Ann. Inst. Statist. Math.
Vol. 46, No. 4, 683–705 (1994)

MINIMUM DISPARITY ESTIMATION FOR CONTINUOUS MODELS: EFFICIENCY, DISTRIBUTIONS AND ROBUSTNESS

AYANENDRANATH BASU[1] AND BRUCE G. LINDSAY[2]

[1] *Department of Mathematics, University of Texas at Austin, Austin, TX 78712-1082, U.S.A.*
[2] *Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.*

(Received August 9, 1993; revised March 3, 1994)

Abstract. A general class of minimum distance estimators for continuous models called minimum disparity estimators are introduced. The conventional technique is to minimize a distance between a kernel density estimator and the model density. A new approach is introduced here in which the model and the data are smoothed with the same kernel. This makes the methods consistent and asymptotically normal independently of the value of the smoothing parameter; convergence properties of the kernel density estimate are no longer necessary. All the minimum distance estimators considered are shown to be first order efficient provided the kernel is chosen appropriately. Different minimum disparity estimators are compared based on their characterizing residual adjustment function (RAF); this function shows that the robustness features of the estimators can be explained by the shrinkage of certain residuals towards zero. The value of the second derivative of the RAF at zero, A_2, provides the trade-off between efficiency and robustness. The above properties are demonstrated both by theorems and by simulations.

Key words and phrases: Disparity, Hellinger distance, Pearson residuals, MLE^*, robustness, efficiency, transparent kernels.

1. Introduction and overview

A pioneering work by Beran (1977) showed that by using minimum Hellinger distance estimators one could obtain robustness properties together with first order efficiency. Further investigation of this idea came from Tamura and Boos (1986) and Simpson (1987, 1989). The present paper continues this line of work, in the process extending the general minimum disparity approach of Lindsay (1994), who extended the range of choice beyond the Hellinger distance. The latter work deals only with multinomial models; here we widen the technique to include continuous models, with emphasis on the multivariate normal where we show how to jointly estimate μ and Σ robustly without loss in first order efficiency. We will refer to our estimators as the minimum disparity estimators (MDEs).

The methods described herein differ from previous work in this area in two key ways:

- Unlike previous Hellinger distance approaches, our procedures are constructed in such a way that we do not require consistency or rate of convergence results for the nonparametric density estimators. However we can still obtain first order efficiency with robustness.
- Additionally our construction of minimum distance procedures allows simple parametric adjustment of possible tradeoffs between efficiency and robustness features, just as one can do with tuning constants in M-estimation. However, unlike M-estimation, the minimum distance methods are first order efficient and applicable to a wide range of models, not just to location scale models.

In addition, we believe our method offers some new insights into the mechanism that enable the minimum Hellinger distance estimator to be simultaneously efficient and robust. One of the most appealing features of M-estimation in the location model is that one can see directly how the method limits the impact of large observations. That is, given residuals $\epsilon_i = y_i - \mu$, one solves for μ in the equation $\sum \psi(\epsilon_i) = 0$ for some function ψ. Since $\psi(\epsilon) = \epsilon$ gives the sample mean (normal theory maximum likelihood estimator) as a solution, other ψ-functions with $|\psi(\epsilon)| \ll |\epsilon|$ for large ϵ, will limit the effect of large ϵ "outliers" on the estimator, relative to maximum likelihood. Our minimum distance estimators have a very similar form, but with a modified definition of residual δ. All our estimators solve an estimating equation depending entirely on a user-selected function $A(\delta)$, such that $A(\delta) = \delta$ gives a solution to the likelihood equations and $A(\delta) \ll \delta$ for large δ implies that large δ outliers have small influence on the parameter estimates. This analysis is distinctly different from the usual influence function approach, as all our estimators have the same influence function. However, it will be shown that the choice of function $A(\cdot)$ still has a dramatic effect on robustness. For each *MDE* there is a corresponding parameter A_2, called the curvature parameter, that is a measure of its robustness in a contaminated model and a measure of its second order efficiency at the model. Negative values of A_2 imply robustness, with larger absolute values implying greater contamination robustness.

In Section 2 we will define the general disparity measure and introduce the *MDE*s for continuous models. A natural analogue of the ordinary maximum likelihood estimator (*MLE*) which emerges in this context is called the *MLE**. In Section 3 we investigate the efficiency of the *MLE** and the other *MDE*s. Section 4 investigates the robustness properties of the *MDE*s. Section 5 presents their influence curve analysis which indicates the asymptotic equivalence of the *MDE*s to the *MLE**. Section 6 provides the asymptotic results. For the sake of keeping a clear focus here, we have emphasized the theoretical and simulation results that verify our claims about efficiency and robustness. In the last section, Section 7, we address in brief a number of other important statistical considerations, including equivariance, standard error calculations and numerical considerations.

2. General disparity measures

It has been conventional to extend density-based minimum distance methods to the continuous case by forming a nonparametric density estimator from the

data, say $f^*(x)$ and then constructing a distance between $f^*(x)$ and the model density $m_\beta(x)$ (indexed by an unknown parameter vector β), such as the squared Hellinger distance

$$\int \left[\sqrt{f^*(x)} - \sqrt{m_\beta(x)}\right]^2 dx.$$

See the formulation of Beran (1977) and Tamura and Boos (1986).

The key idea that we introduce here to simplify the statistical analysis is as follows: first, construct f^* using kernel density estimation, say

$$f^*(x) = \int k(x; t, h) d\hat{F}(t)$$

where $\hat{F}$ is the empirical distribution function obtained from the data and k is a smooth family of kernel functions such as the normal densities with mean t and standard deviation h. The parameter h controls the smoothness of the resulting density, with increasing h corresponding to greater smoothness (e.g., Silverman (1986)). Next apply the same smoothing to the model to get

$$m^*_\beta(x) = \int k(x; t, h) dM_\beta(t), \tag{2.1}$$

where M_β is the model cdf. Now construct a density based "distance" between $f^*(x)$ and $m^*_\beta(x)$, such as the squared Hellinger distance

$$HD(f^*, m^*_\beta) := \int \left[\sqrt{f^*(x)} - \sqrt{m^*_\beta(x)}\right]^2 dx. \tag{2.2}$$

A natural central figure in this new setting is the analogue of the maximum likelihood estimator; we will let MLE^* be the value of β that minimizes the likelihood disparity

$$LD(f^*, m^*_\beta) := \int f^*(x) \ln[f^*(x)/m^*_\beta(x)] dx; \tag{2.3}$$

this is a form of the Kullback-Leibler divergence. In a discrete model with no kernel smoothing, minimizing the LD yields the maximum likelihood estimator (MLE) of β. Thus a central question involves the effect of kernel smoothing on the MLE^*. This is addressed in Subsection 3.1.

To provide further intuition, consider a variant of the idea for data on the real line. Instead of a fixed h kernel estimate suppose that we use a histogram with fixed bin width, say h. The strategy is to compute the empirical probabilities for the bins, and to minimize their distance from the corresponding model based bin probabilities. This variant of smooth-the-model strategy reduces to the countable support minimum distance estimation problem considered by Simpson (1987) and Lindsay (1994). Usually, though, a discretization of this type will entail a loss of information. In this paper we will show that the type of model smoothing

considered by us in equation (2.1) leads in some cases to no loss of information and is thus an improvement over the histogram approach.

In general we will be concerned with estimators based on minimizing a disparity measure

$$\rho_G(f^*, m_\beta^*) = \int G(\delta^*(x)) m_\beta^*(x) dx \tag{2.4}$$

where $\delta^*(x) = (f^*(x) - m_\beta^*(x))/m_\beta^*(x)$ will be called the Pearson residual at x and G is a strictly convex function. Similarly defined disparities in the discrete model were considered by Lindsay (1994), where it is shown that the *MDE*s obtained by minimizing disparity measures of this type generated first order efficient estimators and they are robust for certain functions G. Both (2.2) and (2.3) can be put in the form of (2.4).

Our numerical studies in this paper focus on a particular class of disparity measures called by Lindsay the blended weight Hellinger distances:

$$BWHD_\alpha(f^*, m_\beta^*) = \int \frac{(f^*(x) - m_\beta^*(x))^2}{\left(\alpha\sqrt{f^*(x)} + \bar{\alpha}\sqrt{m_\beta^*(x)}\right)^2} dx, \tag{2.5}$$

$$\bar{\alpha} = 1 - \alpha, \quad \alpha \in [0, 1].$$

From robustness considerations we are more interested in the range $\alpha \in [\frac{1}{3}, 1]$ (see Section 4). Note that $BWHD_\alpha = HD$ at $\alpha = 0.5$ up to a scalar multiple. We have chosen to focus on this class because we can adjust efficiency and robustness properties simply and dramatically by changing α. Another important class of such measures are the Cressie and Read (1984) family of power divergences (generalized to continuous models) defined as

$$\begin{aligned} PD_\lambda(f^*, m_\beta^*) &= \int f^*(x)\{[f^*(x)/m_\beta^*(x)]^\lambda - 1\} dx/\lambda(\lambda+1) \\ &= \int m_\beta^*(x)\{(1+\delta^*)^{\lambda+1} - 1\} dx/\lambda(\lambda+1), \end{aligned}$$

where $\lambda = -2, -1, -1/2, 0$, and 1 generate the Neyman's chi-square, Kullback-Leibler divergence, Hellinger distance, likelihood disparity, and Pearson's chi-square respectively.

Under differentiability of the model, and letting ∇ denote derivatives with respect to β, the minimum disparity estimating equations have the form:

$$-\nabla\rho = \int [G'(\delta^*(x)) f^*(x)/m_\beta^*(x) - G(\delta^*(x))] \nabla m_\beta^*(x) = 0, \tag{2.6}$$

and can alternatively be written as

$$-\nabla\rho = \int A(\delta^*(x)) \nabla m_\beta^*(x) dx = 0 \tag{2.7}$$

for $A(\delta^*) = (1+\delta^*)G'(\delta^*) - G(\delta^*)$. As G is strictly convex, $A(\delta^*)$ is a strictly increasing function of δ^*. Also $A(\delta^*)$ can, without loss of generality, be centered and rescaled so that $A(0) = 0$ and $A'(0) = 1$. (For example, in the *BWHD* family this amounts to dividing the disparity by 2. Such standardizations do not change the estimating properties of the disparities.) This centered and rescaled function $A(\cdot)$ is called the residual adjustment function (*RAF*) corresponding to the disparity measure ρ. Most of the theoretical properties of the *MDE*s are derived using the properties of $A(\cdot)$. Unlike the approach of Beran, the *MDE*s are Fisher consistent.

In (2.6) we use the negative gradient because then the estimating equation of the *MLE** has a form similar to the likelihood score equations. For the likelihood disparity $A(\delta^*) = \delta^*$. The curvature parameter A_2 of the disparity, which is defined to be the second derivative of its residual adjustment function evaluated at $\delta^* = 0$, equals zero for the likelihood disparity. In Section 5 we will provide the theoretical justification of the robustness of the estimators generated by disparities with large negative A_2. For the Hellinger distance $A_2 = -1/2$.

3. Efficiency of the *MDE*s

3.1 *Efficiency of the MLE** *under transparent kernels*

In this section we show that the estimating equation of the *MLE** has the form

$$\sum u^*(X_i, \beta) = 0,$$

so that the mathematical theory of such estimating equations leads directly to the consistency and asymptotic normality of the *MLE** and its asymptotic efficiency can be directly determined. In particular, we will show full efficiency in the normal model with the normal kernel. Since we later show that the other *MDE*s are asymptotically equivalent to the *MLE** at the model, they too are necessarily efficient at the normal model.

We use the notations $\tilde{u}(x,\beta) = \nabla \ln m^*_\beta(x)$ and $u^*(t,\beta) = \int \tilde{u}(x,\beta)k(x;t,h)dx$. Under the assumption that $\int \nabla m^*_\beta(x)dx = 0$ (i.e. the derivative can be taken inside the integral sign), the following result is easily proved.

LEMMA 3.1. *Let f^* and m^*_β be respectively the kernel density estimator obtained from the data and the smoothed model density. Let E_β represent the expectation with respect to m_β. Then the estimating equation of the MLE** *can be written as*

$$-\nabla LD(f^*, m^*_\beta) = \frac{1}{n}\sum u^*(X_i,\beta) = 0. \tag{3.1}$$

Further, $E_\beta[u^*(X,\beta)] = 0$ for all β.

We will call u^* the *MLE** *score function*. Let $u = u(x,\beta) = \nabla \ln m_\beta(x)$ be the *MLE score function*. Since the *MLE** is an M-estimator (most of the other *MDE*s are not so), its efficiency and robustness are easy to study.

How much information is lost by smoothing the original density function with the kernel? It seems intuitively clear that this will depend on the relationship of the kernel to the model. The following example shows that for some models it may be possible to choose kernels which do not lead to any information loss.

Suppose that $X_1, X_2, \ldots, X_n$ are independently distributed with common distribution $MVN(\mu, \Sigma)$, and that the chosen kernel is $MVN(0, h^2 I)$. For the MVN problem m^*_β is $MVN(\mu, \Sigma + h^2 I)$. The score equations for the likelihood disparity are, for the parameters in μ

$$\int (\Sigma + h^2 I)^{-1}(x - \mu) dF^*(x) = 0$$

and for the parameters in Σ

$$\int \{(x - \mu)(x - \mu)^T - (\Sigma + h^2 I)\} dF^*(x) = 0$$

where F^* is the cdf corresponding to f^*. Since the distribution of X under F^* is the convolution of $\hat{F}$ and $MVN(0, h^2 I)$, the solution to these score equations are just the usual maximum likelihood estimators:

$$\hat{\mu} = \bar{X}, \qquad \hat{\Sigma} = \frac{1}{n} \sum (X_i - \bar{X})(X_i - \bar{X})^T.$$

Thus, quite remarkably, the answer does not depend on the bandwidth h at all. Moreover, there is no information loss in this case, and in this sense the kernel is *transparent*.

Formally, the kernel $k(x; t, h)$ is defined to be a *transparent kernel* for m_β if the relation

$$Cu(X, \beta) + D = u^*(X, \beta)$$

holds for all $\beta \in \Omega$. Here β is a p-dimensional parameter, C is a $p \times p$ nonsingular matrix which may depend on β and D is a p-dimensional vector. However, since both u and u^* have expectation zero, D must be 0 and we may define transparency by

$$Cu(X, \beta) = u^*(X, \beta) \tag{3.2}$$

from which the next result follows.

LEMMA 3.2. *Suppose that k is a transparent kernel for the family of models m_β. Then the estimating equations for the MLE* of β are equivalent to the ordinary maximum likelihood score equations for β.*

For which other models do transparent kernels exist? We do not have a general answer to this, but we can offer two more examples. To motivate this result, note that, if $X \sim N(\mu, \sigma^2) = m_\beta(x)$, then $m^*_\beta(x)$ is the density of $Y = X + Z$, where Z is $N(0, h^2)$ and independent of X, and therefore Y is $N(\mu, \sigma^2 + h^2)$. Two other prominent exponential families share this closure under convolution—the

gamma and the Poisson—and if we use convolution smoothing on them, we obtain a transparent kernel.

PROPOSITION 3.1. (i) *If $m_\beta(x)$ is the gamma density, with parameter (β, λ), λ known, then $k(y; x, h) = m_h(y - x)$ is transparent.*

(ii) *If $m_\beta(x)$ is the Poisson density, mean β, then $k(y; x, h) = m_h(y - x)$ is transparent.*

Clearly the asymptotic efficiency of the *MDE*s depends on the availability of transparent kernels. If they are not used, attainment of full asymptotic efficiency will require that the bandwidth $h \to 0$ as $n \to \infty$. However, we should note that even for h held fixed, there can be surprisingly little information lost when using a non-transparent kernel. If we calculate the asymptotic variance (see Theorem 6.1 and Corollary 6.2) for the *MLE** estimator of μ in the $N(\mu, \sigma^2)$ model, at $\mu = 0$ and $\sigma^2 = 1$, where the density estimator

$$f^*(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)$$

is based on the Epanechnikov kernel $k(x) = 0.75(1 - x^2)$ for $|x| < 1$, we find that for $h = 0.1, 0.2, \ldots, 0.9$ the information loss is less than 0.003%. The identical calculation for the uniform kernel $k(x) = 1/2$ for $|x| < 1$ gives a maximum information loss of 0.04% for values of $h = 0.1, 0.2, \ldots, 0.9$. These information calculations were substantiated by a simulation that compared the normal kernel with the Epanechnikov kernel and found that the estimators gave very similar results for both efficiency at the model and robustness under contamination.

We can offer a heuristic explanation for this stability of information under smoothing. For simplicity, suppose that the kernel smoothing is by convolution, so that $k(x; t, h) = k((x - t)/h)/h$ and $m^*_\beta(y)$ is the density of $Y = X + h\epsilon$, where ϵ has density $k(\cdot)$. In this case, we can write

$$\tilde{u}(y, \beta) = E[u(X, \beta) \mid Y = y] \quad \text{and} \quad u^*(x, \beta) = E[\tilde{u}(Y, \beta) \mid X = x].$$

If we were able only to see data from Y's distribution, then we would experience a net loss of information due to the addition of the noise ϵh. Suppose, again for simplicity, that the parameter β is a scalar, so that the relative information in the Y data can be calculated as $corr^2(u(X, \beta), \tilde{u}(Y, \beta))$, which here equals

$$\begin{aligned} E^2[\tilde{u}^2(Y, \beta)]/(E[u^2(X, \beta)]E[\tilde{u}^2(Y, \beta)]) &= E[\tilde{u}^2(Y, \beta)]/E[u^2(X, \beta)] \\ &= 1 - E\{\mathrm{Var}[u(X, \beta) \mid Y]\}/\mathrm{Var}[u(X, \beta)]. \end{aligned}$$

However, the information lost in noise is at least partially recovered in going to u^*. The calculations now give the relative information in u^* as being:

$$\frac{E[\tilde{u}^2(Y, \beta)]}{E[u^2(X, \beta)]} \frac{E[\tilde{u}^2(Y, \beta)]}{E[u^{*2}(X, \beta)]}.$$

Table 1. The means and the standard deviation of the *MHDE*s of μ.

	Data without contamination		Data with contamination	
h	mean	standard deviation	mean	standard deviation
0.1	0.0098	0.1483	0.1139	0.1697
0.2	0.0086	0.1473	0.1523	0.1697
0.3	0.0067	0.1466	0.1766	0.1661
0.4	0.0051	0.1463	0.1967	0.1628
0.5	0.0039	0.1456	0.2142	0.1594
0.6	0.0030	0.1454	0.2293	0.1559
0.7	0.0023	0.1452	0.2423	0.1523
0.8	0.0018	0.1450	0.2533	0.1493
0.9	0.0014	0.1449	0.2624	0.1463
MLE*	0.0001	0.1442	0.3060	0.1315

The first ratio is the information lost due to noise, and is less than one. We can think of it also as the relative loss in variance due to going from $u(X, \beta)$ in the denominator to its conditional expectation $E[u(X, \beta) \mid Y]$ in the numerator. The second ratio, however, is greater than one, and is the relative increase in variance in going from $E[\tilde{u}(Y, \beta) \mid X]$ to $\tilde{u}(Y, \beta)$. In the normal model, with normal kernel, these factors balance each other perfectly, and so no information loss results. However, under no circumstances can we do any worse than the first ratio, the loss in information due to making the data more noisy by convolution.

3.2 *Efficiency of the MDE*

Later in this paper we establish that the minimum disparity estimators are asymptotically equivalent to the *MLE** when the model assumptions are correct. In particular, Lemma 5.1 establishes that they have the same influence function and Corollary 6.2 establishes that under sufficient regularity conditions the estimators have the same limiting distributions. Combining this with the results of Subsection 3.1 we have that the *MDE*'s are fully efficient when a transparent kernel is used. In this subsection we verify these theoretical results by simulation.

We generated 50 pseudo random samples, each of size 50, from the $N(0, 1)$ distribution using the IMSL subroutine in FORTRAN for the generation of standard normal random variables. The kernel function $k(x; t, h)$ was the normal pdf with mean t and standard deviation h, which is transparent at the normal model. We calculated the mean of each sample, which is the *MLE** of μ (as well as the *MLE* of μ) for that particular sample, and then calculated the mean and standard deviation of the *MLE** of μ over the samples. For uncontaminated data the bias will just be sampling error, but the standard deviation will be an important descriptor of the efficiency of the method at the model. The mean and the standard deviation of the *MLE** of μ are 0.0001 and 0.1442 respectively. Table 1 (the no

Table 2. The means and the standard deviation of the Huber estimates of μ.

	Data without contamination		Data with contamination	
b	mean	standard deviation	mean	standard deviation
0.6	−0.0190	0.1717	0.1349	0.1755
0.7	−0.0161	0.1682	0.1398	0.1706
0.8	−0.0143	0.1649	0.1460	0.1661
0.9	−0.0127	0.1615	0.1526	0.1628
1.0	−0.0118	0.1584	0.1595	0.1597
1.1	−0.0108	0.1559	0.1663	0.1575
1.2	−0.0094	0.1533	0.1728	0.1562
1.3	−0.0081	0.1513	0.1789	0.1549
1.4	−0.0070	0.1497	0.1850	0.1536
1.5	−0.0064	0.1479	0.1924	0.1526
*MLE**	0.0001	0.1442	0.3060	0.1315

contamination part) gives the mean and the standard deviation of the corresponding minimum Hellinger distance estimators (*MHDE*s) of μ (for different values of h) assuming that σ^2 is known to be 1. The Newton-Raphson algorithm was used to solve all the optimization problems in this paper. Simpson's 1/3rd rule was used to evaluate the integrals numerically. The *MHDE* is clearly highly efficient at all levels of h.

For the purpose of comparison, we can also determine the robust estimates of μ using Huber's ψ function. Since $\sigma^2 = 1$, to get the robust estimates we have to solve

$$\sum_{i=1}^{n} \psi_b(X_i - \mu) = 0$$

where the function $\psi_b(x) = \min\{b, \max[x, -b]\}$. We choose 10 different values $0.5, 0.6, \ldots, 1.5$ of b. Table 2 (the no contamination part) gives the corresponding values of the means of the Huber estimates of μ and their standard deviations. A comparison of the *MHDE*s and the Huber estimates shows that among estimators that have the same level of bias under contamination (Section 4), the standard deviation of the *MHDE* is consistently smaller than the Huber estimates at the model.

4. Robustness of the minimum disparity estimators: adjustments in trade-off with efficiency

The robustness of the *MDE* can be understood in a large part through an investigation of the form of $A(\delta^*)$. Consider the *BWHD* family introduced in Section 2. In Fig. 1 we have plotted the residual adjustment functions for the set

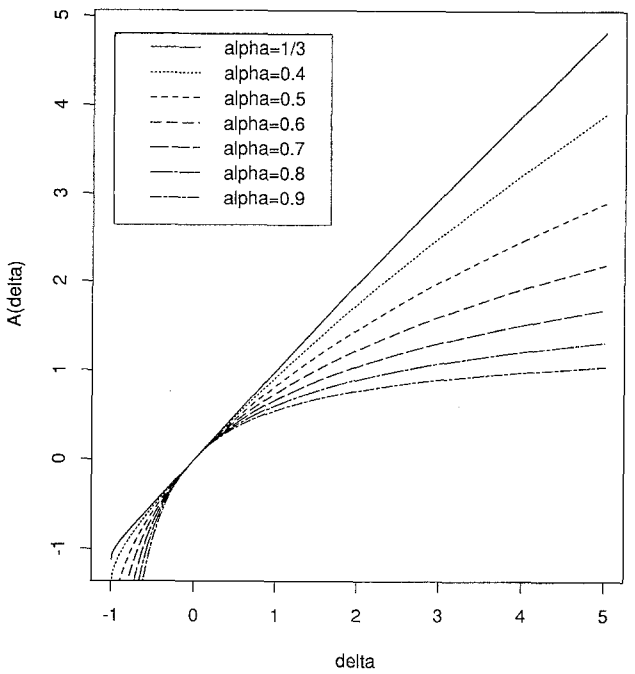

Fig. 1. The residual adjustment functions of the members of the blended weight Hellinger distance family.

of these modified Hellinger distance functions. The residual adjustment functions have the form

$$A_\alpha(\delta^*) = \delta^*(\alpha\sqrt{\delta^*+1}+\bar{\alpha})^{-2} + \frac{\bar{\alpha}}{2}\delta^{*2}(\alpha\sqrt{\delta^*+1}+\bar{\alpha})^{-3}. \tag{4.1}$$

The plot of the residual adjustment functions and equation (4.1) carry the following important information (for more details see Lindsay (1994)):

1. The curvature at 0, $A_2 = A''(0)$ is a measure of the second order efficiency of the method (Lindsay (1994)), with $A_2 = 0$ giving full second order efficiency in the sense of Rao (1961). (For $\alpha = 1/3$ in $BWHD_\alpha$, we also have $A'''_\alpha(0) = 0$, a form of third order efficiency.)

2. The Pearson residuals are bounded below by -1, and -1 occurs only when $f^*(x) = 0$. The value of $A(-1)$ then reflects the impact (relative to maximum likelihood) of having *holes* in the data—sparse data where one would expect more observations.

3. The Pearson residuals can be made arbitrarily large for a fixed value of $f^*(x)$ making $m^*_\beta(x)$ very small. If we interpret $f^*(x)$ as being large in the neighborhood of an observation, then we might say an observation is *surprising* if $f^*(x)$ is large and $m^*_\beta(x)$ is small. A residual adjustment function such as that for the Hellinger distance thus downweights surprising observations relative to maximum likelihood. The residual adjustment functions of the members of the $BWHD$ family in the range $\alpha \in [0, 1/3)$ curve in the opposite direction (at $\delta = 0$) as the

members of the *BWHD* family with $\alpha \in [1/3, 1]$, and thus end up giving *higher* weight to surprising observations compared to maximum likelihood.

4. For the $BWHD_\alpha$ family it can be seen that $A_2 = A''(0) = 1 - 3\alpha$ is also a measure of robustness, with larger negative values of A_2 implying greater robustness against surprising observations. In Section 5 we provide a calculation that demonstrates the general role of the curvature parameter A_2 in asymptotic robustness properties.

5. Suppose that the model has finite Fisher information. For residual adjustment functions of the $BWHD_\alpha$ class, for which $A(\delta^*)/\delta^{1/2} = O(1)$ as $\delta \to \infty$, there is a type of bounded effective influence in the following sense: If the data are contaminated at a fixed positive level ϵ at point y, then the estimator stays bounded as $y \to \infty$, in fact converging to the estimator one would obtain ignoring the point y.

These heuristic remarks concerning robustness will be substantiated first through simulation; the theory will be postponed until Section 5. Consider the 50 samples investigated in Section 3. We assume $\sigma^2 = 1$ and use the $N(t, h^2)$ kernel. Each sample was contaminated by the replacement of 10% of the observations (5 observations in this case) by $y = 3$. We considered the target value of the parameter to be the mean of the normal component ($\mu = 0$), and considered any systematic deviation from zero by our estimators to be their *bias*. Robust estimators should have low bias. Our theoretical analysis of Section 5 leads us to the following predictions. From Corollary 5.1, as the curvature parameter A_2 approaches zero from below, the estimator should simultaneously increase in efficiency at the model and increase in bias under contamination. Secondly, according to Corollary 5.1 and the calculations in Table 3, increasing h will make the *MDE* more like the *MLE* and hence will increase the bias under contamination. The simulation results verified these predictions. The mean and the standard deviation of the *MLE* of μ for the 50 contaminated samples were 0.3060 and 0.1315 respectively. For the blended Hellinger distance A_2 is negative for $\alpha > 1/3$ and increases in absolute magnitude as α increases. Figure 2 is a visual description of the bias of the minimum blended weight Hellinger distance estimator (*MBWHDE*) of μ for three different values of α (1/3, 0.5 and 0.7) and for values of h from 0.1 to 0.9. The effect of α and h on bias is evident in Fig. 2.

In Fig. 3 we graphically represent the effect of α on the mean and the standard deviation of the *MBWHDE* of μ. We fixed the value of h at 0.5. The graph clearly shows that an increase in α reduces the bias while increasing the standard deviation. It also shows that the effect of α on the bias is much greater than on the standard deviation—the price for robustness is small in terms of mean squared error.

The next investigation regards a further theoretical prediction (Section 5) that a large outlier y should have very little impact on the robust *MDE*. We now show that in case of the *MHDE*, in fact, an outlying value fails to affect the estimation at all when it is totally inconsistent with the model. As the contaminating value y grows larger, its effect on the *MHDE* quickly dissipates. Interestingly, this happens even though the *MHDE* does not have a bounded influence function. In Fig. 4 we present the mean *MHDE*s of μ corresponding to different contaminating values of

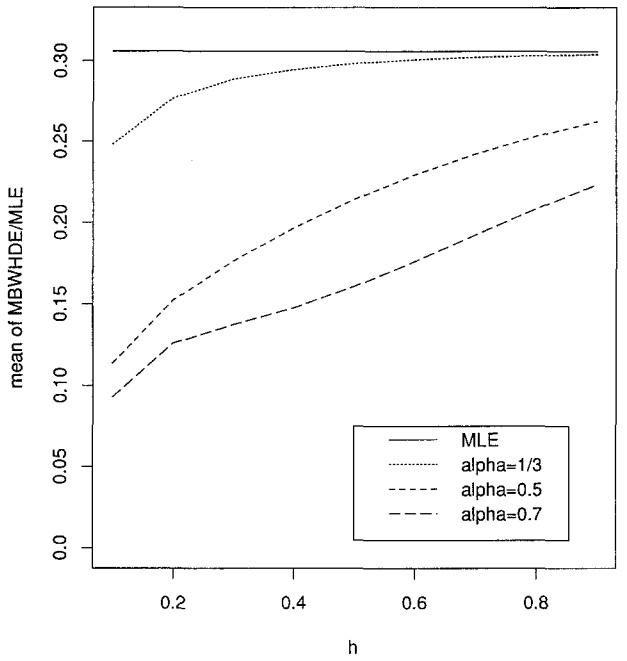

Fig. 2. Average bias under contamination for the maximum likelihood estimator and *MBWHDE*s of μ in the normal model as a function of h.

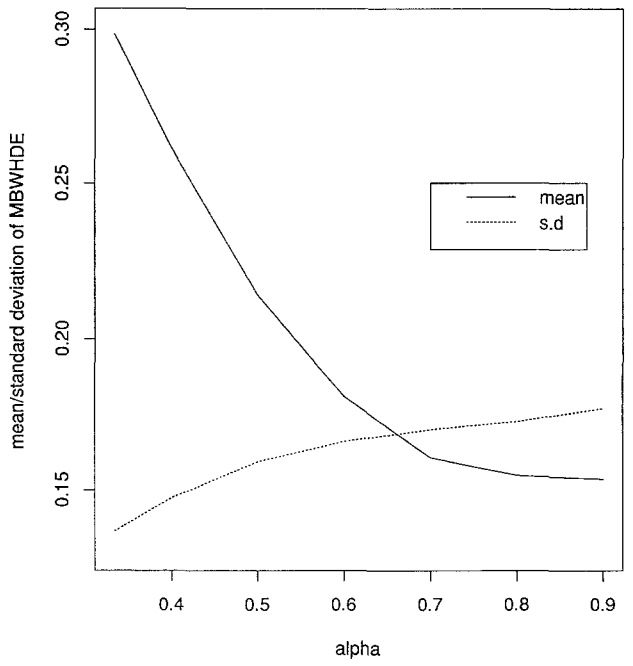

Fig. 3. The bias and standard deviation of the *MBWHDE*s of μ as a function of α.

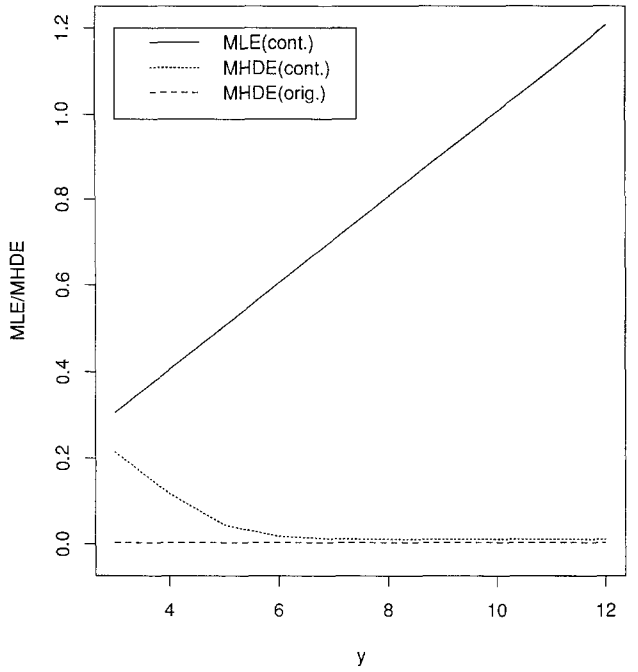

Fig. 4. Average bias for the minimum Hellinger distance estimator and the maximum likelihood estimator of μ as a function of the contaminating value.

y, ranging from 3 to 12. For comparison, we present the mean *MLE*s also. The horizontal dashed line represents the mean *MHDE* in the original sample with no contamination. Figure 4 clearly shows that by the time the contaminant is as large as 7, the mean *MHDE* is practically equivalent to that for the uncontaminated case, showing the limitation of the influence function approach. A similar example can be found in Simpson (1987) for the Poisson model. Some theoretical results for this type of behavior of the *MHDE* are provided in Beran (1977) and Lindsay (1994).

We can obtain similar tradeoffs between bias and variance by using the Huber's ψ estimator for the contaminated data (see Table 2). In these limited simulations the Huber estimators were slightly more efficient than the *MHDE* for the same level of bias. However, there are several points in favor of the methods of this paper. First, there is superior efficiency at the normal model. Secondly, our methods allow for simultaneous efficient estimation of the scale parameter. Finally our approach clearly allows the extension of robust methods to models that do not have location-scale parameters.

Finally we investigate the effect of not knowing σ^2 on the estimation of μ. In Fig. 5 we have presented the values of the mean and the standard deviation of the *MHDE*s of μ over the 50 samples, 10% contaminated by the value 3. We have presented the unknown variance and the known variance cases in the same graph for comparison, using the values $0.2, 0.3, \ldots, 0.9$ for h. In general the bias of the *MHDE* is higher and the standard deviation is lower than the case when

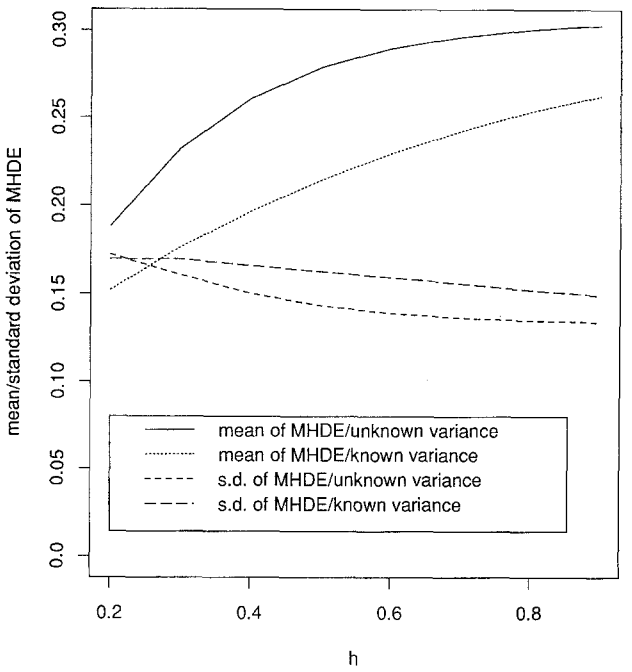

Fig. 5. Average bias and standard deviation of the minimum Hellinger distance estimator of μ in the normal model as a function of h.

the variance is known. It appears that by choosing a considerably more robust estimator one can attain the same level of bias reduction for σ^2 unknown as for σ^2 known, with low cost in variance.

5. Influence curve analysis

In this section we will show that all the *MDE*s under consideration (including the *MLE**) *have the same influence function* at the model. In addition to suggesting their asymptotic equivalence it demonstrates that the influence function can have severe shortcomings as a measure of the effect of contaminations. We also show that a second order prediction of bias exhibits the important role of the estimation curvature parameter A_2 in determining the robustness of the estimator. Our analysis will also clarify the effect of h on robustness. Let T denote the minimum disparity functional. By Fisher consistency, $T(M_\beta) = \beta$. Suppose that the true cdf is $S(x)$, with density $s(x)$ not necessarily in the model, and let $T(S) = \beta^s$. We define the influence function of the functional T at a cdf S by first defining $S_\epsilon(x) = (1-\epsilon)S(x) + \epsilon I[x \geq y]$ and then letting the influence function be $T'(y) = T'(S, y) = \frac{\partial}{\partial \epsilon} T(S_\epsilon)|_{\epsilon=0}$. I represents the indicator function.

Let ∇_j and ∇_{jk} represent the partial derivatives with respect to β_j and β_j, β_k and write $\tilde{u}_j(x, \beta) = \nabla_j \ln m^*_\beta(x)$ and $\tilde{u}_{jk}(x, \beta) = \nabla_{jk} \ln m^*_\beta(x)$. Also let

$$u^*_j(t, \beta) = \int k(x; t, h)\tilde{u}_j(x, \beta)dx = \nabla_j \int \ln m^*_\beta(x) k(x; t, h)dx,$$

$$u^*_{jk}(t,\beta) = \int k(x;t,h)\tilde{u}_{jk}(x,\beta)dx = \nabla_{jk}\int \ln m^*_\beta(x)k(x;t,h)dx.$$

$J^*(\beta)$ is the $p \times p$ matrix whose jk-th element is given by $E_\beta[-u^*_{jk}(X,\beta)]$; it is nonnegative definite as it is the information matrix corresponding to a random variable with pdf $m^*_\beta(x)$. Let $s^*(x) = \int k(x;t,h)s(t)dt$ be the kernel smoothed version of $s(x)$. Let $\delta^*(x) = s^*(x)/m^*_{\beta^s}(x) - 1$. We will define $J^{*s}(\beta^s)$ to be the $p \times p$ matrix whose jk-th element is given by

$$\int A'(\delta^*)\tilde{u}_j(x,\beta^s)\tilde{u}_k(x,\beta^s)s^*(x)dx - \int A(\delta^*)\nabla_{jk}m^*_{\beta^s}(x)dx$$

and let $v^*(t,\beta^s)$ be the p-dimensional vector whose j-th component is

$$\int A'(\delta^*)\tilde{u}_j(x,\beta^s)k(x;t,h)dx - \int A'(\delta^*)\tilde{u}_j(x,\beta^s)s^*(x)dx.$$

Under the above definitions, a straightforward calculation gives the following result.

LEMMA 5.1. *Let $S(x)$ be the true distribution not necessarily in the model. For the minimum disparity functional T, let $T(S) = \beta^s$. Then the influence function of T has the form $T'(y) = [J^{*s}(\beta^s)]^{-1}v^*(y,\beta^s)$. If $S = M_{\beta_0}$ for some β_0, then the above reduces to $T'(y) = [J^*(\beta_0)]^{-1}u^*(y,\beta_0)$. If in addition k is a transparent kernel for the family M_β then we get $T'(y) = [I(\beta_0)]^{-1}u(y,\beta_0)$, where $I(\beta)$ is the Fisher information about β in m_β.*

If T is the minimum disparity functional and $M_{\beta,\epsilon}(x) = (1-\epsilon)M_\beta(x) + \epsilon I[x \geq y]$, then the bias in estimation is $\Delta T = T(M_{\beta,\epsilon}) - T(M_\beta) = T(M_{\beta,\epsilon}) - \beta$. For the estimator to be robust it is necessary that ΔT be small. Expanding the bias in a Taylor series we get that the first order approximation to the bias as $\Delta T = T(M_{\beta,\epsilon}) - \beta \approx \epsilon T'(y)$. For the mean parameter in the one parameter exponential family, in the transparent kernel case, the influence function of all *MDE*s at the model is $T'(y) = (y-\mu)$; thus predicted bias is unbounded. Yet simulation results show that for some disparities like the Hellinger distance the actual bias is much lower than the predicted bias. For the *MLE** the predicted bias is exact, indicating that the actual biases for some disparities are lower than the bias of the *MLE**. A second order expansion of the bias function can help explain this behavior if the second order term in the expansion is large in magnitude and opposite in sign to the first order term, thus balancing its effect. Consider the estimating equation $\int A(\delta^*_\epsilon(x))\nabla m^*_{\beta_\epsilon}(x) = 0$ where $\beta_\epsilon = T(S_\epsilon)$ and $\delta^*_\epsilon(x) = (s^*_\epsilon/m^*_{\beta_\epsilon}) - 1$. For simplicity, we now let S be in the model and look at the case where β is a scalar. Evaluating the second derivative of the above estimating equation at $\epsilon = 0$ we get the following theorem.

THEOREM 5.1. *Let $T''(y) = \frac{\partial^2}{\partial\epsilon^2}T(M_{\beta,\epsilon})|_{\epsilon=0}$. Then for an estimating function of the type $\int A(\delta^*)\nabla m^*_\beta(x)dx$, we have*

$$T''(y) = T'(y)\left[\int \tilde{u}^2(x,\beta)m^*_\beta(x)\right]^{-1}\{f_1(y) + A_2 f_2(y)\},$$

Table 3. Values of $f_2(y)$ for the normal example.

	y					
h	2.0	2.5	3.0	3.5	4.0	4.5
0.10	43.997	146.572	606.299	3119.198	20274.770	168193.646
0.20	17.889	64.779	277.450	1424.326	9065.688	73008.995
0.30	9.419	37.173	163.052	823.770	5052.851	38780.168
0.40	5.458	23.463	104.229	512.146	2987.208	21510.281
0.50	3.325	15.507	69.046	326.917	1794.768	11978.329
0.60	2.095	10.521	46.536	210.925	1082.376	6638.372
0.70	1.356	7.261	31.682	136.914	654.207	3668.817
0.80	0.899	5.075	21.734	89.400	397.326	2033.806
0.90	0.609	3.585	15.019	58.813	243.401	1138.232

where

$$f_1(y) = 2\nabla u^*(y,\beta) - 2E_\beta[\nabla u^*(X,\beta)] + T'(y)E_\beta[\nabla^2 u^*(X,\beta)],$$

and

$$f_2(y) = [u^*(y,\beta)]^{-1}\left[\int \tilde{u}^2(x,\beta)m^*_\beta(x)dx\right]\left[\int k^2(x;y,h)\tilde{u}(x,\beta)(m^*_\beta(x))^{-1}dx\right]$$
$$- 2\int \tilde{u}^2(x,\beta)k(x;y,h)dx + T'(y)\int \tilde{u}^3(x,\beta)m^*_\beta(x)dx.$$

COROLLARY 5.1. *Suppose that all the conditions of Theorem* 5.1 *hold. Assume that the model is a one parameter exponential family, the kernel k is transparent, and β is the mean value parameter. Then $f_1(y)$ is zero and $T''(y) = A_2T'(y)f_2(y)$.*

PROOF. In the one parameter exponential family the transparent kernel gives

$$f_1(y) = C\{2\nabla u(y,\beta) - 2E[\nabla u(X,\beta)] + T'(y)E[\nabla^2 u(X,\beta)]\} \equiv Cg(y,\beta).$$

The quantity $g(y,\beta)$ has been shown to be zero in Lindsay ((1994), Corollary 4). □

In the one parameter exponential family, if $f_2(y)$ is positive and A_2 is negative, the second derivative in the Taylor series approximation of the bias will have sign opposite to the first. For values of ϵ where the first and the second order approximations differ substantially, the second order approximation can predict a much smaller bias than the first.

It is not obvious that $f_2(y)$ is necessarily positive for all y under any model. In Table 3 we present some numerical calculations for the $N(\mu, 1)$ model with the transparent kernel we have used in Sections 3 and 4. We have determined the values of $f_2(y)$ for several choices of h and y, using the true value 0 of μ. The following points deserve mention. First, the entries in Table 3 are all positive.

Secondly, the value of $f_2(y)$ increases in magnitude as the absolute value of y grows large (making it a more surprising value). So for such values of y, the balancing effect of the second order term will be stronger. Finally, the value of $f_2(y)$ decreases with h, indicating stronger robustness for a small value of h. Thus for finite samples, more smoothing will mean higher efficiency and higher bias.

6. Asymptotic properties

In this section we will establish important asymptotic results involving the minimum disparity estimators, namely consistency and asymptotic normality. Subscripts j, k and l will represent the partial derivatives with respect to β_j, β_k and β_l. Also let $\delta_s^*(x) = (s^*(x) - m_\beta^*(x))/m_\beta^*(x)$ be the Pearson residual corresponding to $s^*(x)$ and β^s be the unique value of β which solves the minimum disparity estimating equation. Let $J^*(\beta)$ and $J^{*s}(\beta^s)$ be as in Section 5. Many of the proofs in this section closely follow the methods of Simpson (1987) and Lindsay (1994) and will only be briefly outlined.

LEMMA 6.1. *Provided it exists,* $\mathrm{Var}(f^*(x)) = \frac{1}{n}\lambda(x)$, *where* $\lambda(x)$ *is given by*

$$\lambda(x) = \int k^2(x;t,h)s(t)dt - [s^*(x)]^2.$$

PROOF. Note that $\lambda(x) = \mathrm{Var}(k(x, X_i, h))$. As $f^*(x) = \frac{1}{n}\sum_{i=1}^n k(x, X_i, h)$ has the form of a sample mean, the result follows. □

Assume that the kernel function k is bounded. That is, assume $k(x;t,h) \leq N(h)$, with $N(h) < \infty$, where $N(h)$ may depends on h, but not on x or t. From Lemma 6.1 it follows that $\lambda(x) \leq N(h)s^*(x)$.

LEMMA 6.2. $n^{1/4}(f^{*1/2}(x) - s^{*1/2}(x)) \to 0$ *with probability* 1 *if* $\lambda(x) < \infty$.

PROOF. Using the central limit theorem we get $n^{1/4}(f^*(x) - s^*(x)) \to 0$. The result then follows by looking at a Taylor series expansion of the above. □

DEFINITION. The residual adjustment function $A(\delta^*)$ will be called *regular*, if it is twice differentiable and $A'(\delta^*)$ and $A''(\delta^*)(1+\delta^*)$ are bounded on $[-1, \infty)$.

In the following proofs it will be easier to use the Hellinger residuals rather than the Pearson residuals. We define the Hellinger residual Δ^* as

$$\Delta^* = \frac{f^{*1/2}}{m_\beta^{*1/2}} - 1.$$

The Hellinger residual Δ_s^* is obtained by replacing f^* by s^* in Δ^*. Let $Y_n(x) = n^{1/2}(\Delta^*(x) - \Delta_s^*(x))^2$.

LEMMA 6.3. *For any* $k \in [0, 2]$

(i) $E[Y_n^k] \leq E[|\delta^* - \delta_s^*|]^k n^{k/2} \leq (\lambda^{1/2}(x)/m_\beta^*(x))^k$,
(ii) $E[|\delta^* - \delta_s^*|] \leq (\lambda^{1/2}(x)/m_\beta^*(x))$.

PROOF. The first part of the (i) follows by using the result that for $a, b \geq 0$, $(\sqrt{a} - \sqrt{b})^2 \leq |a - b|$. The second part follows by an application Liapounov's inequality. For part (ii), note that

$$E|\delta^* - \delta_s^*| \leq [m_\beta]^{-1} \frac{1}{n} \sum E|k(x; X_i, h) - s^*(x)|.$$

The result then follows from Liapounov's inequality and Lemma 6.1. □

LEMMA 6.4. $\lim_{n\to\infty} E[Y_n^p] = 0$ *for* $p \in [0, 2)$.

PROOF. From Lemma 6.2, $Y_n \to 0$ in probability. Using Lemma 6.3(i), $\sup_n E[Y_n^p]$ is bounded for $p \in [0, 2)$. The result then follows by Chung ((1974), Theorem 4.5.2). □

Let $a_n(x) = A(\delta^*(x)) - A(\delta_s^*(x))$ and $b_n(x) = (\delta^*(x) - \delta_s^*(x))A'(\delta_s^*(x))$. Also let $\gamma_n = \int n^{1/2}(a_n(x) - b_n(x))\nabla m_\beta^*(x)dx$. At this stage we will need to assume that

$$\int s^{*1/2}(x)|\tilde{u}(x, \beta)|dx < \infty. \tag{6.1}$$

We then have the following result.

LEMMA 6.5. *If A is a regular RAF and* (6.1) *is satisfied,* $E|\gamma_n| \to 0$ *as* $n \to \infty$.

PROOF. Let $\tau_n(x) = n^{1/2}|a_n(x) - b_n(x)|$. From Lemma 23 (Lindsay (1994)), $E(\tau_n(x)) \leq BE[Y_n(x)]$ for $B > 0$. By Lemma 6.4, $E(\tau_n(x)) \to 0$. Now

$$E|\gamma_n| \leq \int E(\tau_n(x))|\nabla m_\beta^*(x)|dx.$$

By assumption (6.1), the integrand in the above equation can be bounded by an integrable function. Thus by dominated convergence theorem the result holds. □

It follows from the last lemma and a simple application of Markov's inequality that $\gamma_n \to 0$ in probability. Next we will use the limiting distribution of $n^{1/2} \int b_n(x)\nabla m_\beta^*(x)dx$ in place of that of $n^{1/2} \int a_n(x)\nabla m_\beta^*(x)dx$. This is justified by the above result.

COROLLARY 6.1. *Suppose that* $V = \mathrm{Var}(\int k(x, X, h)A'(\delta_s^*(x))\tilde{u}(x, \beta)dx)$ *is finite and* (6.1) *is satisfied. Then for a regular RAF*

$$n^{1/2} \int [A(\delta^*) - A(\delta_s^*)]\nabla m_\beta^*(x)dx \to N(0, V).$$

PROOF. The result follows by using Lemma 6.5 and a simple application of the central limit theorem. □

Next we present some regularity conditions. We do this in terms of m^*_β, which makes the conditions much simpler than trying to relate them to the original density m_β directly. Technically however, we should remember that there is a kernel involved and the choice of the kernel should be made in such a way that the following conditions hold.

DEFINITION. We will say that the kernel integrated family of distributions is smooth if the conditions of Lehmann ((1983), p. 409, p. 429) are satisfied with $m^*_\beta(x)$ in place of $f(x)$. Also suppose that the conditions

$$|\tilde{u}_{jkl}(x)| \le M_{jkl}(x), \quad |\tilde{u}_{jk}(x)\tilde{u}_l(x)| \le M_{jk,l}(x), \quad |\tilde{u}_j(x)\tilde{u}_k(x)\tilde{u}_l(x)| \le M_{j,k,l}(x)$$

hold for all j, k and l in a neighborhood ω of β^s and $M_{jkl}(x)$, $M_{jk,l}(x)$ and $M_{j,k,l}(x)$ have finite expectations with respect to $m^*_\beta(x)$ for all β in ω. The true density $s(x)$ will be called compatible with $m_\beta(x)$ if $s(x) > 0$ on the common support of $m_\beta(x)$ and the functions M_{jkl}, $M_{jk,l}$, $M_{j,k,l}$ have finite expectation with respect to $s^*(x)$; in addition (6.1) holds and the integrals $\int s^{*1/2}(x)|\tilde{u}_j(x)||\tilde{u}_k(x)|dx$ and $\int s^{*1/2}(x)|\tilde{u}_{jk}(x)|dx$ are finite for all j and k.

THEOREM 6.1. *Assume that the residual adjustment function $A(\delta^*)$ corresponding to a particular disparity measure ρ is regular, m_β is smooth, $s(x)$ is compatible with m_β and the matrix $J^{*s}(\beta^*)$, as defined in Lemma* 5.1 *is positive definite. Then there exists a consistent sequence of roots β_n to the minimum disparity estimating equations. The asymptotic distribution of $n^{1/2}(\beta_n - \beta^s)$ is MVN with mean* 0 *and variance $[J^{*s}(\beta^s)]^{-1}V_s[J^{*s}(\beta^s)]^{-1}$ where V_s is the quantity V in Corollary* 6.1 *evaluated at $\beta = \beta^s$.*

PROOF. The proof is similar to the proof of Theorem 31 in Lindsay (1994)—which utilizes the techniques of Simpson (1987)—if we replace sums by integrals. □

COROLLARY 6.2. *Assume the conditions of Theorem* 6.1. *In addition suppose that the true distribution $S = M_\beta$ for some $\beta \in \Omega$ and k is a transparent kernel for the model family. Then for the MDE β_n, $n^{1/2}(\beta_n - \beta)$ has an asymptotic normal distribution with mean* 0 *and variance $[I(\beta)]^{-1}$, where $I(\beta)$ is the Fisher information about β in m_β.*

PROOF. When $S = M_\beta$ and k is a transparent kernel for the model family, we get $V_s = \mathrm{Var}_\beta(u^*(X,\beta)) = CI(\beta)C^T$ and $J^{*s}(\beta^s) = J^*(\beta) = CI(\beta)$. The result then follows by substituting these expressions. □

7. Some further issues

7.1 *Invariance of the minimum disparity estimators*

Although one of our themes is that when the kernel is appropriately chosen the selection of the smoothing parameter h plays a minor role in the asymptotic properties of the procedure, it has been seen to be a more important factor in the robustness properties. As such, although we no longer need to force h to go to zero, we will want to choose it in such a way that meets important statistical criteria. In particular, in the normal model we will want the estimators of μ and σ^2 to have the correct equivariance properties under location and scale transformations. In this section we show that if we let h be chosen, call it $\hat{h}$, as a fixed multiple of a scale invariant/location invariant estimator $\hat{\tau}$ of scale, then the resulting estimators have the right transformation properties. The asymptotic properties proved in Section 6 will still hold provided $\hat{h} \to h_0 > 0$ almost surely as $n \to \infty$. If we choose $\hat{\tau}$ to be robust, we then expect to preserve the overall lack of sensitivity of the procedure to outlying observations.

As we will be using different values of h, different parameters, and different variables, we first define some notation. Let $X_1, \ldots, X_n$ be the original variables from a location scale model with mean μ, and scale σ. Let $W_1, \ldots, W_n$ be the transformed variables, where $W = aX + b$. If we use a fixed multiple of a robust scale estimator for h, then the bandwidth $\hat{h}_w$ obtained from the W observations equals $a\hat{h}_x$, a scale change of the bandwidth for the X observations. Let the kernel density estimator be expressed as:

$$f^*(x \mid X, \hat{h}_x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\hat{h}_x} K\left(\frac{x - X_i}{\hat{h}_x}\right).$$

The parameters in the definition of f^* indicate which set of variables and what smoothing parameter has been used. Similarly we can define $f^*(x \mid W, \hat{h}_w)$. Let

$$\beta_1 = \begin{pmatrix} \beta_{11} \\ \beta_{12} \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \qquad \beta_2 = \begin{pmatrix} \beta_{21} \\ \beta_{22} \end{pmatrix} = \begin{pmatrix} a\mu + b \\ a^2\sigma^2 \end{pmatrix}.$$

Let $m^*(x \mid \hat{h}_x, \beta_1)$ be the smoothed model when the smoothing parameter $\hat{h}_x$ is used and the parameter is β_1. Similarly we can find $m^*(x \mid \hat{h}_w, \beta_2)$. Also

$$\delta^*(x \mid X, \hat{h}_x, \beta_1) = \frac{f^*(x \mid X, \hat{h}_x)}{m^*(x \mid \hat{h}_x, \beta_1)} - 1.$$

Similarly $\delta^*(x \mid W, \hat{h}_w, \beta_2)$ can be defined.

PROPOSITION 7.1. *Under the above definitions,*

$$\int G(\delta^*(x \mid X, \hat{h}_x, \beta_1)) m^*(x \mid \hat{h}_x, \beta_1) dx = \int G(\delta^*(x \mid W, \hat{h}_w, \beta_2)) m^*(x \mid \hat{h}_w, \beta_2) dx.$$

PROOF. Let $w = ax + b$. Then

$$f^*(w \mid W, \hat{h}_w) = \frac{1}{n}\sum \frac{1}{\hat{h}_w} K\left(\frac{w - W_i}{\hat{h}_w}\right) \tag{7.1}$$
$$= \frac{1}{n}\sum \frac{1}{a\hat{h}_x} K\left(\frac{ax + b - aX_i - b}{a\hat{h}_x}\right)$$
$$= \frac{1}{a}\left[\frac{1}{n}\sum \frac{1}{\hat{h}_x} K\left(\frac{x - X_i}{\hat{h}_x}\right)\right]$$
$$= \frac{1}{a}[f^*(x \mid X, \hat{h}_x)].$$

Using the X observations, the smoothed model m^* is nothing but the density of the convolution $X + \hat{h}_x Z$ where $\hat{h}_x$ is the bandwidth and Z is a standard normal random variable, independent of X. Similarly, using the W observations, the smoothed model is the density of $aX + b + \hat{h}_w Z = a(X + \hat{h}_x Z) + b$, a location-scale change of the former. Let $w = ax + b$. As the density of a location scale model with location parameter θ and scale parameter τ is of the form $\tau^{-1} f(\tau^{-1}(x - \theta))$, an investigation of this form shows that

$$m^*(w \mid \hat{h}_w, \beta_2) = \frac{1}{a} m^*(x \mid \hat{h}_x, \beta_1). \tag{7.2}$$

Combining (7.1) and (7.2) gives

$$\delta^*(x \mid X, \hat{h}_x, \beta_1) = \delta^*(w \mid W, \hat{h}_w, \beta_2)$$

which upon substitution yields

$$\int G(\delta^*(w \mid W, \hat{h}_w, \beta_2)) m^*(w \mid \hat{h}_w, \beta_2) dw$$
$$= \int G(\delta^*(x \mid X, \hat{h}_x, \beta_1)) \frac{1}{a} m^*(x \mid \hat{h}_x, \beta_1) d(ax + b)$$
$$= \int G(\delta^*(x \mid X, \hat{h}_x, \beta_1)) m^*(x \mid \hat{h}_x, \beta_1) dx. \qquad \square$$

Our desired result is now the following simple corollary.

COROLLARY 7.1. *If β_1 is the parameter value where the minimum is achieved when the X observations and $\hat{h}_x$ is used, then β_2 must be the value where the distance will achieve its minimum when W and $\hat{h}_w$ is used. Thus the estimators will be equivariant.*

7.2 *Estimation of standard error*

A practical implementation of the methods described herein requires a useful method for constructing standard errors and hypothesis tests. In this regard we note that Basu (1993) has turned the disparity measures into test statistics analogous to the likelihood ratio statistics, and developed the corresponding asymptotic

distribution theory. Simpson (1989) and Lindsay (1994) have also discussed tests of hypothesis based on disparity measures. As to standard errors, we have from Theorem 6.1 an explicit formula for the asymptotic variance of the parameter estimates as a function of the true distribution $s(x)$. These can be consistently estimated by using the empirical distribution $\hat{F}(x)$ in place of $S(x)$, effectively replacing $s^*(x)$ by $f^*(x)$ wherever it appears.

Let $\hat{\beta}$ be the minimum disparity estimator of β and $\delta^*(x) = f^*(x)/m^*_{\hat{\beta}}(x) - 1$. Also let $\hat{J}$ represent the $p \times p$ matrix whose jk-th element is given by

$$\int A'(\delta^*(x))\tilde{u}_j(x,\hat{\beta})\tilde{u}_k(x,\hat{\beta})f^*(x)dx - \int A(\delta^*(x))\nabla_{jk}m^*_{\hat{\beta}}(x)dx.$$

In addition, let $\hat{V}$ be the $p \times p$ matrix

$$\frac{1}{n-1}\sum_{i=1}^{n}\hat{v}^*_i\hat{v}^{*T}_i$$

where $\hat{v}^*_i$ is the p-dimensional vector whose j-th element is given by

$$\int A'(\delta^*(x))\tilde{u}_j(x,\hat{\beta})k(x;X_i,h)dx - \int A'(\delta^*(x))\tilde{u}_j(x,\hat{\beta})f^*(x)dx.$$

Then the standard error of the parameter estimates can be estimated by $\hat{J}^{-1}\hat{V}\hat{J}^{-1}$.

7.3 *Numerical considerations*

The minimum disparity estimating equations (2.7) will usually be nonlinear and numerical techniques will be required to solve them. As such, the simplicity and the rate of convergence of the iterative algorithm is of prime importance. Since it involves a large number of numerical integrations and the calculation and inversion of a p-dimensional Hessian matrix, the numerical difficulty associated with the Newton-Raphson algorithm quickly increases as p, the number of parameters, increase. Here we briefly explain an iterative reweighting technique which vastly simplifies the computation of the *MDE*s without sacrificing the speed of convergence. See Basu and Lindsay (1993) for a detailed description of this algorithm and several examples of its application.

Since $\int \nabla m^*_\beta(x)dx = 0$, the estimating equation (2.7) can be rewritten as

$$\int \frac{A(\delta^*(x)) - \lambda}{\delta^*(x)+1}(\delta^*(x)+1)\nabla m^*_\beta(x)dx = 0$$

for any constant λ, or

$$\int w(x)\frac{\nabla m^*_\beta(x)}{m^*_\beta(x)}f^*(x)dx = 0 \tag{7.3}$$

where $w(x) = [A(\delta^*(x)) - \lambda]/(\delta^*(x)+1)$ represents the weights. Note that (7.3) is a weighted version of the estimating equation of the *MLE**. If m^*_β is in the

exponential family, a relation like $\nabla m^*_\beta(x)/m^*_\beta(x) = K(\beta)[S(x,\beta) - \beta]$ is often holds, so that we can compute β by iteratively solving the fixed point equation $\beta = \tau(\beta)$, where

$$\tau(\beta) = \frac{\int w(x)S(x,\beta)f^*(x)dx}{\int w(x)f^*(x)dx}.$$

In general this algorithm converges slower than the Newton Raphson algorithm, but if we choose $\lambda = -1$ (Basu and Lindsay (1993)) the rate of convergence of this method is comparable to the Newton-Raphson method.

Acknowledgements

Professor Lindsay's research was partially supported by the National Science foundation under grant DMS 9106895 and by a Humboldt Senior Scientist Research Award.

References

Basu, A. (1993). Minimum disparity estimation: applications to robust tests of hypotheses, Technical Report, Center for Statistical Sciences, University of Texas at Austin.

Basu, A. and Lindsay, B. G. (1993). The iteratively reweighted estimating equation in minimum distance problems, Technical Report, Center for Statistical Sciences, University of Texas at Austin.

Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models, *Ann. Statist.*, **5**, 445–463.

Chung, K. L. (1974). *A Course in Probability Theory*, Academic Press, New York.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *J. Roy. Statist. Soc. Ser. B*, **46**, 440–464.

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods, *Ann. Statist.* (to appear).

Rao, C. R. (1961). Asymptotic efficiency and limiting information, *Proc. Fourth Berkeley Symp. on Math. Statist. Prob.*, **1**, 531–546.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data, *J. Amer. Statist. Assoc.*, **82**, 802–807.

Simpson, D. G. (1989). Hellinger deviance test: efficiency, breakdown points, and examples, *J. Amer. Statist. Assoc.*, **84**, 107–113.

Tamura, R. N. and Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance, *J. Amer. Statist. Assoc.*, **81**, 223–229.