MINIMUM HELLINGER DISTANCE ESTIMATION OF MIXTURE PROPORTIONS

Wayne A. Woodward, Southern Methodist University
Paul Whitney, Southern Methodist University
Paul W. Eslinger, Battelle, Pacific Northwest Laboratory

SMU/DS/TR-238

Research Sponsored by DARPA
Contract No. F19628-88-K-0042

March 1990

MINIMUM HELLINGER DISTANCE ESTIMATION OF MIXTURE PROPORTIONS

Wayne A. Woodward, Paul Whitney and Paul W. Eslinger

ABSTRACT

Beran (1977) showed that, under certain restrictive conditions, the minimum distance estimator based on the Hellinger distance (MHDE) between a projection model density and a nonparametric sample density is an exception to the usual perception that a robust estimator cannot achieve full efficiency under the true model. We examine the MHDE in the case of estimation of the mixing proportion in the mixture of two normals. We discuss the practical feasibility of employing the MHDE in this setting and examine empirically its robustness properties. Our results indicate that the MHDE obtains 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.

KEY WORDS: Minimum Distance Estimation, Robustness, Simulation, Relative Efficiency

Wayne A. Woodward and Paul Whitney are in the Department of Statistical Science at Southern Methodist University. Paul W. Eslinger is employed by Battelle, Pacific Northwest Laboratory. The research of the first author was partially supported by DARPA contract number F19628-88-K-0042.

1. INTRODUCTION

Several authors have examined the estimation of the proportions p_1, p_2, ... ,p_m in the mixture density

$$f(x) = p_1f_1(x) + p_2f_2(x) + \cdots + p_mf_m(x) \tag{1.1}$$

where the component densities are specified as belonging to some parametric family, usually the normal. Hasselblad (1966), Day (1969), Hosmer (1973), Fowlkes (1979), and Redner and Walker (1984) have examined the use of maximum likelihood (ML) estimation of the parameters in (1.1) under the assumption that the component distributions are normal. Woodward et. al. (1984) investigated the use of minimum distance estimation based on a mixture-of-normals projection family and using Cramér-von Mises distance as an alternative to maximum likelihood. We denote estimates obtained in this manner as MCVMD estimates. They were able to show that the MCVMDE is more robust than the MLE to symmetric departures from the component normality such as the double exponential, $t(4)$, and $t(2)$ distributions. Not surprisingly, however, the MLE was shown to be superior to the MCVMDE when the components were normal.

Intuitively, robust procedures are those which are insensitive to small deviations from the assumptions. Donoho and Liu (1988) have shown that the class of minimum distance estimators has "automatic" robustness properties over neighborhoods of the true model based on the distance functional defining the estimator. However, robust procedures such as minimum distance estimators typically obtain this robustness at the expense of not being optimal at the true model. In fact, Bickel (1978) describes robustness as "paying a price in terms of efficiency at the (true) model in terms of reasonably good maximum MSE over the neighborhood." The behavior of the MCVMDE described above is a good example of this trade-off. However, Beran (1977) has suggested the use of the minimum Hellinger distance (MHD) estimator which has certain robustness properties and is asymptotically efficient at the true model. Although Beran suggested a computational procedure for evaluating the MHDE, he provided very limited empirical evidence concerning its performance as an estimator. Eslinger and Woodward (1990) investigated the use of the MHDE for estimation of the parameters of the normal distribution with unknown location and scale. They demonstrated the practical feasibility of employing the MHDE in the normal setting and demonstrated empirical robustness far outside Hellinger neighborhoods of the true model, and also demonstrated the true model efficiency properties shown theoretically by Beran. Tamura and Boos (1986) have investigated the performance of the MHDE in the estimation of location and covariance in multivariate data. The empirical findings of Eslinger and Woodward and of Tamura and Boos indicate that the MHDE is an attractive estimator.

In this paper we examine the use of MHD estimation in the mixture of two normals whose density is given by

$$f_\theta(x) = \frac{p}{\sqrt{2\pi}\,\sigma_1}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\} + \frac{(1-p)}{\sqrt{2\pi}\,\sigma_2}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right\} \qquad (1.2)$$

where $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, p)'$. We will let $\hat{p}(\mathrm{H})$ and $\hat{p}(\mathrm{L})$ denote the MHD and ML estimates of the parameter p. In Section 2 we provide background material on the MHDE. In Section 3 we discuss its application to (1.2) where p is unknown and the remaining parameters are known while in Section 4 we investigate the case in which all five parameters are unknown.

2. THE MINIMUM HELLINGER DISTANCE ESTIMATOR

Let $X_1, X_2, \ldots, X_n$ denote a random sample from some unknown population with distribution function G. Further, let $\mathcal{F} = \{F_\theta;\ \theta \epsilon \Theta\}$, be a family of distributions, called the projection family or projection model, depending on the (possibly vector valued) parameter θ. We will assume here that the distributions in $\mathcal{F}$ are mixtures of normals with densities of the form (1.2). A minimum distance estimator of θ is a value $\hat{\theta}$ which minimizes the distance between the data distribution and the projection model, usually by minimizing the "distance" between F_θ and G_n where G_n is the empirical distribution function

$$G_n(t) = \tfrac{1}{n} \sum_{i=1}^{n} I(X_i \leq t), \tag{2.1}$$

where I denotes the indicator function. For example the MCVDE is obtained by using Cramér-von Mises distance, w^2, which for distribution functions Q_1 and Q_2 is given by

$$w^2(Q_1, Q_2) = \int_{-\infty}^{\infty} [Q_1(x) - Q_2(x)]^2 \, dQ_2(x) \ , \tag{2.2}$$

to compute the distance between F_θ and G_n.

The Hellinger distance between two absolutely continuous distributions with distribution functions Q_1 and Q_2 is defined to be $|| q_1^{\frac{1}{2}} - q_2^{\frac{1}{2}} ||$ where q_1 and q_2 are the corresponding densities and the notation $|| \bullet ||$ denotes the usual L_2 norm, i.e.

$$|| q_1^{\frac{1}{2}} - q_2^{\frac{1}{2}} || = \left[\int \left(q_1^{\frac{1}{2}} - q_2^{\frac{1}{2}} \right)^2 \right]^{1/2} \tag{2.3}$$

where the integration is with respect to Lebesgue measure on the real line. The MHD estimator of θ is defined as a value of $\hat{\theta}_{\mathrm{H}}$ which minimizes $|| f_\theta^{\frac{1}{2}} - \hat{g}_n^{\frac{1}{2}} ||$ where $\hat{g}_n$ is a suitable nonparametric density estimator. We use the kernel density estimator

$$\hat{g}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h_n} \right) \tag{2.4}$$

based on the Epanechnikov (1969) kernel $w(x) = .75(1-x^2)$ for $| x | \leq 1$.

Parzen (1962) found the h_n which minimizes the integrated mean square error between a kernel density estimator and the true density g. The optimal h_n in this sense is $h_n = \alpha(w)\ \beta(g)\ n^{-1/5}$ where

$$\alpha(w) = \frac{\left[\int w^2(y)dy\right]^2}{\left[\int w(y)\ y^2 dy\right]^{2/5}} \tag{2.5}$$

and

$$\beta(g) = \left[\ \int \left(\frac{\partial^2 g(x)}{\partial\ x^2}\right)^2 dx\right]^{-1/5} . \tag{2.6}$$

For the Epanechnikov kernel $\alpha(w) = 1.71877$, and when $g(x)$ is a $\mathcal{N}(\mu,\sigma^2)$ density, i.e. with mean μ and variance σ^2, then $\beta(g) = 1.364\sigma$. A natural implementation of the Epanechnikov kernel density estimate is to use $h_n = (1.71877)(1.364 s_n)n^{-1/5}$ where s_n is an estimate of scale. In the case in which $g(x)$ is a mixture of normals as in (1.2), $\int \left\{\frac{\partial^2 g(x)}{\partial\ x^2}\right\}^2 dx$ is given by

$$\begin{aligned}\int \left\{\frac{\partial^2 g(x)}{\partial\ x^2}\right\}^2 dx = \int \Bigg\{ & \frac{p^2}{2\sigma_1^5\sqrt{\pi}}\ \phi\left(x;\ \mu_1,\ \sigma_1^2/2\right)\left(z_1-1\right)^2 \\ & + \frac{(1-p)^2}{2\ \sigma_2^5\ /\ \sqrt{\pi}}\ \phi\left(x;\ \mu_2,\ \sigma_2^2\ /\ 2\right)\left(z_2\ -\ 1\right)^2 \\ & + \frac{2p(1-p)}{2\pi\ \sigma_1^3\ \sigma_2^3}\ \mathrm{e}^{-\frac{1}{2}\ (z_1^2\ +\ z_2^2)}(z_1\ -\ 1)(z_2-1)\Bigg\} dx\end{aligned} \tag{2.7}$$

where $z_1 = \frac{x-\mu_1}{\sigma_1}$, $z_2 = \frac{x-\mu_2}{\sigma_2}$, and $\phi(x;\ \mu,\ \sigma^2)$ denotes the normal density function with mean μ and variance σ^2. In our implementation we used $h_n = 1.71877\ \beta(g)n^{-1/5}$ where $\beta(g)$ was obtained using numerical integration to approximate the integral in (2.7). From (2.7) it is seen that in this setting,

$\beta(g)$ depends on all five of the mixture model parameters rather than simply being a function of scale as in the univariate normal setting.

3. MHD ESTIMATION WHEN ONLY p IS UNKNOWN

(a) Theoretical Results

As a first step in examining the use of the MHDE in the mixture-of-normals setting, we consider the case in which $f_\theta(x)$ is given by (1.2) and only p is unknown. In Theorems 3.1 and 3.2 we provide conditions for which the MHD estimator in this setting is consistent and asymptotically normal. The consistency of the MHDE follows from the Hellinger consistency of the kernel density estimator together with the equivalence of the Hellinger metric on the probability distributions and the Euclidean metric on the parameter space, see Theorem 3 in Beran (1977) or Theorem 3.1 in Tamura and Boos (1986). In this section the Tamura and Boos paper will be referred to as TB. Either of these theorems implies the following:

Theorem 3.1. Let $f_\theta(x) = \theta f_1(x) + (1 - \theta)f_2(x)$, where f_1 and f_2 are distinct, continuous densities on $\mathbf{R}$, and let $\theta \in [0,1] = \Theta$. If $\hat{g}_n$ is Hellinger consistent, then the MHDE is consistent.

The asymptotic distribution of $\hat{\theta}_n$ is described in the next theorem, which is a consequence of TB's Theorem 4.1.

Theorem 3.2. Let $f_\theta(x)$ be as in Theorem 3.1, and let $\theta \in (0,1) \subset [0,1] = \Theta$. Denote by $\hat{\theta}_n$ the MHDE of θ based on a random sample of size n from a population with density f_θ. Also suppose:

1. $\int|x|^k f_1(x)dx< \infty$ and $\int|x|^k f_2(x)dx < \infty$ for every $k > 0$.
2. $\lim_{|x|\to\infty} f_i(x) = 0,\ i = 1,2.$

3. f_1 and f_2 satisfy Condition 5 from TB's Theorem 4.1.

4. The bandwidth for the kernel, h_n, satisfies $h_n = an^{-c}$ for some $c \in (0, 1/4)$ and $a>0$.

Then $\sqrt{n}(\hat{\theta}_n - \theta - B_n) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$, where $I(\theta)$ is the information matrix and B_n is given by

$$B_n = 2\ C_n^* \int \psi_\theta \sqrt{f_\theta} \left(\sqrt{\tilde{f}_{2n}} - \sqrt{f_\theta}\right) \quad \text{and} \quad C_n^* \xrightarrow{p} 0$$

where $\mathrm{E}[\hat{g}_n] = \tilde{g}_n$.

As a result of Theorem 3.2 we see that $\hat{\theta}_n$ is asymptotically fully efficient. Our utilization of these results will be to the case in which f_θ is the mixture of normals in (1.2) with μ_1, σ_1, μ_2, and σ_2 known, and as mentioned earlier, we will use the notation $\hat{p}$(H) for $\hat{\theta}_n$.

(b) Implementation Details

The estimates may be obtained by minimizing $-\int f_\theta^{\frac{1}{2}} \hat{g}_n^{\frac{1}{2}}$ over $\theta \in [0,1]$. This minimization was performed using a golden section search as described in Press, et. a. (1986). The starting values for this optimization were obtained by examining the values of the integral over a grid of θ's on [0,1]; the optimization routine was always started in an interval which contained the global minimum of the quantity over the grid values. The integral was estimated using Simpson's rule with a mesh of 201 points over the support of $\hat{g}_n$. The bandwidth of the estimate $\hat{g}_n$ was obtained by plugging into (2.7) the known μ_1, σ_1, μ_2, and σ_2 along with the mixing proportion estimated by the quasi-clustering technique in Woodward et. al. (1984).

(c) Simulation Results

Simulations were run in order to examine empirically the theoretical results of this section using the parameter configurations employed by Woodward, et. al. (1984). Simulations reported in this section and the next are based on mixing proportions .25, .5 and .75. For each of these mixing proportions, we considered mixtures of the densities $f_1(x)$ and $f_2(x)$ where $f_1(x)$ is the density for the random variable $X = aY$ and $f_2(x)$ is the density associated with $X = Y + b$ where $a>0$ and $b>0$. Thus, a is the ratio of scale parameters which we take to be 1 and $\surd 2$ while b was selected to provide the desired overlap between the two distributions. We considered "overlaps", as defined by Woodward, et. al. (1984) of .03 and .1. In this section we consider the case in which Y is normally distributed. For each set of configurations considered, 500 samples of size n=100 were generated from the corresponding mixture distribution, and for each sample considered, the ML and MHD estimates were obtained. In Table 3.1 we present the results of the simulations, showing simulation-based estimates of the bias and MSE given by

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

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

where n_s denotes the number of samples (500 in our case) and $\hat{p}_i$ denotes an estimate of p for the ith sample. In the tables we report $n\hat{\text{MSE}}$ where n is the size of each sample (n = 100 in our case), and in all cases, an approximate standard error of a tabled $n\hat{\text{MSE}}$ is (.0632)($n\hat{\text{MSE}}$). We also table empirical measures of the relative efficiencies of the MHDE with the MLE, i.e.

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

Examination of the table shows that the asymptotic full efficiency with respect to the MLE guaranteed by Theorem 3.2 holds approximately in the current setting with n=100 as evidenced by the fact that all $\hat{\text{E}}$ values are near 1. In Figure 1 we show a normal probability plot of $\hat{p}_i$(H) and $\hat{p}_i$(L), i = 1, ... , 500, obtained in the simulation for the case $p = .25$, ratio of scale parameters = 1 and overlap = .1. There it can be seen that the sampling distribution for each estimator closely approximates a normal curve.

It should be noted that the asymptotic result in Theorem 3.2 is for the case in which the bandwidth h_n is nonstochastic. In our implementation this bandwidth is random since it depends on the starting value estimate of the parameter p. The simulations indicate that the results hold in this case.

4. MHD ESTIMATION WHEN p, μ_1, σ_1, μ_2 AND σ_2 ARE UNKNOWN

We consider in this section the case in which the five parameters p, μ_1, σ_1, μ_2 and σ_2 in (1.2) are all unknown, and we will again compare the MHD estimators with maximum likelihood. It is well-known that the likelihood function is not bounded in this case (see Day 1969), and thus "ML" estimators in this setting are obtained by finding an appropriate local maximum. We will empirically compare the MHD and ML estimators in this setting using a large-scale simulation analysis in which we examine the efficiency and robustness of the estimators.

(a) Implementation Details

Since minimizing $|| f_\theta^{\frac{1}{2}} - \hat{g}_n^{\frac{1}{2}} ||$ is equivalent to maximizing

$$\int f_\theta^{\frac{1}{2}} \hat{g}_n^{\frac{1}{2}} , \tag{4.1}$$

Beran (1977) and Eslinger and Woodward (1990) obtained MHD estimates by using Newton's method to maximize (4.1). One advantage of this approach is the fact that $\hat{g}_n$ is zero outside a finite interval, simplifying the integration in (4.1). Woodward and Eslinger (1983) investigated the corresponding use of Newton's method in the mixture-of-normals case with starting values for the iteration being obtained using the quasi-clustering technique discussed by Woodward et. al. (1984) for obtaining starting values of the mixture model parameters. However, they found that Newton's method in this setting often failed to converge to reasonable estimates, with convergence occurring in less than 80% of the simulated samples from some configurations. Since the MHDE, $\hat{\theta}_{\mathrm{H}}$, is defined to be a value which minimizes the integral

$$I = \int \left\{ f_{\theta}^{\frac{1}{2}} - \hat{g}_{n}^{\frac{1}{2}} \right\}^2, \tag{4.2}$$

we approximated this integral using the trapezoidal rule to obtain

$$\hat{I} = \Delta t_i \sum_{i=1}^{k} a_i \left(f_{\theta}^{\frac{1}{2}}\left(t_i\right) - \hat{g}_{n}^{\frac{1}{2}}\left(t_i\right) \right)^2 \tag{4.3}$$

where $a_1 = a_k = \frac{1}{2}$ and $a_i = 1$ for $i = 2, 3, \ldots, k-1$ for a partition $t_1, t_2, \ldots, t_k$ of $[a,b]$, a finite interval. In our case we took $k = 200$ and $[a,b]$ to be the interval $[X_{(1)} - 3, X_{(n)} + 3]$ where $X_{(j)}$ denotes the jth order statistic. The procedure employed was to minimize the sum-of-squares in (4.3) using IMSL routine ZXSSQ which utilizes the Marquardt-Levenberg algorithm (1963). Using this procedure, the MHD estimates converged in at least 97.8% of the samples for each configuration considered. In the simulations, if convergence to "reasonable" values was not obtained, the starting values were used as the corresponding estimates. Specifically, if any of the conditions $\hat{\sigma}_1 > X_{(n)} - X_{(1)}$, $\hat{\sigma}_2 > X_{(n)} - X_{(1)}$, $\hat{\mu}_1 < X_{(1)} - (X_{(n)} - X_{(1)})/10$ or $\hat{\mu}_2 > X_{(n)} + (X_{(n)} - X_{(1)})/10$ for any estimate, the corresponding estimate was taken to be the starting value. The kernel density estimate $\hat{g}_n$ was obtained using the Epanechnikov kernel. In this case $\beta(g)$ was obtained by substituting the starting values for μ_1, σ_1, μ_2, σ_2, and p into (2.7) and then performing the required integration numerically.

(b) Simulation Results

The MLE and MHDE estimates were examined using simulations based on the basic framework used in Section 3, i.e. we considered the same mixing proportions, ratios of scale parameters and overlaps as considered there. As before, 500 samples of size n=100 were generated from the corresponding mixture distributions, and we considered the cases in which the simulated component densities were normal, $t(4)$ and $t(2)$. For each sample considered, we computed the ML, MHD and MCVMD estimates initialized employing the quasi-clustering technique used by Woodward et. al. (1984). In Table 4.1 the simulation results for simulated mixtures of normal distributions indicate that again, as in the results of Section 3, the MHDE appears to obtain full efficiency at the true model as evidenced by $\hat{E}$ near one in all cases. However, the MCVMD estimators had larger $\hat{MSE}$'s than did the MLE in 9 of the 10 cases with some of the efficiencies near .5. In Table 4.2 we show similar results for samples which were simulated as mixtures of $t(4)$ components. All of the $\hat{E}$'s in this table are greater than one providing evidence that the MCVMDE and MHDE are more robust to the departures from the assumption of normal components than is the MLE. Also, comparison of the $\hat{MSE}$'s for the MHD and MCVMD estimators indicate that the robustness of the MHDE is comparable to that of the MCVMDE in this setting. In Table 4.3 we briefly consider the case in which the component distributions are $t(2)$, i.e. the departure from normality is more extreme. In this setting the performance of the MLE further deteriorates with respect to that of the two minimum distance estimators.

Although theoretical results similar to Theorem 3.1 and 3.2 have not been shown in this case, the simulation results suggest that such results hold. Although our emphasis here has been on the estimation of the mixing proportion, p, the ML and MHD routines used here obtain estimates for all five of the parameters in (1.2). The results for location and scale parameters are similar to those shown here for the mixing proportion when sampling from normal mixtures. In the case of simulations from the non-normal components considered here, the results for the location parameters also exhibited patterns similar to those shown in Tables 4.2-4.3. However, the scale estimates obtained by all three estimation methods often exhibited substantial bias in these non-normal cases.

5. CONCLUDING REMARKS

Our results indicate that the MHDE obtains full efficiency at the true model while performing comparably with the MCVMDE under the symmetric departures from component normality considered. Thus, the MHDE is a very attractive alternative to both the MLE and the MCVMDE in these settings.

The computation of the MHDE in this setting is quite straightforward, yet in the cases considered here, it took from 1.5 to 5 times longer to calculate than the MLE and about 2.5 times longer than the MCVMDE. However, Eslinger and Woodward (1990) have shown that for very large sample sizes, the MHDE can be faster to compute than competing estimators because of the fact that it requires only one pass through the data to evaluate the kernel density estimator at the appropriate grid points for numerical integration.

As would be expected, the performance of the estimators declines as the overlap between the two components increases. The sensitivity to overlap was more extreme in the case in which all five parameters are unknown since the location and scale of each component must then be estimated from the data. Estimation in the case in which all five parameters are unknown can be a difficult problem when there is not substantial separation between the components. In our simulations, the estimators were quite poor at .1 overlap when all five parameters were estimated. In fact, in these cases the starting values often outperformed the maximum likelihood and minimum distance estimators. This behavior has been previously observed by Woodward, et. al. (1984) and Woodward and Gunst (1987).

APPENDIX

Proof of Theorem 3.2: The proof proceeds by verifying conditions 1-7 of TB's Theorem 4.1. We begin by setting up the notation.

Let f_1, f_2 be continuous densities on $\mathbf{R}$, and for $\theta \in [0, 1] = \Theta$ let $f_\theta = \theta f_1 + (1-\theta)f_2$, so that f_θ is a simple mixture of f_1 and f_2. As in TB, we let

$$s_\theta = \sqrt{f_\theta}\ ,$$

$$\dot{s}_\theta = \frac{\partial}{\partial\theta}\, s_\theta = \tfrac{1}{2}\, f_\theta^{-1/2}\ (f_1-f_2)\ I_{\text{supp } f_\theta}\ ,$$

$$\ddot{s}_\theta = -\tfrac{1}{4}\, f_\theta^{-3/2}\ (f_1-f_2)^2\ I_{\text{supp } f_\theta}\ ,$$

and

$$\psi_\theta(x) = -\left[\int \ddot{s}_\theta(y)\, f_\theta^{1/2}(y)\ dy\right]^{-1} \frac{\dot{s}_\theta(x)}{2\, f_\theta^{1/2}(x)}\ I_{\text{supp } f_\theta(x)}$$

$$= \frac{1}{I(\theta)}\ \frac{f_1(x)-f_2(x)}{f_\theta(x)}\ I_{\text{supp } f_\theta(x)}\ ,$$

where $I_{\text{supp } f(x)}$ denotes the indicator function of the support of $f(x)$ and where $I(\theta)$ is the Fisher Information which is in this case equal to

$$\int \frac{(f_1-f_2)^2}{f_\theta}\quad .$$

Note that $I(\theta) > 0$ if f_1 and f_2 are not equal. Also, if $\theta \in (0,1)$, $I(\theta) < \infty$, since

$$\int \frac{(f_1-f_2)^2}{f_\theta} \leq 2 \int \frac{f_1{}^2+f_2{}^2}{f_\theta} \leq 2 \int \frac{f_1{}^2}{\theta f_1} + 2 \int \frac{f_2{}^2}{\theta f_2}$$

$$= 2\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right).$$

Finally, note that we often drop the constant a in the bandwidth $h_n = an^{-c}$; this does not effect the result.

TB.1 The conditions on the kernel are satisfied by the Epanechnikov kernel (symmetric, compact support $= [-1, 1]$). Our condition on h_n implies $nh_n \to \infty$ and $h_n \to 0$.

TB.2 Condition 2.b holds: $\Theta = [0, 1]$ is compact, $f_\theta(x)$ is continuous in θ and $\theta_1 \neq \theta_2 \Rightarrow f_{\theta_1} \neq f_{\theta_2}$ on a set of positive Lebesgue measure.

TB.3 Let $\alpha_n = h_n^{-1}$ and let $X \sim f_\theta$, $X_{f_1} \sim f_1$ and $X_{f_1} \sim f_1$. Then for $t \in [-1, 1]$,

$$n \operatorname{Prob}_\theta \left\{| X - h_n t | > \alpha_n\right\}$$

$$= n\, \theta \operatorname{Prob}_{f_1} \left\{| X_{f_1} - h_n t | > \alpha_n\right\}$$

$$+ n(1-\theta) \operatorname{Prob}_{f_2} \left\{ | X_{f_2} - h_n t | > \alpha_n\right\}$$

$$\leq n\, \theta \operatorname{E}_{f_1} | X_{f_1} - h_n t |^k / \alpha_n^k$$

$$+ n(1-\theta)\, E_{f_2} | X_{f_2} - h_n t |^k / \alpha_n^k .$$

Since $h_n \to 0$ as $n \to \infty$ and E $| X - h_n t |^k$ is a continuous function of $h_n t$ (this is particularly easy to see for k an even integer), $E_{f_1} | X_{f_1} - h_n t |^k$ and $E_{f_2} | X_{f_2} - h_n t |^k \to E_{f_1} | X_{f_1} |^k$ and $E_{f_2} | X_{f_2} |^k$, respectively, uniformly over $t \in [-1,1]$. Thus,

$$n \sup_{t\in[-1,1]} \mathrm{Prob}_{\theta}\left\{ | X - h_n t | > \alpha_n \right\} \le O\left(n\, \alpha_n^{-k}\right) .$$

A choice of k can be made so that $n\alpha_n^{-k} \to 0$ as $n \to \infty$.

TB.4 We examine

$$\left(n^{1/2}\, n^{-c}\right)^{-1} \int_{-n^c}^{n^c} \left| \frac{f_1 - f_2}{f_\theta} \right|$$

$$\le n^{c-\frac{1}{2}} \int_{-n^c}^{n^c} \frac{f_1}{f_\theta} + n^{c-\frac{1}{2}} \int_{-n^c}^{n^c} \frac{f_2}{f_\theta}$$

$$\le n^{c-\frac{1}{2}}\, 2n^c\, \frac{1}{\theta} + n^{c-\frac{1}{2}}\, 2n^c\, \frac{1}{1\text{-}\theta} .$$

This converges to zero since $0 < c < \frac{1}{4}$.

TB.5 We must show

$$\sup_{|x| \le n^c}\ \sup_{t \in [-1,1]} \frac{f_\theta(x + h_n t)}{f_\theta(x)} = O(1) .$$

Note

$$\frac{f_\theta(x + h_n t)}{f_\theta(x)} = \theta\,\frac{f_1(x + h_n t)}{f_\theta(x)} + (1-\theta)\,\frac{f_2(x + h_n t)}{f_\theta(x)}$$

$$\le \frac{f_1(x + h_n t)}{f_1(x)}\,I_{\text{supp } f_1(x)} + \frac{f_2(x + h_n t)}{f_2(x)}\,I_{\text{supp } f_2(x)}.$$

The result follows from our Condition 3.

TB.6

1. $\int \psi_\theta^2\, f_\theta = I(\theta)^{-1} < \infty$, since $f_1 \neq f_2$.

2. $$\int \psi_\theta^2(x + a)\, f_\theta(x)\; dx = I(\theta)^{-2} \int \left(\frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)}\right)^2 f_\theta(x)\; dx$$

$$\le I(\theta)^{-2} \int \left(\frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)}\right)^2 f_\theta(x)\; dx$$

$$\le I(\theta)^{-2} \int \left(\frac{1}{\theta} + \frac{1}{1-\theta}\right)^2 f_\theta(x)\; dx < \infty,$$

independent of *a*.

3. $$\int \Big(\psi_\theta(x + a) - \psi_\theta(x)\Big)^2 f_\theta(x)\; dx$$

$$= I(\theta)^{-2} \int \left(\frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)/f_\theta(x)^{1/4}} - \frac{f_1(x) - f_2(x)}{f_\theta(x)/f_\theta(x)^{1/4}}\right)^2 \sqrt{f_\theta(x)}\; dx$$

$$\leq I(\theta)^{-2} \,||\, f_\theta(x)^{1/4} \left(\frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)} - \frac{f_1(x) - f_2(x)}{f_\theta(x)}\right) ||_\infty^2 \int \sqrt{f_\theta}\,.$$

The integral $\int\sqrt{f_\theta} < \infty$ since the tails of f_1 and f_2 (and so f_θ) decrease faster than n^{-b}, for any $b > 0$. Thus, we need to show the L_∞ norm goes to 0 as $a \to 0$. To see this, note first that both

$$\left| \frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)} \right| \leq \frac{1}{\theta} + \frac{1}{1-\theta}$$

and

$$\left| \frac{f_1(x) - f_2(x)}{f_\theta(x)} \right| \leq \frac{1}{\theta} + \frac{1}{1-\theta}\,.$$

Thus

$$\left| \frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)} \right| f_\theta(x)^{1/4} \leq \left(\frac{1}{\theta} + \frac{1}{1-\theta} \right) f_\theta(x)^{1/4} \to 0$$

as $| x | \to \infty$ by Condition 2. Given $\epsilon > 0$ $\exists$ $M > 0$ such that $\forall$ $| x | > M$ and any $a \in \mathbf{R}$,

$$\left| \frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)} \right| f_\theta(x)^{1/4} < \epsilon/2.$$

For $| x | \leq M$,

$$\left(\frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)} - \frac{f_1(x) - f_2(x)}{f_\theta(x)} \right) f_\theta(x)^{1/4} \quad \to 0$$

as $a \to 0$ uniformly over $| x | \leq M$. So $\exists$ $\delta > 0$ with $|a| < \delta$ implying

$$|| f_\theta^{1/4}(x) \left(\frac{f_1(x+a) - f_2(x+a)}{f_\theta(x+a)} - \frac{f_1(x) - f_2(x)}{f_\theta(x)} \right) ||_\infty < \epsilon \ .$$

TB.7

1. Lemma 1 of Beran

(i) $\frac{\partial}{\partial \theta} s_\theta(x) = \frac{1}{2} f_\theta^{-1/2}(f_1 - f_2)$, which is continuous in θ $\forall$ $x \epsilon$ supp f_θ.

(ii) We need to show $|| \dot{s}_\theta ||$ is continuous. We will show the stronger condition, that $\dot{s}_\theta$ is L_2 continuous.

First note

$$\int \dot{s}_\theta^2 = \tfrac{1}{4} I(\theta) < \infty, \text{ so that } \dot{s}_\theta \ \epsilon \ L_2 \ .$$

We now compare $\dot{s}_\theta$ and $\dot{s}_{\theta+\Delta\theta}$:

$$\int \left(\dot{s}_\theta - \dot{s}_{\theta+\Delta\theta} \right)^2 = (\Delta\theta)^2 \int (f_1 - f_2)^4 \frac{1}{f_\theta f_{\theta+\Delta\theta} \left(\sqrt{f_\theta} + \sqrt{f_{\theta+\Delta\theta}} \right)^2}$$

$$\leq (\Delta\theta)^2 \int (f_1 - f_2)^4 \frac{1}{f_\theta f_{\theta+\Delta\theta} f_\theta}$$

$$\leq (\Delta\theta)^2 \int \frac{f_1 + f_2}{f_\theta} \ \frac{f_1 + f_2}{f_{\theta+\Delta\theta}} \ \frac{(f_1 - f_2)^2}{f_\theta}$$

$$\le (\Delta\theta)^2 \left(\frac{1}{\theta} + \frac{1}{1-\theta}\right)\left(\frac{1}{\theta+\Delta\theta} + \frac{1}{1-(\theta+\Delta\theta)}\right) \int \frac{(f_1-f_2)^2}{f_\theta}$$

which converges to zero as $\Delta\theta \to 0$.

2. Lemma 2 of Beran:

(i) $\ddot{s}_\theta = -\frac{1}{4} f_\theta^{-3/2} (f_1-f_2)^2$, which is continuous in θ $\forall$ $x \in$ supp f_θ.

(ii) To show $\ddot{s}_\theta \in \mathrm{L}_2$ and $||\ddot{s}_\theta||$ is continuous, we will show that $\ddot{s}_\theta$ is in fact L_2 continuous.

First note

$$\int \left(\ddot{s}_\theta\right)^2 = \frac{1}{16} \int \frac{(f_1-f_2)^4}{f_\theta^3}$$

$$\le \frac{1}{16} \int \frac{(f_1-f_2)^2}{f_\theta} \frac{(f_1+f_2)^2}{f_\theta^2}$$

$$\le \frac{1}{16} \left(\frac{1}{\theta} + \frac{1}{1-\theta}\right)^2 \int \frac{(f_1-f_2)^2}{f_\theta} < \infty \, .$$

Next we argue that $\ddot{s}$ is L_2 continuous:

$$\int \left(\frac{(f_1-f_2)^2}{f_\theta^{3/2}} - \frac{(f_1-f_2)^2}{f_{\theta+\Delta\theta}^{3/2}}\right)^2$$

$$\le 2 \int (f_1-f_2)^4 \left(\frac{1}{f_\theta^{3/2}} - \frac{1}{f_\theta f_{\theta+\Delta\theta}^{1/2}}\right)^2$$

$$+ 2 \int (f_1 - f_2)^4 \left(\frac{1}{f_\theta f^{1/2}_{\theta+\Delta\theta}} - \frac{1}{f^{3/2}_{\theta+\Delta\theta}} \right)^2$$

$$= 2 \int \frac{(f_1 - f_2)^4}{f_\theta^2} \left(\frac{\Delta\theta \; (f_1 - f_2)}{\sqrt{f_\theta f_{\theta+\Delta\theta}} \left(\sqrt{f_\theta} + \sqrt{f_{\theta+\Delta\theta}} \right)} \right)^2$$

$$+ 2 \int \frac{(f_1 - f_2)^4}{f_{\theta+\Delta\theta}} \left(\frac{\Delta\theta \; (f_1 - f_2)}{f_\theta f_{\theta+\Delta\theta}} \right)^2$$

From here, one may proceed as in Lemma 1 part (ii).

3. $\theta = t(f_\theta) \; \epsilon \; (0,1)$, since $\theta \; \epsilon \; (0,1)$.

4. $\int \ddot{s}_\theta \, f_\theta^{1/2} = I(\theta)^{-1} < \infty$.

REFERENCES

Beran, R. (1977), "Minimum Hellinger Distance Estimates for Parametric Models," *Annals of Statistics* 5, 445-463.

Bickel, P. J. (1978), "Some Recent Developments in Robust Statistcs," Paper presented at the Fourth Australian Statistical Conference.

Day, N.E. (1969), "Estimating the Components of a Mixture of Normal Distributions," *Biometrika* 56, 463-474.

Donoho, D. L. and Liu, R. C. (1988),"The 'Automatic' Robustness of Minimium Distance Functionals," *Annals of Statistics* 16, 552-586.

Epanechnikov, V. A. (1969), "Non-parametric Estimation of a Multivariate Probability Density," *Theory of Probability and its Applications XIV*, 153-158.

Eslinger, P. W. and Woodward, W. A. (1990), "Minimum Hellnger Distance Estimation for Normal Models," under revision for *Journal of Statistical Computation and Simulation.*

Fowlkes, E. B. (1979), "Some Methods for Studying the Mixture of Two Normal (Lognormal) Distributions," *Journal of the American Statistical Association* 74, 561-575.

Hasselblad, V. A. (1966), "Estimation of Parameters for a Mixture of Normal Distributions," *Technometrics* 8, 431-446.

Hosmer, D. W. (1973), "A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of Two Normal Distributions Under Three Different Types of Samples," *Biometrics* 29, 761-770.

Marquardt, D. W. (1963), "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *Journal of the Society for Industrial Engineers* 11, 431-441.

Parzen, E. (1962), "On Estimation of a Probability Density Function and its Mode," *Annals of Mathematical Statistics* 33, 1065-1076.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge: Cambridge University Press.

Redner, R. A. and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review* 26, 195-239.

Tamura, R. N. and Boos, D. D. (1986), "Minimum Hellinger Distance Estimation for Multivariate Location and Covariance," *Journal of the American Statistical Association* 81, 223-229.

Woodward, W. A. and Eslinger, P. W. (1983), "Minimum Hellinger Distance Estimation of Mixture Model Parameters," NASA Technical Report SR-63-04433, July 1983.

Woodward, W. A. and Gunst, R. F. (1987), "Using Mixtures of Weibull Distributions to Estimate Mixing Proportions," *Computational Statistics and Data Analysis* 5,163-176.

Woodward, W. A., Parr, W. C., Schucany, W. R., and Lindsay, H. (1984), "A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion," *Journal of the American Statistical Association* 79, 590-598.

Table 3.1 Simulation Results for Mixtures of Normal Components With Only p Unknown

Sample Size = 100
Number of Samples = 500

			.10 Overlap			.03 Overlap		
p	Ratio of Scale Factors (a)	Estimator	Biâs	nMŜE	Ê	Biâs	nMŜE	Ê
.25	1	MHDE	.011	.297	1.04	−.000	.212	.99
		MLE	−.003	.310		−.002	.209	
.50	1	MHDE	.000	.309	1.11	.003	.281	1.00
		MLE	.000	.343		.002	.280	
.25	$\sqrt{2}$	MHDE	.010	.311	.96	.000	.207	.98
		MLE	.001	.299		−.001	.203	
.50	$\sqrt{2}$	MHDE	.002	.315	1.05	.003	.302	.99
		MLE	−.002	.332		.001	.300	
.75	$\sqrt{2}$	MHDE	−.010	.297	1.07	−.000	.216	1.03
		MLE	−.002	.319		−.001	.222	

Table 4.1 Simulation Results for Mixtures of Normal Components
With All 5 Parameters Unknown

Sample Size = 100
Number of Samples = 500

	Ratio of		.10 Overlap			.03 Overlap		
p	Scale Factors (a)	Estimator	Biâs	nMŜE	Ê	Biâs	nMŜE	Ê
.25	1	MHDE	.064	4.723	1.06	.006	.435	1.03
		MCVMDE	.142	8.944	.56	.028	1.029	.44
		MLE	.063	5.003		.088	.449	
.50	1	MHDE	.009	2.733	1.16	.005	.403	1.02
		MCVMDE	−.009	3.683	.86	.004	.440	.94
		MLE	.007	3.158		.004	.412	
.25	$\sqrt{2}$	MHDE	−.006	2.005	1.06	−.003	.383	1.25
		MCVMDE	.080	5.228	.40	.019	.831	.58
		MLE	−.005	2.117		.005	.479	
.50	$\sqrt{2}$	MHDE	−.021	2.005	1.29	−.006	.376	1.07
		MCVMDE	.005	2.951	.88	−.000	.393	1.02
		MLE	−.014	2.584		−.002	.402	
.75	$\sqrt{2}$	MHDE	−.073	4.660	1.07	−.003	.396	1.29
		MCVMDE	−.119	7.742	.64	−.022	1.020	.50
		MLE	−.077	4.993		−.002	.512	

Table 4.2 Simulation Results for Mixtures of $t(4)$ Components
With All 5 Parameters Unknown

Sample Size = 100
Number of Samples = 500

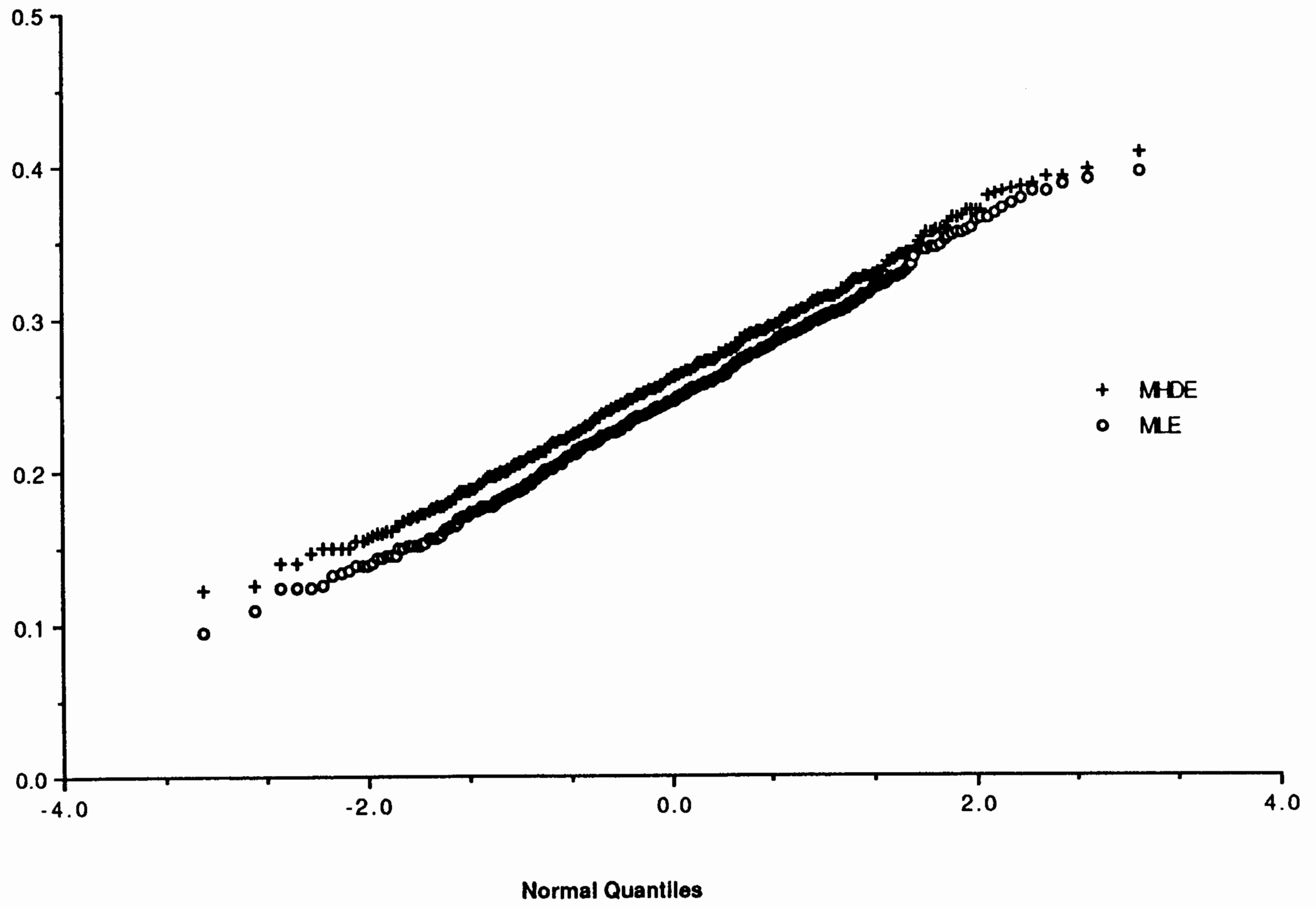
			.10 Overlap			.03 Overlap		
p	Ratio of Scale Factors (a)	Estimator	Biâs	nMŜE	Ê	Biâs	nMŜE	Ê
.25	1	MHDE	.056	4.862	1.18	.015	.297	2.77
		MCVMDE	.066	4.144	1.38	.023	.428	1.92
		MLE	.069	5.725		.035	.823	
.50	1	MHDE	.002	3.489	1.56	.000	.314	1.51
		MCVMDE	.003	1.855	2.94	.001	.301	1.57
		MLE	.024	5.457		.003	.473	
.25	$\sqrt{2}$	MHDE	.076	4.348	1.17	.014	.404	2.48
		MCVMDE	.095	4.968	1.02	.031	.652	1.54
		MLE	.090	5.080		.046	1.003	
.50	$\sqrt{2}$	MHDE	.039	3.300	1.52	−.003	.250	1.82
		MCVMDE	.025	1.978	2.54	−.000	.254	1.80
		MLE	.024	5.030		.009	.456	
.75	$\sqrt{2}$	MHDE	−.031	4.780	1.77	−.012	.273	1.90
		MCVMDE	−.055	4.045	2.10	−.019	.396	1.31
		MLE	−.078	8.483		−.014	.519	

Table 4.3 Simulation Results for Mixtures of $t(2)$ Components
With All 5 Parameters Unknown

Sample Size = 100
Number of Samples = 500

			.10 Overlap			.03 Overlap		
p	Ratio of Scale Factors (a)	Estimator	Biâs	nMŜE	Ê	Biâs	nMŜE	Ê
.25	1	MHDE	.123	6.996	1.14	.013	.257	6.05
		MCVMDE	.079	3.745	2.13	.024	.328	4.74
		MLE	.097	7.962		.069	1.555	
.50	1	MHDE	−.007	4.547	2.20	−.002	.285	2.96
		MCVMDE	.006	1.172	8.55	−.002	.282	2.99
		MLE	−.003	10.016		.004	.843	

Figure 1: Normal Probability Plots of MLE and MHDE Estimates


Normal Quantiles
+ MHDE
o MLE