MINIMUM HELLINGER DISTANCE ESTIMATION IN A SEMIPARAMETRIC MIXTURE MODEL

by

SIJIA XIANG

B.S., Zhejiang Normal University, China, 2010

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2012

Approved by:

Major Professor
Weixin Yao

Copyright

Abstract

In this report, we introduce the minimum Hellinger distance (MHD) estimation method and review its history. We examine the use of Hellinger distance to obtain a new efficient and robust estimator for a class of semiparametric mixture models where one component has known distribution while the other component and the mixing proportion are unknown. Such semiparametric mixture models have been used in biology and the sequential clustering algorithm. Our new estimate is based on the MHD, which has been shown to have good efficiency and robustness properties. We use simulation studies to illustrate the finite sample performance of the proposed estimate and compare it to some other existing approaches. Our empirical studies demonstrate that the proposed minimum Hellinger distance estimator (MHDE) works at least as well as some existing estimators for most of the examples considered and outperforms the existing estimators when the data are under contamination. A real data set application is also provided to illustrate the effectiveness of our proposed methodology.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my appreciation to Dr. Weixin Yao, my major professor, for all his knowledge, guidance and suggestions. I would also like to thank Dr. Nora Bello and Dr. Weixing Song for their willingness to serve on my committee and for their valuable insight.

I would like to thank all my friends for their help and support. I would also like to thank everyone in the department for their kindness. It is my pleasure to study in this department.

Finally, I would also like to thank my family for their endless love, support, understanding and encouragement. Thanks to my parents, who have always supported whatever I wanted to do in my life. Thank you also to my grandparents, who always check to see how I am doing and offer their support.

Chapter 1

Introduction

1.1 Background

The statistical problem which motivates the minimum Hellinger distance (MHD) estimation can be described as follows. Random variables $X_1, X_2, ..., X_n$ are observed, and we postulate that the $\{X_i\}$ are independent and identically distributed (i.i.d) with density function f. If f belongs to a specified parametric family $\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ then θ may be estimated using well-known likelihood procedures. However, we recognize that lack of information, data contamination, and other factors beyond our control make it virtually certain that the model is not strictly correct. Also, assuming f belongs strictly to $\{f_\theta\}$ ignores the possibility of departures from the parametric model.

In practice, for many parametric family of interest, the maximum likelihood estimator (MLE) of θ has full asymptotic efficiency among regular estimators. In general, however, it has long been known that MLE does not possess the property of stability under small perturbations in the underlying model. As a result, robust estimator, like M-estimator has been developed, but many of them achieve the robustness at some cost in first-order efficiency. This is not true for minimum Hellinger distance estimator (MHDE), first introduced by Beran (1977). In fact, Lindsay (1994) has shown that MLE and MHDE are members of a large class of efficient estimators with various second-order efficiency properties, and MHDE has been shown to have excellent robustness properties in parametric models such as

resistance to outliers and robustness with respect to model misspecification (Beran (1977)).

By using the minimum Hellinger distance approach, we assume that f is either in $\{f_\theta\}$ or close to a member of $\{f_\theta\}$, and the MHDE of θ is defined as the value of the parameter that minimize the Hellinger distance between the parametric model and a nonparametric density estimator of f. That is, if we use $\hat{\theta}$ to denote the MHDE, then $\hat{\theta}$ is defined by

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \left\| f_\theta^{1/2} - f_n^{1/2} \right\|, \tag{1.1}$$

where $\|f_1 - f_2\| = (\int [f_1(x) - f_2(x)]^2 \, dx)^{1/2}$ denotes the L_2-norm and f_n is a nonparametric density estimator of f, such as the kernel density estimator, based on the observations $X_1, X_2, ..., X_n$.

From the definition, it is interesting to note that the MHDE $\hat{\theta}$ is related heuristically to the maximum likelihood estimator of θ. When n is sufficiently large, the MLE should be close to θ, the true parameter, and the nonparametric density estimator f_n should be close to f_θ. Finding the MLE amounts to maximizing the integral $\int \log f_\theta(x) dF_n(x)$ over $\theta \in \Theta$, where F_n is the empirical distribution function of the data. Note that

$$\begin{aligned}
\int f_n(x) \log\left[\frac{f_\theta(x)}{f_n(x)}\right] dx &= 2\int f_n(x) \log\left[1 + \left(\frac{f_\theta^{1/2}(x)}{f_n^{1/2}(x)} - 1\right)\right] dx \\
&\approx 2\int f_n(x) \left[\left(\frac{f_\theta^{1/2}(x)}{f_n^{1/2}(x)} - 1\right) - \frac{1}{2}\left(\frac{f_\theta^{1/2}(x)}{f_n^{1/2}(x)} - 1\right)^2\right] dx \\
&= -2\left\| f_\theta^{1/2} - f_n^{1/2} \right\|^2
\end{aligned}$$

thus, it is not unreasonable to expect that the MHDE $\hat{\theta}$ is asymptotically efficient under f_θ.

On the other hand, since

$$\left\| f_\theta^{1/2} - f_n^{1/2} \right\|^2 \le \int \|f_\theta(x) - f_n(x)\| dx \le 2\left\| f_\theta^{1/2} - f_n^{1/2} \right\|,$$

the topology induced on the space of probability measures by the Hellinger metric is the same as that induced by the L_1-norm. It is known that the L_1-norm induces a robust topology, therefore, the MHDE could be expected to be robust as well.

1.2 Development of Minimum Hellinger Distance (MHD) estimation

1.2.1 Parametric models

Beran (1977) defined and studied the minimum Hellinger distance estimator for parametric model, and has shown MHDE to have excellent robustness properties in parametric models such as resistance to outliers and robustness with respect to model misspecification.

Associated with the MHDE $\hat{\theta}$, a functional T was defined. The continuity and differentiability of functional and the conditions for the existence of MHDE was studied in the following theorem by Beran.

Let $\mathscr{F}$ denote the set of all densities with respect to Lebesgue measure on the real line. The functional T is defined on $\mathscr{F}$ such that for every $g \in \mathscr{F}$,

$$\left\| f_{T(g)}^{1/2} - g^{1/2} \right\| = \min_{\theta \in \Theta} \left\| f_{\theta}^{1/2} - g^{1/2} \right\|, \tag{1.2}$$

and the MHDE $\hat{\theta}$ is defined as $T(f_n)$.

Theorem 1.2.1. *(Beran(1977)) Suppose that Θ is a compact subset of $\mathbb{R}^p$, $\theta_1 \neq \theta_2$ implies $f_{\theta_1} \neq f_{\theta_2}$ on a set of positive Lebesgue measure, and for almost every x, $f_\theta(x)$ is continuous in θ. Then*

(i) For every $g \in \mathscr{F}$, there exists $T(g) \in \Theta$ satisfying (1.2).

(ii) If $T(g)$ is unique, the functional T is continuous at g in the Hellinger topology.

(iii) $T(f_\theta) = \theta$ uniquely for every $\theta \in \Theta$.

For notational convenience, let $s_t = f_t^{1/2}$. With further assumptions on s_t, the functional T becomes differentiable, a property that is fundamental for further developments. For specified $t \in \Theta \subseteq \mathbb{R}^p$, we will typically assume that there exist a $p \times 1$ vector $\dot{s}(x)$ with components in L_2 and a $p \times p$ matrix $\ddot{s}(x)$ with components in L_2 such that for every $p \times 1$

real vector e of unit Euclidean length and for every scalar α in a neighborhood of zeros,

$$s_{t+\alpha e}(x) = s_t(x) + \alpha e^T \dot{s}_t(x) + \alpha e^T u_\alpha(x) \tag{1.3}$$

$$\dot{s}_{t+\alpha e}(x) = \dot{s}_t(x) + \alpha \ddot{s}_t(x)e + \alpha v_\alpha(x)e \tag{1.4}$$

where $u_\alpha(x)$ is $p \times 1$, $v_\alpha(x)$ is $p \times p$, and the components of u_α and of v_α individually tend to zero in L_2 as $\alpha \to 0$.

Theorem 1.2.2. *(Beran(1977)) Suppose that (1.3) and (1.4) hold for every $t \in int(\Theta)$, $T(g)$ exists, is unique and lies in $int(\theta)$, $\int \ddot{s}_{T(g)} g^{1/2}(x)dx$ is a nonsingular matrix, and the functional T is continuous at g in the Hellinger topology. Then for every sequence of densities g_n converging to g in the Hellinger metric,*

$$\begin{aligned} T(g_n) = T(g) &+ \int \rho_g(x)[g_n^{1/2}(x) - g^{1/2}(x)]dx \\ &+ a_n \int \dot{x}_{T(g)}(x)[g_n^{1/2}(x) - g^{1/2}(x)dx, \end{aligned} \tag{1.5}$$

where

$$\rho_g(x) = -\frac{\dot{s}_{T(g)}(x)}{\int \ddot{s}_{T(g)}(x) g^{1/2}(x)dx}$$

and a_n is a real $p \times p$ matrix which tends to zero as $n \to \infty$. In particular, for $g = f_\theta$,

$$\begin{aligned} \rho_{f_\theta}(x) &= -\frac{\dot{s}_\theta(x)}{\int \ddot{s}_\theta(x) s_\theta(x)dx} \\ &= -\frac{\dot{s}_\theta(x)}{\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)dx}. \end{aligned}$$

Next the large sample behavior of $T(f_n)$ is examined, where f_n is a kernel density estimator

$$f_n(x) = \frac{1}{nh_nS_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_nS_n}\right), \tag{1.6}$$

where K is a smooth density function, bandwidth h_n are positive constants such that $h_n \to 0$ as $n \to \infty$, and $S_n = S_n(X_1, ..., X_n)$ is a robust scale estimator. $\{X_i\}$ are i.i.d random variables with density f.

With further assumptions on the bandwidths and kernels, the consistency of the MHDE $\hat{\theta}$ follows from the continuity of functionals in the Hellinger topology.

Theorem 1.2.3. *(Beran(1977)) Suppose*

(i) K *is absolutely continuous and has compact support;* K' *is bounded.*

(ii) f *is uniformly continuous.*

(iii) $\lim_{n\to\infty} h_n = 0, \lim_{n\to\infty} n^{1/2}h_n = \infty.$

(iv) As $n \to \infty$, $s_n \xrightarrow{p} s$, *a positive finite constant depending on* f.

Then $\left\| f_n^{1/2} - f^{1/2} \right\| \xrightarrow{p} 0$ *as* $n \to \infty$. *If* T *is a functional continuous at* f *in the Hellinger metric, then* $T(f_n) \xrightarrow{p} T(f)$

In the next theorem, Beran showed that under stronger assumptions, $T(f_n)$ has an asymptotically normal distribution about $T(f)$.

Theorem 1.2.4. *(Beran(1977)) Suppose*

(i) K *is symmetric about 0 and has compact support.*

(ii) K *is twice absolutely continuous;* K'' *is bounded.*

(iii) T *satisfy (1.5) and* ρ_g *has compact support* K *on which it is continuous.*

(iv) $f > 0$ *on* K; f *is twice absolutely continuous and* f'' *is bounded.*

(v) $\lim_{n\to\infty} n^{1/2}h_n = \infty, \lim_{n\to\infty} n^{1/2}h_n^2 = 0.$

(vi) There exists a positive finite constant s *depending on* f *such that* $n^{1/2}(s_n - s)$ *is bounded in probability.*

Then

$$\sqrt{n}\left[T(f_n) - T(f)\right] \xrightarrow{D} N\left(0, \frac{\int \rho_f(x)\rho_f^T(x)dx}{4}\right).$$

In particular, if $f = f_\theta$, *then*

$$\sqrt{n}\left[T(f_n) - \theta\right] \xrightarrow{D} N\left(0, \frac{1}{4\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)dx}\right).$$

To appreciate the robustness of MHDE upon a Hellinger metric model of data contamination, theoretical results showed that the MHDE was minimax robust in a small Hellinger metric neighborhood of the given family, and the local minimax robustness at f_θ entailed asymptotic efficiency at f_θ, but not conversely. On the other hand, in order to examine the behavior of T under a mixture model for gross errors, the α-influence curve was introduced.

Let δ_z denote the uniform density on the interval $(z-\varepsilon, z+\varepsilon)$, where $\varepsilon > 0$ is very small, and let $f_{\theta,\alpha,z} = (1-\alpha)f_\theta + \alpha\delta_z$ for $\theta \in \Theta$, $\alpha \in [0,1)$, and real z. Here, the density $f_{\theta,\alpha,z}$ models an experiment where independent observations distributed according to f_θ are mixed with approximately $100\alpha\%$ gross errors located near z. For every $\alpha \in (0,1)$, the difference quotient, named α-influence curve

$$IC_{t,\alpha}(z) = \frac{T(f_{\theta,\alpha,z}) - \theta}{\alpha}$$

is a bounded continuous function of z such that

$$\lim_{z\to\infty} \frac{T(f_{\theta,\alpha,z}) - \theta}{\alpha} = 0.$$

Hence, the functional T is robust at f_θ against $100\alpha\%$ contamination by gross errors at arbitrary real z, whether or not the influence function of T is irrelevant to the matter.

1.2.2 Mixture of two normals

Based on Beran (1977)'s work, Woodward et al. (1995) examined the MHDE in the case of estimation of the mixing proportion in the mixture of two normals, discussed the practical feasibility of employing the MHDE in this setting and examined empirically its robustness properties. Their results indicated that the MHDE obtained full efficiency at the true model while performing comparably with the minimum distance estimator based on Cramér-von Mises distance under the symmetric departures from component normality considered.

Finite Mixture Model has been a hot topic during the past years. The classic paper on mixture models is by the famous biometrician Pearson (1894), where he used a moment based method to fit a mixture of two heteroscedastic normal components in the paper. A few

years later, Charlier and Wicksell (1924) extended Pearson's work to the bivariate normal component case and Doetsch (1928) used it in the case of more than two univariate normal components.

The mixture of two normal components has density

$$f_\theta(x) = \frac{p}{\sqrt{2\pi}\sigma_1} exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\} + \frac{1-p}{\sqrt{2\pi}\sigma_2} exp\left\{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right\},$$

where $\boldsymbol{\theta} = (\mu_1, \sigma_1, \mu_2, \sigma_2, p)'$.

At the first step, they considered the case in which $f_\theta(x)$ is a mixture of known densities, which implies that $\theta = p$. Since the kernel density estimator is Hellinger consistent and the Hellinger metric on the probability distributions is equivalent to the Euclidean metric on the parameter space, implying Theorem 1.2.3, the MHDE $\hat{p}$ is consistent. Similarly, by implying Theorem 1.2.4, the MHDE $\hat{p}$ has an asymptotic normal distribution and is asymptotically fully efficient.

Next, they considered the case in which the five parameters $p, \mu_1, \sigma_1, \mu_2$ and σ_2 are all unknown, meaning that $\boldsymbol{\theta} = (p, \mu_1, \sigma_1, \mu_2, \sigma_2)'$.

Following Beran (1977), minimizing $\left\| f_\theta^{1/2} - f_n^{1/2} \right\|$ is equivalent to maximizing $\int f_\theta^{1/2} f_n^{1/2}$. However, due to convergence issue, Woodward et al. (1995) approximated this integral by the trapezoidal rule to obtain

$$\hat{I} = \Delta t_i \sum_{i=1}^{k} a_i \left(f_\theta^{1/2}(t_i) - f_n^{1/2}(t_i) \right)^2,$$

where $a_1 = a_k = 1/2$ and $a_i = 1$ for $i = 2, 3, ..., k-1$ for a partition $t_1, t_2, ..., t_k$ of $[a, b]$, a finite interval.

In order to examine the property of MHDE, a stimulation study was conducted to compare MHDE and MLE, and the results were based on Bias, MSE, and the relative

efficiencies

$$\widehat{Bias} = \frac{1}{n_s}\sum_{i=1}^{n_s}(\widehat{p}_i - p)$$

$$\widehat{MSE} = \frac{1}{n_s}\sum_{i=1}^{n_s}(\widehat{p}_i - p)^2$$

$$\widehat{E} = \frac{\widehat{MSE}(MLE)}{\widehat{MSE}(MHDE)}.$$

The study showed that the MHDE appeared to obtain full efficiency at the true model as evidenced by $\hat{E}$ near one in all cases. The probability plots indicated that the normality of the MHDE was very similar to that of the MLE. When checking the results for samples which were simulated as mixtures of $t(4)$ component, all of the $\hat{E}$'s were greater than one providing evidence that the MHDE was more robust to the departures from the assumption of normal components than was the MLE. Further study with the component of $t(2)$ showed that the more the mixed models departed from normality, the better the MHDE was.

1.2.3 Multivariate location and covariance

Tamura and Boos (1986) extended the research of Beran (1977) from univariate to multivariate estimation and added an important new robustness result. The idea of the breakdown point of an estimator originates from Hampel (1971) and may be interpreted as the smallest fraction of bad data that can cause an estimator to give an arbitrarily bad answer. Donoho (1982) has proposed the following definition of the breakdown point.

Let X be a given data set of size n, and let Y be a contaminating data set of size $m \leq n$. An estimator t is said to break down if, by appropriate choice of $Y_1, ..., Y_m$, the difference $t(X \cup Y) - t(X)$ can be made as large as desired. If m^* denotes the smallest number of contamination points for which t breaks down, then the breakdown point $\epsilon^*(t, X)$ of t at X is

$$\frac{m^*}{n + m^*}.$$

Thus, if an estimator t_1 has a larger breakdown point than an estimator t_2, then t_1 is more

robust than t_2, since it can handle a larger fraction of bad data.

Tamura and Boos (1986) showed that the asymptotic breakdown point of the MHDE for location and scatter was greater than or equal to 1/4 regardless of the data dimension, meaning that, roughly speaking, at least one quarter of the data could be badly damaged or arbitrarily changed without destroying the estimator.

In the paper, they mainly focused on parametric families within the class of elliptically symmetric distributions with density function of the form

$$f(x) \propto \frac{\Psi\{(x-\mu)'\Sigma^{-1}(x-\mu)\}}{|\Sigma|^{1/2}}$$

so that $\boldsymbol{\theta} = (\mu, \Sigma)$.

Choosing the nonparametric density estimator f_n properly, the multivariate MHDE's are independent of the coordinate system, that is, $\hat{\mu}$ is affine equivariant and $\hat{\Sigma}$ is affine covariant. Applying Theorem 1.2.3, strong consistency is easily obtained. The asymptotic normality is not so simple in multivariate case, but if the nonparametric density estimator f_n is a kernel estimator, then under some strict restrictions, the MHDE's have asymptotic normal distributions.

To measure the robustness of the MHDE, Beran (1977) introduced the α-influence curve. Unfortunately, however, the result is the consequence of assuming a compact parameter space and it appears that the $IC_{t,\alpha}(z)$ would have to be plotted for numerous values of α in each situation of interest in order to see how the MHDE handles contamination. Instead, Tamura and Boos (1986) gave a general bound on the amount of contamination that the MHDE could handle when estimating location and scatter, showing that the asymptotic breakdown point of the MHDE was bounded below by 1/4. This is more favorable compared to the M-estimator of Maronna (1976), which has a breakdown upper bound of $1/(k+1)$. Thus, for high-dimensional data, the MHDE should have better robustness properties than the M-estimators.

1.2.4 Count data

Simpson (1987) studied the MHDE in the context of discrete data, where the model was allowed to have countably infinite support. An improved breakdown bound of 1/2 was obtained at the model.

For count data, the most commonly used f_n is the empirical density function

$$f_n(x) = N_x/n, x = 0, 1, ...,$$

where N_x is the frequency of x among $X_1, X_2, ..., X_n$.

Since $f_n \geqslant 0$ and $\int f_n = 1$,

$$\left\| f_n^{1/2} - f_\theta^{1/2} \right\|^2 = 2 - 2 \int f_n^{1/2} f_\theta^{1/2}.$$

Then, by definition, the MHDE maximizes $\rho_{n,\theta} = \sum_{x=0}^{\infty} f_n^{1/2}(x) f_\theta^{1/2}(x)$, which yields the standardized estimation equation

$$\rho_{n,\theta}^{-1} = \sum_{x=0}^{\infty} f_n^{1/2}(x) f_\theta^{1/2}(x) l_\theta(x) = 0,$$

where $l_\theta(x)$ is the gradient of $log f_\theta(x)$.

Beran (1977) characterized the existence and the continuity of T in the continuous case for compact Θ. Simpson (1987) extended Beran's existence and continuity result and showed that the result also applied if Θ was embedded in a compact space $\bar{\Theta}$. After applying the smoothness conditions on the model, the asymptotic normality for a discrete distribution with countable support was derived under a readily verified condition on the model.

In order to appreciate the robustness, Simpson (1987) compared the breakdown properties of the MHDE and a sequential outlier screen. Finally an improved breakdown bound of 1/2 was obtained for the MHDE.

1.2.5 Poisson mixtures

Finite Poisson mixtures are used to describe data that are overdispersed and hence can't be fitted by a simple Poisson distribution. Based upon Simpson (1987), Karlis and Xekalaki

(1998) derived MHDE for finite Poisson mixtures, and proved it to be both efficient and robust. To facilitate computation, they provided an iterative algorithm.

For k-finite Poisson mixtures, the empirical density function is still the most commonly used $f_n(x)$, and

$$f_\theta(x) = \sum_{i=1}^{k} p_i \frac{e^{-\lambda_i}\lambda_i^x}{x!}, x = 0, 1, ...,$$

where $\boldsymbol{\theta} = (p_1, p_2, ..., p_{k-1}, \lambda_1, \lambda_2, ..., \lambda_k)$, $\lambda_i > 0$, $i = 1, 2, ..., k$ and $p_i \in (0, 1)$ for $i = 1, 2, ..., k$ with $\sum_{i=1}^{k} p_i = 1$.

To compare MHDE and MLE, Karlis and Xekalaki (1998) studied the estimation equations for both estimates. For parameter $\boldsymbol{\theta}$, the estimating equation for MLE is

$$\sum_{x=0}^{\infty} \frac{f_n(x)}{f_\theta(x)} \frac{\partial f_\theta(x)}{\partial \theta_i} = 0,$$

while the estimating equation for MHDE is

$$\sum_{x=0}^{\infty} \left[\frac{f_n(x)}{f_\theta(x)}\right]^{1/2} \frac{\partial f_\theta(x)}{\partial \theta_i} = 0.$$

If the model is well specified and the sample size is large, the square root of $f_n(x)/f_\theta(x)$ should be close to itself, and thus, we would expect MHDE and MLE to behave similarly. On the other hand, in the case of outliers, for values of x for which the ratio is large, the MHDE gives less weight to the estimation by taking the square root, and thus, not so sensitive to outliers.

Simulation study showed that, for contaminated models, MLE usually modeled the contamination with an additional component. Since mixture models were very often not appropriately specified, including the case where the number of components not being assigned prior to analysis, the MHDE was more reliable in such case.

1.2.6 Finite mixtures of Poisson regression models

Lu et al. (2003) extended the MHDE approach from the finite mixtures of Poisson distributions to the finite mixtures of Poisson regressions for count data.

Let $(y_i, t_i, \boldsymbol{x}_i), i = 1, ..., n$ denote observations, where y_i is the observation value of the ith response variable Y_i, t_i is a non-negative quantity representing the time or extent of exposure, and $\boldsymbol{x}_i$ is the observed value of random covariate vector of dimension $p+1$ corresponding to the regression part of the model. A finite mixture of poisson regression model is defined as

$$f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i) = \sum_{j=1}^{k} \alpha_j g(y_i; log(\lambda_{ij}))$$

$$g(y;\gamma) = \frac{1}{y!} exp[y\gamma - e^{\gamma}], y = 0, 1, ...,$$

where α_j denotes the proportion of the jth component with $\sum_{j=1}^{k} \alpha_j = 1$, k is the number of components, $g(y;\gamma)$ is the Poisson probability distribution with mean $\lambda = e^{\gamma} > 0$, and $\lambda_{ij} = t_i \lambda_j(\boldsymbol{x}_i)$ with

$$log(\lambda_j(\boldsymbol{x}, \boldsymbol{\beta}_j)) = \beta_{j0} + \beta_{j1} x_1 + \cdots + \beta_{jp} x_p = \boldsymbol{x}^T \boldsymbol{\beta}_j, j = 1, ..., k.$$

Here, $\boldsymbol{x} = (1, x_1, ..., x_p)^T$, $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, ..., \beta_{jp})^T \in R^{1+p}$, β_{jl} is the regression coefficient for the l^{th} covariate x_l and j^{th} component, and $\boldsymbol{\theta} = (\alpha_1, ..., \alpha_{k-1}, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_k^T)^T$.

The same as before, Lu et al. (2003) used the empirical probability function as f_n

$$f_n(y) = \frac{N_y}{n}, y = 0, 1, 2, ...,$$

assuming that the sample size is sufficiently large, and N_y is the frequency of y among $Y_1, ..., Y_n$. If

$$f_{\boldsymbol{\theta}}(y) = \int f_{\boldsymbol{\theta}}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) d\boldsymbol{x} = \sum_{j=1}^{k} \alpha_j \int g(y; \boldsymbol{x}^T \boldsymbol{\beta}_j) f_X(\boldsymbol{x}) d\boldsymbol{x} \tag{1.7}$$

is known, except for parameter $\boldsymbol{\theta}$, then

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\| f_{\boldsymbol{\theta}}^{1/2} - f_n^{1/2} \right\|.$$

If, however, $f_X(\boldsymbol{x})$ is unknown, or the integration in (1.7) is complex due to the high dimension of $\boldsymbol{X}$, then replace $f_{\boldsymbol{\theta}}(y)$ by a consistent estimator

$$f_{\boldsymbol{\theta},n}(y) = \frac{1}{n} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i) = \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{\alpha_j}{n} g(y; \boldsymbol{x}_i^T \boldsymbol{\beta}_j)$$

and the MHDE of $\boldsymbol{\theta}$ is defined as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\| f_{\boldsymbol{\theta},n}^{1/2} - f_n^{1/2} \right\|.$$

Evidence from Monte Carlo simulations suggested that MHDE is a viable alternative to the maximum likelihood estimator when the mixture components were not well separated or the model parameters were near zero.

1.2.7 Nonparametric mixture model

Assuming that data from the distributions F and G as well as the mixture distribution $\lambda F(x)+(1-\lambda)G(x)$ are available, Karunamuni and Wu (2009) used the minimum Hellinger distance approach to estimate the mixture proportion λ, where F and G are two unknown distributions, and $\lambda F(x) + (1-\lambda)G(x)$ is known as a nonparametric mixture.

More specifically, they assumed that they observed three independent samples

$$\begin{aligned} &X_1, ..., X_{n_0} \overset{iid}{\sim} F, \\ &Y_1, ..., Y_{n_0} \overset{iid}{\sim} G, \\ &Z_1, ..., Z_{n_0} \overset{iid}{\sim} \lambda F + (1-\lambda)G, \end{aligned}$$

and then, the problem was to estimate the mixture parameter λ, treating F and G as nuisance parameters.

In order to employ the MHD technique of Beran (1987), they defined a parametric family of densities

$$M_\lambda(x) = \lambda f(x) + (1-\lambda)g(x), \tag{1.8}$$

where f and g denote the density functions of F and G, respectively. Then, they defined adaptive kernel density estimators of f ad g, based on data $X_1, ..., X_{n_0}$ and $Y_1, ..., Y_{n_0}$:

$$\begin{aligned} \tilde{f}(x) &= \frac{1}{n_0 S_{n_0} h_{n_0}} \sum_{i=1}^{n_0} K_1\left(\frac{x - X_i}{S_{n_0} h_{n_0}}\right) \\ \tilde{g}(x) &= \frac{1}{n_1 S_{n_1} h_{n_1}} \sum_{i=1}^{n_1} K_2\left(\frac{x - Y_i}{S_{n_1} h_{n_1}}\right), \end{aligned}$$

where K_1 and K_2 were two smooth density functions, bandwidths h_{n_0} and h_{n_1} were positive constants such that $h_{n_i} \to 0$ as $n_i \to \infty$, $i = 0, 1$, and $S_{n_0} = S_{n_0}(X_1, ..., X_{n_0})$ $S_{n_1} = S_{n_1}(Y_1, ..., Y_{n_0})$ were robust scale statistics. Replace f and g from (1.8) with $\tilde{f}$ and $\tilde{g}$, a parametric mixture model was defined as:

$$\tilde{M}_\lambda(x) = \lambda \tilde{f}(x) + (1 - \lambda)\tilde{g}(x).$$

Next, a kernel density estimator based on the Z_i's was defined:

$$\hat{M}(x) = \frac{1}{n_2 S_{n_2} h_{n_2}} \sum_{i=1}^{n_2} K\left(\frac{x - Z_i}{S_{n_2} h_{n_2}}\right),$$

where K, h and S were defined similarly. Then, the MHDE $\hat{\lambda}$ is the minimizer of the Hellinger distance between $\tilde{M}_\lambda$ and $\hat{M}$.

Similar to Beran (1987), the MHDE was proved to be consistent, asymptotic normally distributed, and have good efficiency and robustness properties.

1.2.8 Two-sample semiparametric model

Over the past few years, semiparametric models have continued to receive increasing attention from both practical and theoretical point of views, due to its wide application, primarily in biostatics and econometrics. Wu et al. (2010) investigated the estimation problem of parameters in a two-sample semiparametric model. Let $X_1, ..., X_n$ be a sample from a population with distribution function G and density function g, and $Z_1, ..., Z_n$ be another sample, independent of $X_i's$, with distribution function H and density function $h(x) = exp[\alpha + r(x)\boldsymbol{\beta}]g(x)$, where α an $\boldsymbol{\beta}$ are unknown parameters of interest and g is an unknown density. Define $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$, then,

$$X_1...., X_n \overset{iid}{\sim} g(x),$$
$$Z_1...., Z_m \overset{iid}{\sim} h_\theta(x),$$

where $h_{\boldsymbol{\theta}}(x) = g(x)exp[(1, \boldsymbol{r}(x))\boldsymbol{\theta}]$, $\boldsymbol{r}(x) = (r_1(x), ..., r_p(x))$ is a $1 \times p$ vector of continuous functions of x on $\mathbb{R}$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a $p \times 1$ parameter vector, and α is a normalizing parameter that makes $h_{\boldsymbol{\theta}}(x)$ integrate to 1.

Based on $X_1, ..., X_n$ and $Z_1, ..., Z_m$, Wu et al. (2010) first defined the kernel density estimators of g and h_θ:

$$g_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K_0 \left(\frac{x - X_i}{b_n} \right),$$
$$h_m(x) = \frac{1}{mb_m} \sum_{j=1}^{m} K_1 \left(\frac{x - Z_j}{b_m} \right),$$

where K_0 and K_1 were symmetric density functions, bandwidths b_n and b_m were positive constants such that $b_n \to 0$ as $n \to \infty$ and $b_m \to 0$ as $m \to \infty$.

Applying the plug-in rule, they used the estimator g_n in the place of g and constructed a parametric model as:

$$\hat{h}_{\boldsymbol{\theta}}(x) = exp[(1, \boldsymbol{r}(x))\boldsymbol{\theta}] g_n(x).$$

Note that $\hat{h}_{\boldsymbol{\theta}}$ is a parametric density function with the unknown parameter being $\boldsymbol{\theta}$. Then, the MHDE $\hat{\boldsymbol{\theta}}$ is the minimizer of the Hellinger distance between the parametric density $\hat{h}_{\boldsymbol{\theta}}$ and the nonparametric density estimator h_m.

The approach here is in line with Beran (1987), thus it is not difficult to prove that the MHDE is consistent, asymptotic normally distributed, and has good efficiency and robustness properties.

Chapter 2

MHD Estimation in a Semiparametric Mixture Model

2.1 Introduction

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

$$g(x) = \pi f_0(x;\xi) + (1-\pi) f(x-\mu), \forall x \in \mathbb{R}, \tag{2.1}$$

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

Bordes *et al.* (2006) studied the case when ξ is assumed to be known, i.e., the first component density is completely known, and model (2.1) becomes

$$g(x) = \pi f_0(x) + (1-\pi) f(x-\mu), \forall x \in \mathbb{R}. \tag{2.2}$$

The model (2.2) is motivated by the problem of detection of differentially expressed genes under two or more conditions in microarray data. We build a test statistic for each gene and then observe the response of thousands of genes, which corresponds in practice to thousands of observations from statistical tests. Under the null hypothesis, due to a lack of difference in expression, the test statistic is assumed to have a known distribution, say f_0, and the samples obtained in the way mentioned earlier should come from a mixture of two distributions: the known distribution f_0, that is under null hypotheses, and the other

distribution, $f(\cdot - \mu)$, the unknown distribution of the test statistics under the alternative hypothesis. The probability that a gene comes from the null component of the mixture distribution (2.2) conditionally on the observation can be estimated if we can estimate the parameters π, μ and f. Consequently, we can classify each gene to a component by using a classification criterion, and therefore distinguish the genes differentially expressed from the genes non-differentially expressed. Please see Bordes *et al*. (2006) for more detail about the application of model (2.2) to Microarray data analysis.

Song *et al*. (2010) studied another special case of model (2.1)

$$g(x) = \pi\phi(x; 0, \sigma) + (1 - \pi)f(x), \forall x \in \mathbb{R}, \tag{2.3}$$

where $\phi(x; 0, \sigma)$ is a normal density with mean 0 and *unknown* standard deviation σ and $f(x)$ is an unknown density. The model (2.3) was motivated by a sequential clustering algorithm, proposed by Song and Nicolae (2009). Unlike most clustering algorithms, the sequential clustering algorithm doesn't require specifying the number of clusters and allows some objects not to be assigned to any clusters. The algorithm works by finding a local center of a cluster first, and then identifying whether a object belongs to that cluster or not based on some penalty score. If we assume that the objects belonging to the cluster come from a normal distribution with known mean (such as 0) 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 can be considered as estimating the mixing proportion in model (2.3). This estimation of the mixing proportion will be repeated whenever a new cluster is considered.

Note that the semiparametric mixture model (2.1) is not generally identifiable. Bordes *et al*. (2006) has shown that model (2.2) is not generally identifiable if we don't put any restriction on unknown density $f(x)$, but identifiability can be achieved through some sufficient conditions. One important condition is that $f(\cdot)$ is symmetric about 0. Then, they proposed an estimation procedure based on the symmetry of the unknown component f. Song *et al*. (2010) also addressed the problems of unidentifiability and noticed that model

(2.3) was not generally identifiable. In addition, due to the additional unknown parameter σ in the first component, Song *et al.* (2010) mentioned that it was hard to find the conditions to avoid unidentifiability of model (2.3) and proposed to use simulation studies to check the performance of the proposed estimators.

In this report, we mainly focus on the estimation part of model (2.1) and propose a new estimator for model (2.1) based on Minimum Hellinger Distance, which has been shown to have good efficiency and robustness properties (see, for example, Beran, 1977 and Lindsay, 1994). Please refer to Bordes *et al.* (2006) and Song *et al.* (2010) for some detailed discussions about the identifiability of model (2.1). A simple and effective algorithm is also given to find the proposed estimator. Using simulation studies, we illustrate the finite sample performance of the proposed estimate and compare it to the estimators proposed by Bordes *et al.* (2006) and Song *et al.* (2010). Our empirical studies demonstrate that the proposed MHDE works at least as well as some existing estimators for most of the examples considered and outperforms the existing estimators when the data are under contamination.

2.2 Review of Existing Methods

2.2.1 Estimating by symmetrization

Bordes *et al.* (2006) proposed an inference procedure based on the symmetry of the unknown component of model (2.2). Let $X_1, ..., X_n$ be random variables from model (2.2) and G be the cumulative distribution function (cdf) of model (2.2), i.e.

$$G(x) = \pi F_0(x) + (1-\pi)F(x-\mu), \forall x \in \mathbb{R}, \tag{2.4}$$

where G, F_0, and F are the corresponding cdfs of g, f_0 and f. Assuming that the G is uniquely defined in (2.4), then

$$F(x) = \frac{1}{1-\pi}((G(x+\mu) - \pi F_0(x+\mu)), \forall x \in \mathbb{R}. \tag{2.5}$$

Let

$$H_1(x;\pi,\mu,G) = \frac{1}{1-\pi}G(x+\mu) + (1-\frac{1}{1-\pi})F_0(x+\mu),$$
$$H_2(x;\pi,\mu,G) = 1 - \frac{1}{1-\pi}G(\mu-x) + (\frac{1}{1-\pi}-1)F_0(\mu-x).$$

Since f is assumed to be symmetric, $F(x) = 1-F(-x)$, for all $x \in \mathbb{R}$. Then, $H_1(\cdot;\pi_0,\mu_0,G) = H_2(\cdot;\pi_0,\mu_0,G)$, where π_0 and μ_0 are the unknown true values of π and μ. Consequently, if d is a distance measure, such as L^2-norm, between two functions, then we have $d(H_1(\cdot;\pi_0,\mu_0,G), H_2(\cdot;\pi_0,\mu_0,G)) = 0$, where

$$d(\pi,\mu) = \|H_1 - H_2\|_2 = \left(\int |H_1(x;\pi,\mu,G) - H_2(x;\pi,\mu,G)|^2 dx\right)^{1/2}.$$

Since G is unknown, it is estimated by

$$\hat{G}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x), \forall x \in \mathbb{R},$$

where $I(\cdot)$ is the indicator function. Replace G by G_n, we get an empirical version d_n of d defined by $d_n(\pi,\mu) = d(H_1(\cdot;\pi,\mu,G_n), H_2(\cdot;\pi,\mu,G_n))$. Bordes *et al.* (2006) proposed to estimate π and μ of model (2.2) by minimizing $d_n(\pi,\mu)$.

2.2.2 EM-type estimator

Let

$$Z_i = \begin{cases} 1, & \text{if } X_i \text{ is from the first component;} \\ 0, & \text{otherwise.} \end{cases}$$

Song *et al.* (2010) proposed an EM-type estimator for model (2.3).

E-step In the $(k+1)^{th}$ step, compute the conditional expectation of Z_i given parameters of the k^{th} step and data, i.e.,

$$Z_i^{(k+1)} = E(Z_i|\pi^{(k)},\sigma^{(k)},X_i) = \frac{\pi^{(k)}\phi_{\sigma^{(k)}}(X_i)}{\widehat{g(X_i)}}, \tag{2.6}$$

where

$$\widehat{g(x)} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \tag{2.7}$$

and K is a kernel function, such as Gaussian kernel, and h is the bandwidth.

M-step The values of the parameters are updated in the M-step as follows.

$$\pi^{(k+1)} = \frac{\sum_{i=1}^{n} Z_i^{(k+1)}}{n},$$
$$\sigma^{(k+1)} = \sqrt{\frac{\sum_{i=1}^{n} Z_i^{(k+1)} X_i^2}{\sum Z_i^{(k+1)}}}.$$

In addition, Song *et al*. (2010) also recommended to use

$$Z_i^{(k+1)} = \frac{2\pi^{(k)}\phi_{\sigma^{(k)}}(X_i)}{\pi^{(k)}\phi_{\sigma^{(k)}}(X_i) + \widehat{g(X_i)}}$$

truncated to 1 when it is greater than 1, in the E step, to stabilize the Z-values.

2.2.3 Maximizing π-type estimator

Song *et al*. (2010) demonstrated that the EM-type estimator introduced in Section 2.2.2 is biased when two component densities overlap significantly based on their simulation studies. Therefore, they proposed an alternative estimator, by finding the maximum mixing proportion π that satisfies the following condition:

$$\pi\phi_\sigma(x_i) \leq \widehat{g(x_i)}, i = 1, ..., n.$$

Therefore, the estimator for π is

$$\hat{\pi} = \max_{\sigma} \min_{x_i} \frac{\widehat{g(x_i)}}{\phi_\sigma(x_i)},$$

where $\widehat{g(x_i)}$ has the same definition as in (2.7), and the estimator for σ is

$$\hat{\sigma} = \arg\max \min_{x_i} \frac{\widehat{g(x_i)}}{\phi_\sigma(x_i)}.$$

Please refer to Song *et al*. (2010) for more detailed explanation about this method.

2.3 New Estimate Based on MHD

In this section, we propose an alternative estimator of the general semiparametric mixture model (2.1) based on Minimum Hellinger Distance (MHD) due to its good efficiency and

robustness properties. Note that model (2.2) considered by Bordes *et al.* (2006) and model (2.3) considered by Song *et al.* (2010) are just the special cases of the model (2.1).

First, we introduce the general results of MHDE for semiparametric models. Let $(\mathscr{X}, \mathscr{S}, v)$ be a measure space and $\mathscr{H}$ be a semiparametric model of v-densities of the form

$$\mathscr{H} = \{h_{\theta,f} : \theta \in \Theta, f \in \mathscr{F}\},$$

where Θ is a compact subset of $\mathbb{R}^p$ and $\mathscr{F}$ is an arbitrary set of infinite dimension. Let $\mathscr{G}$ be a class of v-densities that contains $\mathscr{H}$. For member a of $L_2(v)$ we denote the $L_2(v)$-norm of a as $\|a\|$. For any members g_1 and g_2 of $\mathscr{G}$, the Hellinger distance between them is defined by

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

The functional T is defined on $\mathscr{G}$ such that for every $g \in \mathscr{G}$,

$$\left\| h_{T(g),f}^{1/2} - g^{1/2} \right\| = \inf_{\theta \in \Theta} \left\| h_{\theta,f}^{1/2} - g^{1/2} \right\|, \tag{2.8}$$

where T is referred to as the MHD functional and assumed to be continuous for the Hellinger distance metric d_H. Assume that $\mathscr{H}$ is identifiable, and therefore, T is Fisher consistent: $T(h_{\theta,f}) = \theta$ for any $\theta \in \Theta$ and any $f \in \mathscr{F}$. Let $X_1, X_2, ..., X_n$ be a sample of independently and identically distributed $\mathscr{X}$ random variables with density $h_0 = h_{\theta_0,f_0}$ where $\theta_0 \in int(\Theta)$ and $f_0 \in \mathscr{F}$. Then the MHDE of θ_0 is defined as $T(h_n)$, where h_n is a $\mathscr{G}$-valued estimator of h_0 based on the sample $X_1, X_2, ..., X_n$.

Next, we apply the MHD estimation method to model (2.1). Let

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

where

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

Assume we have a sample of $X_1, X_2, ..., X_n$ from a population with density $h_{\theta,f} \in \mathscr{H}$, and a nonparametric density estimation of $h_{\theta,f}$ to be denoted by $\hat{g}$. We define functional $\hat{f}$ of t and $\hat{g}$ as

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

and then the MHDE of θ is defined as

$$\hat{\theta}(\hat{g}) = arg \min_{\theta \in \Theta} \left\| h_{\theta,\hat{f}(\theta,\hat{g})}^{1/2} - \hat{g}^{1/2} \right\|.$$

Suppose the initial estimates of $\theta = (\pi, \sigma, \mu)$ and f are $\theta^{(0)} = (\pi^{(0)}, \sigma^{(0)}, \mu^{(0)})$ and $f^{(0)}$. Then the proposed MHDE is calculated by iterating the following two steps.

Step 1 For fixed $\pi^{(k)}, \sigma^{(k)}$ and $\mu^{(k)}$, find $f^{(k+1)}$ which minimizes

$$\left\| [\pi^{(k)} \phi_{\sigma^{(k)}}(\cdot) + (1 - \pi^{(k)}) f^{(k+1)}(\cdot - \mu^{(k)})]^{1/2} - \hat{g}^{1/2}(\cdot) \right\|.$$

It turns out (Wu *et al.* 2011) that the solution is

$$f^{(k+1)}(x) = \begin{cases} \frac{\alpha}{1-\pi^{(k)}} \hat{g}(x + \mu^{(k)}) - \frac{\pi^{(k)}}{1-\pi^{(k)}} \phi_{\sigma^{(k)}}(x + \mu^{(k)}), & \text{if } x \in M, \\ 0, & \text{if } x \in M^C, \end{cases} \tag{2.9}$$

where $M = \{x : \alpha \hat{g}(x) \geq \pi^{(k)} \phi_{\sigma^{(k)}}(x)\}$ and

$$\alpha = \frac{1}{\int_M \hat{g}(x) dx} \left\{ \pi^{(k)} \int_M \phi_{\sigma^{(k)}}(x) dx + (1 - \pi^{(k)}) \right\}.$$

If we further assume $f(\cdot)$ is symmetric about 0, i.e., $f(x) = f(-x)$, then we can symmetrize $f^{(k+1)}(x)$ by

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

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

$$\left\| [\pi^{(k+1)} \phi_{\sigma^{(k+1)}}(\cdot) + (1 - \pi^{(k+1)}) f^{(k+1)}(\cdot - \mu^{(k+1)})]^{1/2} - \hat{g}^{1/2}(\cdot) \right\|. \tag{2.10}$$

Chapter 3

Simulation Studies and Real Data Application

3.1 Simulation studies

In this section, we investigate the performance of the proposed MHDE, Maximizing-π type estimator and EM-type estimator (Song *et al.*(2010)), and the Symmetrized estimator (Bordes *et al.*(2006)) in the case of σ unknown and σ known.

The initial model (2.3) Song *et al.* (2010) considered did not have the location parameter μ in the second component. After we have $\hat{\pi}$ and $\hat{\sigma}$, we can simply estimate μ by

$$\hat{\mu} = \frac{\sum_{i=1}^{n}(1-\hat{Z}_i)X_i}{\sum_{i=1}^{n}(1-\hat{Z}_i)}, \tag{3.1}$$

where $\hat{Z}_i$ is

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

We use both the true values and the estimates from normal mixture models to be the initial estimates $\theta^{(0)} = (\pi^{(0)}, \sigma^{(0)}, \mu^{(0)})$ and choose the one that produces the smaller value in (2.10).

3.1.1 σ unknown

In this section, we simulate 200 samples of n i.i.d. random variables from a population with density function (2.1), where (π, σ, μ) are unknown parameters and f is an unknown density

that is symmetric about zero. We consider the following cases:

Case I: $X \sim 0.3N(0,1) + 0.7N(1.5,1) \Rightarrow (\pi, \sigma, \mu) = (0.3, 1, 1.5)$

Case II: $X \sim 0.3N(0,1) + 0.7N(3,1) \Rightarrow (\pi, \sigma, \mu) = (0.3, 1, 3)$

Case III: $X \sim 0.3N(0,1) + 0.7U(2,4) \Rightarrow (\pi, \sigma, \mu) = (0.3, 1, 3)$

Case IV: $X \sim 0.7N(0,4) + 0.3N(3,1) \Rightarrow (\pi, \sigma, \mu) = (0.7, 2, 3)$

Case V: $X \sim 0.85N(0,4) + 0.15N(3,1) \Rightarrow (\pi, \sigma, \mu) = (0.85, 2, 3)$

Case I, Case II and Case III are the models used by Song *et al.*(2010) to show the performance of their Maximizing-π type and EM-type estimators, where Case I represents the situation when two components are close and Case II represents the situation when two components are apart. Case IV and Case V are suggested by Bordes *et al.*(2006) to show the performance of their semiparametric EM algorithm. In addition, we also consider a set of contaminated model by adding 2% outliers from $U(10,20)$ to the original set of models.

To estimate the unknown parameters (π, σ, μ), we consider the following methods: a) The MHD estimator; b) Modified Maximizing-π type and EM-type estimator proposed by Song *et al.*(2010), estimating μ after π, σ have been estimated; c) Modified Symmetrized estimator proposed by Bordes *et al.*(2006) to incorporate the variance σ.

To assess the performance, we look at both the mean and the mean squared error (MSE) of each estimate, where

$$\text{mean}(\hat{\boldsymbol{\theta}}) = \bar{\hat{\boldsymbol{\theta}}} = \frac{1}{m}\sum_{t=1}^{m}\hat{\boldsymbol{\theta}}_t,$$

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = \frac{1}{m}\sum_{t=1}^{m}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta})^2.$$

For the five cases considered, Table 3.1 and Table 3.2 report the mean and MSE of the parameter estimates based on the four methods when $n = 250$ and $n = 1000$. Table 3.3 and Table 3.4 report the result when models are under 2% contamination from $U(10,20)$.

From the tables, we can see that Maximizing-π type estimator works better when the two components have the same parametric distribution, and perform relatively well under a mild contamination. The EM-type estimator performs quite well when the other component is not normally distributed, but performs poorly when the two normal components are close. The Symmetrized estimator outperforms the two methods suggested by Song *et al.* (2010) in all cases when there's no contamination, but is not robust under contaminations. The MHD estimator that we proposed provides satisfactory results when there's no contamination, but perform much better under severe contamination. Therefore, the MHD estimator is more robust than the rest methods.

Table 3.1: *Average (MSE) of Point Estimates Over 200 Repetitions When* $n = 250$

Case	TRUE	MHDE	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	0.257(0.014)	0.364(0.006)	0.602(0.093)	0.252(0.015)
	$\sigma : 1$	1.058(0.021)	0.899(0.075)	1.157(0.032)	1.020(0.033)
	$\mu : 1.5$	1.436(0.051)	1.720(0.059)	1.921(0.186)	1.421(0.049)
II	$\pi : 0.3$	0.295(0.001)	0.272(0.003)	0.393(0.011)	0.298(0.001)
	$\sigma : 1$	1.046(0.013)	1.330(0.912)	1.377(0.191)	0.999(0.021)
	$\mu : 3$	2.995(0.010)	2.871(0.054)	3.121(0.022)	2.983(0.011)
III	$\pi : 0.3$	0.263(0.002)	0.257(0.004)	0.305(0.002)	0.302(0.001)
	$\sigma : 1$	0.939(0.013)	1.609(1.741)	1.163(0.100)	1.013(0.022)
	$\mu : 3$	2.994(0.001)	2.767(0.085)	2.931(0.009)	3.001(0.002)
IV	$\pi : 0.7$	0.692(0.003)	0.632(0.009)	0.821(0.016)	0.686(0.007)
	$\sigma : 2$	2.036(0.023)	2.023(0.035)	2.142(0.028)	2.009(0.032)
	$\mu : 3$	3.108(0.054)	2.563(0.269)	3.153(0.067)	2.930(0.140)
V	$\pi : 0.85$	0.836(0.003)	0.774(0.010)	0.910(0.004)	0.774(0.028)
	$\sigma : 2$	2.093(0.027)	2.069(0.035)	2.046(0.011)	2.027(0.048)
	$\mu : 3$	3.115(0.205)	2.088(1.024)	2.778(0.266)	2.427(0.981)

Table 3.2: *Average (MSE) of Point Estimates Over 200 Repetitions When* $n = 1000$

Case	TRUE	MHDE	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	0.281(0.005)	0.353(0.004)	0.601(0.091)	0.280(0.005)
	$\sigma : 1$	1.040(0.008)	0.853(0.028)	1.177(0.034)	1.025(0.011)
	$\mu : 1.5$	1.481(0.017)	1.736(0.059)	1.923(0.181)	1.476(0.018)
II	$\pi : 0.3$	0.299(0.001)	0.263(0.002)	0.399(0.010)	0.300(0.001)
	$\sigma : 1$	1.017(0.003)	0.956(0.007)	1.407(0.176)	0.998(0.005)
	$\mu : 3$	3.009(0.002)	2.958(0.005)	3.151(0.025)	3.003(0.002)
III	$\pi : 0.3$	0.271(0.001)	0.253(0.003)	0.311(0.001)	0.301(0.001)
	$\sigma : 1$	0.949(0.005)	0.971(0.007)	1.177(0.044)	1.005(0.004)
	$\mu : 3$	2.997(0.001)	2.878(0.017)	2.969(0.002)	2.999(0.001)
IV	$\pi : 0.7$	0.692(0.001)	0.631(0.006)	0.825(0.016)	0.696(0.001)
	$\sigma : 2$	2.002(0.006)	1.949(0.013)	2.172(0.032)	1.999(0.006)
	$\mu : 3$	3.058(0.017)	2.654(0.153)	3.161(0.035)	2.982(0.015)
V	$\pi : 0.85$	0.847(0.001)	0.783(0.006)	0.922(0.005)	0.825(0.010)
	$\sigma : 2$	2.053(0.009)	1.995(0.008)	2.087(0.010)	2.008(0.031)
	$\mu : 3$	3.099(0.042)	2.255(0.633)	3.135(0.060)	2.820(0.293)

3.1.2 σ known

Next, we consider the cases when the variance σ^2 is assumed to be known:

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: $X \sim 0.85N(0,4) + 0.15N(3,1) \Rightarrow (\pi, \mu) = (0.85, 3)$

In order to estimate the unknown parameters (π, μ), we consider the following methods: a) Symmetrized estimator proposed by Bordes *et al.*(2006); b) Modified Maximizing-π

Table 3.3: *Average (MSE) of Point Estimates Over 200 Repetitions When* $n = 250$ *under* 2% *contamination from* $U(10, 20)$

Case	TRUE	MHDE	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	0.192(0.024)	0.360(0.006)	0.592(0.087)	0.136(0.038)
	$\sigma : 1$	1.103(0.056)	0.985(0.184)	1.155(0.031)	0.784(0.116)
	$\mu : 1.5$	1.355(0.070)	2.197(0.550)	2.585(1.277)	1.323(0.067)
II	$\pi : 0.3$	0.289(0.001)	0.267(0.003)	0.387(0.009)	0.251(0.005)
	$\sigma : 1$	1.056(0.014)	1.306(0.843)	1.400(0.204)	0.805(0.062)
	$\mu : 3$	2.989(0.012)	3.245(0.115)	3.525(0.316)	2.953(0.016)
III	$\pi : 0.3$	0.275(0.001)	0.227(0.008)	0.277(0.002)	0.258(0.003)
	$\sigma : 1$	0.943(0.012)	2.125(3.379)	1.081(0.055)	0.797(0.056)
	$\mu : 3$	2.992(0.001)	2.932(0.060)	3.207(0.073)	2.971(0.004)
IV	$\pi : 0.7$	0.676(0.004)	0.611(0.012)	0.802(0.011)	0.623(0.013)
	$\sigma : 2$	2.010(0.018)	2.035(0.041)	2.138(0.028)	1.787(0.078)
	$\mu : 3$	3.118(0.064)	3.406(0.435)	4.339(2.125)	2.968(0.084)
V	$\pi : 0.85$	0.823(0.006)	0.752(0.014)	0.887(0.002)	0.736(0.038)
	$\sigma : 2$	2.052(0.029)	2.069(0.034)	2.041(0.010)	1.807(0.099)
	$\mu : 3$	3.215(0.228)	3.715(1.406)	4.963(4.889)	2.870(0.460)

type and EM-type estimator proposed by Song *et al.*(2010), assuming σ to be known but estimating μ after π have been estimated; c) the MHD estimator, but assume σ to be known.

Table 3.5 and Table 3.6 report the mean and MSE of the parameter estimates based on the four methods when $n = 250$ and $n = 1000$. Table 3.7 and Table 3.8 report the result when models are under 2% contamination from $U(10, 20)$. From the tables, we can see that the Symmetrized estimator and the MHD estimator perform better than the Maximizing-π type and EM-type estimator in all cases, especially in Case IV and Case V which are suggested by Bordes *et al.*(2006). When the sample is contaminated by outliers, the MHD estimator and the Maximizing-π type estimator provide better estimates than the EM-type and the Symmetrization estimator, and therefore are more robust.

Table 3.4: *Average (MSE) of Point Estimates Over 200 Repetitions When* $n = 1000$ *under* 2% *contamination from* $U(10, 20)$

Case	TRUE	MHDE	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	0.217(0.015)	0.349(0.003)	0.591(0.085)	0.089(0.051)
	$\sigma : 1$	1.099(0.026)	0.872(0.022)	1.178(0.033)	0.904(0.050)
	$\mu : 1.5$	1.384(0.039)	2.206(0.515)	2.568(1.162)	1.242(0.085)
II	$\pi : 0.3$	0.288(0.001)	0.258(0.002)	0.392(0.009)	0.250(0.003)
	$\sigma : 1$	1.025(0.003)	0.969(0.007)	1.422(0.189)	0.801(0.045)
	$\mu : 3$	2.992(0.002)	3.299(0.099)	3.537(0.297)	2.953(0.005)
III	$\pi : 0.3$	0.279(0.001)	0.247(0.003)	0.304(0.001)	0.258(0.002)
	$\sigma : 1$	0.960(0.004)	0.967(0.006)	1.185(0.050)	0.806(0.042)
	$\mu : 3$	2.996(0.001)	3.208(0.049)	3.302(0.099)	2.980(0.001)
IV	$\pi : 0.7$	0.683(0.001)	0.621(0.008)	0.810(0.012)	0.641(0.004)
	$\sigma : 2$	1.981(0.004)	1.955(0.013)	2.178(0.034)	1.813(0.042)
	$\mu : 3$	3.094(0.020)	3.493(0.324)	4.386(2.005)	3.024(0.012)
V	$\pi : 0.85$	0.831(0.001)	0.769(0.008)	0.903(0.003)	0.780(0.008)
	$\sigma : 2$	2.013(0.004)	1.992(0.007)	2.083(0.009)	1.833(0.034)
	$\mu : 3$	3.193(0.064)	3.909(1.093)	5.559(6.866)	3.038(0.068)

Figure 3.1 contains the MSE of μ in the five σ unknown cases over 200 repetitions when the sample size is 1000, and Figure 3.2 and Figure 3.3 contain the MSE of μ and π in the five cases σ known over 200 repetitions when the sample size is 1000 and under 2% contamination from $U(10, 20)$. From the plots, we can see that almost all the four estimators considered perform well in case II and case III. The EM-type estimator performs poorly in case I, and is the worst in case IV and V when the model is under contamination. The Symmetrized estimator is sensitive to contamination, especially in case IV and V, no matter σ known or not. Comparatively, the Maximizing-π type estimator is more robust, but doesn't perform well in case IV and V when data is not under contamination. From the plots, we can see that the MHD estimator perform well in all cases, and is robust when data is under

contamination.

Table 3.5: *Average (MSE) of Point Estimates Over 200 Repetitions When* $n = 250$

Case	TRUE	MHDE	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	0.210(0.028)	0.328(0.005)	0.569(0.074)	0.220(0.021)
	$\mu : 1.5$	1.390(0.084)	1.662(0.041)	1.972(0.231)	1.393(0.060)
II	$\pi : 0.3$	0.291(0.001)	0.242(0.005)	0.334(0.002)	0.299(0.001)
	$\mu : 3$	3.007(0.007)	2.882(0.027)	3.057(0.009)	2.996(0.009)
III	$\pi : 0.3$	0.259(0.003)	0.229(0.006)	0.284(0.001)	0.299(0.001)
	$\mu : 3$	2.999(0.001)	2.812(0.043)	2.918(0.010)	2.999(0.002)
IV	$\pi : 0.7$	0.691(0.003)	0.592(0.018)	0.802(0.012)	0.683(0.009)
	$\mu : 3$	3.131(0.067)	2.382(0.501)	3.063(0.069)	2.905(0.159)
V	$\pi : 0.85$	0.810(0.014)	0.729(0.021)	0.902(0.003)	0.809(0.011)
	$\mu : 3$	3.217(0.444)	1.866(1.503)	2.677(0.349)	2.655(0.625)

3.2 Real Data Application

Iris data (used by Song *et at.* (2010)) is perhaps the best known database to be found in the pattern recognition literature. It is first introduced by Fisher (1936), and is referenced frequently to this day. The data set contains four attributes: sepal length (in cm), sepal width (in cm), petal length (in cm), and petal width (in cm), and there are 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2 and the latter are not linearly separable from each other.

We want to find the clusters for the data. After applying the research algorithm for centers of clusters by Song *et al.* (2010), observation 8 is selected as the center of the first cluster. We adjust all observations by subtracting observation 8 from all observations. As discussed by Song *et al.* (2010), the proportion of observations that belong to a cluster can be considered as estimating the mixing proportion in the two-component mixture model.

Principal component analysis shows that the first principal component accounts for

Table 3.6: *Average (MSE) of Point Estimates Over 200 Repetitions When* $n = 1000$

Case	TRUE	MHDE	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	0.291(0.005)	0.280(0.003)	0.563(0.069)	0.276(0.005)
	$\mu : 1.5$	1.503(0.016)	1.583(0.017)	1.959(0.213)	1.469(0.015)
II	$\pi : 0.3$	0.294(0.001)	0.245(0.004)	0.339(0.002)	0.297(0.001)
	$\mu : 3$	3.006(0.002)	2.917(0.016)	3.093(0.010)	2.998(0.002)
III	$\pi : 0.3$	0.272(0.001)	0.239(0.005)	0.296(0.001)	0.300(0.001)
	$\mu : 3$	2.997(0.001)	2.847(0.029)	2.956(0.002)	2.998(0.001)
IV	$\pi : 0.7$	0.692(0.001)	0.585(0.020)	0.804(0.011)	0.693(0.001)
	$\mu : 3$	3.045(0.013)	2.446(0.400)	3.174(0.039)	2.970(0.017)
V	$\pi : 0.85$	0.843(0.001)	0.749(0.016)	0.911(0.004)	0.843(0.002)
	$\mu : 3$	3.172(0.063)	2.071(1.043)	3.019(0.067)	2.934(0.104)

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. The first principal component loading vector is $(0.36, -0.08, 0.86, 0.35)$, which implies that petal length contains most of the information. Therefore, we apply each of the four estimating method discussed above to the first principal component as well as Petal Length.

Table 3.9 lists the estimators of proportion on petal length and the first principal component. Compared to the true proportion of 1/3, the MHD estimator and the maximizing π-type estimators performs quite well compared to the other estimators. Figure 3.4 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 center of it.

Table 3.7: *Average (MSE) of Point Estimates Over 200 Repetitions When* $n = 250$ *under* 2% *contamination from* $U(10, 20)$

Case	TRUE	MHDE	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	0.21(0.026)	0.332(0.006)	0.563(0.071)	0.120(0.043)
	$\mu : 1.5$	1.398(0.085)	2.113(0.434)	2.543(1.146)	1.276(0.081)
II	$\pi : 0.3$	0.281(0.001)	0.235(0.006)	0.327(0.002)	0.256(0.003)
	$\mu : 3$	2.991(0.007)	3.213(0.076)	3.415(0.202)	2.956(0.012)
III	$\pi : 0.3$	0.279(0.001)	0.227(0.007)	0.285(0.001)	0.272(0.002)
	$\mu : 3$	2.996(0.001)	3.119(0.043)	3.245(0.086)	2.989(0.003)
IV	$\pi : 0.7$	0.680(0.005)	0.578(0.021)	0.786(0.009)	0.398(0.164)
	$\mu : 3$	3.149(0.096)	3.162(0.296)	4.149(1.594)	2.254(1.137)
V	$\pi : 0.85$	0.797(0.025)	0.719(0.023)	0.884(0.002)	0.539(0.140)
	$\mu : 3$	3.220(0.513)	3.358(1.000)	4.859(4.597)	1.907(1.785)

Table 3.8: *Average (MSE) of Point Estimates Over 200 Repetitions When* $n = 1000$ *under* 2% *contamination from* $U(10, 20)$

Case	TRUE	MHDE	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	0.254(0.007)	0.276(0.003)	0.555(0.065)	0.060(0.059)
	$\mu : 1.5$	1.444(0.019)	2.009(0.284)	2.548(1.119)	1.187(0.103)
II	$\pi : 0.3$	0.286(0.001)	0.243(0.004)	0.332(0.001)	0.257(0.002)
	$\mu : 3$	3.001(0.002)	3.257(0.081)	3.444(0.204)	2.966(0.005)
III	$\pi : 0.3$	0.281(0.001)	0.234(0.005)	0.289(0.001)	0.265(0.002)
	$\mu : 3$	2.999(0.001)	3.179(0.044)	3.299(0.096)	2.989(0.001)
IV	$\pi : 0.7$	0.681(0.001)	0.572(0.023)	0.789(0.008)	0.389(0.149)
	$\mu : 3$	3.067(0.013)	3.203(0.257)	4.252(1.628)	2.171(1.165)
V	$\pi : 0.85$	0.831(0.001)	0.738(0.018)	0.895(0.002)	0.503(0.134)
	$\mu : 3$	3.177(0.067)	3.574(0.836)	5.275(5.478)	1.534(2.329)

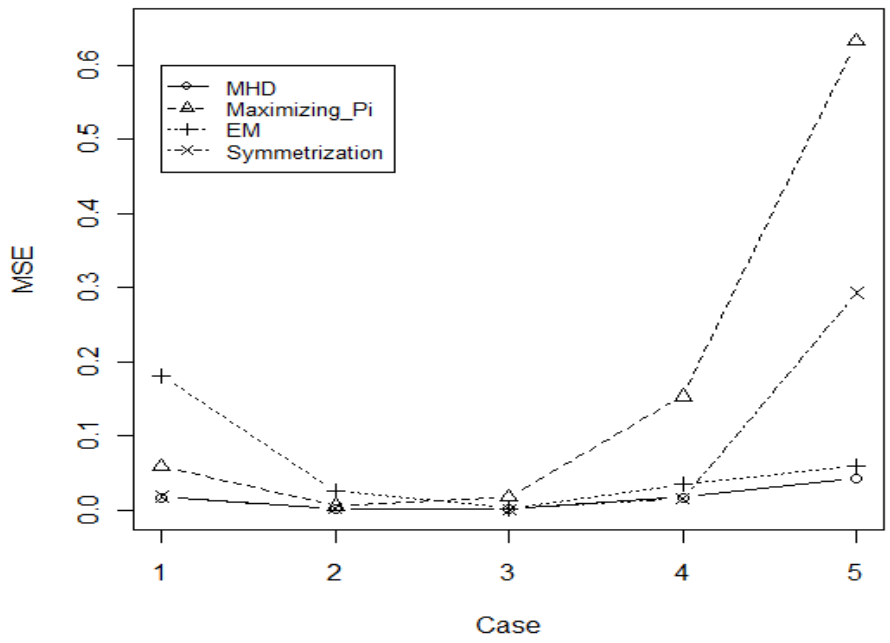

Figure 3.1: *MSE of μ In The Five Cases Considered Over 200 Repetitions When $n = 1000$ (σ Unknown)*

Table 3.9: *Estimators of mixing proportion in Iris data*

Variable	MHDE	Maximizing π-type	EM-type	Symmetrization
Petal Length	0.251	0.266	0.446	0.628
Principal Component	0.320	0.327	0.289	0.399

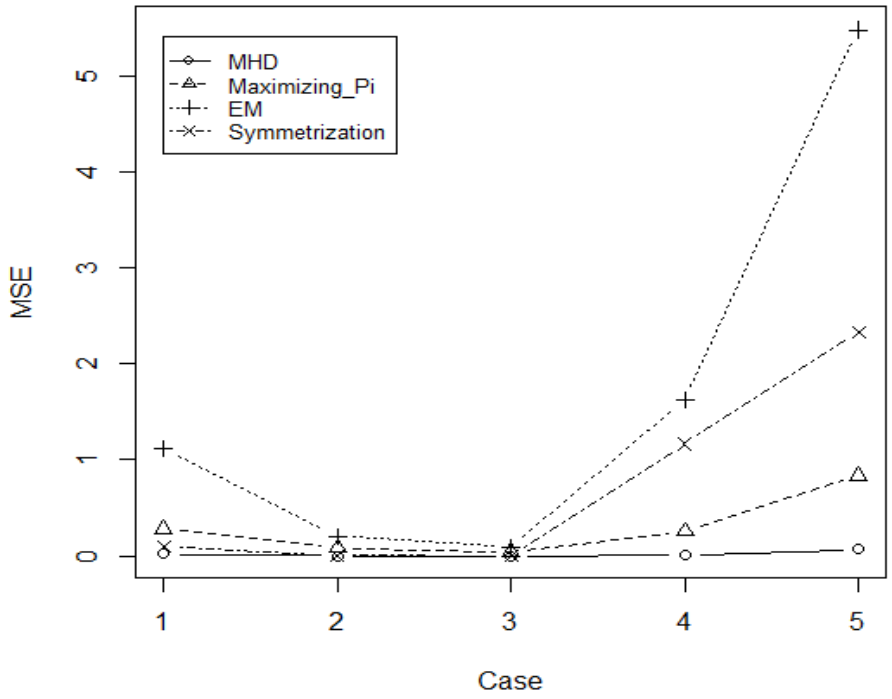

Figure 3.2: *MSE of μ In The Five Cases Considered Over 200 Repetitions When $n = 1000$ under 2% Contamination From $U(10, 20)$ (σ Known)*

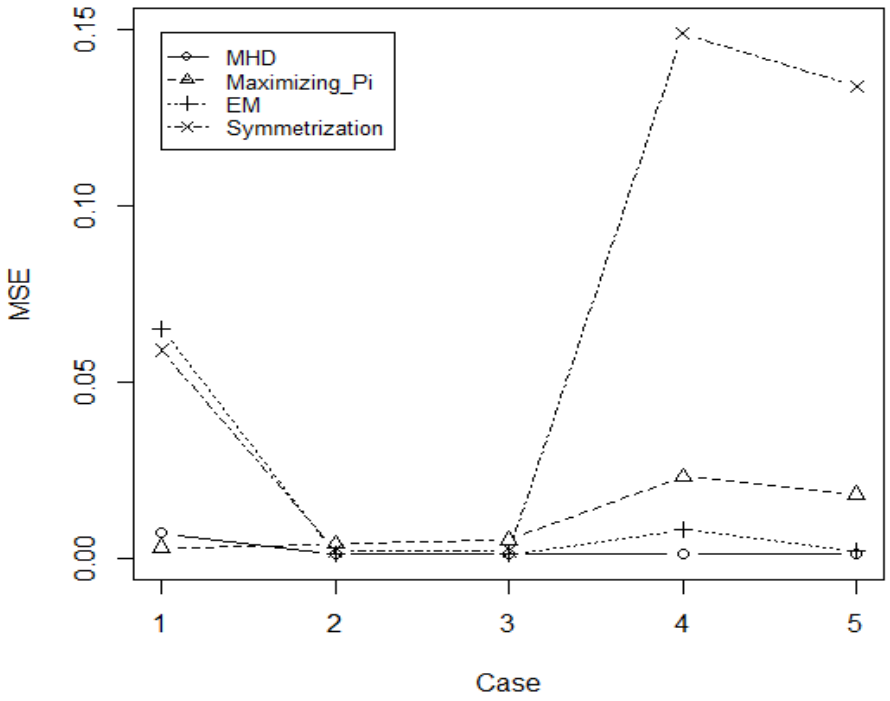

Figure 3.3: *MSE of π In The Five Cases Considered Over 200 Repetitions When $n = 1000$ under 2% Contamination From $U(10, 20)$ (σ Known)*

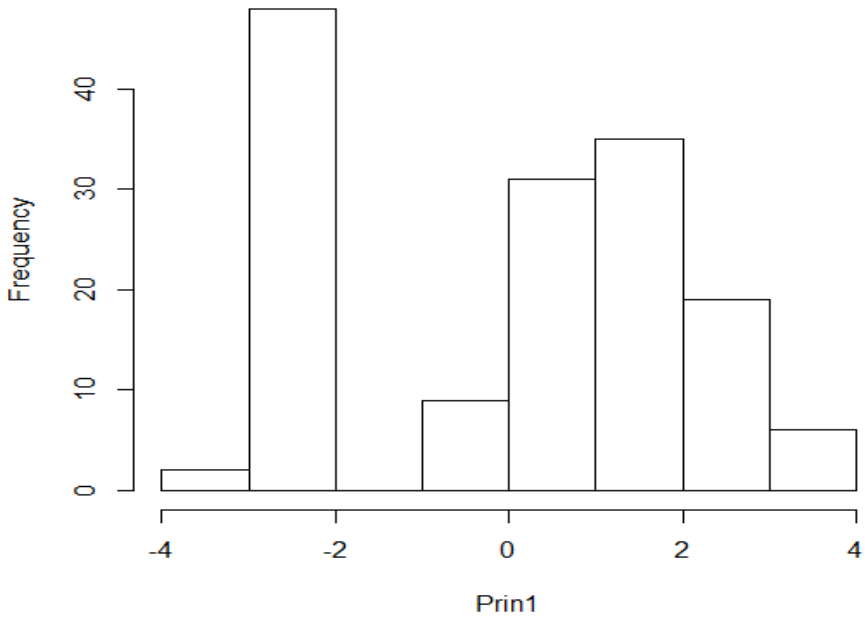

Figure 3.4: *Histogram of the first principal component in Iris data*

Chapter 4

Discussion

In this report, we introduce the Minimum Hellinger Distance estimator and review its history. We introduce a new semiparametric mixture model that completes the recent semiparametric finite mixture models introduced by Bordes *et al.*(2006) and Song *et al.*(2010). We briefly introduce the estimators suggested by Bordes *et al.*(2006) and Song *et al.*(2010), and propose a minimum Hellinger distance estimator, which has been shown to have good efficiency and robustness properties. Simulation study shows that the MHDE performs comparably to the other estimators when no contamination and outperforms them when data are under contamination.

We indicate two fields of application for our model. First, microarray data analysis, which was the initial motivation of the introduction of model (2.2) (see Bordes *et al.*(2006)). Secondly, sequential clustering algorithm, which was the initial motivation of the introduction of model (2.3) (see Song *et al.*(2010)). A real data set application considering sequential clustering algorithm is also provided to illustrate the effectiveness of our proposed methodology.

More work remains to be down on the theoretical properties of the proposed estimator, and application of the MHDE to other models, like mixture of regression models.

Bibliography

[1] Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of statistics*, 5, 445-463.

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

[3] Charlier, C. V. L., and Wicksell, S. D. (1924), On the Dissection of Frequency Functions. *Arkiv fiir Mathematik, Astronomi och Fysik*, BD 18, 6.

[4] Cutler, A. and CorderoBrana, OI. (1998). Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91, 1716-1723.

[5] Doetsch, G.(1928). Die elimination des dopplereffekts auf spektroskopische feinstrukturen und exakte bestimmung der komponenten. *Zeitschrift für Physik* 49, 705-730.

[6] Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. *unpublished qualifying paper, Harvard University, Statistics Dept.*

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

[8] Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42, 1887-1896.

[9] Izenman, A.J. (1991). Recent developments in nonparametric density-estimation. *Journal of the American Statistical Association*, 86, 205-224.

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

[27] Karunamuni, R.J. and Wu, J. (2009). Minimum Hellinger distance estimation in a nonparametric mixture model. *Journal of Statistical Planning and Inference*, 139, 1118-1133.

[12] Lindsay, B. G. (1977). Efficiency versus robustness: The case for minimum Hellinger distance estimation and related Methods. *Annals of statistics*, 22, 1081-1114.

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

[14] Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4, 51-67.

[15] N'dri, A and Hili, O (2011). Hellinger distance estimation of stationary Gaussian strongly dependent processes. *Comptes Rendus Mathematique*, 349, 991-994.

[16] Pak, R. (1996). Minimum Hellinger distance estimation in simple linear regression models; distribution and efficiency. *Statistics & probability letters*, 26, 263-269.

[17] Pearson, K. (1894). Contributions to the mathematical theory of evolution. Philosophical. *Transactions of the Royal Society of London* A 185 , 71-110.

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

[19] Simpson, D. (1989). Hellinger deviance tests - efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, 84, 107-113.

[20] Song, J. and Nicolae, D. (2009). A sequential clustering algorithm with applications to gene expression data. *Journal of the Korean Statistical Society*, 38, 175-184.

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

[22] Tamura, R. N. and Boos, D. D. (1983). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81, 223-229.

[23] Xiang, L., Yau, K.K.W., Hui, Y.V. and Lee, A.H. (2008). Minimum Hellinger distance estimation for k-component poisson mixture with random effects. *Biometrics*, 64, 508-518.

[24] Yang, S. (2008). Minimum Hellinger distance estimation of parameter in the random censorship model. *Analysis of statistics*, 19, 579-602.

[25] Ying, Z. (1992). Minimum Hellinger-type distance estimation for censored-data. *Analysis of statistics*, 20, 1361-1390.

[26] Woodward, W. A., Whitney, P. and Eslinger, P. W. (1995). Minimum Hellinger distance estimation of mixture proportions. *Journal of Statistical Planning and Inference*, 48, 303-319.

[27] Wu, J. and Karunamuni, R.J. (2009). On minimum Hellinger distance estimation. *JThe Canadian Journal of Statistics*, 37, 514-533.

[28] Wu, J., Schick, A. and Karunamuni, R.J. (2011). Profile Hellinger distance estimation. Technical Report.

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

Appendix A

Matlab Code

```
n=250; k=2; m=200; p=0.3; mu=1.5; sigma=1; sigma2=1; mhde1prop(m)=0; mhde1sig(m)=0;
mhde1mu(m)=0; mhde2prop(m)=0; mhde2sig(m)=0; mhde2mu(m)=0; tempprop(m)=0;
tempsig(m)=0; tempmu(m)=0; symmprop(m)=0; symmsig(m)=0; symmmu(m)=0; semiem-
prop(m)=0; semiemsig(m)=0; semipiprop(m)=0;
semipisig(m)=0; semipimu(m)=0; semiemmu(m)=0; mhde=[]; sym=[];
numitersemi=[]; semiemtrueprop=[]; semiemtruesig=[]; semiemtruemu=[];

for i=1:m
n1=binornd(n,p);
x1=normrnd(0,sigma,1,n1);x2=normrnd(mu,sigma2,1,n-n1);x=[x1,x2]';
temp=mixonekn(x);
mhdeest=mhdem1(x,temp,p,sigma,mu);
mhde1prop(i)=mhdeest.pi;
mhde1sig(i)=mhdeest.sigma;
mhde1mu(i)=mhdeest.mu;
mhde(i,:)=[mhdeest.initialtrue,mhdeest.numiter];
symmest=symm2(x,temp,p,sigma,mu);
symmprop(i)=symmest.pi;
symmsig(i)=symmest.sigma;
```

```
symmmu(i)=symmest.mu;
sym(i,:)=[symmest.initialtrue,symmest.numiter];
semiest=semisong(x,temp,p,sigma,mu);
semiemprop(i)=semiest.emprop;
semiemsig(i)=semiest.emsig;
semiemmu(i)=semiest.emmu;
semiemtrueprop(i)=semiest.emtrueprop;
semiemtruesig(i)=semiest.emtruesig;
semiemtruemu(i)=semiest.emtruemu;
semipiprop(i)=semiest.piprop;
semipisig(i)=semiest.pisig;
semipimu(i)=semiest.pimu;
numitersemi(i,:)=[semiest.emnumiter,semiest.emtruenumiter];
end

resmhde1.prop=[mean(mhde1prop),sqrt(var(mhde1prop)),mean((mhde1prop-p).^2)];
resmhde1.sig=[mean(mhde1sig),sqrt(var(mhde1sig)),mean((mhde1sig-sigma).^2)];
resmhde1.mu=[mean(mhde1mu),sqrt(var(mhde1mu)),mean((mhde1mu-mu).^2)]

ressemipi.prop=[mean(semipiprop),sqrt(var(semipiprop)),mean((semipiprop-p).^2)];
ressemipi.sig=[mean(semipisig),sqrt(var(semipisig)),mean((semipisig-sigma).^2)];
ressemipi.mu=[mean(semipimu),sqrt(var(semipimu)),mean((semipimu-mu).^2)]

ressemiem.prop=[mean(semiemprop),sqrt(var(semiemprop)),mean((semiemprop-p).^2)];
ressemiem.sig=[mean(semiemsig),sqrt(var(semiemsig)),mean((semiemsig-sigma).^2)];
ressemiem.mu=[mean(semiemmu),sqrt(var(semiemmu)),mean((semiemmu-mu).^2)]
```

```
ressemiemtrue.prop=[mean(semiemtrueprop),sqrt(var(semiemtrueprop)),mean((semiemtrueprop-
p).^2)];
ressemiemtrue.sig=[mean(semiemtruesig),sqrt(var(semiemtruesig)),mean((semiemtruesig-sigma).^2)];
ressemiemtrue.mu=[mean(semiemtruemu),sqrt(var(semiemtruemu)),mean((semiemtruemu-mu).^2)]

ressymm.prop=[mean(symmprop),sqrt(var(symmprop)),mean((symmprop-p).^2)];
ressymm.sig=[mean(symmsig),sqrt(var(symmsig)),mean((symmsig-sigma).^2)];
ressymm.mu=[mean(symmmu),sqrt(var(symmmu)),mean((symmmu-mu).^2)]



% Function to calculate MHDE
function[out]=mhdem1(x,ini,p,sigma,mu)
%x: the observations.
%ini: the initial values for mu and prop.
%h: the bandwidth for density estimate. h=1.06*n^(-1/5) by default.
%acc: stopping rule.
stopiter=30;
k=2;
n=length(x);
h=kdebw(x,2^14);
true=[p,sigma,mu];
if exist('ini')==0
    ini=mixnveq(x',k); end
if ini.mu(2)¿ini.mu(1)
    prop=ini.pi(1);mu=ini.mu(2);sigma=ini.sigma(1);
else
```

```
    prop=ini.pi(2);mu=ini.mu(1);sigma=ini.sigma(2);
end
est=[prop,sigma,mu];
out.tempprop=prop;out.tempmu=mu;out.tempsig=sigma;

xgridmin=min(x)-5*h;xgridmax=max(x)+5*h;lxgrid=100;
xgrid=linspace(xgridmin,xgridmax,lxgrid);hspace=(xgridmax-xgridmin)/lxgrid;
acc=10^(-5)/hspace;

%nonparametric estimator
deng=@(t) mean(exp(-(repmat(x,1,length(t))-repmat(t(:)',n,1)).^2/2/h^2))/h/sqrt(2*pi);
dengx=deng(xgrid).^(1/2);

%% calculate the MHDE using temp
dif=acc+1;numiter=0;fval=10^10;%preest=est;
%denf=@(t) normpdf(t,0,sigma);
while dif>acc && numiter<stopiter
numiter=numiter+1; pfval=fval;
denf1=@(t) normpdf(t,0,sigma);
%Find alpha and M by iteration
difa=1;step=0;a=1;
while difa> 10^(-3) && step<20
prea=a;step=step+1;mfun=@(t) a*deng(t)>prop*denf1(t);
temp=@(t) denf1(t).*mfun(t);temp1=@(t) deng(t).*mfun(t);
a=min((prop*quadl(temp,xgridmin,xgridmax)+1-prop)/max(quadl(temp1,xgridmin,xgridmax),1-
prop),1);
difa=abs(prea-a);
```

```
end
if a>0.99
a=1;
end
%% Given theta update f
denfmu=((max(a*deng(xgrid)-prop*denf1(xgrid),0))+(max(a*deng(2*mu-xgrid)-prop*denf1(2*mu-xgrid),0)))/2/(1-prop);
%assume f is symmetric 0
preest=est;
%% Given f, update theta
denf=@(t) interpcut([xgrid-mu,mu-xgrid],[denfmu,denfmu],t);

obj=@(t)sum(((min(0.95,max(t(1),0.05))*normpdf(xgrid,0,min(std(x),max(0.1*std(x),t(2))))+(1-min(0.95,max(t(1),0.05)))*denf(xgrid-min(max(x),max(0,t(3))))).^(1/2)-dengx).^2);
[est,fval]=fminsearch(obj,preest);
est=min([est;0.95,std(x),max(x)]); est=max([est;0.05,std(x)*0.1,0]);

dif=pfval-fval;
if dif<0
est=preest;fval=pfval;
end
prop=est(1);sigma=est(2);mu=est(3);
end
res.fval=fval; res.pi=prop; res.sigma=sigma; res.mu=mu; res.numiter=numiter;

%% calculate the MHDE using true
dif=acc+1; numiter=0; fval=10^10; est=true; prop=true(1); sigma=true(2); mu=true(3);
```

```
while dif>acc && numiter<stopiter
numiter=numiter+1; pfval=fval;
denf1=@(t) normpdf(t,0,sigma);
%Find alpha and M by iteration
difa=1;step=0;a=1;
while difa> 10^(−3)&&step < 20
prea=a;step=step+1;mfun=@(t) a*deng(t)>prop*denf1(t);
temp=@(t) denf1(t).*mfun(t);temp1=@(t) deng(t).*mfun(t);
a=min((prop*quadl(temp,xgridmin,xgridmax)+1-prop)/max(quadl(temp1,xgridmin,xgridmax),1-
prop),1);
difa=abs(prea-a);
end
if a>0.99
a=1;
end

%% Given theta update f
denfmu=((max(a*deng(xgrid)-prop*denf1(xgrid),0))+(max(a*deng(2*mu-xgrid)-prop*denf1(2*mu-
xgrid),0)))/2/(1-prop);
preest=est;
%% Given f, update theta
denf=@(t) interpcut([xgrid-mu,mu-xgrid],[denfmu,denfmu],t);
obj=@(t)sum(((min(0.95,max(t(1),0.05))*normpdf(xgrid,0,min(std(x),max(0.1*std(x),t(2))))+(1-
min(0.95,max(t(1),0.05)))*
denf(xgrid-min(max(x),max(0,t(3))))).^(1/2)-dengx).^2);
[est,fval]=fminsearch(obj,preest);
est=min([est;0.95,std(x),max(x)]); est=max([est;0.05,std(x)*0.1,0]);
```

```
dif=pfval-fval;
if dif<0
est=preest;fval=pfval;
end
prop=est(1);sigma=est(2);mu=est(3);
end
if res.fval<fval out.pi=res.pi; out.sigma=res.sigma; out.mu=res.mu; out.numiter=res.numiter;
out.initialtrue=0;
else
out.pi=prop;out.sigma=sigma;out.mu=mu;out.numiter=numiter;out.initialtrue=1;
end
```

```
% Function to calculate symm function[out]=symm2(x,temp,p,sig,mu) %Bordes, L. et.al,
2006. Semiparametric Estimation of a Two-component
%Assume sigma unknown
%Mixture Model where One Component is Known.
%Estimating the Euclidean parameter by symmetrization
%x: the observations.
%h: the bandwidth for density estimate. h=1.06*n^(-1/5) by default.
%acc: stopping rule.
stopiter=30;
k=2;
n=length(x);
h=kdebw(x,2^14);
```

```
true=[1-p,sig,mu];

if exist('temp')==0
temp=mixnveq(x',k);
end
if temp.mu(2)>temp.mu(1)
prop=temp.pi(2);mu=temp.mu(2);sigma=temp.sigma(1);
else
prop=temp.pi(1);mu=temp.mu(1);sigma=temp.sigma(2);
end
est=[prop,sigma,mu];

xgridmin=min(x)-5*h; xgridmax=max(x)+5*h; lxgrid=100;
xgrid=linspace(xgridmin,xgridmax,lxgrid); hspace=(xgridmax-xgridmin)/lxgrid;
acc=10^(-8)/hspace;
%nonparametric estimator
denGn=@(t) 1-mean((repmat(x,1,length(t))-repmat(t(:)',n,1))>0);
denGnx=denGn(xgrid);

%% Estimating using temp
dif=acc+1; numiter=0; fval=10^10; exitflag=1;
while dif>acc && numiter<stopiter && exitflag
numiter=numiter+1; pfval=fval;
preest=est;
obj=@(t) sum((denGn(xgrid+min(max(x),max(0,t(3))))/min(0.95,max(t(1),0.05))+(1-1/min(0.95,
max(t(1),0.05)))*normcdf(xgrid+min(max(x),max(0,t(3))),0,min(std(x),max(0.1*std(x),t(2))))
-(1-denGn(min(max(x),max(0,t(3)))-xgrid)/min(0.95,max(t(1),0.05))+(1/min(0.95,max(t(1),0.05))
```

```
-1)*normcdf(min(max(x),max(0,t(3)))-xgrid,0,min(std(x),max(0.1*std(x),t(2))))).^2);
[est,fval,exitflag]=fminsearch(obj,preest);
dif=pfval-fval;
if dif<0
est=preest;fval=pfval;
else
est=min([est;0.95,std(x),max(x)]); est=max([est;0.05,std(x)*0.1,0]);
end
if exitflag<1
est=preest;fval=pfval;
end
end
res.fval=fval; res.pi=est(1); res.sigma=est(2); res.mu=est(3); res.numiter=numiter; res.dif=dif;

%% Estimating using true
est=true;
dif=acc+1;numiter=0;fval=10^10;exitflag=1;
while dif>acc && numiter<stopiter && exitflag
numiter=numiter+1; pfval=fval;
preest=est;
obj=@(t) sum((denGn(xgrid+min(max(x),max(0,t(3))))/min(0.95,max(t(1),0.05))+(1-1/min(0.95,
max(t(1),0.05)))*normcdf(xgrid+min(max(x),max(0,t(3))),0,min(std(x),max(0.1*std(x),t(2))))
-(1-denGn(min(max(x),max(0,t(3)))-xgrid)/min(0.95,max(t(1),0.05))+(1/min(0.95,max(t(1),0.05))
-1)*normcdf(min(max(x),max(0,t(3)))-xgrid,0,min(std(x),max(0.1*std(x),t(2))))).^2);
[est,fval,exitflag]=fminsearch(obj,preest);
dif=pfval-fval;
if dif<0
```

```
est=preest;fval=pfval;
else
est=min([est;0.95,std(x),max(x)]); est=max([est;0.05,std(x)*0.1,0]);
end
if exitflag<1
est=preest;fval=pfval;
end
end
if res.fval<fval
out.pi=1-res.pi; out.sigma=res.sigma; out.mu=res.mu; out.numiter=res.numiter; out.initialtrue=0;
out.dif=res.dif;
else
out.pi=1-est(1); out.sigma=est(2); out.mu=est(3); out.numiter=numiter; out.initialtrue=1;
out.dif=dif;
end

% Function for estimators from Song's paper
function[out]=semisong(x,temp,p,sigma,mu)
%Song,S., et.al, 2010. Estimating the mixing proportion in a semiparametric
%mixture model.
%Initial method by authors
%Estimating the Euclidean parameter by symmetrization
%x: the observations.
%h: the bandwidth for density estimate. h=1.06*n^(-1/5) by default.
%acc: stopping rule.
k=2;n=length(x);
h=kdebw(x,2^14);
```

```
true=[p,sigma];
acc=10^(-4);stopiter=50;
if exist('temp')==0
temp=mixnveq(x',k);
end
if temp.mu(2)>temp.mu(1)
prop=temp.pi(1);sigma=temp.sigma(1);
else
prop=temp.pi(2);sigma=temp.sigma(2);
end
est=[prop,sigma];

denm=@(t) mean(exp(-(repmat(x,1,length(t))-repmat(t(:)',n,1)).^2/2/h^2))/h/sqrt(2*pi);

%%EM-type estimator % use EM algorithm to calculate the mle
dif=acc+1;numiter=0;fval=10^10;%preest=est;
z(n)=0;
while dif>acc && numiter<stopiter
numiter=numiter+1;
preest=est;
%% E-step
denf0=@(t) normpdf(t,0,preest(2));
z=min(1,(2*preest(1)*denf0(x)')./(preest(1)*denf0(x)'+denm(x)));

%% M-step
est(1)=mean(z);
est(2)=sqrt(z*(x.^2)/(z*ones(n,1)));
```

```
dif=max(abs(est(1)-preest(1)),abs(est(2)-preest(2)));
end

mu=(ones(1,n)-z)*x/(n-sum(z));
out.emprop=est(1); out.emsig=est(2); out.emmu=mu; out.emnumiter=numiter;

%%using true initial value
%EM-type estimator
% use EM algorithm to calculate the mle
dif=acc+1; numiter=0; fval=10^10; z(n)=0;est=true;
while dif>acc && numiter<stopiter
numiter=numiter+1;
preest=est;
%% E-step denf0=@(t) normpdf(t,0,preest(2));
z=min(1,(2*preest(1)*denf0(x)')./(preest(1)*denf0(x)'+denm(x)));

%% M-step
est(1)=mean(z);
est(2)=sqrt(z*(x.^2)/(z*ones(n,1)));
dif=max(abs(est(1)-preest(1)),abs(est(2)-preest(2)));
end
mu=(ones(1,n)-z)*x/(n-sum(z));
out.emtrueprop=est(1); out.emtruesig=est(2); out.emtruemu=mu;
out.emtruenumiter=numiter;

%% maximizing pi-type estimator
m=38;zz(n)=0;
```

```
labx=repmat(x,1,m);
sigma=0.3:0.1:4;
laby=repmat(sigma,n,1);
denmx=repmat(denm(x)',1,m);
densig=exp(-labx.^2./laby.^2/2)./laby/sqrt(2*pi);
z=denmx./densig;
val=min(z);
prop=max(val);
loc=find(val==prop);
sig=0.3+(loc-1)*0.1;
denf1=@(t) normpdf(t,0,sig);
zz=(2*prop*denf1(x)')./(prop*denf1(x)'+denm(x));
mu=(ones(1,n)-zz)*x/(n-sum(zz));
out.piprop=prop;out.pisig=sig;out.pimu=mu;
```