},$$

where

$$\hat{Z}_i = \frac{2\hat{\pi}\phi_{\hat{\sigma}}(X_i)}{\hat{\pi}\phi_{\hat{\sigma}}(X_i) + \hat{h}(X_i)}.$$

We first compare the performance of different estimators under model (2). Suppose $(X_1, \ldots, X_n)$ are generated from one of the following five cases:

Case I: $X \sim 0.3N(0, 1) + 0.7N(1.5, 1) \Rightarrow (\pi, \mu) = (0.3, 1.5)$,
Case II: $X \sim 0.3N(0, 1) + 0.7N(3, 1) \Rightarrow (\pi, \mu) = (0.3, 3)$,
Case III: $X \sim 0.3N(0, 1) + 0.7U(2, 4) \Rightarrow (\pi, \mu) = (0.3, 3)$,
Case IV: $X \sim 0.7N(0, 4) + 0.3N(3, 1) \Rightarrow (\pi, \mu) = (0.7, 3)$,
Case V: $X \sim 0.85N(0, 4) + 0.15N(3, 1) \Rightarrow (\pi, \mu) = (0.85, 3)$.

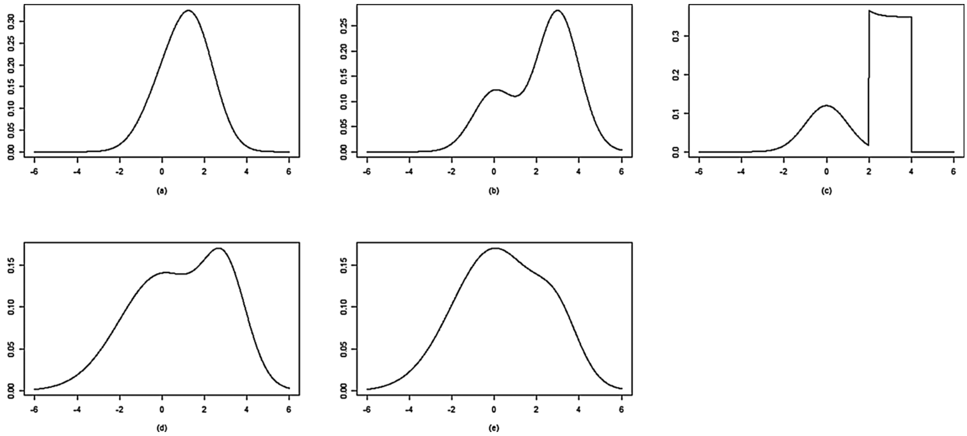

Figure 1: Density plots of: (a) Case I; (b) Case II; (c) Case III; (d) Case IV and (e) Case V.

Figure 1 shows the density plots of the five cases. Cases I, II and III are the models used by Song, Nicolae, & Song (2010) to show the performance of their Maximizing-π type and EM-type estimators. Case I represents the situation when two components are close, and Case II represents the situation when two components are apart. Cases IV and V are suggested by Bordes, Delmas, & Vandekerkhove (2006) to show the performance of their semiparametric EM algorithm. In addition, we also consider the corresponding contaminated models by adding 2% outliers from $U(10, 20)$ to the above five models.

Tables 1, 2 and 3 report the bias and MSE of the parameter estimates of (π, μ) for the four methods when $n = 100$, $n = 250$ and $n = 1,000$, respectively, based on 200 repetitions. Tables 4, 5 and 6 report the respective results for $n = 100$, $n = 250$ and $n = 1,000$ when the data are under 2% contamination from $U(10, 20)$. The best values are highlighted in bold. From the six tables, we can see that the MPHD estimator has better overall performance than the Maximizing-π type, the EM-type and the Symmetrization estimators, especially when sample size is large. When the sample is not contaminated by outliers, the MPHD estimator and the Symmetrization estimator

Table 1: Bias (MSE) of point estimates for model (2) over 200 repetitions with $n = 100$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	−0.092(0.030)	0.057(0.011)	0.271(0.078)	**0.003(0.009)**
	$\mu : 1.5$	−0.113(0.118)	0.196(0.070)	0.465(0.239)	**0.020(0.026)**
II	$\pi : 0.3$	−**0.014(0.003)**	−0.052(0.005)	**0.027(0.003)**	−**0.002(0.003)**
	$\mu : 3$	−**0.000(0.021)**	−0.123(0.038)	**0.020(0.017)**	−0.009(0.025)
III	$\pi : 0.3$	−0.046(0.005)	−0.108(0.014)	−0.045(0.005)	**0.001(0.003)**
	$\mu : 3$	−**0.008(0.004)**	−0.341(0.138)	−0.212(0.058)	−**0.002(0.006)**
IV	$\pi : 0.7$	−**0.044(0.015)**	−0.131(0.025)	**0.086(0.010)**	−0.089(0.028)
	$\mu : 3$	0.173(0.247)	−0.697(0.659)	−**0.053(0.177)**	−0.326(0.465)
V	$\pi : 0.85$	−0.094(0.041)	−0.147(0.030)	**0.039(0.003)**	−0.106(0.024)
	$\mu : 3$	**0.109(1.145)**	−1.375(2.298)	−**0.697(1.136)**	−0.742(1.184)

TABLE 2: Bias (MSE) of point estimates for model (2) over 200 repetitions with $n = 250$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi: 0.3$	−0.090(0.028)	**0.028(0.005)**	0.269(0.074)	−0.080(0.021)
	$\mu: 1.5$	−0.110(0.084)	**0.162(0.041)**	0.472(0.231)	**−0.107(0.060)**
II	$\pi: 0.3$	**−0.009(0.001)**	−0.058(0.005)	0.034(0.002)	**−0.001(0.001)**
	$\mu: 3$	**0.007(0.007)**	−0.118(0.027)	0.057(0.009)	**−0.004(0.009)**
III	$\pi: 0.3$	−0.041(0.003)	−0.071(0.006)	**−0.016(0.001)**	**−0.001(0.001)**
	$\mu: 3$	**−0.001(0.001)**	−0.188(0.043)	−0.082(0.010)	**−0.001(0.002)**
IV	$\pi: 0.7$	**−0.009(0.003)**	−0.108(0.018)	0.102(0.012)	−0.017(0.009)
	$\mu: 3$	**0.131(0.067)**	−0.618(0.501)	**0.063(0.069)**	−0.095(0.159)
V	$\pi: 0.85$	**−0.040(0.014)**	−0.121(0.021)	**0.052(0.003)**	−0.041(0.011)
	$\mu: 3$	**0.217(0.444)**	−1.134(1.503)	**−0.323(0.349)**	−0.345(0.625)

are very competitive and perform better than other estimators. When the sample is contaminated by outliers, the MPHD estimator performs much better and therefore is more robust than the other three methods. We also observe that when the sample is contaminated by outliers, among the Maximizing-π type, the EM-type and the Symmetrization estimators, the EM-type estimator tends to give better mixing proportion estimates than the other two.

Next, we also evaluate how the MPHD estimator performs under model (3), *where the variance σ^2 is assumed to be unknown*, and compare it with other methods using the same five cases as in Tables 1–6.

Tables 7, 8 and 9 report the bias and MSE of the parameter estimates for $n = 100$, $n = 250$ and $n = 1,000$, respectively, when there are no contaminations. Based on these three tables,

TABLE 3: Bias (MSE) of point estimates for model (2) over 200 repetitions with $n = 1,000$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi: 0.3$	**−0.009(0.005)**	**−0.020(0.003)**	0.263(0.069)	−0.024(0.005)
	$\mu: 1.5$	**0.003(0.016)**	0.083(0.017)	0.459(0.213)	**−0.031(0.015)**
II	$\pi: 0.3$	**−0.006(0.001)**	−0.055(0.004)	0.039(0.002)	**−0.003(0.001)**
	$\mu: 3$	**0.006(0.002)**	−0.083(0.016)	0.093(0.010)	**−0.002(0.002)**
III	$\pi: 0.3$	**−0.028(0.001)**	−0.061(0.005)	**−0.004(0.001)**	**0.000(0.001)**
	$\mu: 3$	**−0.003(0.001)**	−0.153(0.029)	−0.044(0.002)	**−0.002(0.001)**
IV	$\pi: 0.7$	**−0.008(0.001)**	−0.115(0.020)	0.104(0.011)	**−0.007(0.001)**
	$\mu: 3$	**0.045(0.013)**	−0.554(0.400)	0.174(0.039)	**−0.030(0.017)**
V	$\pi: 0.85$	**−0.007(0.001)**	−0.101(0.016)	0.061(0.004)	**−0.007(0.002)**
	$\mu: 3$	**0.172(0.063)**	−0.929(1.043)	**0.019(0.067)**	−0.066(0.104)

TABLE 4: Bias (MSE) of point estimates for model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 100$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	−0.124(0.036)	**0.060(0.010)**	0.267(0.075)	−0.063(0.014)
	$\mu : 1.5$	−0.163(0.128)	0.692(0.629)	1.079(1.348)	**−0.031(0.015)**
II	$\pi : 0.3$	−0.029(0.005)	−0.055(0.006)	**0.018(0.004)**	−0.300(0.090)
	$\mu : 3$	**−0.011(0.046)**	0.252(0.136)	0.398(0.228)	−3.000(9.000)
III	$\pi : 0.3$	**−0.034(0.003)**	−0.108(0.015)	−0.048(0.005)	**−0.032(0.004)**
	$\mu : 3$	**−0.011(0.004)**	−0.034(0.080)	0.104(0.091)	−0.014(0.009)
IV	$\pi : 0.7$	**−0.054(0.020)**	−0.133(0.027)	**0.081(0.009)**	−0.200(0.083)
	$\mu : 3$	**0.152(0.389)**	0.172(0.668)	1.141(2.123)	−0.582(0.867)
V	$\pi : 0.85$	−0.125(0.071)	−0.158(0.033)	**0.024(0.002)**	−0.217(0.080)
	$\mu : 3$	0.048(1.364)	**−0.007(1.314)**	1.373(4.337)	−0.910(1.444)

we can see that when there are no contaminations, the MPHD estimator and the Symmetrization estimator perform better than the Maximizing-π type estimator and the EM-type estimator. Tables 10, 11 and 12 report the results when models are under 2% contamination from $U(10, 20)$ for $n = 100$, $n = 250$ and $n = 1,000$, respectively. From these three tables, we can see that the MPHD estimator performs much better again than the other three methods.

To see the comparison and difference better, we also plot in Figures 2–4 the results reported in Tables 6 and 9. Figure 2 contains the MSE of point estimates of μ that are presented in Table 9 for model (3) (σ unknown) and Figures 3 and 4 contain the MSEs of point estimates of μ and π, respectively, that are presented in Table 6 for model (2) (σ known), under 2% contamination from

TABLE 5: Bias (MSE) of point estimates for model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 250$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	−0.090(0.026)	**0.032(0.006)**	0.263(0.071)	−0.180(0.043)
	$\mu : 1.5$	**−0.102(0.085)**	0.613(0.434)	1.043(1.146)	**−0.224(0.081)**
II	$\pi : 0.3$	**−0.019(0.001)**	−0.065(0.006)	0.027(0.002)	−0.044(0.003)
	$\mu : 3$	**−0.009(0.007)**	0.213(0.076)	0.415(0.202)	−0.044(0.012)
III	$\pi : 0.3$	**−0.021(0.001)**	−0.073(0.007)	**−0.015(0.001)**	−0.028(0.002)
	$\mu : 3$	**−0.004(0.001)**	0.119(0.043)	0.245(0.086)	−0.011(0.003)
IV	$\pi : 0.7$	**−0.020(0.005)**	−0.122(0.021)	0.086(0.009)	−0.302(0.164)
	$\mu : 3$	**0.149(0.096)**	0.162(0.296)	1.149(1.594)	−0.746(1.137)
V	$\pi : 0.85$	−0.053(0.025)	−0.131(0.023)	**0.034(0.002)**	−0.311(0.140)
	$\mu : 3$	**0.220(0.513)**	0.358(1.000)	1.859(4.597)	−1.093(1.785)

TABLE 6: Bias (MSE) of point estimates for model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1,000$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	−0.460(0.007)	−**0.024(0.003)**	0.255(0.065)	−0.240(0.059)
	$\mu : 1.5$	−**0.056(0.019)**	0.509(0.284)	1.048(1.119)	−0.313(0.103)
II	$\pi : 0.3$	−**0.014(0.001)**	−0.057(0.004)	**0.032(0.001)**	−0.043(0.002)
	$\mu : 3$	**0.001(0.002)**	0.257(0.081)	0.444(0.204)	−0.034(0.005)
III	$\pi : 0.3$	−**0.019(0.001)**	−0.066(0.005)	−**0.011(0.001)**	−0.035(0.002)
	$\mu : 3$	−**0.001(0.001)**	0.179(0.044)	0.299(0.096)	−**0.011(0.001)**
IV	$\pi : 0.7$	−**0.019(0.001)**	−0.128(0.023)	0.089(0.008)	−0.311(0.149)
	$\mu : 3$	**0.067(0.013)**	0.203(0.257)	1.252(1.628)	−0.829(1.165)
V	$\pi : 0.85$	−**0.019(0.001)**	−0.112(0.018)	0.045(0.002)	−0.347(0.134)
	$\mu : 3$	**0.177(0.067)**	0.574(0.836)	2.275(5.478)	−1.466(2.329)

$U(10, 20)$. From the three plots, we can see that all four estimators perform well in Cases II and III. The EM-type estimator performs poorly in Case I, and is the worst estimate of μ in Cases IV and V when data are contaminated. The Symmetrization estimator is sensitive to contamination, especially in Cases IV and V, no matter σ is known or not. Comparatively, the Maximizing-π

TABLE 7: Bias (MSE) of point estimates for model (3) over 200 repetitions with $n = 100$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	−0.058(0.021)	0.110(0.021)	0.302(0.097)	−**0.047(0.015)**
	$\sigma : 1$	0.052(0.045)	0.758(2.207)	**0.143(0.042)**	−**0.047(0.071)**
	$\mu : 1.5$	−0.057(0.082)	0.098(0.095)	0.463(0.242)	−**0.055(0.061)**
II	$\pi : 0.3$	−**0.008(0.004)**	0.062(0.017)	0.082(0.014)	−**0.006(0.004)**
	$\sigma : 1$	**0.095(0.041)**	1.821(5.180)	0.331(0.252)	**0.012(0.056)**
	$\mu : 3$	−**0.014(0.025)**	−0.341(0.216)	0.081(0.031)	−0.032(0.030)
III	$\pi : 0.3$	−0.051(0.005)	0.024(0.011)	−0.042(0.006)	−**0.009(0.003)**
	$\sigma : 1$	−**0.101(0.030)**	2.258(6.708)	−**0.028(0.105)**	−0.031(0.045)
	$\mu : 3$	−**0.021(0.005)**	−0.436(0.223)	−0.187(0.049)	−**0.008(0.008)**
IV	$\pi : 0.7$	−**0.014(0.011)**	−0.060(0.012)	0.114(0.016)	−0.054(0.018)
	$\sigma : 2$	0.101(0.047)	0.195(0.161)	**0.120(0.034)**	**0.039(0.065)**
	$\mu : 3$	0.100(0.201)	−0.537(0.504)	**0.019(0.175)**	−0.320(0.511)
V	$\pi : 0.85$	−**0.028(0.009)**	−0.076(0.014)	**0.042(0.003)**	−0.159(0.078)
	$\sigma : 2$	0.098(0.043)	0.179(0.100)	−**0.006(0.021)**	−0.118(0.247)
	$\mu : 3$	**0.275(0.432)**	−1.080(1.719)	−0.622(1.088)	−0.845(1.717)

TABLE 8: Bias (MSE) of point estimates for model (3) over 200 repetitions with $n = 250$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	**−0.043(0.014)**	**0.064(0.006)**	0.302(0.093)	−0.048(0.015)
	$\sigma : 1$	**0.058(0.021)**	−0.101(0.075)	0.157(0.032)	**0.020(0.033)**
	$\mu : 1.5$	**−0.064(0.051)**	0.220(0.059)	0.421(0.186)	**−0.079(0.049)**
II	$\pi : 0.3$	**−0.005(0.001)**	−0.028(0.003)	0.093(0.011)	**−0.002(0.001)**
	$\sigma : 1$	**0.046(0.013)**	0.330(0.912)	0.377(0.191)	**−0.001(0.021)**
	$\mu : 3$	**−0.005(0.010)**	−0.129(0.054)	0.121(0.022)	−0.017(0.011)
III	$\pi : 0.3$	−0.037(0.002)	−0.043(0.004)	0.005(0.002)	**0.002(0.001)**
	$\sigma : 1$	**−0.061(0.013)**	0.609(1.741)	0.163(0.100)	**0.013(0.022)**
	$\mu : 3$	**−0.006(0.001)**	−0.233(0.085)	−0.069(0.009)	**0.001(0.002)**
IV	$\pi : 0.7$	**−0.008(0.003)**	−0.068(0.009)	0.121(0.016)	−0.014(0.007)
	$\sigma : 2$	**0.036(0.023)**	0.023(0.035)	0.142(0.028)	**0.009(0.032)**
	$\mu : 3$	**0.108(0.054)**	−0.437(0.269)	0.153(0.067)	**−0.070(0.140)**
V	$\pi : 0.85$	**−0.014(0.003)**	−0.076(0.010)	0.060(0.004)	−0.076(0.028)
	$\sigma : 2$	0.093(0.027)	0.069(0.035)	**0.046(0.011)**	**0.027(0.048)**
	$\mu : 3$	**0.115(0.205)**	−0.912(1.024)	−0.222(0.266)	−0.573(0.981)

TABLE 9: Bias (MSE) of point estimates for model (3) over 200 repetitions with $n = 1,000$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	**−0.019(0.005)**	**0.053(0.004)**	0.301(0.091)	−0.020(0.005)
	$\sigma : 1$	**0.040(0.008)**	−0.147(0.028)	0.177(0.034)	**0.025(0.011)**
	$\mu : 1.5$	**−0.019(0.017)**	0.236(0.059)	0.423(0.181)	−0.024(0.018)
II	$\pi : 0.3$	**−0.001(0.001)**	−0.037(0.002)	0.099(0.010)	**0.000(0.001)**
	$\sigma : 1$	**0.017(0.003)**	−0.044(0.007)	0.407(0.176)	**−0.002(0.005)**
	$\mu : 3$	**0.009(0.002)**	−0.042(0.005)	0.151(0.025)	**0.003(0.002)**
III	$\pi : 0.3$	**−0.029(0.001)**	−0.047(0.003)	**0.011(0.001)**	**0.001(0.001)**
	$\sigma : 1$	−0.051(0.005)	−0.029(0.007)	0.177(0.044)	**0.005(0.004)**
	$\mu : 3$	**−0.003(0.001)**	−0.122(0.017)	−0.031(0.002)	**−0.001(0.001)**
IV	$\pi : 0.7$	**−0.008(0.001)**	−0.069(0.006)	0.125(0.016)	**−0.004(0.001)**
	$\sigma : 2$	**0.002(0.006)**	−0.051(0.013)	0.172(0.032)	**−0.001(0.006)**
	$\mu : 3$	0.058(0.017)	−0.346(0.153)	0.161(0.035)	**−0.018(0.015)**
V	$\pi : 0.85$	**−0.003(0.001)**	−0.067(0.006)	0.072(0.005)	−0.025(0.010)
	$\sigma : 2$	0.053(0.009)	**−0.005(0.008)**	0.087(0.010)	0.008(0.031)
	$\mu : 3$	**0.099(0.042)**	−0.745(0.633)	0.135(0.060)	−0.180(0.293)

TABLE 10: Bias (MSE) of point estimates for model (3), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 100$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	−0.104(0.025)	**0.102(0.018)**	0.295(0.093)	−0.132(0.031)
	$\sigma : 1$	**0.132(0.090)**	0.680(1.919)	**0.133(0.046)**	−0.213(0.150)
	$\mu : 1.5$	−0.148(0.088)	0.591(0.560)	1.115(1.507)	**−0.137(0.068)**
II	$\pi : 0.3$	**−0.022(0.005)**	0.051(0.016)	0.067(0.011)	−0.062(0.010)
	$\sigma : 1$	**0.081(0.034)**	1.755(5.036)	0.301(0.235)	−0.244(0.121)
	$\mu : 3$	**−0.025(0.036)**	0.053(0.180)	0.467(0.323)	−0.079(0.051)
III	$\pi : 0.3$	**−0.036(0.003)**	**0.019(0.012)**	−0.036(0.005)	−0.046(0.006)
	$\sigma : 1$	**−0.061(0.019)**	2.229(6.635)	**0.025(0.102)**	−0.201(0.076)
	$\mu : 3$	**−0.022(0.004)**	−0.116(0.114)	0.144(0.085)	−0.034(0.009)
IV	$\pi : 0.7$	**−0.033(0.017)**	**−0.066(0.013)**	**0.099(0.013)**	−0.110(0.033)
	$\sigma : 2$	**0.088(0.058)**	0.184(0.147)	**0.104(0.032)**	−0.152(0.110)
	$\mu : 3$	**0.103(0.262)**	0.449(0.928)	1.209(2.263)	−0.226(0.354)
V	$\pi : 0.85$	−0.045(0.023)	−0.084(0.014)	**0.024(0.002)**	−0.198(0.106)
	$\sigma : 2$	0.145(0.082)	0.222(0.135)	**−0.013(0.027)**	−0.172(0.199)
	$\mu : 3$	**0.379(2.637)**	0.646(2.505)	1.235(3.351)	**−0.501(1.258)**

type estimator is more robust, but it does not perform well in Cases IV and V when data are not under contamination. However, the MPHD estimator performs well in all cases.

4. REAL DATA APPLICATION

Example 1(Iris data). We illustrate the application of the new estimation procedure to the sequential clustering algorithm using the Iris data, which are perhaps one of the best known data sets in pattern recognition literature. Iris data were first introduced by Fisher (1936) and are referenced frequently to this day. These data contain four attributes: sepal length (in cm), sepal width (in cm), petal length (in cm) and petal width (in cm), and there are three classes of 50 instances each, where each class refers to a type of Iris plant. One class is linearly separable from the other two and the latter are not linearly separable from each other.

Assuming the class indicators are unknown, we want to recover the three clusters in the data. After applying the search algorithm for centres of clusters by Song, Nicolae, & Song (2010), observation 8 is selected as the centre of the first cluster. We adjust all observations by subtracting observation 8 from each observation. As discussed by Song, Nicolae, & Song (2010), the proportion of observations that belong to a cluster can be considered as the mixing proportion in the two-component semiparametric mixture model (3).

Principal component analysis shows that the first principal component accounts for 92.46% of the total variability, so it would seem that the Iris data tend to fall within a 1-dimensional subspace of the 4-dimensional sample space. Figure 5 is a histogram of the first principal component. From the histogram, we can see that the first cluster is separated from the rest of the data, with observation 8 (first principal component score equals −2.63) being the centre of it. The first

TABLE 11: Bias (MSE) of point estimates for model (3), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 250$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	−0.108(0.024)	**0.060(0.006)**	0.292(0.087)	−0.164(0.038)
	$\sigma : 1$	0.103(0.056)	**−0.015(0.184)**	**0.155(0.031)**	−0.216(0.116)
	$\mu : 1.5$	**−0.145(0.070)**	1.697(0.550)	1.085(1.277)	**−0.177(0.067)**
II	$\pi : 0.3$	**−0.011(0.001)**	−0.033(0.003)	0.087(0.009)	−0.049(0.005)
	$\sigma : 1$	**0.056(0.014)**	0.306(0.843)	0.400(0.204)	−0.195(0.062)
	$\mu : 3$	**−0.011(0.012)**	0.245(0.115)	0.525(0.316)	−0.047(0.016)
III	$\pi : 0.3$	**−0.025(0.001)**	−0.073(0.008)	−0.723(0.002)	−0.042(0.003)
	$\sigma : 1$	**−0.057(0.012)**	1.125(3.379)	0.081(0.055)	−0.203(0.056)
	$\mu : 3$	**−0.008(0.001)**	−0.068(0.060)	0.207(0.073)	−0.029(0.004)
IV	$\pi : 0.7$	**−0.024(0.004)**	−0.089(0.012)	0.102(0.011)	−0.077(0.013)
	$\sigma : 2$	**0.010(0.018)**	0.035(0.041)	0.138(0.028)	−0.213(0.078)
	$\mu : 3$	**0.118(0.064)**	0.406(0.435)	1.339(2.125)	**−0.032(0.084)**
V	$\pi : 0.85$	**−0.027(0.006)**	−0.098(0.014)	**0.037(0.002)**	−0.114(0.038)
	$\sigma : 2$	0.052(0.029)	0.069(0.034)	**0.041(0.010)**	−0.193(0.099)
	$\mu : 3$	**0.215(0.228)**	0.715(1.406)	1.963(4.889)	**−0.130(0.460)**

principal component loading vector is $(0.36, -0.08, 0.86, 0.35)$, which implies that the petal length contains most of the information. We apply each of the four estimation methods discussed above to the first principal component. Note, however, that the leading principal components are not necessary to have better clustering information than other components. Some cautious are needed when using principal components in clustering applications.

Similar to Song, Nicolae, & Song (2010), in Table 13, we report the estimates of proportion based on the first principal component. Noting that the true proportion is $1/3$, we can see that the MPHD and the Symmetrization estimators perform better than the other two estimators.

Example 2 (Breast cancer data). Next, we illustrate the application of the new estimation procedure to multiple hypothesis testing using the breast cancer data from Hedenfalk et al. (2001), who examined gene expressions in breast cancer tissues from women who were carriers of the hereditary BRCA1 or BRCA2 gene mutations, predisposing to breast cancer. The breast cancer data were downloaded from “http://research.nhgri.nih.gov/microarray/NEJM_Supplement/" and contains gene expression ratios derived from the fluorescent intensity (proportional to the gene expression level) from a tumour sample divided by the fluorescent intensity from a common reference sample (MCF-10A cell line). The ratios were normalized (or calibrated) such that the majority of the gene expression ratios from a pre-selected internal control gene set was around 1.0, but no log-transformation was used. The data set consists of 3,226 genes on $n_1 = 7$ BRCA1 arrays and $n_2 = 8$ BRCA2 arrays. If any gene had one or more measurement exceeding 20, then this gene was eliminated (Storey & Tibshirani, 2003). This left 3,170 genes. The p-values were calculated based on permutation tests (Storey & Tibshirani, 2003). We then transform the p-values via the probit transformation to z-score, given by $z_i = \Phi^{-1}(1 - p_i)$ (McLachlan & Wockner, 2010). Figure 6 displays the fitted densities, and Table 14 lists the parameter estimates of the four

TABLE 12: Bias (MSE) of point estimates for model (3), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1,000$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	−0.083(0.015)	**0.049(0.003)**	0.291(0.085)	−0.211(0.051)
	$\sigma : 1$	0.099(0.026)	**−0.128(0.022)**	0.178(0.033)	**−0.096(0.050)**
	$\mu : 1.5$	**−0.116(0.039)**	0.706(0.515)	1.068(1.162)	−0.258(0.085)
II	$\pi : 0.3$	**−0.012(0.001)**	−0.042(0.002)	0.092(0.009)	−0.05(0.003)
	$\sigma : 1$	**0.025(0.003)**	−0.031(0.007)	0.422(0.189)	−0.199(0.045)
	$\mu : 3$	**−0.008(0.002)**	0.299(0.099)	0.537(0.297)	−0.047(0.005)
III	$\pi : 0.3$	**−0.021(0.001)**	−0.053(0.003)	**0.004(0.001)**	−0.042(0.002)
	$\sigma : 1$	**−0.040(0.004)**	**−0.033(0.006)**	0.185(0.050)	−0.194(0.042)
	$\mu : 3$	**−0.004(0.001)**	0.208(0.049)	0.302(0.099)	**−0.02(0.001)**
IV	$\pi : 0.7$	**−0.017(0.001)**	−0.079(0.008)	0.110(0.012)	−0.059(0.004)
	$\sigma : 2$	**−0.019(0.004)**	−0.045(0.013)	0.178(0.034)	−0.187(0.042)
	$\mu : 3$	0.094(0.020)	0.493(0.324)	1.386(2.005)	**0.024(0.012)**
V	$\pi : 0.85$	**−0.019(0.001)**	−0.081(0.008)	0.053(0.003)	−0.070(0.008)
	$\sigma : 2$	**0.013(0.004)**	**−0.008(0.007)**	0.083(0.009)	−0.167(0.034)
	$\mu : 3$	**0.193(0.064)**	0.909(1.093)	2.559(6.866)	**0.038(0.068)**

methods discussed in the article. MPHD estimator shows that among the 3170 genes examined, around 29% genes are differentially expressed between those tumour types, which is close to the 33% from Storey & Tibshirani (2003) and 32.5% from Langaas, Lindqvist, & Ferkingstad (2005).

Let

$$\hat{\tau}_0(z_i) = \hat{\pi}\phi_{\hat{\sigma}}(z_i)/[\hat{\pi}\phi_{\hat{\sigma}}(z_i) + (1 - \hat{\pi})\hat{f}(z_i - \hat{\mu})]$$

be the classification probability that the ith gene is not differentially expressed. Then we select all genes with $\hat{\tau}_0(z_i) \leq c$ to be differentially expressed. The threshold c can be selected by controlling the false discovery rate (FDR, Benjamini & Hochberg, 1995). Based on McLachlan, Bean, & Jones (2006), the FDR can be estimated by

$$\widehat{FDR} = \frac{1}{N_r}\sum_i \hat{\tau}_0(z_i) I_{[0,c_0]}\hat{\tau}_0(z_i),$$

where $N_r = \sum_i I_{[0,c_0]}\hat{\tau}_0(z_i)$ is the total number of found differentially expressed genes and $I_A(x)$ is the indicator function, which is one if $x \in A$ and is zero otherwise. Table 15 reports the number of selected differentially expressed genes (N_r) and the estimated false discovery rate (FDR) for different threshold c values based on MPHD estimate. For comparison, we also include the results of McLachlan & Wockner (2010), which assumes a two-component mixture of heterogeneous normals (MLE) for z_is.

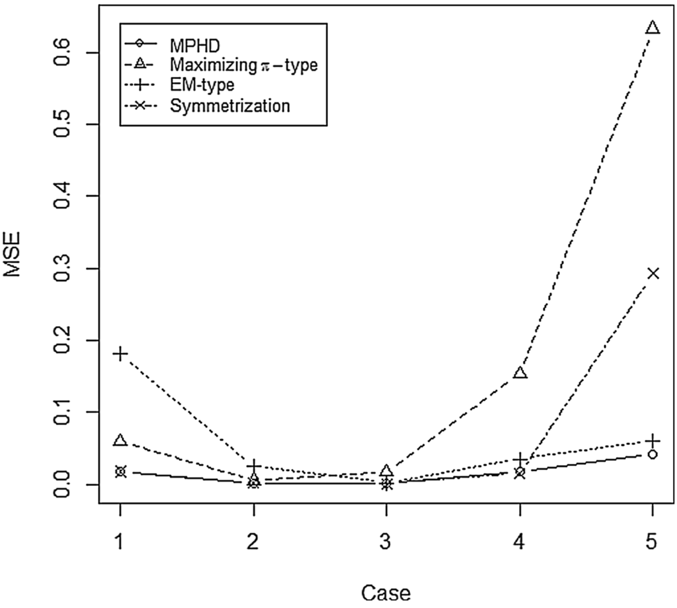

FIGURE 2: MSE of point estimates of μ of model (3), over 200 repetitions with $n = 1,000$.

5. DISCUSSION

In this paper, we proposed a MPHD estimator for a class of semiparametric mixture models and investigated its existence, consistency and asymptotic normality. Simulation study shows that the MPHD estimator outperforms existing estimators when data are under contamination, while it performs competitively to other estimators when there is no contamination.

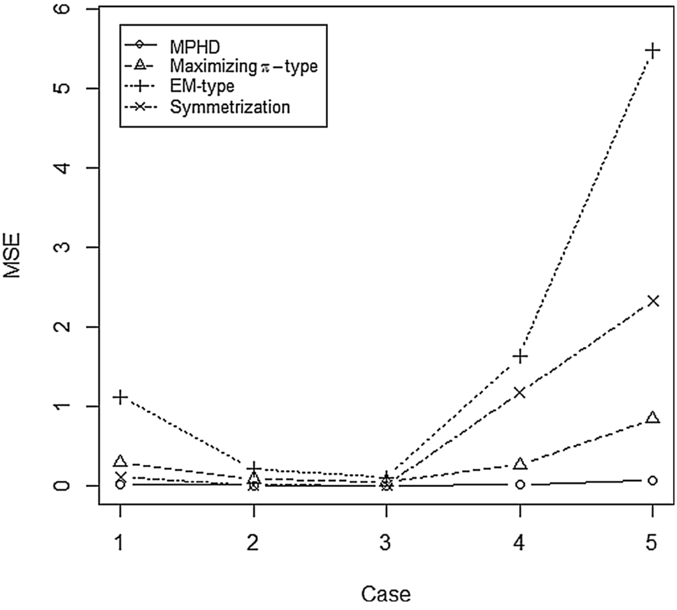

FIGURE 3: MSE of point estimates of μ of model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1,000$.

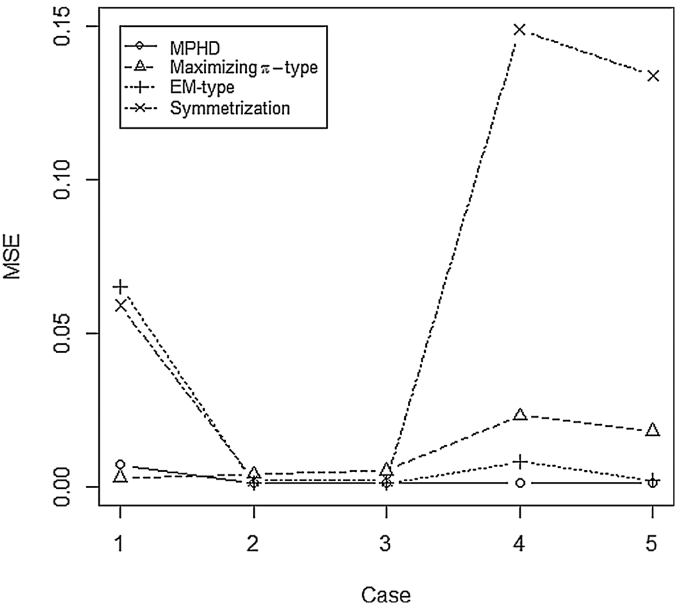

FIGURE 4: MSE of point estimates of π of model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1,000$.

We indicated two fields of application of the model. The first is microarray data analysis, which is the initial motivation of introducing model (2) (see Bordes, Delmas, & Vandekerkhove, 2006). The second is sequential clustering algorithm, which is the initial motivation of introducing model (3) (see Song, Nicolae, & Song, 2010). Two real data applications are also provided to illustrate the effectiveness of the proposed methodology.

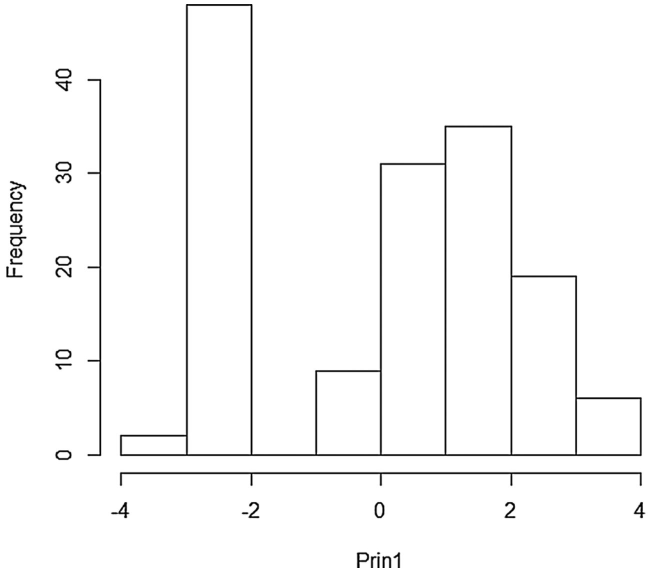

FIGURE 5: Histogram of the first principal component in the Iris data.

TABLE 13: Estimates of first principal component in Iris data.

Variable	True value	MPHD	Maximizing π-type	EM-type	Symmetrization
π	0.3000	0.3195	0.3986	0.2896	0.3266
σ	0.2208	0.2457	4.0000	0.1629	0.2055
μ	3.9469	3.9526	2.6240	3.6979	3.9077

In this article, we only considered the asymptotic results for model (2), since its identifiability property has been established by Bordes, Delmas, & Vandekerkhove (2006). When the first component of the general model (1) has normal distribution, empirical studies demonstrated the success of proposed MPHD estimator. We conjecture that the asymptotic results of MPHD also apply to the more general model (1) when it is identifiable. However, it requires further research to find sufficient conditions for the identifiability of model (1). In addition, more work remains to be done on the application of MPHD estimation in regression settings such as mixture of regression models.

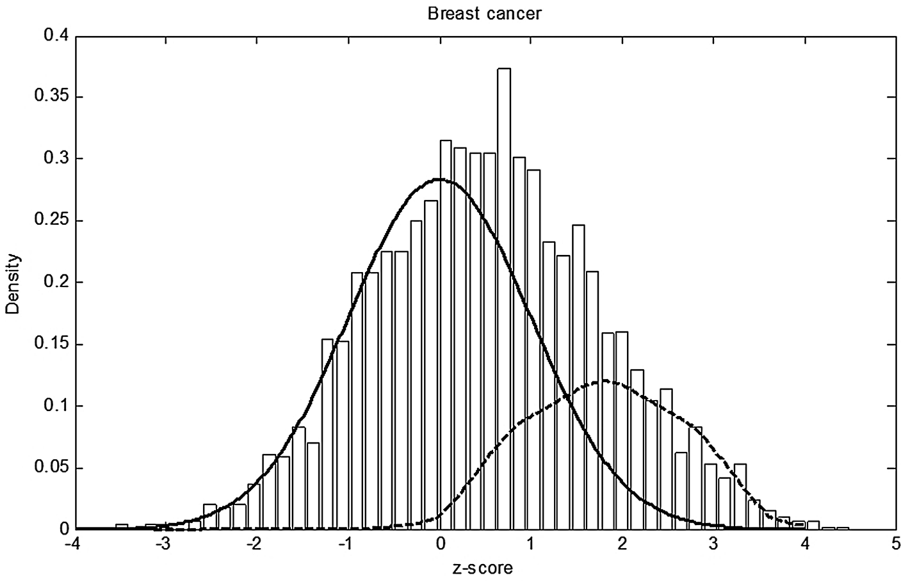

FIGURE 6: Breast cancer data: plot of fitted two-component mixture model with theoretical $N(0, 1)$ null and non-null component (weighted respectively by $\hat{\pi}$ and $(1 - \hat{\pi})$) imposed on histogram of z-score.

TABLE 14: Parameter estimates for the breast cancer data.

Variable	MPHD	Maximizing π-type	EM-type	Symmetrization
π	0.7109	0.6456	0.8365	0.5027
σ	1.0272	1	1.1441	1.0773
μ	1.8027	1.6756	1.9366	1.0765

TABLE 15: Estimated FDR for various levels of the threshold c applied to the posterior probability of nondifferentially expression for the breast cancer data.

	MLE		MPHD	
c	N_r	$\widehat{\mathrm{FDR}}$	N_r	$\widehat{\mathrm{FDR}}$
0.1	143	0.06	179	0.052
0.2	338	0.11	320	0.093
0.3	539	0.16	477	0.144
0.4	743	0.21	624	0.193
0.5	976	0.27	780	0.244

APPENDIX

The proofs of Theorems 1, 2 and 3 are presented in this section.

Proof of Theorem 1. The method of proof is similar to that of Theorem 2.1 of Beran (1977). (1) Let $d(t) = \left\| h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})} - h^{1/2}_{\boldsymbol{\theta},f} \right\|$. For any sequence $\{t_n : t_n \in \Theta, t_n \to t \text{ as } n \to \infty\}$,

$$
\begin{aligned}
|d^2(t_n) - d^2(t)| &= \left| \int (h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) - h^{1/2}_{\boldsymbol{\theta},f}(x))^2 \mathrm{d}x - \int (h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x) - h^{1/2}_{\boldsymbol{\theta},f}(x))^2 \mathrm{d}x \right| \\
&= 2 \left| \int (h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x)) h^{1/2}_{\boldsymbol{\theta},f}(x) \mathrm{d}x \right| \\
&\le 2 \left\| h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})} - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})} \right\| .
\end{aligned}
$$

Since $\int h_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x)\mathrm{d}x = \int h_{t,f(t,h_{\boldsymbol{\theta},f})}(x)\mathrm{d}x = 1$, we have

$$
\begin{aligned}
\left\| h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})} - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})} \right\|^2 &= \int \left[h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x) \right]^2 \mathrm{d}x \\
&\le \int \left| h_{t,f(t,h_{\boldsymbol{\theta},f})}(x) - h_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) \right| \mathrm{d}x \\
&= 2 \int \left[h_{t,f(t,h_{\boldsymbol{\theta},f})}(x) - h_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) \right]^+ \mathrm{d}x.
\end{aligned}
$$

Also, $[h_{t,f(t,h_{\boldsymbol{\theta},f})}(x) - h_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x)]^+ \le h_{t,f(t,h_{\boldsymbol{\theta},f})}(x)$, and $h_{t,f(t,h_{\boldsymbol{\theta},f})}(x)$ is continuous in t for every x. Thus, by the Dominated Convergence Theorem, $\|h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})} - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}\| \to 0$ as $n \to \infty$. So, $d(t_n) \to d(t)$ as $n \to \infty$, that is, d is continuous on Θ and achieves a minimum for $t \in \Theta$.

(2) By assumption, $h_{\boldsymbol{\theta},f}$ is identifiable. Immediately, we have $T(h_{\boldsymbol{\theta},f}) = \boldsymbol{\theta}$ uniquely.

(3) Let $d_n(t) = \|h^{1/2}_{t,f(t,h_n)} - h_n^{1/2}\|$ and $d(t) = \|h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})} - h^{1/2}_{\boldsymbol{\theta},f}\|$. By Minkowski's inequality,

$$\begin{aligned}|d_n(t) - d(t)| &= \left| \left[\int (h^{1/2}_{t,f(t,h_n)}(x) - h_n^{1/2}(x))^2 \mathrm{d}x \right]^{1/2} - \left[\int (h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x) - h^{1/2}_{\boldsymbol{\theta},f}(x))^2 \mathrm{d}x \right]^{1/2} \right| \\ &\le \left\{ \int \left[h^{1/2}_{t,f(t,h_n)}(x) - h_n^{1/2}(x) - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x) + h^{1/2}_{\boldsymbol{\theta},f}(x) \right]^2 \mathrm{d}x \right\}^{1/2} \\ &\le \left\{ 2 \int \left[h^{1/2}_{t,f(t,h_n)}(x) - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x) \right]^2 \mathrm{d}x + 2 \int \left[h_n^{1/2}(x) - h^{1/2}_{\boldsymbol{\theta},f}(x) \right]^2 \mathrm{d}x \right\}^{1/2}\end{aligned}$$

Consequently,

$$\begin{aligned}\sup_{t\in\Theta} |d_n(t) - d(t)| \le \Bigg\{ & 2 \sup_{t\in\Theta} \int \left[h^{1/2}_{t,f(t,h_n)}(x) - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x) \right]^2 \mathrm{d}x \\ & + 2 \int \left[h_n^{1/2}(x) - h^{1/2}_{\boldsymbol{\theta},f}(x) \right]^2 \mathrm{d}x \Bigg\}^{1/2}, \end{aligned} \tag{10}$$

and the right-hand side of (10) goes to zero as $n \to \infty$ by assumptions. Then with $\boldsymbol{\theta}_0 = T(h_{\boldsymbol{\theta},f})$ and $\boldsymbol{\theta}_n = T(h_n)$, we have $d_n(\boldsymbol{\theta}_0) \to d(\boldsymbol{\theta}_0)$ and $d_n(\boldsymbol{\theta}_n) - d(\boldsymbol{\theta}_n) \to 0$ as $n \to \infty$.

If $\boldsymbol{\theta}_n \nrightarrow \boldsymbol{\theta}_0$, then there exists a subsequence $\{\boldsymbol{\theta}_m\} \subseteq \{\boldsymbol{\theta}_n\}$ such that $\boldsymbol{\theta}_m \to \boldsymbol{\theta}' \neq \boldsymbol{\theta}_0$, implying that $\boldsymbol{\theta}' \in \Theta$ and $d(\boldsymbol{\theta}_m) \to d(\boldsymbol{\theta}')$ by the continuity of d. From the above result, we have $d_m(\boldsymbol{\theta}_m) - d_m(\boldsymbol{\theta}_0) \to d(\boldsymbol{\theta}') - d(\boldsymbol{\theta}_0)$. By the definition of $\boldsymbol{\theta}_m$, $d_m(\boldsymbol{\theta}_m) - d_m(\boldsymbol{\theta}_0) \le 0$, and therefore, $d(\boldsymbol{\theta}') - d(\boldsymbol{\theta}_0) \le 0$. However, by the definition of $\boldsymbol{\theta}_0$ and the uniqueness of it, $d(\boldsymbol{\theta}') > d(\boldsymbol{\theta}_0)$. This is a contradiction, and therefore $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$. ■

Proof of Theorem 2. Let H_n denote the empirical cdf of $X_1, X_2, ..., X_n$, which are assumed i.i.d. with density $h_{\boldsymbol{\theta},f}$ and cdf H. Let

$$\tilde{h}_n(x) = (c_n s_n)^{-1} \int K((c_n s_n)^{-1}(x - y)) \mathrm{d}H(y).$$

Let $B_n(x) = n^{1/2}[H_n(x) - H(x)]$, then

$$\begin{aligned}\sup_x |\hat{h}_n(x) - \tilde{h}_n(x)| &= \sup_x n^{-1/2}(c_n s_n)^{-1} \left| \int K((c_n s_n)^{-1}(x - y)) \mathrm{d}B_n(y) \right| \\ &\le n^{-1/2}(c_n s_n)^{-1} \sup_x |B_n(x)| \int |K'(x)| \mathrm{d}x \xrightarrow{p} 0. \end{aligned} \tag{11}$$

Suppose $[a, b]$ is an interval that contains the support of K , then

$$\begin{aligned}\sup_x |\tilde{h}_n(x) - h_{\boldsymbol{\theta},f}(x)| &= \sup_x \left| \int K(t) h_{\boldsymbol{\theta},f}(x - c_n s_n t) \mathrm{d}t - h_{\boldsymbol{\theta},f}(x) \right| \\ &= \sup_x \left| h_{\boldsymbol{\theta},f}(x - c_n s_n \xi) \int K(t) \mathrm{d}t - h_{\boldsymbol{\theta},f}(x) \right|, \text{ with } \xi \in [a, b] \\ &\le \sup_x \sup_{t\in[a,b]} |h_{\boldsymbol{\theta},f}(x - c_n s_n t) - h_{\boldsymbol{\theta},f}(x)| \xrightarrow{p} 0 \end{aligned} \tag{12}$$

From (11) and (12), we have

$$\sup_x |\hat{h}_n(x) - h_{\boldsymbol{\theta},f}(x)| \xrightarrow{p} 0.$$

From an argument similar to the proof of Theorem 1, $\|\hat{h}_n^{1/2}(x) - h_{\boldsymbol{\theta},f}^{1/2}(x)\| \xrightarrow{p} 0$ and $\sup_{t\in\Theta} \|h_{t,f(t,\hat{h}_n)} - h_{t,f(t,h_{\boldsymbol{\theta},f})}\| \to 0$ as $n \to \infty$. By Theorem 1, $T(\hat{h}_n) \xrightarrow{p} T(h_{\boldsymbol{\theta},f})$ as $n \to \infty$. ■

Proof of Theorem 3. Let

$$D(\boldsymbol{\theta}, g) = \int \dot{s}_{\boldsymbol{\theta},g}(x) g^{1/2}(x) \mathrm{d}x =< \dot{s}_{\boldsymbol{\theta},g}, g^{1/2} >,$$

and it follows that $D(T(h_{\boldsymbol{\theta},f}), h_{\boldsymbol{\theta},f}) = 0$, $D(T(\hat{h}_n), \hat{h}_n) = 0$, and therefore

$$\begin{aligned} 0 &= D(T(\hat{h}_n), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), h_{\boldsymbol{\theta},f}) \\ &= [D(T(\hat{h}_n), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), \hat{h}_n)] + [D(T(h_{\boldsymbol{\theta},f}), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), h_{\boldsymbol{\theta},f})]. \end{aligned}$$

Since the map $\boldsymbol{\theta} \mapsto s_{\boldsymbol{\theta},g}$ satisfies (8) and (9), $D(\boldsymbol{\theta}, g)$ is differentiable in $\boldsymbol{\theta}$ with derivative

$$\dot{D}(\boldsymbol{\theta}, g) =< \ddot{s}_{\boldsymbol{\theta},g}, g^{1/2} >$$

that is continuous in $\boldsymbol{\theta}$. Then,

$$D(T(\hat{h}_n), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), \hat{h}_n) = (T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f}))\dot{D}(T(h_{\boldsymbol{\theta},f}), \hat{h}_n) + o_p(T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f})).$$

With $\boldsymbol{\theta} = T(h_{\boldsymbol{\theta},f})$,

$$\begin{aligned} D(T(h_{\boldsymbol{\theta},f}), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), h_{\boldsymbol{\theta},f}) &= < \dot{s}_{\boldsymbol{\theta},\hat{h}_n}, \hat{h}_n^{1/2} > - < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} > \\ &= 2 < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > + < \dot{s}_{\boldsymbol{\theta},\hat{h}_n} - \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} \\ &\quad - h_{\boldsymbol{\theta},f}^{1/2} > + < \dot{s}_{\boldsymbol{\theta},\hat{h}_n}, h_{\boldsymbol{\theta},f}^{1/2} > - < \hat{h}_n^{1/2}, \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}} > \\ &= 2 < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > +[< \dot{s}_{\boldsymbol{\theta},\hat{h}_n}, h_{\boldsymbol{\theta},f}^{1/2} > \\ &\quad - < \hat{h}_n^{1/2}, \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}} >] + O(\|\dot{s}_{\boldsymbol{\theta},\hat{h}_n} \\ &\quad - \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}\| \cdot \|\hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2}\|) \\ &= 2 < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > + o_p(\|\hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2}\|). \end{aligned}$$

Applying the algebraic identity

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

we have that

$$\begin{aligned} n^{1/2} < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > &= n^{1/2}\int \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(x)\frac{\hat{h}_n(x) - h_{\boldsymbol{\theta},f}(x)}{2h_{\boldsymbol{\theta},f}^{1/2}(x)}\mathrm{d}x + R_n \\ &= n^{1/2}\int \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(x)\frac{\hat{h}_n(x)}{2h_{\boldsymbol{\theta},f}^{1/2}(x)}\mathrm{d}x + R_n \\ &= n^{1/2}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{\dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(X_i)}{2h_{\boldsymbol{\theta},f}^{1/2}(X_i)} + o_p(1) + R_n \end{aligned}$$

with $|R_n| \leq n^{1/2}\int \frac{|\dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(x)|}{2h_{\boldsymbol{\theta},f}^{3/2}(x)}[\hat{h}_n(x) - h_{\boldsymbol{\theta},f}(x)]^2\mathrm{d}x \xrightarrow{p} 0$. Since $< \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} >$ is assumed to be invertible, then

$$T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f}) = -\big[< \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} >^{-1} + o_p(1)\big]\frac{1}{n}\sum_{i=1}^{n}\frac{\dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(X_i)}{h_{\boldsymbol{\theta},f}^{1/2}(X_i)} + o_p(n^{-1/2})$$

and therefore, the asymptotic distribution of $n^{1/2}(T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f}))$ is $N(0, \Sigma)$ with variance matrix Σ defined by

$$\Sigma = < \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} >^{-1} < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \dot{s}^{T}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}} > < \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} >^{-1} .$$

■

ACKNOWLEDGEMENTS

The authors would like to thank the editors and two anonymous referees for their valuable comments and suggestions, which greatly improved this article.

BIBLIOGRAPHY

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate—A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, 5, 445–463.

Bordes, L., Delmas, C., & Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where one-component is known. *Scandinavian Journal of Statistics*, 33, 733–752.

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, 2010, 2916–2957.

Donoho, D. L. & Liu, D. C. (1988). The automatic robustness of minimum distance functionals. *Annals of Statistics*, 16(2), 552–586.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenic*, 7, Part II, 179–188.

García-Escudero, L. A., Gordaliza, A., & Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2), 434–449.

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3), 1324–1345.

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2), 89–109.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, B., Borg, A., & Trent, J. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344, 539–548.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1), 112–147.

Langaas, M., Lindqvist, B. H., & Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 555–572.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance estimation and related methods. *Annals of Statistics*, 22, 1081–1114.

McLachlan, G. J., Bean, R. W., & Jones, L. B. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22, 1608–1615.

McLachlan, G. J. & Wockner, L. (2010). Use of mixture models in multiple hypothesis testing with applications in bioinformatics. In *Classification as a Tool for Research: Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft fr Klassifikation*, Locarek-Junge, H. & Weihs, C., editors. Springer-Verlag, Heidelberg, pp. 177–184.

Punzo, A. & McNicholas, P. D. (2013). Outlier detection via parsimonious mixtures of contaminated Gaussian distributions. ArXiv:1305.4669.

Song, J. & Nicolae, D. L. (2009). A sequential clustering algorithm with applications to gene expression data. *Journal of the Korean Statistical Society*, 38, 175–184.

Song, S., Nicolae, D. L., & Song, J. (2010). Estimating the mixing proportion in a semiparametric mixture model. *Computational Statistics and Data Analysis*, 54, 2276–2283.

Storey, J. D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 3889–3894.

Swanepoel, J. W. H. (1999). The limiting behavior of a modified maximal symmetric 2s-spacing with applications. *Annals of Statistics*, 27, 24–35.

Wu, J. & Karunamuni, R. J. (2014). Profile Hellinger distance estimation. *Statistics Invited for revision.*

Received 29 June 2013
Accepted 10 February 2014