Minimum negative exponential disparity estimation in parametric models

Ayanendranath Basu[a], Sahadeb Sarkar[b,*], A.N. Vidyashankar[c]

[a]*Applied Statistics Unit, Indian Statistical Institute, Calcutta 700 035, India*
[b]*Department of Statistics, Oklahoma State University, Stillwater, OK 74078, USA*
[c]*STATcomp Inc., 221 Washington St., Suite 101, Waukegan, IL 60085, USA*

Received 4 January 1995; revised 1 December 1995

Abstract

Works of Lindsay (1994) and Basu and Sarkar (1994a) provide heuristic arguments and some empirical evidence that the minimum negative exponential disparity estimator (MNEDE), like the minimum Hellinger distance estimator (MHDE) (Beran, 1977), is a robust alternative to the usual maximum likelihood estimator when data contain outliers. In this paper we establish the robustness properties of the MNEDE and prove that it is asymptotically fully efficient under a specified regular parametric family of densities. Also our simulation results show that unlike the MHDE the MNEDE is robust not only against outliers, but also against inliers, defined as values with less data than expected.

AMS subject classification: primary 62F35; 62F10; secondary 62F12

Keywords: Minimum disparity estimation; Residual adjustment function; Hellinger distance; Asymptotic efficiency; Robustness; Maximum likelihood estimator; Outliers; Inliers

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population having a continuous probability distribution which is a member of a specified parametric family of densities $\{f_\theta: \theta \in \Theta\}$, where Θ is a subset of $\mathbb{R}$. For ease of presentation, we discuss our results for a scalar parameter θ, but our results hold for a vector valued parameter. Let $\mathscr{G}$ denote the class of continuous densities topologized by the L_2-norm, and let $\mathscr{F}_\theta = \{f_\theta: \theta \in \Theta\}$ be a parametric subclass of $\mathscr{G}$ and we are interested in estimating θ. In parametric estimation, there are two issues of fundamental importance which the

statistician has to deal with. The first is the efficiency of the estimator when the model is correctly specified, and the second is the robustness of the estimator under deviations from the true model. Unfortunately, these two are usually competing concepts. The maximum likelihood estimator, which has full asymptotic efficiency among regular estimators, usually has poor robustness properties; on the other hand, the class of robust M-estimators achieve their robustness at the cost of first-order efficiency at the model (e.g., Hampel et al., 1986).

The conflicting concepts of robustness and efficiency are at least partially reconciled by some density-based minimum 'distance' estimators such as the Hellinger distance and its relatives (Beran, 1977; Stather, 1981; Tamura and Boos, 1986; Simpson, 1987, 1989; Lindsay, 1994). The pioneering work by Beran (1977) first showed that the minimum Hellinger distance estimator can simultaneously achieve first-order efficiency and robustness properties. Note that asymptotic efficiency under the model implies that the estimator must have the same influence function as the maximum likelihood estimator – – which can be potentially unbounded. The above-mentioned references show that the minimum Hellinger distance estimator (MHDE) enjoys robustness properties in spite of this. Beran has showed that under gross error contaminations and compact parameter spaces the minimum Hellinger distance functional stays bounded. Tamura and Boos (1986) showed that the affine-invariant MHDE for multivariate location and covariance has a breakdown point of at least $\frac{1}{4}$. This result is important because it is independent of the dimension of the parameter space. In contrast, the breakdown point of affine-invariant M-estimator is at most $1/(d+1)$, where d is the dimension of parameter space; in addition, the M-estimator is not fully efficient at the model. Simpson (1987) showed that the MHDE has 50% breakdown at some discrete models such as the Poisson. Lindsay (1994) gave a general result for the breakdown points for the class of *minimum disparity estimators* (a subclass of density-based minimum distance estimators) in terms of a characterizing function of the distance called the *residual adjustment function* (RAF) in discrete models. We will discuss the RAFs and their role in estimation in Section 2.

Throughout this paper the range of all the integrals will be $\mathbb{R}$. Let $\hat{g}_n$ denote a nonparametric density estimator of the true density g based on the data $X_1, X_2, \ldots, X_n$. It is usually obtained by the kernel density estimation method as

$$\hat{g}_n(x) = \int w(x; y, h_n)\, \mathrm{d}G_n(y),$$

where w is a smooth family of kernel functions like the normal densities with mean y and variance h_n^2, and G_n is the empirical distribution function. The MHDE is obtained by minimizing the squared Hellinger distance between $\hat{g}_n$ and f_θ

$$\mathrm{HD}(\hat{g}_n, \theta) = \int [(\hat{g}_n(x))^{1/2} - (f_\theta(x))^{1/2}]^2\, \mathrm{d}x.$$

The purpose of this paper is to establish that another density-based minimum distance estimator of θ, namely the minimum negative exponential disparity estimator

(MNEDE) is asymptotically as efficient as the MLE at the model and robust under data contamination. Properties of general disparities have been discussed by Lindsay (1994) in detail for discrete models. The MNEDE is obtained by minimizing the negative exponential disparity $D(\hat{g}_n, \theta)$ between $\hat{g}_n$ and f_θ over Θ, where $f_\theta \in \mathscr{F}_\Theta$, $g \in \mathscr{G}$,

$$D(g,\theta) = \int \{\exp[-\delta(g,\theta,x)] - 1\} f_\theta(x)\,dx \tag{1.1}$$

and

$$\delta(g,\theta,x) = \left(\frac{g(x)}{f_\theta(x)} - 1\right). \tag{1.2}$$

Following Lindsay (1994) we will call $\delta(g,\theta,x)$ the *Pearson residual* at x. Moreover, unlike the MHDE, the MNEDE is robust against *inliers* (Lindsay, 1994) which is demonstrated by our numerical results (Section 5, Tables 5 and 7). Some evidence of the robustness of the MNEDE is provided by Lindsay (1994) and Basu and Sarkar (1994a). The form of the disparity in (1.1) is natural in the sense that in this case the disparity is nonnegative taking the value zero if and only if $g = f_\theta$ (Proposition 1 below). However, the development of the procedure will be facilitated (Sections 2.1 and 2.2 below) by considering the disparity

$$D_M(g,\theta) = \int \{\exp[-\delta(g,\theta,x)] - 2\} f_\theta(x)\,dx. \tag{1.3}$$

Note that (1.1) and (1.3) differ only by a constant and have the same estimating properties because (1.1) and (1.3) are minimized at the same value of θ, i.e., one can equivalently consider the MNEDE to be the minimizer of (1.3).

Instead of a geometric interpretation, in which an observation can be called an outlier if it is far away from a bulk of the data, we will use a probabilistic interpretation to characterize an observation as an outlier. In this case an observation will be called an outlier if it would be a very unlikely one if the fitted model were true. Such a probabilistic outlier can also be called a *surprising* observation (Lindsay, 1994). Note that the Pearson residuals defined in Eq. (1.2) will all have values zero if the data exactly fit the model. Large positive values of the Pearson residuals correspond to probabilistic outliers (surprising observations). Density-based minimum distance estimators such as the MHDE or MNEDE are robust in the sense that they reduce the impact of surprising observations on the parameter estimates (we will discuss this in more detail in Section 2). The MNEDE, in addition, reduces the impact of Pearson inliers on the parameter estimates also, something which the MHDE fails to do. Inliers correspond to negative values of the Pearson residual $\delta(g,\theta,x)$, i.e., values with less observations than expected under the model — and have received little attention in the robustness literature. Lindsay (1994), however, recognized that the treatment of Pearson inliers by some popular estimators such as the MHDE can be the source of a problem; inliers can cause a larger bias in the MHDE relative to the MLE. This point is further discussed in Sections 2 and 5 below. Some empirical evidence of the

problem caused by inliers in discrete models have been provided in Harris and Basu (1994) and Basu et al. (1996) where it has been shown that an adjustment for the extreme inlying cells can often lead to significant improvements in the small sample performance of the procedures based on the Hellinger distance.

Although in this paper we consider the MNEDE, it is possible to extend all the ideas to a general class of disparities having bounded RAFs (defined in Section 2 below), which provides a class of procedures for obtaining robust and asymptotically efficient estimators. A natural question is how do these procedures compare among themselves in terms of robustness and efficiency. There has not been any comprehensive study on this, although some comparisons of the minimum Hellinger distance estimator and the minimum negative exponential disparity estimator, based on the approach of model smoothing as defined in (2.6) below, is provided in Basu and Sarkar (1994a).

The rest of the paper is organized as follows. In Section 2 we discuss previous works on the negative exponential disparity and provide some rationale (Section 2.2) as to why the complex looking negative exponential disparity leads to an estimator which is fully efficient and robust against both outliers and inliers. Sections 3 and 4, respectively, contain the asymptotic efficiency and robustness results for the MNEDE. In Section 5 we present results of a Monte Carlo study comparing the efficiency and robustness of the MNEDE to those of the MHDE and MLE. Some conclusive remarks are given in Section 6.

2. The negative exponential disparity

2.1. Discrete models

For discrete models, Lindsay (1994) has introduced the MNEDE as a member of the general class of density-based minimum distance estimators. Here we briefly discuss minimum disparity estimation in discrete models. Let $\{f_\theta(x)\}$ represent a family of probability mass functions having a countable support and indexed by θ. The discrete case does not involve kernel density estimation. Given a random sample $\{X_1, X_2, \ldots, X_n\}$ from a distribution g, define $\hat{g}_n(x)$ to be the observed proportion of X_i's taking the value x and let

$$\delta(\hat{g}_n, \theta, x) = [\hat{g}_n(x) - f_\theta(x)]/f_\theta(x)$$

denote the 'Pearson' residual at the value x, which depends on the data and the parameter θ. When there is no scope for confusion, we will write $\delta(\hat{g}_n, \theta, x)$ simply as $\delta(x)$. Let G be a three times differentiable, strictly convex function with $G(0) = 0$. Then, the nonnegative disparity measure ρ corresponding to G is defined as

$$\rho(\hat{g}_n, \theta) \equiv \sum_x G(\delta(x)) f_\theta(x). \tag{2.1}$$

The value of θ that minimizes (2.1) is called the minimum disparity estimator. When $G(\delta) = (\delta + 1)\log(\delta + 1)$, the disparity

$$\mathrm{LD}(\hat{g}_n, \theta) = \sum_x \hat{g}_n(x)[\log(\hat{g}_n(x)) - \log(f_\theta(x))] \tag{2.2}$$

is called the likelihood disparity, and its minimizer is the maximum likelihood estimator (MLE) of θ. Note that (2.2) is a form of Kullback–Leibler divergence. On the other hand, $G(\delta) = [(\delta + 1)^{1/2} - 1]^2$ generates the squared Hellinger distance and

$$G(\delta) = [e^{-\delta} - 1] \tag{2.3}$$

generates the negative exponential disparity. (The alternative form of the negative exponential disparity in (1.3) corresponds to $G(\delta) = [e^{-\delta} - 2]$.) Other examples of disparities include the Pearson's chi-square, Neyman's chi-square, the power divergence family (Cressie and Read, 1984), the blended weight Hellinger distance family (Lindsay, 1994; Basu and Sarkar, 1994b, Shin et al., 1995) and the blended weight chi-square family (Shin et al., 1996).

Let ∇ represent the gradient with respect to θ. Under differentiability of the model, the minimum disparity estimating equation takes the form

$$-\nabla\rho = \sum_x A(\delta(x))\nabla f_\theta(x) = 0,$$

where

$$A(\delta) = (\delta + 1)[\dot{G}(\delta)] - G(\delta) \tag{2.4}$$

and $\dot{G}(\delta)$ denotes the first derivative of $G(\delta)$. The function $A(\delta)$ is an increasing function on $[-1, \infty)$ and can be redefined, without changing the estimating properties of the disparity, so that $A(0) = 0$ and $\dot{A}(0) = 1$, where $\dot{A}(\delta)$ denotes the first derivative of $A(\delta)$. This function $A(\delta)$ is called the residual adjustment function (RAF) of the disparity and plays a key role in determining the theoretical properties of the estimators. For the likelihood disparity the RAF is linear with $A(\delta) = \delta$, and after the above standardization, for the Hellinger distance $A(\delta) = 2[(\delta + 1)^{1/2} - 1]$ and for the negative exponential disparity

$$A(\delta) = 2 - (2 + \delta)e^{-\delta}. \tag{2.5}$$

Note that this corresponds exactly to the modified definition of the negative exponential disparity D_M in (1.3) with the associated function G given by

$$G(\delta) = (e^{-\delta} - 2).$$

2.2. Development of the negative exponential disparity

The graphs of the RAFs for the Hellinger distance and the negative exponential disparity, together with that of the likelihood disparity which corresponds to $A(\delta) = \delta$, are provided in Fig. 5 of Lindsay (1994). The figure shows that the RAFs for

both the Hellinger distance and negative exponential disparity have strong downweighting effects on surprising observations (large positive Pearson residuals), i.e., $A(\delta) \ll \delta$ for large positive δ. The curvature parameter of the disparity, which is the second derivative of the RAF at zero, denoted by A_2, has been used by Lindsay as a measure of the trade-off between robustness and second-order efficiency. Large negative values of A_2 correspond to robustness properties, since in this case the RAF quickly becomes flat compared to that of the likelihood disparity, thus providing higher downweighting effect for surprising observations relative to the likelihood disparity (note that for the Hellinger distance $A_2 = -0.5$). On the other hand, when the curvature parameter is zero, the estimator is second-order efficient in the sense of Rao (1961). However, looking at the negative side of the δ axis in Fig. 5 of Lindsay (1994), one can see that the Hellinger distance fails to shrink the effect of large *negative* residuals, in fact it magnifies them. If one wishes the estimation procedure to be robust against both inliers and outliers (relative to the maximum likelihood estimation), in the sense that $|A(\delta)| \ll |\delta|$ for large (in magnitude) δ values, the corresponding RAF must have curvature parameter zero, since it must cross $A(\delta) = \delta$ at $\delta = 0$; thus the procedure must be second-order efficient. The third derivative of this RAF, if not itself zero, must be negative so that large Pearson residuals (both positive and negative) shrink towards zero. These are precisely the considerations which led to the development of the negative exponential disparity, for which the second derivative of the RAF at zero is zero and its third derivative at zero is negative one (see Lindsay, 1994, Section 7.2). In fact, the negative exponential disparity can be thought to be generated starting from the RAF in (2.5) via Eq. (15) of Lindsay (1994), who showed that for any given differentiable increasing function $A(\delta)$, one can construct an associated disparity measure ρ. Thus, the negative exponential disparity provides downweighting for all residuals (positive and negative) while the corresponding estimator is second-order efficient at the model.

2.3. Continuous models

To generalize Lindsay's (1994) work to continuous models $\{f_\theta: \theta \in \Theta\}$, Basu and Lindsay (1994) also apply the same smoothing to the model density f_θ that is applied to the data to define

$$\hat{f}_\theta(x) \equiv \int w(x; y, h_n)\, \mathrm{d}F_\theta(y),$$

where F_θ is the cumulative distribution function of f_θ. Then, to obtain the estimator of θ for a disparity measure ρ corresponding to a convex, thrice differentiable function G with $G(0) = 0$, Basu and Lindsay minimize

$$\int G\left(\frac{\hat{g}_n(x) - \hat{f}_\theta(x)}{\hat{f}_\theta(x)}\right) \hat{f}_\theta(x)\, \mathrm{d}x \tag{2.6}$$

instead of minimizing

$$\rho(\hat{g}_n, \theta) \equiv \int_{-\infty}^{\infty} G\left(\frac{\hat{g}_n(x) - f_\theta(x)}{f_\theta(x)}\right) f_\theta(x)\,dx \tag{2.7}$$

with respect to θ while keeping the bandwidth h_n of the kernel function constant. For example, in case of the negative exponential disparity their approach minimizes

$$\int_{-\infty}^{\infty} \left\{\exp\left[-\left(\frac{\hat{g}_n(x)}{\hat{f}_\theta(x)} - 1\right)\right] - 1\right\} \hat{f}_\theta(x)\,dx,$$

which is the negative exponential disparity between $\hat{g}_n$ and $\hat{f}_\theta$. Note that $\hat{g}_n(x)$ is an unbiased estimator of $\hat{f}_\theta(x)$. In the Basu–Lindsay approach the minimum disparity estimators are robust and generally consistent for a given value of the bandwidth of the kernel function which is not varied with the sample size n. One does not need to let h_n go to zero as $n \to \infty$, as is conventionally done. There is no loss in efficiency due to the smoothing of the model, if suitable kernels called *transparent kernels*, like the normal kernel for the normal model, are used (Basu and Lindsay, 1994; Basu and Sarkar, 1994c). If, however, a transparent kernel is not used, the minimum disparity estimators are asymptotically normal, but no longer enjoy full asymptotic efficiency. To overcome this problem, in this paper we combine the ideas of Beran (1977), Tamura and Boos (1986) and Lindsay (1994) and establish the asymptotic efficiency and robustness of the MNEDE irrespective of the transparency of the kernel. The numerical results presented in Section 5 also show the insensitivity of the MNEDE to inliers, a property that the MHDE does not share.

3. Asymptotic efficiency

In this section, under some regularity conditions, we first show the existence and consistency of the MNEDE and then establish that the MNEDE, like the MHDE, is asymptotically as efficient as the MLE under $\mathscr{F}_\theta$. Under differentiability of the model $f_\theta(x)$ with respect to θ, let $u(\theta, x) = \partial \log f_\theta(x)/\partial\theta$ and let

$$I(\theta) = \int_{-\infty}^{\infty} u^2(\theta, x) f_\theta(x)\,dx \tag{3.1}$$

denote the Fisher Information. We first consider the problem of existence of the MNEDE. Define the negative exponential disparity functional $T: \mathscr{G} \to \Theta$ as $T(g) = \theta_g$ where θ_g satisfies

$$D(g, \theta_g) = \min_{\theta} D(g, \theta), \tag{3.2}$$

provided such a θ_g exists, where $D(\cdot,\cdot)$ is as defined in (1.1). Note that θ_g also minimizes $D_M(g, \theta)$. Since $T(g)$ may be multiple valued we will use the notation $T(g)$ to indicate any one of the possible values chosen arbitrarily. Our first proposition

gives conditions for the existence of θ_g. Unless mentioned otherwise, all the limits we consider below will be as $n \to \infty$.

Proposition 1. *Assume that*

(a) *the parameter space Θ is compact;*

(b) *for $\theta_1 \neq \theta_2$ $f_{\theta_1}(x) \neq f_{\theta_2}(x)$ on a set of positive Lebesgue measure;*

(c) *$f_\theta(x)$ is continuous in θ for almost all x (with respect to the Lebesgue measure).*

Then, (i) *for any $g \in \mathcal{G}$ there exists a $\theta_g \in \Theta$ such that $T(g) = \theta_g$, and* (ii) *for any $\theta^* \in \Theta$, $T(f_{\theta^*}) = \theta^*$ is unique.*

Proof. Fix $g \in \mathcal{G}$. We will show that $D(g, \theta)$ is continuous in θ. From assumption (a) it would then follow that there exists a $\theta_g \in \Theta$ such that $D(g, \theta_g) = \min_\theta D(g, \theta)$. We now proceed to establish continuity of $D(g, \theta)$ in θ. Let $\theta_n \to \theta$ as $n \to \infty$. We show that $D(g, \theta_n) \to D(g, \theta)$ as $n \to \infty$. Consider

$$D(g, \theta_n) = \exp(1) \int \exp[-g(x)/f_{\theta_n}(x)] f_{\theta_n}(x)\,dx - 1.$$

Note that

$$\exp[-g(x)/f_{\theta_n}(x)] f_{\theta_n}(x) \leqslant f_{\theta_n}(x)$$

and by assumption (c) $\exp[-g(x)/f_{\theta_n}(x)] f_{\theta_n}(x)$ converges to $\exp[-g(x)/f_\theta(x)] f_\theta(x)$ while $f_{\theta_n}(x)$ converges to $f_\theta(x)$. Since $\lim_{n\to\infty} \int f_{\theta_n}(x)\,dx = \int f_\theta(x)\,dx = 1$, by a generalized version of the dominated convergence theorem (Royden, 1968, p. 89) $\int \exp[-g(x)/f_{\theta_n}(x)] f_{\theta_n}(x)\,dx$ converges to $\int \exp[-g(x)/f_\theta(x)] f_\theta(x)\,dx$ and hence $D(g, \theta_n) \to D(g, \theta)$ as $n \to \infty$.

To prove (ii), note that alternatively for $G^*(\delta) = (e^{-\delta} - 1 + \delta)$, instead of G in (2.3), we can write

$$D(f_{\theta^*}, \theta) = \int G^*(\delta(f_{\theta^*}, \theta, x)) f_\theta(x)\,dx.$$

Now $G^*(\delta)$ is a nonnegative and strictly convex function with $\delta = 0$ as the unique point of minimum. Therefore, for every θ, $D(f_{\theta^*}, \theta) \geqslant 0$ and $D(f_{\theta^*}, \theta) = 0$ if and only if $\delta(f_{\theta^*}, \theta, x)$ is equal to zero on the support of the distribution f_{θ^*}, which, by assumption (b), is true if and only if $\theta = \theta^*$. Therefore, $D(f_{\theta^*}, \theta)$ is uniquely minimized at $\theta = \theta^*$. □

Using the line of reasoning used in the proof of Proposition 1 it is also possible to give sufficient conditions for the existence of a minimum disparity estimator for a general disparity measure ρ defined in (2.7). This can be done by assuming, for instance, that the RAF for the disparity is bounded.

We can also apply Proposition 1 to a location-scale family

$$\left\{ f_\theta = \frac{1}{\sigma} f\left(\frac{1}{\sigma}(x - \mu)\right) : \theta = (\mu, \sigma) \in (-\infty, \infty) \times (0, \infty) \right\}, \quad f \text{ continuous},$$

where the parameter space is not compact. This is because (μ,σ) can be reparameterized as $\beta=(\beta_1,\beta_2)$, where (β_1,β_2) is defined by $\mu=\tan(\beta_1)$, $\sigma=\tan(\beta_2)$, $(\beta_1,\beta_2)\in\Theta_*=(-\pi/2,\pi/2)\times(0,\pi/2)$, and the arguments in the last paragraph after the proof of Theorem 1 in Beran (1977, pp. 447–448) are applicable.

Having established the existence of the MNEDE in the previous proposition, we now turn to study some large sample properties which are closely connected to the continuity of the functional T. We have the following result on the continuity of T.

Proposition 2. *Let g_0 be any fixed density in $\mathscr{G}$ and let $\{g_n\}$ be a sequence of densities in $\mathscr{G}$. If $T(g_0)$ is unique then under the assumptions of Proposition 1 the functional T is continuous at g_0 in the sense that if $g_n\to g_0$ in L_1 then $T(g_n)$ converges to $T(g_0)$.*

Proof. Let $g_n\to g_0$ in L_1. We will show that $T(g_n)\to T(g_0)$. By continuity of $D(g_n,\theta)$, $n\geqslant 0$, there exists θ_n such that $D(g_n,\theta_n)=\min_\theta D(g_n,\theta)$. Thus establishing continuity of T is equivalent to showing that $\theta_n\to\theta_0$. We now claim that it is enough to show that

$$\sup_\theta |D(g_n,\theta)-D(g_0,\theta)|\to 0. \tag{3.3}$$

To see this, note that if θ_n does not converge to θ_0, by compactness of Θ, there exists $\theta_*\neq\theta_0\in\Theta$ and a subsequence θ_{n_k} such that $\theta_{n_k}\to\theta_*$ and hence by the continuity of $D(g_0,\theta)$ we have $D(g_0,\theta_{n_k})\to D(g_0,\theta_*)$. Also from (3.3) and the definition of θ_n for $n\geqslant 0$, it follows that $D(g_0,\theta_n)\to D(g_0,\theta_0)$. Therefore, $D(g_0,\theta_*)=D(g_0,\theta_0)$, contradicting the uniqueness of T. Finally, to establish (3.3) note that for a fixed x, letting $y=g_n(x)$ and $y_0=g_0(x)$, by the mean value theorem for the function $\exp[-y/f_\theta(x)]$ we have

$$\exp\left[-\frac{g_n(x)}{f_\theta(x)}\right]-\exp\left[-\frac{g_0(x)}{f_\theta(x)}\right]=[g_n(x)-g_0(x)]\left\{\frac{1}{f_\theta(x)}\exp\left[-\frac{g_n^*(x)}{f_\theta(x)}\right]\right\},$$

where $g_n^*(x)$ is a point between $y=g_n(x)$ and $y_0=g_0(x)$. Note that $g_n^*(x)/f_\theta(x)\geqslant 0$ and hence $\exp[-g_n^*(x)/f_\theta(x)]\leqslant 1$ for all n, for all x and for all θ. Therefore,

$$\begin{aligned}|D(g_n,\theta)-D(g_0,\theta)| &\leqslant \exp(1)\int\left|\exp\left[-\frac{g_n(x)}{f_\theta(x)}\right]-\exp\left[-\frac{g_0(x)}{f_\theta(x)}\right]\right| f_\theta(x)\,dx\\ &=\exp(1)\int\left|[g_n(x)-g_0(x)]\left\{\frac{1}{f_\theta(x)}\exp\left[-\frac{g_n^*(x)}{f_\theta(x)}\right]\right\}\right| f_\theta(x)\,dx\\ &\leqslant \exp(1)\int|g_n(x)-g_0(x)|\,dx.\end{aligned}$$

Therefore, $\sup_\theta|D(g_n,\theta)-D(g_0,\theta)|\leqslant\exp(1)\int|g_n(x)-g_0(x)|\,dx\to 0$. This completes the proof of (3.3). □

In the rest of the paper we assume that the assumptions of Proposition 1 hold.

Remark 1. Let the true density $f_{\theta_0} \in \mathscr{F}_\Omega$. Define the kernel density estimate of $f_{\theta_0}(x)$ by

$$\hat{g}_n(x) \equiv \frac{1}{nh_n} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h_n}\right), \tag{3.4}$$

where w is any nonnegative Borel measurable function such that $\int w(x)\,dx = 1$. Assume that $J_n \equiv \int |\hat{g}_n(x) - f_{\theta_0}(x)|\,dx$ converges to 0 almost surely (a.s.). Then, by Proposition 2, $\hat{\theta}_n = T(\hat{g}_n)$ converges to $\theta_0 = T(f_{\theta_0})$ a.s. Hence, $f_{\hat{\theta}_n}(x) \to f_{\theta_0}(x)$ a.s. for every x, and by Glick's theorem (Devroye and Gyorfi, 1985, p. 10) $\int |f_{\hat{\theta}_n}(x) - f_{\theta_0}(x)|\,dx$ converges to 0 a.s. To obtain the rate of convergence of J_n to zero, we can apply Theorem 1 in Chap. 1 of Devroye and Gyorfi (1985) to get $P(J_n > \varepsilon) \leqslant e^{-rn}$, where $r > 0$ does not depend on the true density f_{θ_0}.

Again by Theorem 1 in Chap. 1 of Devroye (1987), a necessary and sufficient condition for $\int |\hat{g}_n(x) - f_{\theta_0}(x)|\,dx \to 0$ a.s. is that $h_n \to 0$ and $nh_n \to \infty$, and from the previous paragraph it follows that $\int |f_{\hat{\theta}_n}(x) - f_{\theta_0}(x)|\,dx \to 0$ a.s. However, a more interesting question is whether the negative exponential disparity between $f_{\hat{\theta}_n}$ and f_{θ_0} converges to 0 a.s. In fact, an application of the dominated convergence theorem yields the following result.

Proposition 3. *Let $\hat{g}_n$ be as defined in* (3.4) *and assume that $h_n \to 0$ and $nh_n \to \infty$. Then $D(f_{\hat{\theta}_n}, \theta_0)$ converges to 0 a.s.*

Remark 2. In practice, one often uses the automatic kernel density estimator (see Devroye, 1987) of the form

$$\tilde{g}_n(x) \equiv (nc_n s_n)^{-1} \sum_{i=1}^{n} w[(c_n s_n)^{-1}(x - X_i)],$$

with bandwidth $c_n s_n$, where $s_n = s(X_1, \ldots, X_n)$ is a robust scale estimator and c_n is a sequence of positive constants. This allows one to choose the amount of smoothing as a function of the amount of variation in the data, and may help prevent smoothing the data too much or too little. In this case under the assumptions that $c_n \to 0$, $nc_n \to \infty$ and $n^{1/2}(s_n - s)$ is bounded a.s. where s is a finite, positive constant, we have $\int |\tilde{g}_n(x) - f_{\theta_0}(x)|\,dx \to 0$ a.s. (Devroye and Gyorfi, 1985, Chap. 5, Theorem 3). Then it follows from Proposition 2 that $f_{\tilde{\theta}_n}(x) \to f_{\theta_0}(x)$ a.s. for every x where $\tilde{\theta}_n$ denotes $T(\tilde{g}_n)$. Then, as in Proposition 3, one can show that $D(f_{\tilde{\theta}_n}, \theta_0)$ converges to zero a.s.

Having established the consistency of the MNEIDE we now proceed to establish its asymptotic normality. In what follows, we will assume that the model $f_\theta(x)$ is twice continuously differentiable with respect to θ, and $D(g, \theta)$ and $D_M(g, \theta)$ can be twice differentiated with respect to θ under the integral sign. Since $A(\delta)$ and $\dot{A}(\delta)(1 + \delta)$ are bounded for the negative exponential disparity, a set of sufficient conditions for the above are: For any $\theta \in \Theta$, $\varepsilon > 0$, and $\theta' \in (\theta - \varepsilon, \theta + \varepsilon)$,

(i) $|\dot{f}_{\theta'}(x)| < K_\theta(x)$, $\int K_\theta(x)\,dx < \infty$;

(ii) $|\dot{f}'_{\theta'}(x)| < L_\theta(x)$, $\int L_\theta(x)\,dx < \infty$;

(iii) $|u^2(\theta', x) f_{\theta'}(x)| < M_\theta(x)$, $\int M_\theta(x)\,dx < \infty$.

Theorem 1. *Let the true density belong to $\mathscr{F}_\theta$ and be denoted by f_θ. Let $\hat{\theta}_n \equiv T(\hat{g}_n)$ where $\hat{g}_n$ is the kernel density estimate defined in (3.4). Assume the following:*

(a) *For any sequence of estimators $\{\varphi_n\}$ converging to $\theta_0 \in \Theta$ in probability, $\int |\ddot{f}_{\varphi_n}(x) - \ddot{f}_{\theta_0}(x)|\,dx$ converges to zero in probability.*

(b) *For any sequence of estimators $\{\varphi_n\}$ converging to $\theta_0 \in \Theta$ in probability, $\int |u^2(\varphi_n, x) f_{\varphi_n}(x) - u^2(\theta_0, x) f_{\theta_0}(x)|\,dx$ converges to zero in probability.*

(c) $I(\theta_0) < \infty$, *and* $\int [u^2(\theta_0, x + a) f_{\theta_0}(x) - u^2(\theta_0, x) f_{\theta_0}(x)]\,dx \to 0$ *as* $|a| \to 0$.

(d) $\limsup_{n \to \infty} \sup_{y \in \mathscr{A}_n} \int |\dot{f}^{(2)}_{\theta_0}(x + y) u_{\theta_0}(x)|\,dx < \infty$,

where $\mathscr{A}_n = \{y : y = h_n z, z \in S\}$ and $\dot{f}^{(2)}_{\theta_0}(x)$ denotes the second derivative of $f_{\theta_0}(x)$ with respect to x.

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

(f) *w is symmetric about 0, has compact support S, and is twice continuously differentiable.*

For the following assumptions let $\{\alpha_n\}$ denote a sequence of positive real numbers going to infinity. In condition (h), $\chi(\cdot)$ denotes the indicator function.

(g) $n \sup_{t \in S} P(|X_1 - h_n t| > \alpha_n) \to 0$.

(h) $n^{-1/2} h_n^{-1} (\int |u(\theta_0, x)| \chi(|x| \leqslant \alpha_n)\,dx) \to 0$.

(i) $M_n = \sup_{x \leqslant x_n} \sup_{t \in S} \{f_{\theta_0}(x + h_n t)/f_{\theta_0}(x)\} = O(1)$.

Then, $n^{1/2}(\bar{\theta}_n - \theta_0)$ converges in distribution to $N[0, I^{-1}(\theta_0)]$.

Proof. First observe that by condition (e) and Remark 1 we have

$$\hat{g}_n(x) \to f_{\theta_0}(x) \text{ a.s.} \quad \text{for every } x,$$

and

$$\int |\hat{g}_n(x) - f_{\theta_0}(x)|\,dx \to 0 \text{ a.s.}$$

Let $\dot{D}_M(\hat{g}_n, \theta)$ and $\ddot{D}_M(\hat{g}_n, \theta)$ denote the first and second derivatives of $D_M(\hat{g}_n, \theta)$ with respect to θ. Since $\hat{\theta}_n$ minimizes $D_M(\hat{g}_n, \theta)$ over Θ, the Taylor series expansion of $\dot{D}_M(\hat{g}_n, \hat{\theta}_n)$ around θ_0 yields:

$$0 = \dot{D}_M(\hat{g}_n, \hat{\theta}_n) = \dot{D}_M(\hat{g}_n, \theta_0) + (\hat{\theta}_n - \theta_0) \ddot{D}_M(\hat{g}_n, \theta_n^*),$$

where θ_n^* is a point between θ_0 and $\hat{\theta}_n$. Hence,

$$n^{1/2}(\hat{\theta}_n - \theta_0) = [\ddot{D}_M(\hat{g}_n, \theta_n^*)]^{-1} [-n^{1/2} \dot{D}_M(\hat{g}_n, \theta_0)].$$

It now suffices to prove that

$$\ddot{D}_M(\hat{g}_n, \theta_n^*) \xrightarrow{P} I(\theta_0) \tag{3.5}$$

and

$$-n^{1/2} \dot{D}_M(\hat{g}_n, \theta_0) \xrightarrow{L} N(0, I(\theta_0)). \tag{3.6}$$

We first prove (3.5). Note that

$$\ddot{D}_M(\hat{g}_n, \theta_n^*) = -\int A(\delta(\hat{g}_n, \theta_n^*, x))\ddot{f}_{\theta_n^*}(x)\,dx + \int \dot{A}(\delta(\hat{g}_n, \theta_n^*, x))[1 + \delta(\hat{g}_n, \theta_n^*, x)]u^2(\theta_n^*, x)f_{\theta_n^*}(x)\,dx,$$

where $\dot{A}(\delta) = (1 - \delta)\exp[-\delta]$. Now

$$\int A(\delta(\hat{g}_n, \theta_n^*, x))[\ddot{f}_{\theta_n^*}(x) - \ddot{f}_{\theta_0}(x)]\,dx \leqslant \int |A(\delta(\hat{g}_n, \theta_n^*, x))||\ddot{f}_{\theta_n^*}(x) - \ddot{f}_{\theta_0}(x)|\,dx$$

$$\leqslant 2\int |\ddot{f}_{\theta_n^*}(x) - \ddot{f}_{\theta_0}(x)|\,dx \xrightarrow{P} 0 \tag{3.7}$$

using assumption (a) and the fact that for the negative exponential disparity $\sup_\delta |A(\delta)| \leqslant 2$. Also

$$\int \dot{A}(\delta(\hat{g}_n, \theta_n^*, x))[1 + \delta(\hat{g}_n, \theta_n^*, x)][u^2(\theta_n^*, x)f_{\theta_n^*}(x) - u^2(\theta_0, x)f_{\theta_0}(x)]\,dx$$

$$\leqslant \int |\dot{A}(\delta(\hat{g}_n, \theta_n^*, x))[1 + \delta(\hat{g}_n, \theta_n^*, x)]||u^2(\theta_n^*, x)f_{\theta_n^*}(x) - u^2(\theta_0, x)f_{\theta_0}(x)|\,dx$$

$$\leqslant B_1 \int |u^2(\theta_n^*, x)f_{\theta_n^*}(x) - u^2(\theta_0, x)f_{\theta_0}(x)|\,dx \xrightarrow{P} 0 \tag{3.8}$$

by assumption (b), and the fact that for the negative exponential disparity $\sup_\delta |\dot{A}(\delta)(1 - \delta)| \leqslant B_1$ where B_1 is a positive constant. Now using the dominated convergence theorem we have

$$\int A(\delta(\hat{g}_n, \theta_n^*, x))\ddot{f}_{\theta_0}(x)\,dx \xrightarrow{P} 0$$

and hence by (3.7)

$$\int A(\delta(\hat{g}_n, \theta_n^*, x))\ddot{f}_{\theta_n^*}(x)\,dx \xrightarrow{P} 0.$$

By assumption (c) and the dominated convergence theorem it follows that

$$\int \dot{A}(\delta(\hat{g}_n, \theta_n^*, x))[1 + \delta(\hat{g}_n, \theta_n^*, x)]u^2(\theta_0, x)f_{\theta_0}(x)\,dx \xrightarrow{P} \int u^2(\theta_0, x)f_{\theta_0}(x)\,dx,$$

and, hence,

$$\int \dot{A}(\delta(\hat{g}_n, \theta_n^*, x))[1 + \delta(\hat{g}_n, \theta_n^*, x)]u^2(\theta_n^*, x)f_{\theta_n^*}(x)\,dx \xrightarrow{P} \int u^2(\theta_0, x)f_{\theta_0}(x)\,dx$$

by (3.8). Therefore, we have (3.5).

Next we prove (3.6). Note that

$$n^{1/2}\dot{D}_M(\hat{g}_n,\theta_0) = n^{1/2}\int A(\delta(\hat{g}_n,\theta_0,x))\dot{f}_{\theta_0}(x)\,dx.$$

Therefore, it is enough to prove that

$$n^{1/2}\int \delta(\hat{g}_n,\theta_0,x)\dot{f}_{\theta_0}(x)\,dx \xrightarrow{L} N(0,I(\theta_0)) \tag{3.9}$$

and

$$n^{1/2}\int [A(\delta(\hat{g}_n,\theta_0,x)) - \delta(\hat{g}_n,\theta_0,x)]\dot{f}_{\theta_0}(x)\,dx \xrightarrow{P} 0. \tag{3.10}$$

Observe that

$$n^{1/2}\int \delta(\hat{g}_n,\theta_0,x)\dot{f}_{\theta_0}(x)\,dx = n^{1/2}\int [\hat{g}_n(x) - f_{\theta_0}(x)]u(\theta_0,x)\,dx$$

and hence (3.9) follows from Beran (1977, Eqs. (3.12), (3.13), p. 452). Next we show (3.10). Note that

$$A(\delta(\hat{g}_n,\theta_0,x)) - \delta(\hat{g}_n,\theta_0,x) = A\left(\left(\left(\frac{\hat{g}_n(x)}{f_{\theta_0}(x)}\right)^{1/2}\right)^2 - 1\right) - \left[\left\{\left(\frac{\hat{g}_n(x)}{f_{\theta_0}(x)}\right)^{1/2}\right\}^2 - 1\right]$$

$$\leqslant B_2\left[\left(\frac{\hat{g}_n(x)}{f_{\theta_0}(x)}\right)^{1/2} - 1\right]^2 \tag{3.11}$$

for some $B_2 > 0$ since $|A(t^2-1) - (t^2-1)| \leqslant B_2(t-1)^2$ for every nonnegative t (see Lindsay, 1994, p. 1107). Thus,

$$\left|n^{1/2}\int [A(\delta(\hat{g}_n,\theta_0,x)) - \delta(\hat{g}_n,\theta_0,x)]\dot{f}_{\theta_0}(x)\,dx\right|$$

$$\leqslant n^{1/2}\int |A(\delta(\hat{g}_n,\theta_0,x)) - \delta(\hat{g}_n,\theta_0,x)|\,|\dot{f}_{\theta_0}(x)|\,dx$$

$$\leqslant n^{1/2}B_2\int [(g_n(x))^{1/2} - (f_{\theta_0}(x))^{1/2}]^2\,|u(\theta_0,x)|\,dx.$$

Now we consider

$$n^{1/2}\int [(\hat{g}_n(x))^{1/2} - (f_{\theta_0}(x))^{1/2}]^2\,|u(\theta_0,x)|\,dx. \tag{3.12}$$

It is bounded by the sum of two terms:

$$2n^{1/2}\int[(\hat{g}_n(x))^{1/2}-(E\{\hat{g}_n(x)\})^{1/2}]^2|u(\theta_0,x)|\,dx$$

and

$$2n^{1/2}\int[(E\{\hat{g}_n(x)\})^{1/2}-(f_{\theta_0}(x))^{1/2}]^2|u(\theta_0,x)|\,dx.$$

The first term represents the Hellinger deviation of the estimator $\hat{g}_n$ from its mean and its convergence to zero in probability has been established by Tamura and Boos (1986, p. 226) using conditions (g)–(i). The second term represents the bias in the Hellinger metric and its convergence to zero in probability follows from conditions (d) and (e). This completes the proof. □

Condition (a) in Theorem 1 says that the second derivative of $f_\theta(x)$ with respect to θ is L_1 continuous in probability at the model while condition (b) says that $E_\theta[u^2(\theta,x)]$ (expectation being computed with respect to the density $f_\theta(x)$) is continuous in probability at the mdel. Conditions (a) and (b) are simple continuity conditions and are satisfied, for example, by the distributions belonging to the exponential family. The second part of condition (c) is satisfied, for instance, if $u(\theta_0,x)$ is uniformly continuous in x on compact sets. The conditions (g)–(i) have been used and discussed by Tamura and Boos (1986, Theorem 4.1).

The problem of the bandwidth selection is very important and it has been studied by several authors. See, for example, Härdle et al. (1988), Marron (1989), and Hall and Marron (1991). For our problem we need $h_n \sim n^{-(1/2-\delta)}$ where $0 < \delta < \frac{1}{4}$, which is dictated by condition (e) of Theorem 1.

4. Robustness

We now study the robustness properties of the MNEDE. We do this by examining the behavior of the functional T defined by (3.2) under a mixture model for gross errors. Our approach lies in studying the α-influence curves of T, as was done by Beran (1977) for the MHDE. Beran showed that to assess the robustness of a functional with respect to the gross-error model it is necessary to examine the α-influence curve rather than the influence curve, except when the influence curve provides a uniform approximation to the α-influence curve. For this reason we study the α-influence curve for the MNEDE. The results are summarized in the following.

Theorem 2. *Let $f_{\theta,\alpha,z} \equiv (1-\alpha)f_\theta + \alpha\eta_z$, where η_z denotes the uniform density on the interval $(z-\varepsilon, z+\varepsilon)$, where $\varepsilon > 0$ is small, $\theta \in \Theta$, $\alpha \in (0,1)$, $z \in \mathbb{R}$. Then*

(i) *for every $\alpha \in (0,1)$ and every $\theta \in \Theta$, under the assumptions of Proposition 1, and under the condition that $T(f_{\theta,\alpha,z})$ is unique for all z, $T(f_{\theta,\alpha,z})$ is a bounded, continuous*

function of z such that

$$\lim_{|z|\to\infty} T(f_{\varepsilon,\alpha,z}) = \theta; \tag{4.1}$$

$$\text{(ii)}\quad \lim_{\varepsilon\to 0} \varepsilon^{-1}[T(f_{\varepsilon,\alpha,z}) - \theta] = [I(\theta)]^{-1}\int [\eta_z(x)u(\theta,x)]\,dx, \tag{4.2}$$

where $I(\theta)$ is as defined in (3.1).

Proof. Let θ_z denote $T(f_{\varepsilon,\alpha,z})$. We first show (4.1), i.e., $\theta_z \to \theta$ as $|z| \to \infty$. Suppose not, then without loss of generality, by going to a subsequence if necessary, we may assume that $\theta_z \to \theta_1 \neq \theta$ as $|z| \to \infty$. Observe that

$$D(f_{\varepsilon,\alpha,z}, \theta_z) \leqslant D(f_{\varepsilon,\alpha,z}, t) \tag{4.3}$$

for every $t \in \Theta$. Then by a generalized version of the dominated convergence theorem (Royden, 1968, p. 89)

$$D(f_{\varepsilon,\alpha,z}, \theta_z) \to D((1-\alpha)f_\theta, \theta_1) \tag{4.4}$$

as $|z| \to \infty$. By (4.3) and (4.4) we have

$$D((1-\alpha)f_\theta, \theta_1) \leqslant D((1-\alpha)f_\theta, t) \quad \forall t \in \Theta. \tag{4.5}$$

Now consider

$$D^*(\alpha, f_\theta, t) = \int (\exp[-(1-\alpha)\delta(f_\theta, t, x)] - 1 + (1-\alpha)\delta(f_\theta, t, x))f_t(x)\,dx$$

where δ is as defined in (1.2). Since

$$G^*(\delta) = (\exp[-(1-\alpha)\delta] - 1 + (1-\alpha)\delta)$$

is a nonnegative and strictly convex function of δ with $\delta = 0$ as the unique point of minimum, $D^*(\alpha, f_\theta, t) > 0$ unless $\delta(f_\theta, t, x) = 0$ on a set of Lebesgue measure zero which, by assumption (b) of Proposition 1, is true if and only if $t = \theta$. Since $\theta_1 \neq \theta$,

$$D^*(\alpha, f_\theta, \theta_1) > D^*(\alpha, f_\theta, \theta).$$

Because $D((1-\alpha)f_\theta, t)$ is a strictly increasing function of $D^*(\alpha, f_\theta, t)$, this implies

$$D((1-\alpha)f_\theta, \theta_1) > D((1-\alpha)f_\theta, \theta)$$

which contradicts (4.5). This concludes the proof of (4.1). The continuity of θ_z follows from Proposition 2, and the boundedness of $\{\theta_z; z \in \mathbb{R}\}$ follows from the compactness of Θ.

Now we prove part (ii). Note that since θ_z minimizes $D(f_{\varepsilon,\alpha,z}, t)$ over Θ, the Taylor series expansion of $\dot{D}(f_{\varepsilon,\alpha,z}, \theta_z)$ around θ gives:

$$0 = \dot{D}(f_{\varepsilon,\alpha,z}, \theta_z) = \dot{D}(f_{\varepsilon,\alpha,z}, \theta) + (\theta_z - \theta)\ddot{D}(f_{\varepsilon,\alpha,z}, \theta_z^*),$$

where θ_z^* is a point between θ and θ_z. Therefore, we have

$$\frac{\theta_z - \theta}{\alpha} = -\frac{\alpha^{-1}\dot{D}(f_{\theta,\alpha,z}, \theta)}{\ddot{D}(f_{\theta,\alpha,z}, \theta_z^*)}, \tag{4.6}$$

where

$$\dot{D}(f_{\theta,\alpha,z}, \theta) = -\int A(\delta(f_{\theta,\alpha,z}, \theta, x))\dot{f}_\theta(x)\,dx,$$

and

$$\ddot{D}(f_{\theta,\alpha,z}, \theta_z^*) = -\int A(\delta(f_{\theta,\alpha,z}, \theta_z^*, x))\ddot{f}_{\theta_z^*}(x)\,dx + \int \dot{A}(\delta(f_{\theta,\alpha,z}, \theta_z^*, x))[1 + \delta(f_{\theta,\alpha,z}, \theta_z^*, x)]u^2(\theta_z^*, x)f_{\theta_z^*}(x)\,dx,$$

where $A(\cdot)$ is the RAF. Now

$$\lim_{z \to 0} \ddot{D}(f_{\theta,\alpha,z}, \theta_z^*) = I(\theta)$$

and

$$\lim_{z \to 0} \alpha^{-1}\dot{D}(f_{\theta,\alpha,z}, \theta) = -\int \eta_z(x)u(\theta, x)\,dx,$$

the last limit follows by L'Hospital's rule. Therefore, by (4.6) the result follows. □

The limit in part (ii) of Theorem 2 viewed as a function of z, defines the influence curve of the functional T at f_θ. The influence curve of the MNEDE indicates its asymptotic efficiency at the model. Clearly, the right-hand side of Eq. (4.2) can be an unbounded function of z; however, from part (i) of Theorem 2, for every $\alpha \in (0, 1)$, $\alpha^{-1}[T(f_{\theta,\alpha,z}) - \theta]$, called the α-influence curve of T, is a bounded, continuous function of z and

$$\lim_{|z| \to \infty} \alpha^{-1}[T(f_{\theta,\alpha,z}) - \theta] = 0.$$

Thus, T is robust against $100\alpha\%$ contamination by gross errors at arbitrary z.

Since the influence curve of T is unbounded, it does not provide a uniform approximation to the bounded α-influence curves. Therefore, the robustness of T, with respect to the gross-error model, cannot be assessed just by examining the influence curve, indicating its limitation in this context.

5. Simulation results

In our simulation study we have compared MHDE and MNEDE to the MLE in terms of efficiency and robustness against outliers as well as inliers. We have done the computations for the normal model and also for the binomial model. For comparison

Huber's M-estimates for the location parameter have also been provided in the normal model.

5.1. The normal model

Under the normal model $N(\mu, \sigma^2)$ with both parameters unknown, we compared the performance of the MNEDE of μ with the MHDE, the MLE and Huber's M-estimator (Huber, 1964) when the data were generated from a variety of contaminated normal distributions. For computing the MNEDE and MHDE we used the *Epanechnikov* kernel to get an automatic kernel density estimate from the data of the form:

$$f^*(x) = \frac{1}{nc_n s_n} \sum_{i=1}^{n} w\left(\frac{x - X_i}{c_n s_n}\right),$$

where $w(x) = 0.75(1 - x^2)$ for $|x| \leqslant 1$, and 0 otherwise. The simulations were performed using the values 0.5, 0.6, 0.7, 0.8, 0.9 for c_n and the scale estimate s_n was set equal to $1.48 \times \text{median}(|X_i - \text{median}(X_i)|)$. The numerical integrations were performed with Simpson's one-third rule, and the Newton–Raphson algorithm was used to solve for the roots of the estimating equations. The estimates

$$\hat{\mu}^{(0)} = \text{median}(X_i), \qquad \hat{\sigma}^{(0)} = 1.48 \times \text{median}(|X_i - \hat{\mu}^{(0)}|)$$

were used as the starting values of μ and σ. Huber's M-estimate of the location parameter μ was computed by solving $\sum_i \psi(\sigma^{-1}(X_i - \mu)) = 0$ using Huber's ψ-function with the tuning constant $b = 1.345$ where $\psi(x) = x$ if $|x| \leqslant b$, $\psi(x) = b$ if $x > b$, $\psi(x) = -b$ if $x < -b$. During the iterative computation of the M-estimate of μ the scale parameter was estimated as $\hat{\sigma} = 1.48 \times \text{median}(|X_i - \text{median}(X_i)|)$ and kept fixed. All the results presented in this section are based on 1000 replications, with samples of size 50.

We considered the following cases: The data were generated from (i) pure $N(0, 1)$ distribution, (ii) $0.9\ N(0, 1) + 0.1\ N(3, 1)$, (iii) $0.9\ N(0, 1) + 0.1\ N(0, 25)$, and (iv) $0.5\ N(0, 1) + 0.5\ U(-1, 1)$. In each case, $\mu = 0$ is our target parameter. Case (ii) puts a small normal contamination with its mean being at a point three standard deviations away from the target parameter, whereas case (iii) creates a much heavier tail relative to the true distribution. Case (iv), on the other hand, studies the effect of making the tails lighter compared to the true distribution. The results are presented in Tables 1–4. On the basis of these limited simulations we see that the MHDE and MNEDE have good performances at the model (having small mean square errors) while being reasonably good under contamination. Also note that the MHDE and MNEDE are applicable to general parametric models and not restricted to location-scale models.

The choice of c_n in computing $f^*(x)$ is a delicate issue. Based on our remark after Theorem 1 regarding the choice of the bandwidth, for the uncontaminated model we

Table 1
Empirical mean and mean square error (in parentheses) of the MHDE and MNEDE of the location parameter for data generated under N(0, 1) for sample size 50 for different values of c_n

c_n	HD	NED
0.5	0.0006 (0.0208)	0.0011 (0.0210)
0.6	0.0007 (0.0207)	0.0010 (0.0208)
0.7	0.0008 (0.0206)	0.0010 (0.0207)
0.8	0.0008 (0.0205)	0.0010 (0.0206)
0.9	0.0009 (0.0204)	0.0010 (0.0206)

Mean (mean square error) of MLE = 0.0017 (0.0203).

Mean (mean square error) of Huber's M-estimate = 0.0010 (0.0214).

Table 2
Empirical mean and mean square error (in parentheses) of the MHDE and MNEDE of the location parameter for data generated under 0.9N(0, 1) + 0.1N(3, 1) for sample size 50 for different values of c_n

c_n	HD	NED
0.5	0.2260 (0.0907)	0.1849 (0.0746)
0.6	0.2343 (0.0939)	0.1931 (0.0777)
0.7	0.2408 (0.0965)	0.2005 (0.0806)
0.8	0.2462 (0.0987)	0.2077 (0.0833)
0.9	0.2510 (0.1007)	0.2146 (0.0859)

Mean (mean square error) of MLE = 0.2990 (0.1241).

Mean (mean square error) of Huber's M-estimate = 0.2011 (0.0717).

Table 3
Empirical mean and mean square error (in parentheses) of the MHDE and MNEDE of the location parameter for data generated under 0.9N(0, 1) + 0.1N(0, 25) for sample size 50 for different values of c_n

c_n	HD	NED
0.5	0.0031 (0.0257)	0.0032 (0.0247)
0.6	0.0031 (0.0259)	0.0031 (0.0248)
0.7	0.0031 (0.0260)	0.0030 (0.0249)
0.8	0.0032 (0.0261)	0.0030 (0.0250)
0.9	0.0032 (0.0262)	0.0030 (0.0251)

Mean (mean square error) of MLE = 0.0044 (0.0287).

Mean (mean square error) of Huber's M-estimate = 0.0030 (0.0246).

Table 4
Empirical mean and mean square error (in parentheses) of the MHDE and MNEDE of the location parameter for data generated under 0.5N(0, 1) + 0.5U(−1, 1) for sample size 50 for different values of c_n

c_n	HD	NED
0.5	0.0013 (0.0133)	0.0023 (0.0134)
0.6	0.0012 (0.0132)	0.0023 (0.0133)
0.7	0.0011 (0.0131)	0.0023 (0.0132)
0.8	0.0011 (0.0131)	0.0023 (0.0132)
0.9	0.0011 (0.0131)	0.0022 (0.0132)

Mean (mean square error) of MLE = 0.0009 (0.0133).

Mean (mean square error) of Huber's M-estimate = 0.0009 (0.0131).

need c_n to be of the order $n^{-(1/2-\delta)}$, where $0 < \delta < \frac{1}{4}$. However, at this point we are not able to come up with a good estimate of δ for a general sample size. One way to find a best choice of δ is to define a criterion that relates the consistency and the MSE of the density estimator and the MSE of the MNEDE. Alternatively, for a given sample size n, one can choose c_n such that under the normal model the MNEDE of the mean μ and its MSE roughly match the MLE of μ and its MSE, respectively. Beran (1977, p. 461) discussed a similar criterion for the choice of c_n in computing the minimum Hellinger distance estimates of the location and scale parameters in the normal model.

We also studied the performance of the estimators under the contaminated model

$$f_\varepsilon(x) = (1-\varepsilon)\mathrm{N}(0,1) + \varepsilon \mathrm{U}(2, 2.1), \tag{5.1}$$

to illustrate the performance of the MLE, MHDE and MNEDE against inliers. We used the actual density of the contaminated distribution in (5.1) rather than a density estimate obtained from the data. Note that a positive value of the contaminating proportion ε in (5.1) generates an outlier and a negative value generates an inlier. Let $f(x)$ denote the true density function $N(0,1)$. Let T_{ML}, T_{HD} and T_{NED} denote the functionals defined on $\mathcal{G}$, the set of all densities with respect to the Lebesgue measure, corresponding to the maximum likelihood, Hellinger distance and negative exponential disparity estimation methods respectively. Let

$$\Delta T_{ML} = T_{ML}(f_\varepsilon) - T_{ML}(f),$$

$$\Delta T_{HD} = T_{HD}(f_\varepsilon) - T_{HD}(f),$$

$$\Delta T_{NED} = T_{NED}(f_\varepsilon) - T_{NED}(f).$$

Note that ΔT_{ML}, ΔT_{HD} and ΔT_{NED} measure potential biases in estimation introduced by an ε-contamination. We have used $\varepsilon = 0.2$, 0.1, 0.05, 0.01,0.005, 0.001, 0.0005, 0.0003, 0.0001, −0.001, −0.002, −0.003, −0.004. From Table 5 we see that the negative exponential disparity produces less bias than Hellinger distance for contaminations above 0.01. With the value of ε increasing in magnitude in the negative direction, the MHDE has increasingly more bias than the MLE, wheras the MNEDE has increasingly less bias than the MLE.

5.2. *The binomial model*

In this section we demonstrate the efficiency and robustness properties of the MNEDE in the discrete case. Computation of minimum disparity estimators in the discrete models, unlike in the continuous case, does not require kernel density estimation and has been described in Section 2. We considered the binomial(m, p) model and chose $m = 12$ and $p = 0.1$ as the true parameter values.

In Table 6 we present the empirical efficiencies of the estimators MHDE and MNEDE compared to the MLE for sample sizes 25, 50, 100, 250, 500 and 1000, based on 5000 replications in each case. The efficiency of the MHDE and MNEDE compared to the MLE has been estimated by the ratio

$$\frac{\text{MSE(MLE)}}{\text{MSE(MHDE)}}, \qquad \frac{\text{MSE(MLE)}}{\text{MSE(MNEDE)}},$$

respectively, where MSE denotes the mean square error. The standard error of the estimated efficiency has been estimated by the approximate formula for the variance of the ratio of two-dependent random variables obtained by a second-order Taylor series expansion. The numbers in parentheses in Table 6 represent the standard errors for the empirical efficiencies. Table 6 shows that under the uncontaminated model the MNEDE performs a little better than MHDE in terms of both empirical mean and its standard error.

Table 5
Relative bias for Hellinger distance and negative exponential disparity under the contaminated model $(1-\varepsilon)N(0,1)+\varepsilon U[2,2.1]$ for different values of ε

ε	$\Delta T_{HD}/\Delta T_{ML}$	$\Delta T_{NED}/\Delta T_{ML}$
0.2	0.332	0.083
0.1	0.403	0.134
0.05	0.496	0.244
0.01	0.744	0.768
0.005	0.835	0.911
0.001	0.955	0.994
0.0005	0.976	0.997
0.0003	0.985	0.998
0.0001	0.994	0.999
−0.001	1.055	0.992
−0.002	1.122	0.968
−0.003	1.220	0.920
−0.004	1.383	0.838

Table 6
Empirical efficiency under the binomial (12, 0.1) model

Sample size	MHDE	MNEDE
25	0.862 (0.010)	0.949 (0.007)
50	0.902 (0.009)	0.976 (0.005)
100	0.932 (0.008)	0.980 (0.004)
250	0.958 (0.006)	0.993 (0.003)
500	0.970 (0.005)	0.993 (0.002)
1000	0.985 (0.004)	1.000 (0.002)

Next we consider the performance of the estimators under contamination. Let $f(x)$ denote the true binomial probability mass function, $x = 0, 1, 2, \ldots, 12$. Following Lindsay's (1994, pp. 1083, 1102–1104) idea we have considered a contaminated version $f_\varepsilon(x)$ of $f(x)$ to assess robustness of the estimators against outliers as well as inliers, where

$$f_\varepsilon(x) \equiv (1-\varepsilon)f(x) + \varepsilon\chi_y(x), \tag{5.2}$$

ε is the contaminating proportion, y is the contaminating value and χ_y is the indicator function at y, i.e., $\chi_y(x) = 1$ for $x = y$ and $\chi_y(x) = 0$ for $x \neq y$. A positive value of ε generates an outlier at $x = y$ and a negative value generates an inlier. Let T_{ML}, T_{HD}, T_{NED}, ΔT_{ML}, ΔT_{HD} and ΔT_{NED} be defined as in Section 5.1, using the contaminated binomial model $f_\varepsilon(x)$ of (5.2). We have used $\varepsilon = 0.2, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0003, 0.0001, -0.0001, -0.0002, -0.0003, -0.0004$ and the value $y = 6$ which has a probability of approximately 0.0004911 under the true model. From Table 7 we see that the negative exponential disparity produces less bias than Hellinger distance for contamination level above 0.001. With the value of ε increasing in magnitude in the negative direction, the MHDE has increasingly more bias than the MLE, whereas the MNEDE has increasingly less bias than the MLE, demonstrating robust performance of the MNEDE against inliers as well.

In this section we have exhibited the performance of the MHDE and MNEDE (relative to the MLE) in the presence of inliers for the binomial model. Further insight into the behavior of these estimators in the presence of inliers can be obtained by examining the count data models such as Poisson and geometric. In such count data models for any finite sample size the problem of inliers is bound to crop up in the form of empty cells, which represent the extreme cases of inliers. For more discussion on this see Harris and Basu (1994).

Table 7
Relative bias for Hellinger distance and negative exponential disparity for the contaminated model $(1-\varepsilon)\text{Bin}(12,0.1) + \varepsilon I_{\{x\}}(x)$ for different values of ε

ε	$\Delta T_{\text{HD}}/\Delta T_{\text{ML}}$	$\Delta T_{\text{NED}}/\Delta T_{\text{ML}}$
0.2	0.132	0.007
0.1	0.158	0.012
0.05	0.201	0.022
0.01	0.370	0.103
0.005	0.472	0.204
0.001	0.735	0.734
0.0005	0.840	0.898
0.0003	0.884	0.956
0.0001	0.955	0.994
−0.0001	1.056	0.993
−0.0002	1.128	0.967
−0.0003	1.227	0.916
−0.0004	1.390	0.833

6. Conclusive remarks

The MNEDE, like the MHDE, is a very attractive robust estimator since it attains its robustness properties without sacrificing first-order efficiency at the model. In this paper we have established the asymptotic efficiency properties of the MNEDE and presented some robustness results using the α-influence curve. A particularly nice feature of the MNEDE is the robustness it provides against inliers, a property that the MHDE does not share. On the whole, the MNEDE appears to be a promising estimator and a major competitor of the MHDE within the class of robust first-order efficient estimators. We are currently investigating further robustness features of the MNEDE for both outliers and inliers for continuous models such as the strong breakdown point for outliers. Defining a similar feature for inliers seems to be a formidable problem at this stage. We are also studying the problem of generalizing our results to the case when we have a sample from a stationary ergodic Markov chain.

Acknowledgements

The research of the first author was supported in part by the URI Summer Research Grant and a Mathematics Department Research Award, University of Texas at Austin, Austin, TX 78712. The research of the second author was supported by a Grant from the College of Arts and Sciences at Oklahoma State University. Most of the research of the third author was done while he was visiting the Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago,

Chicago, IL 60607. The authors wish to thank the associate editor and the referee for many helpful suggestions that led to the present improved version of the paper.

References

Basu, A., I.R. Harris and S. Basu (1996). Tests of hypotheses in discrete models based on the penalized Hellinger distance. *Statist. Probab. Lett.* **27**, 367–373.

Basu, A. and B.G. Lindsay (1994). Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **46**, 683–705.

Basu, A. and S. Sarkar (1994a). The trade-off between robustness and efficiency and the effect of model smoothing in minimum disparity inference. *J. Statist. Comput. Simul.* **50**, 173–185.

Basu, A. and S. Sarkar (1994b). On disparity based goodness-of-fit tests for multinomial models. *Statist. Probab. Lett.* **19**, 307–312.

Basu, A. and S. Sarkar (1994c). Minimum disparity estimation in the errors-in-variables model. *Statist. Probab. Lett.* **20**, 69–73.

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

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

Devroye, L. (1987). *A course in Density Estimation*. Birkhauser, Boston.

Devroye, L. and L. Gyorfi (1985). *Nonparametric Density Estimations: The L₁ View*. Wiley, New York.

Hall, P. and J.S. Marron (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields* **90**, 149–173.

Hardle, W., P. Hall and J.S. Marron (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83**, 86–95.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383–393.

Hampel, F.R., E.M. Ronchetti, P.M. Rousseeuw and W.A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Harris, I.R. and A. Basu (1994). Hellinger distance as a penalized log likelihood. *Commun. Statist. Simula.* **23**, 1097–1113.

Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Huber, P.J. (1972). Robust statistics: a review. *Ann. Math. Statist.* **43**, 1041–1067.

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

Marron, J.S. (1989). Comments on a data based bandwidth selector. *Comput. Statist. Data Analy.* **8**, 155–170.

Rao, C.R. (1961). Asymptotic efficiency and limiting information. *Proc. 4th Berkeley Symp.* Vol. 1, 531–546.

Royden, H.L. (1968). *Real analysis*. The Macmillan Company, New York.

Shin, D.W., A. Basu and S. Sarkar (1995). Comparisons of the blended weight Hellinger distance based goodness-of-fit test statistics. *Sankhya*, Series **B 57**, 365–376.

Shin, D.W., A. Basu and S. Sarkar (1996). Small sample comparisons for the blended weight chi-square goodness-of-fit test statistics. *Commun. Statist. – Theory Methods* **25**, 211–226.

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 tests: Efficiency, breakdown points and examples. *J. Amer. Statist. Assoc.* **84**, 107–113.

Stather, C.R. (1981). Robust statistical inference using Hellinger distance methods. Unpublished Ph.D. Dissertation, La Trobe University, Melbourne, Australia.

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