ROBUST INFERENCE IN PARAMETRIC MODELS USING THE FAMILY OF GENERALIZED NEGATIVE EXPONENTIAL DISPARITIES

SUBIR K. BHANDARI[1], AYANENDRANATH BASU[2] AND SAHADEB SARKAR[3]*

Indian Statistical Institute and Indian Institute of Management Calcutta

Summary

We examine robust estimators and tests using the family of generalized negative exponential disparities, which contains the Pearson's chi-square and the ordinary negative exponential disparity as special cases. The influence function and α-influence function of the proposed estimators are discussed and their breakdown points derived. Under the model, the estimators are asymptotically efficient, and are shown to have an asymptotic breakdown point of 50%. The proposed tests are shown to be equivalent to the likelihood ratio test under the null hypothesis, and their breakdown points are obtained. The competitive performance of the proposed estimators and tests relative to those based on the Hellinger distance is illustrated through examples and simulation results. Unlike the Hellinger distance, several members of this family of generalized negative exponential disparities generate estimators which also possess excellent inlier-controlling capability. The corresponding tests of hypothesis are shown to have better power breakdown than the Hellinger deviance test in the cases examined.

Key words: breakdown point; efficiency; Hellinger distance; minimum disparity estimation; negative exponential disparity; robustness.

1. Introduction

Consider the general setting of inference in a parametric class of distributions $\mathcal{F}_\Theta = \{F_\theta, \theta \in \Theta \subseteq I\!R^p\}$. Let G be the true distribution belonging to $\mathcal{G}$, the class of all distributions having probability density functions (pdfs) with respect to a dominating measure such as the Lebesgue or the counting measure. Assume that $\mathcal{F}_\Theta \subset \mathcal{G}$. In reality, G is often close to, but not exactly in, the model $\mathcal{F}_\Theta$. Classical methods of inference, such as those based on maximum likelihood (ML), can be arbitrarily perturbed by deviations from the assumed model, although they are often optimal when the assumed model is correct. On the other hand, classical robust estimators such as the M-estimators necessarily sacrifice first order efficiency to achieve robustness for most parametric models (Hampel *et al.*, 1986).

Beran (1977) showed that the apparent conflict between efficiency and robustness is at least partially reconciled by using minimum Hellinger distance estimation. Among others,

Received May 2002; revised July 2004; accepted August 2005.

*Author to whom correspondence should be addressed.

[1]Stat-Math Unit, Indian Statistical Institute, Kolkata 700 108, India.

[2]Applied Statistics Unit, Indian Statistical Institute, Kolkata 700 108, India.

[3]Operations Management Group, Indian Institute of Management Calcutta, Kolkata 700 104, India.
e-mail: sahadeb@iimcal.ac.in

Acknowledgments. The research of the third author was conducted in part while he was visiting the Department of Mathematics at the University of Louisiana at Lafayette, Louisiana, USA. The authors are grateful to two anonymous referees, an Associate Editor and the Editor for their valuable suggestions which led to a substantially improved version of the paper.

Tamura & Boos (1986) and Simpson (1987, 1989) pursued this line of research. Lindsay (1994) and Basu & Lindsay (1994) discussed a class of density based divergences, called disparities, which includes the Hellinger distance (HD). This class contains the negative exponential disparity (NED), an excellent competitor to the HD in generating robust statistics (Lindsay, 1994; Basu *et al.*, 1997). Jeong & Sarkar (2000) generalized the NED to construct a class of multinomial goodness-of-fit tests. In this paper we investigate efficiency and robustness properties of parametric inference procedures based on this family of generalized negative exponential disparities (GNED). Several members of this family appear to perform very well in combining robustness with full asymptotic efficiency.

The distinguishing feature of our proposed estimators is their ability to combine full asymptotic efficiency with strong robustness properties, which is also the feature of minimum disparity estimators in general. Under standard regularity conditions, the minimum disparity estimators are first order efficient and have the same influence function as the maximum likelihood estimator (MLE) under the model, which is often unbounded. Hence the robustness of these estimators cannot be explained through their influence functions. However, several other factors do indicate their robustness. First, many minimum disparity methods exhibit a markedly dampened response to observations inconsistent with the model in their estimating equations, and strongly downweight outlying observations. Second, the second order analysis of bias prediction (Lindsay, 1994) shows how the higher order terms reduce the predicted bias more than the first order influence function analysis. Third, the α-influence functions (Beran, 1977) of the estimators are often bounded, continuous functions of the contaminating point. Finally, this approach often leads to high breakdown points in both parametric estimation and testing of hypotheses.

A related approach to robust inference is the weighted likelihood method. The weight functions are based upon disparities and provide a natural downweighting to probabilistic outliers. The weighted likelihood estimation procedure was developed by Markatou *et al.* (1997, 1998). Its extension to the hypothesis testing scenario was discussed in Agostinelli & Markatou (2001). While we do not deal explicitly with weighted likelihood estimation in this paper, one can also use the disparities considered herein for generating weights to produce robust analogs of the likelihood equations.

Some of the properties of our proposed procedures follow routinely from previous results. As such they have only been stated here without proofs; details can be found in Bhandari *et al.* (2000). To remove ambiguity, we clearly spell out the original methodological contributions of the current paper: We have (a) utilized the GNED family for the first time (except the NED) for robust inference; (b) established the breakdown points of the corresponding estimators; (c) derived asymptotic distributions of the tests of hypotheses and (d) obtained breakdown points of the tests of hypotheses.

A natural question concerning the applicability of the new methods is: "What does one gain by using these methods over the HD which has already been widely studied and is known to generate robust and efficient inference procedures?" The outlier resistant properties of the HD are offset in part by its poor handling of inliers (defined in Section 2) which can severely affect its small sample performance (see Simpson, 1989; Lindsay, 1994; Harris & Basu, 1994; Basu *et al.*, 1996). As such the HD requires additional accessories such as an artificial empty cell penalty (Harris & Basu, 1994; Basu *et al.*, 1996) to shore up its small sample performance. Several GNED family members, on the other hand, control the inliers naturally and have substantially better small sample performance (eg. compare Table 1 of

TABLE 1

Relative biases for estimators based on the GNED and HD under the contaminated model $(1-\alpha)Bin(12, 0.1) + \alpha\iota_6(x)$ *for different values of* α. *[Bold entries indicate smaller relative biases for the minimum GNED estimators than the HD estimators].*

		GNED with λ					
α	HD	0.5	0.75	1.0	1.25	1.50	2.0
+0.2	0.132	**0.024**	**0.011**	**0.007**	**0.005**	**0.004**	**0.002**
+0.1	0.158	**0.038**	**0.019**	**0.012**	**0.008**	**0.006**	**0.004**
+0.05	0.201	**0.069**	**0.035**	**0.022**	**0.016**	**0.012**	**0.008**
+0.01	0.370	**0.324**	**0.162**	**0.103**	**0.074**	**0.057**	**0.038**
+0.005	0.472	0.617	**0.321**	**0.204**	**0.146**	**0.112**	**0.075**
+0.001	0.735	1.161	0.915	**0.734**	**0.598**	**0.497**	**0.359**
+0.0005	0.830	1.146	1.012	0.897	**0.799**	**0.716**	**0.581**
+0.0003	0.884	1.109	1.028	0.956	0.889	**0.828**	**0.723**
+0.0001	0.955	1.045	1.020	0.994	0.970	**0.945**	**0.902**
−0.0001	1.056	**0.945**	**0.967**	**0.993**	**1.017**	**1.042**	1.097
−0.0002	1.128	**0.877**	**0.921**	**0.967**	**1.016**	**1.069**	1.185
−0.0003	1.227	**0.797**	**0.853**	**0.916**	**0.986**	**1.062**	1.239
−0.0004	1.390	**0.701**	**0.763**	**0.833**	**0.913**	**1.004**	**1.224**

Harris & Basu (1994) with Table 7 of the current paper). Also, Table 1 of this paper shows that several GNED family members may have smaller bias under inliers and moderate outliers than the HD. Several of these tests also have better power breakdown than the HD methods in the cases we have examined (see Examples 1 and 2). The methods are competitive with HD in our real data examples. Thus we believe that some of the methods proposed here can have more general applicability in many situations over the HD.

The rest of the paper is organized as follows: Section 2 reviews minimum disparity estimation, introduces the minimum GNED estimators and discusses their asymptotic efficiency. Section 3 studies other related properties including breakdown. Corresponding robust tests are discussed in Section 4. Examples and simulation results are presented in Section 5. Section 6 gives concluding remarks. Proofs are given in the Appendix.

Hereafter the corresponding lower case letters denote the pdfs of the cumulative distribution functions (cdfs), e.g., the pdfs of G, F_θ and G_n will be g, f_θ and g_n respectively.

2. Minimum disparity estimation

For a random sample $X_1, X_2, \ldots, X_n$ from a distribution G, let

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

be a nonparametric density estimator of g, where w is a smooth kernel function and h_n is the bandwidth. In practice, one may use the automatic kernel density estimator with $h_n = c_n s_n$, where s_n is a robust scale estimator and $c_n > 0$ is an appropriate constant depending on n. For discrete models, take g_n to be the empirical density function, defined as $g_n(x) =$ the proportion of sample values equal to x for any x in the sample space. Define the Pearson residual at x as $\delta(x) = \delta(g_n, f_\theta, x) = (g_n(x) - f_\theta(x))/f_\theta(x)$. A value x in the sample space is called an outlier if it has a large positive Pearson residual $\delta(x)$; it is called an inlier if $\delta(x)$ is negative. Let C be a real-valued, thrice differentiable, convex function on $[-1, \infty)$ with

$C(0) = 0$. The *disparity* D_C (between g_n and f_θ) is defined as

$$D_C(g_n, f_\theta) = \int C(\delta) f_\theta, \tag{2}$$

where the integral is with respect to the dominating measure. Examples of disparities include the likelihood disparity (LD) and the (two times, squared) Hellinger distance (HD), defined by $LD(g_n, f_\theta) = \int [g_n \log(g_n/f_\theta) - (f_\theta - g_n)]$ with $C_{LD}(\delta) = (\delta + 1)\log(\delta + 1) - \delta$ and $HD(g_n, f_\theta) = 2\int (g_n^{1/2} - f_\theta^{1/2})^2$ with $C_{HD}(\delta) = 2[(\delta + 1)^{1/2} - 1]^2$ respectively. The LD, in the discrete case, is minimized by the MLE of θ. The HD is minimized by the minimum Hellinger distance estimator (MHDE).

Let ∇ denote the gradient of a function of θ. Let $a'(x)$ and $a''(x)$ be the first two derivatives of a real valued function $a(x)$. The corresponding minimum disparity estimator (MDE) minimizes $D_c(g_n, f_\theta)$ over θ. Under differentiability, its estimation equation is

$$-\nabla D_C = \int A(\delta) \nabla f_\theta = 0, \tag{3}$$

where $A(\delta) = (\delta + 1)\, C'(\delta) - C(\delta)$. The function $A(\delta)$ is increasing on $[-1, \infty)$, and can be redefined, without affecting the estimating properties of the disparity D_C, to satisfy $A(0) = 0$ and $A'(0) = 1$. This standardized function $A(\delta)$ is called the residual adjustment function (RAF) of the disparity; it determines how strongly the large outlying observations (manifesting themselves as large positive values of δ) are downweighted. One would expect better robustness properties for an estimator if its RAF shrinks the effect of large δ values more towards zero. The RAF for the LD and HD are, respectively, $A(\delta) = \delta$ and $A(\delta) = 2[(\delta + 1)^{1/2} - 1]$. The property $A''(0) = 0$ leads to the second order efficiency of the estimator (Lindsay, 1994).

2.1. Generalized negative exponential disparity estimator

The NED corresponds to $C(\delta) = \exp(-\delta) - 1$ in (2), or equivalently, to $C(\delta) = \exp(-\delta) - 1 + \delta$. Consider the family of generalized negative exponential disparities $\{\text{GNED}_\lambda\}$ (Jeong & Sarkar, 2000) defined by (2) with

$$C_\lambda(\delta) = \begin{cases} (e^{-\lambda\delta} - 1 + \lambda\delta)/\lambda^2, & \text{if } \lambda > 0 \\ \delta^2/2, & \text{if } \lambda = 0. \end{cases}$$

Note that $C_0(\delta) = \lim_{\lambda \to 0+} C_\lambda(\delta)$, GNED_0 is the Pearson's chi-square (PCS) and GNED_1 is the NED. For $\lambda > 0$, $\text{GNED}_\lambda(g, f_\theta)$ is bounded below by zero (achieved when $g = f_\theta$) and bounded above by $(e^\lambda - 1)/\lambda^2$ (achieved when g and f_θ are singular).

We will write MGNEDE_λ, or simply MGNEDE, for the minimum generalized negative exponential disparity estimator obtained by minimizing $\text{GNED}_\lambda(g_n, f_\theta)$ over Θ. Define the GNED_λ estimation functional $T_\lambda : \mathcal{G} \to \Theta$ as $T_\lambda(G) = \theta_{\lambda,G}$ satisfying

$$GNED_\lambda(g, f_{T_\lambda(G)}) = \min_{\theta \in \Theta} GNED_\lambda(g, f_\theta), \tag{4}$$

provided such a minimum exists. By definition, the MGNEDE_λ of θ is $T_\lambda(G_n)$.

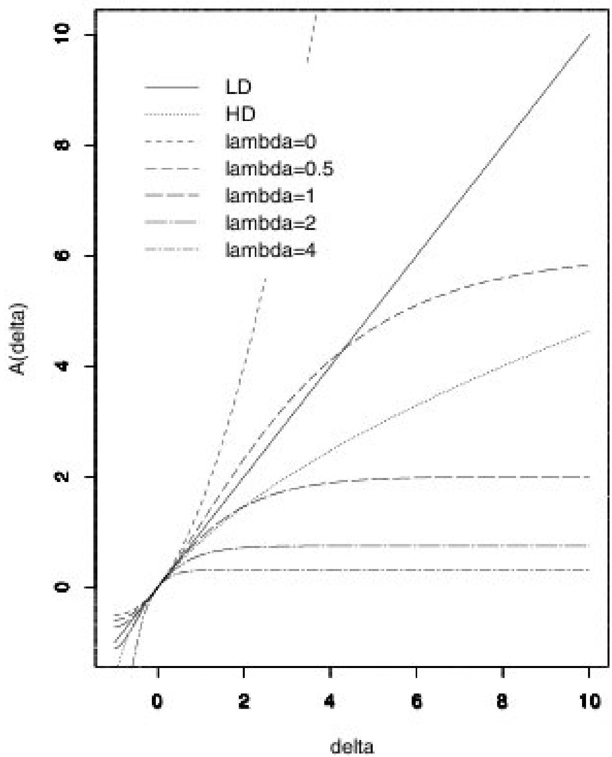

Figure 1. The RAFs of the LD, HD and GNED$_\lambda$ for $\lambda = 0, 0.5, 1, 2, 4$.

The RAF of the GNED$_\lambda$ is given by

$$A_\lambda(\delta) = \begin{cases} \dfrac{(\lambda+1) - ((\lambda+1)+\lambda\delta)e^{-\lambda\delta}}{\lambda^2}, & \text{if } \lambda > 0 \\ \delta + \dfrac{\delta^2}{2}, & \text{if } \lambda = 0. \end{cases} \quad (5)$$

We display a plot of $A_\lambda(\delta)$ against δ for $\lambda = 0, 0.5, 1, 2, 4$ in Figure 1, along with the RAFs of the LD and HD. For comparison, we use the $A_{LD}(\delta)$ as reference. The behavior of other methods can be described by how their RAFs depart from linearity. The NED is second order efficient, and its RAF is close to that of the LD in the neighborhood of $\delta = 0$. Thus it downweights small outliers mildly, but downweights large outliers more strongly than the HD. Other values of λ provide different degrees of downweighting for large positive δ; the degree increases with λ. For $\lambda < 1$, the GNED$_\lambda$ provides less downweighting around $\delta = 0$ than LD, but eventually its RAF drops below $A_{LD}(\delta)$ for larger values of δ, except when $\lambda = 0$. For the latter case (PCS), $A(\delta) = \delta + (\delta^2/2)$ magnifies the effect of large δ outliers rather than shrinking them. As a result this estimator is poorer than the MLE in terms of robustness. In addition, the development of the asymptotic results for the PCS case requires a different treatment than the other disparities in this class. As such we have excluded the PCS from our discussion, and the theoretical results in the rest of the paper assume $\lambda > 0$.

Under discrete models, the MGNEDE$_\lambda$ is first order efficient (Lindsay, 1994). In the continuous models, existence, consistency, and asymptotic normality results of these estimators follow with slight modifications of the proofs of Basu et al. (1997).

3. Robustness

3.1. Influence and α-Influence Functions

The influence function of T_λ at $G \in \mathcal{G}$ is given by $IF_{\lambda,G}(z) = W_\lambda^{-1} Q_\lambda$, where

$$Q_\lambda = A'_\lambda(\delta(z))u_{\theta_{\lambda,G}}(z) - \int A'_\lambda(\delta)u_{\theta_{\lambda,G}}dG,$$
$$W_\lambda = \int A'_\lambda(\delta)u_{\theta_{\lambda,G}}u^\top_{\theta_{\lambda,G}}dG - \int A_\lambda(\delta)\nabla^2 f_{\theta_{\lambda,G}}dG,$$

and $\delta(x) = \delta(g, f_{\theta_{\lambda,G}}, x) = (g(x) - f_{\theta_{\lambda,G}}(x))/f_{\theta_{\lambda,G}}(x)$ and $A_\lambda(\cdot)$ is as in (5). Thus, if G is a model point F_θ, then the influence function of T_λ reduces to $I^{-1}(\theta)u_\theta(z)$, the influence function of the MLE. While this suggests that the MGNEDE$_\lambda$ is asymptotically fully efficient at the model, it is potentially unbounded – thus failing to capture its robustness.

Consider the gross-error model (Beran, 1977), when G is the model point for a continuous distribution F_θ with pdf f_θ. Let $f_{\theta,\alpha,z} = (1-\alpha)f_\theta + \alpha\eta_z$, where η_z denotes the uniform density on the interval $(z-\epsilon, z+\epsilon)$, $\epsilon > 0$ is a very small, fixed number, $\theta \in \Theta$, $\alpha \in (0, 1)$ and $z \in \mathbb{R}$. It can be shown (Bhandari *et al.*, 2000, Theorem 2) that the α-influence function of T_λ, defined by the difference quotient $\phi^\alpha_{\lambda,\theta}(z) = (T_\lambda(F_{\theta,\alpha,z}) - \theta)/\alpha$, can be bounded even when the influence function $\mathrm{IF}_{\lambda,F_\theta}$ or its generalized form $\phi_{\lambda,\theta}(z) = \lim_{\alpha\to 0}\phi^\alpha_{\lambda,\theta}(z)$ is not. In such a case the functional is robust against $100\alpha\%$ contamination by gross errors at arbitrary z. Note that while the influence function of T_λ fails to capture its robust behavior, its α-influence function can do it successfully.

3.2. Breakdown point analysis

The breakdown point of a statistical functional is roughly the smallest fraction of contamination in the data that may cause an arbitrarily extreme value in the estimate. Here we establish the breakdown point of the functional $T_\lambda(G)$ under the following set-up. For $\alpha \in (0, 1)$, consider the contamination model

$$H_{\alpha,m} = (1-\alpha)G + \alpha K_m, \quad m \geq 1,$$

where G is the true distribution and $\{K_m\}$ is a sequence of contaminating distributions; $h_{\alpha,m}$, g and k_m are the corresponding densities. For a given contamination sequence $\{K_m\}$ we will say that there is breakdown in T_λ for α level of contamination if

$$\lim_{m\to\infty} |T_\lambda(H_{\alpha,m})| = \infty. \tag{6}$$

We will examine the smallest α for which there exists a sequence $\{K_m\}$ satisfying (6). Our analysis is based on the following assumptions on the model and the contamination sequence, necessary to determine the disparities under extreme forms of contaminations.

Assumptions. *The true density g, the model density $\{f_\theta\}$, and the contamination densities $\{K_m\}$ satisfy the following*:

A1. $\int \min\{g, k_m\} \to 0$ *as* $m \to \infty$; *that is the contamination distribution becomes asymptotically singular to the true distribution.*

A2. $\int \min\{f_\theta, k_m\} \to 0$ *as* $m \to \infty$ *uniformly for* $|\theta| \le c, >$ *for any fixed c; that is the contamination distribution becomes asymptotically singular to specified models.*

A3. $\int \min\{g, f_{\theta_m}\} \to 0$ *as* $m \to \infty$, *if* $|\theta_m| \to \infty$ *as* $m \to \infty$; *that is large values of* θ *give model distributions that become asymptotically singular to the true distribution.*

Contamination sequences that satisfy assumptions **A1** and **A2** are called outlier sequences. Intuitively, outlier sequences represent the worst possible type of contamination sequences, so it seems appropriate to study the breakdown properties of the functionalunder such sequences. Assumption **A3** formalizes the expected behavior of the model.

Minimizing $\text{GNED}_\lambda(g, f_\theta)$ with respect to θ is equivalent to minimizing

$$\rho_\lambda(g, f_\theta) = \int e^{-\lambda g/f_\theta} f_\theta. \tag{7}$$

Let $\rho_\lambda(h_{\alpha,m}, f_\theta)$ be minimized at $\theta = T_\lambda(H_{\alpha,m})$. We then have the following result.

Theorem 1. *Assume that the true distribution G, the model* $\{F_\theta\}$, *and the contamination sequence* $\{K_m\}$ *satisfy conditions* **A1–A3**. *Let* θ_α^* *be the minimizer of* $\rho_\lambda((1-\alpha)g, f_\theta)$, b_α *satisfy* $\rho_\lambda((1-\alpha)g, f_{\theta_\alpha^*}) = e^{-\lambda(1-\alpha)b_\alpha}$, *and* $\alpha^* = \inf\{\alpha : e^{-\lambda(1-\alpha)b_\alpha} \ge e^{-\lambda\alpha}\}$. *Then, for any outlier sequence* $\{K_m\}$, $\lim sup_{m\to\infty} |T_\lambda(H_{\alpha,m})| < \infty$ *whenever* $e^{-\lambda(1-\alpha)b_\alpha} < e^{-\lambda\alpha}$, *in particular, if* $\alpha < \alpha^*$. *Furthermore, when* $G = F_{\theta_0}$ *belongs to the model, if* $\alpha < 1/2$ *then* $\lim sup_{m\to\infty} |T_\lambda(H_{\alpha,m})| < \infty$ *for any outlier sequence and the minimizer of* $\rho_\lambda(h_{\alpha,m}, f_\theta)$ *in the limit as* $m \to \infty$ *is* θ_0.

Theorem 1 shows that for a general G, α^* is the breakdown point of T_λ. In addition the proof (given in the Appendix) also shows that as long as $\alpha < \alpha^*$, θ_α^* minimizes the divergence $\rho_\lambda(h_{\alpha,m}, f_\theta)$ in the limit as $m \to \infty$. In the special case when $G = F_{\theta_0}$, T_λ achieves a breakdown point of 1/2; as long as $\alpha < 1/2$ the minimizer of $\rho_\lambda(h_{\alpha,m}, f_\theta)$ in the limit as $m \to \infty$ is θ_0 itself, so that the contamination has no limiting impact at all.

3.3. Bias study and inlier analysis

Using several contaminated distributions for a particular model, we demonstrate here that the MGNEDE can perform more favorablythan the MHDE in terms of relative bias. Consider the binomial model and the contaminated binomial density $f_\alpha(x) = (1-\alpha)f + \alpha\iota_c(x)$ where f is the true density, α the contaminating proportion, c the contaminating value, and $\iota_c(x)$ the indicator function at c. Let F and F_α be the corresponding distributions. An outlier is generated at $x = c$ if $\alpha > 0$ and an inlier if $\alpha < 0$. Let T_{ML}, T_{HD} and T_λ be the estimation functionals for the ML, HD and GNED_λ methods respectively. Let

$$\Delta T_{ML} = T_{ML}(F_\alpha) - T_{ML}(F), \Delta T_{HD} = T_{HD}(F_\alpha) - T_{HD}(F), \Delta T_\lambda = T_\lambda(F_\alpha) - T_\lambda(F)$$

measure potential biases in estimation due to α amount of contamination.

We use $f = Bin(12, p)$ density with $p = 0.1$, $c = 6$ and α = 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. We compute the biases relative to the ML method in estimating p for HD and GNED_λ, defined as $\Delta T_{HD}/\Delta T_{ML}$ and $\Delta T_\lambda/\Delta T_{ML}$, respectively. We present the results in Table 1; bold entries represent smaller relative biases for the MGNEDE than the HD. Table 1 leads to very interesting observations:

(a) most of the MGNEDEs perform substantially better than the MHDE against inliers. Even the strongly outlier resistant estimators for $\lambda = 1.5, 2.0$ perform comparably or better than the MHDE against inliers, while being far better than the latter under all the outlier situations considered here. In particular, $\text{MGNEDE}_{1.5}$ beats the MHDE for every single value of α; (b) the estimators in the middle range of the scale ($\lambda = 1.0, 1.25$) comfortably beat the MHDE under all inlier situations, and under all moderate and large values of α ($\alpha > 0.0005$). Only under very small outliers (when the bias is very small anyway) does the MHDE beat the MGNEDEs in this range; (c) for the estimators in the lower range of the scale ($\lambda = 0.5, 0.75$) the MHDE is comparatively even poorer under inlier situations, and is again substantially worse than the MGNEDE when $\alpha > 0.005$; (d) overall, the MHDE is outperformed by all members of the GNED_λ family considered for $\alpha > 0.005$, and by every member of the GNED_λ family considered here, except $\lambda = 2.0$, under all inlier situations ($\alpha < 0$).

This superior performance of the GNED based methods under the presence of inliers raises the following question. Can one have an idea of how often inliers show up in practice, so that one can assess the amount of gain in choosing an appropriate GNED-based method over the HD. While the general answer to this is clearly a difficult one, we provide a numerical study in Section 5.2 which demonstrates that even when the data are generated by the pure model there are enough inliers in the data to really retard the performance of the HD-based methods than those based on GNED.

4. Generalized negative exponential disparity tests

Because of the lack of robustness of the likelihood ratio test (LRT), alternative robust tests have received a lot of attention in the literature. One such robust alternative is Simpson's (1989)Hellinger deviance test (HDT). Here, we show that analogs of the LRT, based upon GNED_λ, have the same asymptotic behavior as the LRT under the null hypothesis. However, unlike the LRT, these tests have good breakdown and strong outlier resistant properties and compete very well with the HDT.

4.1. Definition

Under the parametric set-up of Section 1, suppose that the hypotheses of interest are H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta - \Theta_0$, where $\Theta_0 \subset \Theta$. Let $\theta_{\lambda,G} = T_\lambda(G)$ be as defined in (4) and the functional $T_{\lambda,0} : \mathcal{G} \to \Theta_0$ be defined as $T_{\lambda,0}(G) = \theta^*_{\lambda,G} \in \Theta_0$ which satisfies

$$\text{GNED}_\lambda(g, f_{\theta^*_{\lambda,G}}) = \min_{\theta\in\Theta_0} \text{GNED}_\lambda(g, f_\theta),$$

provided such a $\theta^*_{\lambda,G}$ exists. We will write θ_G for $\theta_{\lambda,G}$ and θ^*_G for $\theta^*_{\lambda,G}$ for brevity. For a random sample of size n with kernel density estimate g_n, denote the estimators $T_{\lambda,0}(G_n)$ and $T_\lambda(G_n)$ under the null hypothesis and under no restriction by $\hat{\theta}^*$ and $\hat{\theta}$, respectively. Define the generalized negative exponential disparity test (GNEDT_λ) statistic as

$$\text{GNEDT}_\lambda = 2n(\text{GNED}_\lambda(g_n, f_{\hat{\theta}^*}) - \text{GNED}_\lambda(g_n, f_{\hat{\theta}})) = 2n\frac{e^\lambda}{\lambda^2} N_\lambda(G_n) \tag{8}$$

where, with ρ_λ as in (7),

$$N_\lambda(G) = \rho_\lambda(g, f_{\theta^*_G}) - \rho_\lambda(g, f_{\theta_G}). \tag{9}$$

Below we will write GNEDT_1 simply as NEDT.

4.2. Asymptotic distribution

Let Θ_0 be a set of $r \leq p$ restrictions on Θ defined by $R_i(\theta) = 0$, $1 \leq i \leq r$. We assume that the parameter space, under H_0, can be described through a parameter $\nu = (\nu^1, \ldots, \nu^{p-r})^T$ with $p - r$ components, i.e., $\theta = b(\nu)$ where b: $I\!R^{p-r} \to I\!R^p$. Then $\hat{\theta}^* = b(\hat{\nu})$, where $\hat{\nu}$ is the MGNEDE$_\lambda$ in the ν-formulation of the model based on the sample of size n. Let $G = F_{\theta_0}$, where θ_0 is the true value of the parameter. Under H_0 let ν_0 be the true value of the ν parameter. It can be shown that under H_0, $\hat{\nu}$ and $\hat{\theta}^*$ are consistent estimators of ν_0 and θ_0 respectively. Let $J(\nu_0)$ be the information matrix under the ν-formulation, and $I(\theta_0)$ be that under no restrictions. The LRT has the form

$$LRT = 2\left[\log\left(\Pi_{i=1}^n f_{\hat{\theta}_{ML}}(X_i)\right) - \log\left(\Pi_{i=1}^n f_{\hat{\theta}^*_{ML}}(X_i)\right)\right]$$

where $\hat{\theta}_{ML}$ and $\hat{\theta}^*_{ML}$ represent the MLEs of θ under no restrictions and under H_0, respectively. Asymptotically the LRT has a $\chi^2(r)$ distribution under H_0. Theorem 2 below, proved in the Appendix, establishes the same for the GNEDT$_\lambda$.

Theorem 2. *Assume the corresponding vector-parameter generalizations of the regularity conditions (R1)–(R3) of Serfling (1980, Section 4.2.2) on the model family $\mathcal{F}_\Theta$, and the regularity conditions assumed for the asymptotic normality of the MGNEDE$_\lambda$ (Bhandari et al., 2000 Theorem 1). Then, the statistic GNEDT$_\lambda$ has the same asymptotic distribution as the LRT under H_0 as $n \to \infty$, which is the $\chi^2(r)$ distribution.*

Although the null behavior of the GNEDT$_\lambda$ is similar to that of the LRT, the breakdown properties of the GNEDT$_\lambda$ are substantially stronger than that of the LRT. In the next section we determine the actual power and level breakdown points.

4.3. Breakdown results for tests

Let $G \in \mathcal{G}$ be the true distribution and consider the parametric hypotheses H_0 and H_1 and the test statistic GNEDT$_\lambda$ of Section 4.1. In this section we give a breakdown point analysis of GNEDT$_\lambda$. Let $H = (1-\alpha)G + \alpha K$ be a contaminated version of G with pdf $h = (1-\alpha)g + \alpha k$. To define the breakdown point of the test functional we focus on the smallest α for which the corresponding P-value of the test attains its maximum or minimum possible value. See He *et al.*, (1990) for a comprehensive discussion of power and level breakdown functions under more general conditions.

We study the functionals associated with the test statistic GNEDT$_\lambda$. Let $N_\lambda(\cdot)$ and $\rho_\lambda(\cdot,\cdot)$ be as defined in (9) and (7) respectively. Assume that $\mathcal{G}$ is convex. Define

$$N_{\lambda,\min} = \inf_{F\in\mathcal{G}} N_\lambda(F),\; N_{\lambda,\max} = \sup_{F\in\mathcal{G}} N_\lambda(F),$$

$$\alpha_0(G; N_\lambda) = \inf\left\{\alpha : \sup_{K\in\mathcal{G}} N_\lambda((1-\alpha)G + \alpha K) = N_{\lambda,\max}\right\},$$

$$\alpha_1(G; N_\lambda) = \inf\left\{\alpha : \inf_{K\in\mathcal{G}} N_\lambda((1-\alpha)G + \alpha K) = N_{\lambda,\min}\right\}.$$

The quantities $\alpha_0(G;N_\lambda)$ [$\alpha_1(G;N_\lambda)$] are the level [power] breakdown points of the test and represent the smallest fraction of data contamination where some suitable contaminating distribution $K \in \mathcal{G}$ causes the P-value to become the minimum [maximum] possible. Under hypothesis H_s, $s = 0, 1$, an $\alpha > \alpha_s$ can lead to an incorrect decision. Usually, α_1 is considered to be more informative and we study this in Theorem 3 below. Note that $N_{\lambda,\min} = 0$ is attained for $G = F_{\theta_0}$, $\theta_0 \in \Theta_0$. For the rest of the section, we write $\theta_G = T_\lambda(G)$, $\theta_H = T_\lambda(H)$, and $\theta^*_G = T_{\lambda,0}(G)$, $\theta^*_H = T_{\lambda,0}(H)$, suppressing λ.

Theorem 3. *For G, $H = (1-\alpha)G + \alpha K$, and the functional N_λ as defined above,*

$$\alpha_1(G; N_\lambda) \geq \frac{N_\lambda(G)}{(\lambda+1) - e^{-\lambda}}. \tag{10}$$

The proof is in the Appendix. Consider the case where $\rho_\lambda(g, f_{\theta_G}) = e^{-\lambda}$, and $\rho_\lambda(g, f_{\theta^*_G}) = 1$, i.e., G belongs to the model family but is singular to the model distributions under H_0. This is the situation where the test should have the maximum power, and should require the largest fraction of data contamination to cause power breakdown. In this case, the lower bound of the breakdown point in (10) turns out to be

$$\frac{1 - e^{-\lambda}}{(\lambda+1) - e^{-\lambda}}. \tag{11}$$

For the NEDT, this is approximately equal to 0.38, which is smaller than 0.5, a bound attained by the HDT in this situation (Simpson, 1989). However, a direct calculation (Lemma 1, proved in the Appendix) shows that a breakdown point of 0.5 is indeed attainable for the MGNEDE$_\lambda$ in this case. Thus the bound in (10) is a crude and and not a sharp lower bound. We expect that it should be possible to derive a sharper bound, but do not have a proof at this point.

Lemma 1. *If $\rho_\lambda(g, f_{\theta_G}) = e^{-\lambda}$ and $\rho_\lambda(g, f_{\theta^*_G}) = 1$, the lower bound of the power breakdown point for the GNED$_\lambda$ tests is 0.5.*

Next we give a result (proven in the Appendix) regarding $\alpha_0(G; N_\lambda)$, the level breakdown point. Notice that $N_{\lambda,\max} \leq 1 - e^{-\lambda}$, with equality if

$$\sup_\theta \rho_\lambda(f_\theta, f_{\theta^*_{F_\theta}}) = 1. \tag{12}$$

The above condition implies that there exists a density $g = f_\theta$ in the model family for which the first and second terms in the expression of N_λ simultaneously attain the maximum and minimum of ρ_λ respectively.

Theorem 4. *Suppose that* (12) *holds. If $\rho_\lambda(g, f_{\theta^*_G}) < 1$, then $\alpha_0(G; N_\lambda) = 1$.*

Thus we cannot drive the P-value of the test to zero under the conditions of Theorem 4, unless 100% of the data are contaminated. We now consider two illustrative examples.

Example 1 (Gaussian mean with nuisance variance). Let F_θ be the $N(\eta, \sigma^2)$ cdf with $\theta = (\eta, \sigma^2)$, and the hypotheses of interest be $H_0 : \eta = 0$ vs. $H_1 : \eta \neq 0$, σ^2 unknown. We have $N_\lambda(F_\theta) = \rho_\lambda(f_\theta, f_{\theta^*_{F_\theta}}) - \rho_\lambda(f_\theta, f_{\theta_{F_\theta}}) = \rho_\lambda(f_\theta, f_{\theta^*_{F_\theta}}) - e^{-\lambda}$. In Table 2 we present the calculated bounds for $\alpha_1(F_\theta; N_\lambda)$ for θ = (3, 1) and (5, 1) for various λ values. For θ = (3, 1) and (5, 1), the breakdown point bounds for Simpson's (1989, p.111) HDT are 0.127 and 0.208 respectively while that for the NEDT are 0.198 and 0.256 and the maximum bounds occur in a neighborhood of $\lambda = 1.10$ and 0.90 respectively. Note that (a) for θ = (3, 1) and (5, 1), the GNEDT breakdown points are higher than the HDT breakdown points for all $\lambda \in [0.5, 2]$, and (b) for different values of θ the best breakdown occurs at different values of λ.

TABLE 2

Power breakdown bounds $\alpha_1(F_\theta;\, N_\lambda)$ of GNEDT$_\lambda$ for testing $H_0 : \eta = 0$ vs. $H_0 : \eta \neq 0$ under the $N(\eta,\, \sigma^2)$ model with unknown σ^2 for various λ values. [Power breakdown bounds for Simpson's HDT are also shown for ease of comparison.]

		GNEDT with λ									
(η, σ^2)	HDT	0.25	0.50	0.75	0.90	1.00	1.10	1.25	1.50	2.00	4.00
(3, 1)	0.127	0.116	0.169	0.192	0.196	0.198	0.198	0.197	0.191	0.176	0.116
(5, 1)	0.208	0.181	0.239	0.255	0.257	0.256	0.254	0.249	0.239	0.217	0.143

Now observe that $\sup_\theta \rho_\lambda(f_\theta, f_{\theta^*_{F_0}}) = 1$. Thus (12) holds and also $\rho_\lambda(f_\theta, f_{\theta^*_{F_0}}) < 1$. Therefore, $\alpha_0(F_\theta;N_\lambda) = 1$ for all λ.

Example 2. (Poisson mean). Let F_θ be the cdf of $Poi(\theta)$, and the hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Here $N_\lambda(F_\theta) = \rho_\lambda(f_\theta, f_{\theta^*_{F_0}}) - \rho_\lambda(f_\theta, f_{\theta_{F_0}}) = \rho_\lambda(f_\theta, f_{\theta^*_{F_0}}) - e^{-\lambda}$. Now,

$$\rho_\lambda(f_\theta, f_{\theta^*_{F_0}}) = \sum e^{-\lambda[e^{(\theta_0-\theta)}(\frac{\theta}{\theta_0})^x]} e^{-\theta_0} \frac{\theta_0{}^x}{x!}.$$

We computed the numerical bounds for $\alpha_1(F_\theta;N_\lambda)$ for different values of θ for λ in $[10^{-5}$, 4.0]. For brevity, we have not reported the actual numbers in a graph or a table. For $\theta = 4$, for example, the breakdown point bound for Simpson's HDT is 0.134, that for the NEDT as given by Theorem 3 is 0.226 and the maximum bound of 0.2482 is attained for λ in a neighborhood of 0.45. For a fixed θ, the GNEDT breakdown points were observed to be higher than those of the HDT for an interval of λ values depending on θ. The value of λ at which the best breakdown occurs changes with θ.

Note that again here $\sup_\theta \rho_\lambda(f_\theta, f_{\theta^*_{F_0}}) = 1$. Thus (12) holds and also $\rho_\lambda(f_\theta, f_{\theta^*_{F_0}}) < 1$. Hence, $\alpha_0(F_\theta;N_\lambda) = 1$ for all λ.

5. Numerical studies

5.1. Examples

Here we demonstrate the performance of the estimators and tests on two real datasets.

Example 3. For Newcomb's light speed data (Stigler, 1977), Table 3 presents the values of the MHDE and MGNEDE$_\lambda$ of μ and σ for various values of λ under the normal model, as well as MLEs for the full data (ML), and those after deleting the two obvious outliers -40, and -2 $(ML + D)$. We have used the automatic kernel density function with the Epanechnikov kernel ($w(x) = 0.75(1 - x^2)$, if $|x| < 1$, and $w(x) = 0$, otherwise), $c_n = 0.5$, and

TABLE 3

MGNEDE$_\lambda$s of μ and σ for the Newcomb data under a normal model.

λ	0.25	0.50	0.75	1.00	1.25	1.50	2.00	4.00	$ML + D$	ML
$\hat{\mu}$	27.74	27.74	27.73	27.63	27.61	27.52	27.39	27.12	27.75	26.21
$\hat{\sigma}$	5.23	5.14	5.02	4.88	4.73	4.57	4.34	3.71	5.04	10.66

MHDE of μ = 27.74, and MHDE of σ = 4.97.

TABLE 4

The signed generalized negative exponential disparity test statistics and associated P-values for various values of λ for the two sample drosophila data. [Corresponding values for the signed LRT and signed HDT are also shown.]

	With all observations		Outlier deleted	
λ	Statistic	P-value	Statistic	P-value
0.25	1.559	0.059	1.558	0.060
0.50	1.234	0.108	1.223	0.111
0.75	0.831	0.203	0.808	0.209
1.00	0.462	0.322	0.439	0.330
1.25	0.273	0.392	0.258	0.398
1.50	0.204	0.419	0.194	0.423
2.00	0.168	0.433	0.162	0.435
4.00	0.157	0.437	0.152	0.439
LRT	2.959	0.002	1.099	0.136
HDT	0.698	0.242	0.694	0.244

$s_n = (0.6745)^{-1}$ median $(|X_i -$ median $(X_i)|)$. Notice that these estimators exhibit strong outlier resistance properties even for quite small values of λ.

Example 4. We apply the proposed disparity tests on a two sample problem with two sets of chemical mutagenicity data involving *drosophila*, also analyzed by Simpson (1989, p.112). The responses are the numbers of daughters with recessive lethal mutations among flies (*drosophila*) exposed to chemicals and subject to control conditions. The responses are modelled as random samples from $Poi(\theta_1)$ (control) and $Poi(\theta_2)$ (exposed) distributions respectively. The hypotheses of interest are $H_0 : \theta_1 \geq \theta_2$ against $H_1 : \theta_1 < \theta_2$. Let $n_i =$ number of observations in group i, $i = 1, 2$, and $n = n_1 + n_2$.

Let $d_i(x)$ denote the fraction of x-values in the ith sample, $i = 1, 2$, $x = 0, 1, 2, \ldots$. Let $\delta_i(x) = (d_i(x) - f_{\theta_i}(x))/f_{\theta_i}(x)$ where $f_{\theta_i}(x)$ is the $Poi(\theta_i)$ density at x. To define the two sample version of the test, let

$$\mathrm{GNED}_{O,\lambda}(d_1, d_2, f_{\theta_1}, f_{\theta_2}) = n^{-1} \sum_{i=1}^{2} n_i \sum_{x=0}^{\infty} \frac{e^{-\lambda\delta_i(x)} - 1 + \lambda\delta_i(x)}{\lambda^2} f_{\theta_i}(x)$$

denote the overall generalized negative exponential disparity in the discrete model (see Simpson, 1989, p. 112; Sarkar & Basu, 1995, p. 359). We then minimize $\mathrm{GNED}_{O,\lambda}$ with respect to $\theta = (\theta_1, \theta_2)^\top$ under the null space and under the unconstrained space to obtain their estimates $(\hat\theta^*_{1,\lambda}, \hat\theta^*_{2,\lambda})$ and $(\hat\theta_{1,\lambda}, \hat\theta_{2,\lambda})$ respectively. Compute the test statistic as

$$t_n = 2n\left[\mathrm{GNED}_{O,\lambda}\left(d_1, d_2, f_{\hat\theta^*_{1,\lambda}}, f_{\hat\theta^*_{2,\lambda}}\right) - \mathrm{GNED}_{O,\lambda}\left(d_1, d_2, f_{\hat\theta_{1,\lambda}}, f_{\hat\theta_{2,\lambda}}\right)\right].$$

The one-sample asymptotics extend to the present case if $n_1/n_2 \to a \in (0, 1)$ as $n \to \infty$.

For testing H_0: $\theta_1 \geq \theta_2$ against H_1: $\theta_1 < \theta_2$ a signed disparity test is appropriate. In the GNEDT case the test statistic is $t_n{}^{1/2}\mathrm{sign}(\hat\theta_{1,\lambda} - \hat\theta_{2,\lambda})$ and it is asymptotically equivalent to the signed LRT. For the full and reduced (after removing the two large counts for the treated group) data, the signed disparity statistics and the associated P-values are given in Table 4. Exclusion of the twolargest counts from the data has little impact on the robust methods, and in either case the robust tests provide similar degree of support for the null hypothesis that

TABLE 5

Observed distribution of the number of cases of peritonitis for 390 patients.

No. of cases	0	1	2	3	4	5	6	7	8	9	10	11	12
observed freq	199	94	46	23	17	4	4	1	0	0	1	0	1

the mean number of recessive lethal daughters in the control group is at least as large as that in the treated group. The conclusions, however, are opposite when one uses the signed LRT with and without the outliers.

Example 5. A data set on the incidence of peritonitis on 390 kidney patients was provided by Professor P.W.M. John (*pers. comm.*, 1995) (Table 5). A look at the data suggests that the observed frequencies (O_k, $k = 1, 2 \ldots$) of the number of cases (k) may be well modelled by the *geometric*(θ) distribution with θ around $\frac{1}{2}$. The sample size is fairly large, and it appears that although there are a few moderately large values, there are no extreme outliers. For an estimate $\hat{\theta}$, the expected frequencies are then obtained as $E_k = n\hat{\theta}\,(1-\hat{\theta})^k$. Treating it as a parametric density estimation problem, we assess the goodness-of-fit of the plug in predictive density corresponding to the estimate $\hat{\theta}$ with the log likelihood ratio statistic (Aitchison, 1975) which is given for these data as

$$G^2 = 2\sum_{k=0}^{\infty} O_k \log(O_k/E_k).$$

The G^2 value for the predictive density corresponding to the MLE is 11.84. The G^2 values for the MGNEDE$_\lambda$ for $\lambda = 0.75$, 0.9, 1.0, 1.1, 1.25 are, respectively, 11.90, 11.96, 12.01, 12.06, and 12.14. The G^2 value for the MHDE equals 12.48. In this example, the GNED based methods clearly give a better predictive fit than the HD.

Example 6. This is an example in discriminant analysis from the field of speech recognition. The dataset consists of 10 classes of two-dimensional measurement vectors. This was created by Peterson & Barney (1952) by a spectographic analysis of vowels in words formed by 'h' followed by a vowel and then followed by 'd'. A number of people were asked to speak the words and the first two *formant* frequencies of 10 vowels were split in two sets. The final data consisted of a training set having 338 cases and a test set having 333 cases. The *formants* are the two lowest resonant frequencies of a speaker's vocal tract. See Peterson & Barney (1952) and Bose (2003) for a more extensive discussion of the nature and construction of the data set.

We assumed a multivariate normal model for these data, and attempted to classify the observations in both the training set and test set using the traditional Bayes' quadratic discriminant rule, where the parameter estimates of the mean vectors and the covariance matrix are obtained (using the training set data alone) by the following methods: ML, MGNED$_\lambda$ for several values of λ, and MHD. The results are presented in Table 6. It is evident from the table that the use of the MLE leads to the worst performance among the cases considered here. The Bayes' linear discriminant analysis results, using the MLE, are also presented in the table for comparison. The use of robust estimators generates comparatively better error rates. Although the difference is not dramatic, the methods based on the GNED$_\lambda$ achieve better misclassification error rates than the HD-based methods for all the cases looked at

TABLE 6

Classification Example.

Method	Misclassification Error Percentage	
	Training Data	Test Data
ML (*LDA*)	28.40	26.13
ML (*QDA*)	21.60	21.02
MGNED ($\lambda = 0.75$)	20.41	19.22
MGNED ($\lambda = 1.00$)	20.41	19.52
MGNED ($\lambda = 1.25$)	20.41	19.22
MHD	21.60	19.82

in this example. Bose (2003) has tested several nonparametric methods on this dataset. Our results show that the misclassification error rates achieved using the MGNEDE (with $\lambda = 1.25$, say) are quite competitive with those obtained by the nonparametric methods used in Bose.

For estimating the unknown density of the data sets we used a multiplicative Epanechnikov kernel in all the cases of this experiment. The smoothing parameters for each component were obtained as $h_n = 1.06\hat{\sigma}n^{-1/5}$, where we used the robust estimate of scale $\hat{\sigma} = MAD/0.6745$, where MAD represents the median absolute deviation.

5.2. Simulation results

We conducted a simulation study for the Poisson model $Poi(\theta)$, and computed the empirical mean, variance and mean square error (MSE), against the target value θ, of the MLE, MHDE and MGNEDE$_\lambda$ for various λ values under different contaminated models of the form $\alpha Poi(\theta) + (1-\alpha)Poi(\theta_c)$. We have chosen $\lambda = 0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 2.00, 4.00$, $(\theta, \theta_c) = (2,12), (5,15)$, and $\alpha = 0$ and 0.10. Sample sizes considered are $n = 20$, 50 and 100. Our computations are based on five thousand replications. For brevity, we have presented only part of these findings in Table 7. More details can be found in Bhandari *et al.* (2000).

A universal recommendation of a most suitable value of λ to cover all possible scenarios can not be made at this point based on our limited numerical studies. The choice of λ for a particular problem will depend on specific needs. However, $\lambda = 1$ appears to be a quite reasonable choice because its performance is close to that of the best in almost all the cases we have looked at. In a broader sense there is not too much to choose from among the estimators in the range $\lambda \in [0.75, 1.25]$, and further research will be necessary for identifying an 'optimal' λ in a given case. In general, smaller values of λ in the above range tend to do better in terms of small sample efficiency, while the larger ones improve robustness. In this study the MHDE is clearly substantially less efficient than the MGNEDEs in the above range. The minimum PCS estimator (for $\lambda = 0$) is highly nonrobust, performing poorly even in comparison with the MLE.

Finally, we present a numerical study which is devised keeping in mind the inlier question. Is it really a practical problem to be concerned about at all? In the following we demonstrate that even when data come from the pure model, the inlier problem can cause serious damage to the methods based on the Hellinger distance. This study, based on the Poisson model $Poi(\theta)$, tested the null hypothesis $H_0 : \theta = 3$ against $\theta \neq 3$ using the HDT,

TABLE 7

Empirical mean, variance and mean square error of the MGNEDE_λ*, MHDE and MLE of* θ *for sample sizes 20, 50 and 100 in the Poisson model under pure and contaminated* $Poi(2)$ *distribution.*

λ	$n=20$			$n=50$			$n=100$		
	Mean	Var	MSE	Mean	Var	MSE	Mean	Var	MSE
				$Poi(2)$					
0	2.135	0.122	0.135	2.095	0.053	0.062	2.067	0.029	0.033
0.25	2.086	0.105	0.112	2.056	0.044	0.047	2.035	0.021	0.023
0.50	2.042	0.101	0.102	2.032	0.043	0.044	2.018	0.021	0.021
0.75	2.006	0.101	0.101	2.013	0.043	0.043	2.006	0.021	0.021
1.00	1.975	0.102	0.103	1.997	0.044	0.044	1.995	0.021	0.021
1.25	1.948	0.104	0.107	1.982	0.044	0.045	1.984	0.021	0.022
1.50	1.923	0.106	0.112	1.967	0.045	0.046	1.975	0.022	0.022
2.00	1.877	0.113	0.128	1.940	0.046	0.050	1.956	0.022	0.024
4.00	1.713	0.152	0.234	1.826	0.054	0.085	1.874	0.026	0.042
MHDE	1.857	0.107	0.127	1.921	0.040	0.046	1.952	0.021	0.023
MLE	1.998	0.096	0.096	2.001	0.042	0.042	1.997	0.020	0.020
				$0.9\ Poi(2) + 0.1\ Poi(12)$					
0	4.932	2.542	11.143	5.555	1.189	13.829	5.750	0.679	14.739
0.25	2.226	0.221	0.272	2.192	0.087	0.123	2.196	0.045	0.083
0.50	2.108	0.150	0.162	2.091	0.062	0.071	2.096	0.030	0.039
0.75	2.060	0.141	0.144	2.050	0.057	0.060	2.061	0.027	0.031
1.00	2.024	0.140	0.141	2.023	0.056	0.056	2.041	0.027	0.028
1.25	1.995	0.143	0.143	2.001	0.055	0.055	2.025	0.026	0.027
1.50	1.970	0.146	0.147	1.982	0.055	0.055	2.012	0.026	0.026
2.00	1.926	0.152	0.158	1.950	0.057	0.059	1.991	0.027	0.027
4.00	1.772	0.186	0.238	1.837	0.067	0.094	1.914	0.030	0.038
MHDE	1.862	0.125	0.144	1.934	0.051	0.055	1.973	0.025	0.026
MLE	3.037	0.624	1.699	2.997	0.250	1.243	3.006	0.125	1.137

NEDT and LRT at the nominal level 0.05. Five thousand independent random samples of size n were generated from $Poi(3)$ for $n = 20, 21, 22, \ldots, 100$. We computed the empirical levels of the three tests based on chi-square critical values, and have presented them here as a function of the sample size in Figure 2. Notice that the observed levels of the test statistics for HDT severely overestimate the true level. Inference based on a test that cannot hold its level will not be reliable. This phenomenon, in connection with the HDT, was also observed by Simpson (1989). Basu et al. (1996) have noted that this limitation of the HDT is at least partially due to the presence of inliers, particularly empty cells. The NEDT or the LRT does not appear to have this problem. Other members of the GNED family such as those corresponding to $\lambda = 0.75$ or 1.25 also produced satisfactory results like the NED, but those curves have not been added to the figure here to present a cleaner contrast.

6. Final remarks

Inference procedures based on the minimized Hellinger distance still remain the standard in density based minimum divergence inference. In this paper we have studied alternatives based on the GNED. They share the positive theoretical properties of the Hellinger distance based methods such as bounded α-influence function, full asymptotic efficiency, high

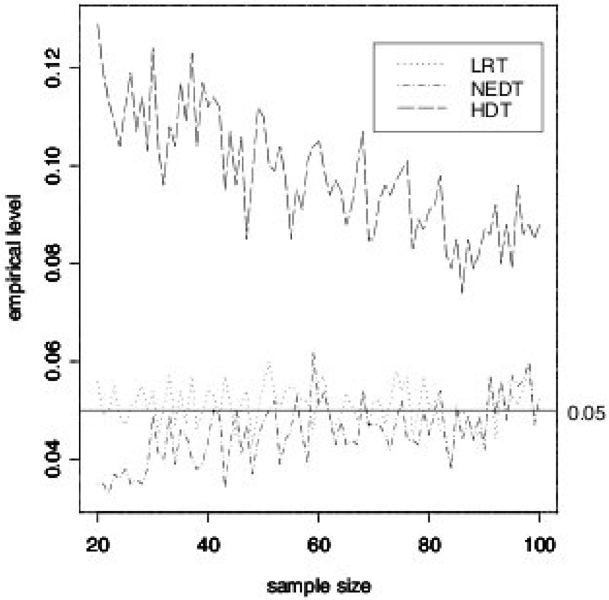

Figure 2. Empirical level of the HDT, NEDT and LRT for various sample sizes.

breakdown points, etc. But, in addition, the proposed methods have many other strengths over the methods based on the Hellinger distance. They appear to outperform the latter when the model is correct in small simulated samples – an effect in part of the inlier cells. Similar results are seen in examples where the model fits the data well. These methods were observed to produce better power breakdowns under some commonly occurring models. In a variety of examples the methods were seen to be extremely competitive or better than those based on the HD. While proposing a most suitable method within the observed class of the GNEDs will evidently require more research, we firmly believe that these methods present a rich class of practical alternatives to minimum divergence inference based on the Hellinger distance.

7. Appendix

Proof of Theorem 1. Fix $\lambda > 0$, and $\alpha \in (0, 1)$. Write t_m for $T_\lambda(H_{\alpha,m})$. Suppose, if possible, breakdown occurs, that is there exists an outlier sequence $\{K_m\}$ such that $|t_m| \to \infty$ as $m \to \infty$. Then we have,

$$\begin{aligned}
\left|\rho_\lambda(h_{\alpha,m}, f_{t_m}) - \rho_\lambda(\alpha k_m, f_{t_m})\right| &\leq \int_{\{x:\, f_{t_m}(x) \leq g(x)\}} \left| e^{-\lambda\left(\frac{(1-\alpha)g+\alpha k_m}{f_{t_m}}\right)} - e^{-\lambda\left(\frac{\alpha k_m}{f_{t_m}}\right)} \right| f_{t_m} \\
&\quad + \int_{\{x:\, f_{t_m}(x) > g(x)\}} \left| e^{-\lambda\left(\frac{(1-\alpha)g+\alpha k_m}{f_{t_m}}\right)} - e^{-\lambda\left(\frac{\alpha k_m}{f_{t_m}}\right)} \right| f_{t_m} \\
&\leq 2 \int_{\{x:\, f_{t_m}(x) \leq g(x)\}} f_{t_m} + \int_{\{x:\, f_{t_m}(x) > g(x)\}} \frac{\lambda(1-\alpha)g}{f_{t_m}} f_{t_m}
\end{aligned}$$

$\to 0$ as $m \to \infty$ by assumption **A3**, and hence $\left|\rho_\lambda(h_{\alpha,m}, f_{t_m}) - \rho_\lambda(\alpha k_m, f_{t_m})\right| \to 0$ as $m \to \infty$. Now, $\rho_\lambda(\alpha k_m, f_{t_m}) \geq e^{-\lambda\alpha}$ using Jensen's inequality. Hence, if a breakdown causing outlier sequence exists, we get $\liminf_{m\to\infty} \rho_\lambda(h_{\alpha,m}, f_{t_m}) \geq e^{-\lambda\alpha}$.

We will get a contradiction to our assumption of the existence of an outlier sequence $\{K_m\}$ for which breakdown occurs at contamination level α if we can show that there exists a bounded sequence $\{\theta_m\}$ of parameter values such that

$$\limsup_{m\to\infty} \rho_\lambda\big(h_{\alpha,m}, f_{\theta_m}\big) < e^{-\lambda\alpha} \leq \liminf_{m\to\infty} \rho_\lambda\big(h_{\alpha,m}, f_{t_m}\big) \tag{13}$$

since in this case the sequence $\{t_m\}$ cannot minimize $\rho_\lambda(h_{\alpha,m}, f_\theta)$ for every m. For any arbitrary bounded sequence $\{\theta_m\}$, using assumption **A2**, we obtain,

$$\left|\rho_\lambda\big(h_{\alpha,m}, f_{\theta_m}\big) - \rho_\lambda\big((1-\alpha)g, f_{\theta_m}\big)\right| \to 0 \text{ as } m \to \infty \tag{14}$$

and hence for such a sequence

$$\limsup_{m\to\infty} \rho_\lambda(h_{\alpha,m}, f_{\theta_m}) = \limsup_{m\to\infty} \rho_\lambda\big((1-\alpha)g, f_{\theta_m}\big) \geq \rho_\lambda\big((1-\alpha)g, f_{\theta_\alpha^*}\big) \geq e^{-\lambda(1-\alpha)}, \tag{15}$$

where the last inequality follows by Jensen's inequality. In particular, if we choose $\{\theta_m = \theta_\alpha^*\}$, a bounded sequence of identical values, we get

$$\limsup_{m\to\infty} \rho_\lambda\big(h_{\alpha,m}, f_{\theta_\alpha^*}\big) = \limsup_{m\to\infty} \rho_\lambda\big((1-\alpha)g, f_{\theta_\alpha^*}\big) = \rho_\lambda\big((1-\alpha)g, f_{\theta_\alpha^*}\big) \geq e^{-\lambda(1-\alpha)}. \tag{16}$$

Then combining (15) and (16) we have, for any bounded sequence $\{\theta_m\}$,

$$\limsup_{m\to\infty} \rho_\lambda(h_{\alpha,m}, f_{\theta_m}) \geq \limsup_{m\to\infty} \rho_\lambda\big(h_{\alpha,m}, f_{\theta_\alpha^*}\big) = \rho_\lambda\big((1-\alpha)g, f_{\theta_\alpha^*}\big) \geq e^{-\lambda(1-\alpha)}. \tag{17}$$

Since by definition $\rho_\lambda((1-\alpha)g, f_\theta) \leq 1$ for all θ, we have $\rho_\lambda((1-\alpha)g, f_{\theta_\alpha^*}) \in [e^{-\lambda(1-\alpha)}, 1]$, and hence $\exists\ 0 \leq b_\alpha \leq 1$ (and b_α not dependent on m) such that

$$\limsup_{m\to\infty}\ \rho_\lambda(h_{\alpha,m}, f_{\theta^*}) = \rho_\lambda((1-\alpha)g, f_{\theta^*}) = e^{-\lambda(1-\alpha)b_\alpha}. \tag{18}$$

Given any α, we have a contradiction, by (13) and (18), to the assumption that there exists an outlier sequence which causes breakdown in T_α if $e^{-\lambda(1-\alpha)b_\alpha} < e^{-\lambda\alpha}$. In particular breakdown does not occur when $\alpha < \alpha^*$.

We have shown that if $\alpha < \alpha^*$, for any outlier sequence $\{K_m\}$ there exists a bounded sequence of parameter values which makes the divergence $\rho_\lambda(h_{\alpha,m}, f_\theta)$ smaller in the limit than a divergent sequence. By (17), we have $\limsup_{m\to\infty} \rho_\lambda(h_{\alpha,m}, f_{\theta_m}) \geq \limsup_{m\to\infty} \rho_\lambda(h_{\alpha,m}, f_{\theta_\alpha^*})$ for any bounded sequence $\{\theta_m\}$. This implies that θ_α^* is the minimizer of the divergence $\rho_\lambda(h_{\alpha,m}, f_\theta)$ in the limit as $m \to \infty$ as long as $\alpha < \alpha^*$.

Next we show that $b_\alpha = 1$ when $G = F_{\theta_0} \in \mathcal{F}_\Theta$. Let $g = f_{\theta_0}$, and θ_α^* is now the minimizer of $\rho_\lambda((1-\alpha)f_{\theta_0}, f_\theta)$. By (14) we then have

$$\limsup_{m\to\infty} \rho_\lambda(h_{\alpha,m}, f_{\theta_\alpha^*}) = \limsup_{m\to\infty} \rho_\lambda((1-\alpha)f_{\theta_0}, f_{\theta_\alpha^*}) \leq \limsup_{m\to\infty} \rho_\lambda((1-\alpha)f_{\theta_0}, f_{\theta_0}) = e^{-\lambda(1-\alpha)}. \tag{19}$$

Thus, by the above inequality and (16),

$$e^{-\lambda(1-\alpha)} \leq \limsup_{m\to\infty} \rho_\lambda\big(h_{\alpha,m}, f_{\theta_\alpha^*}\big) = \rho_\lambda\big((1-\alpha)f_{\theta_0}, f_{\theta_\alpha^*}\big) \leq e^{-\lambda(1-\alpha)}.$$

Hence $\rho_\lambda((1-\alpha)f_{\theta_0}, f_{\theta_\alpha^*}) = e^{-\lambda(1-\alpha)} = \rho_\lambda((1-\alpha)f_{\theta_0}, f_{\theta_0})$. Therefore, $\theta^* = \theta_0, b_\alpha = 1$ and there is no breakdown in the T_λ functional when $e^{-\lambda(1-\alpha)} < e^{-\lambda\alpha} \Leftrightarrow \alpha < 1/2$ for any outlier sequence $\{K_m\}$. Also in this case (i.e., $g = f_{\theta_0}$ and $\alpha < 1/2$) the minimizer of $\rho_\lambda(h_{\alpha,m}, f_\theta)$ in the limit as $m \to \infty$ is θ_0.

Proof of Theorem 2. We prove that $2n(\text{GNED}_\lambda(g_n, f_{\hat\theta^*}) - \text{GNED}_\lambda(g_n, f_{\hat\theta})) - LRT = o_p(1)$ under H_0. By a Taylor series expansion we have

$$2n(\text{GNED}_\lambda(g_m, f_{\hat\theta^*}) - \text{GNED}_\lambda(g_n, f_{\hat\theta}))$$

$$= n(\hat\theta^* - \hat\theta)^\top I(\theta_0)(\hat\theta^* - \hat\theta) + n(\hat\theta^* - \hat\theta)^T(\text{GNED}''_\lambda(g_n, \tilde\theta) - I(\theta_0))(\hat\theta^* - \hat\theta) \tag{20}$$

with $\tilde\theta$ lying between $\hat\theta^*$ and $\hat\theta$. Now, under the assumed conditions, one can show that

$$n^{1/2}(\hat\theta_{ML} - \hat\theta^*_{ML}) = n^{1/2}(\hat\theta - \hat\theta^*) + o_p(1), \tag{21}$$

$$n^{1/2}(\hat\theta - \hat\theta^*) = O_p(1), n^{1/2}(\hat\theta_{ML} - \hat\theta^*_{ML}) = O_p(1), \tag{22}$$

$$(\text{GNED}''_\lambda(g_n, \tilde\theta) - I(\theta_0)) = o_p(1), \tag{23}$$

$$n^{1/2}(\hat\theta - \theta_0) = I^{-1}(\theta_0)n^{1/2}\left(\frac{1}{n}\sum_{i=1}^n u_{\theta_0}(X_i)\right) + o_p(1) \tag{24}$$

by utilizing the steps in the proof of Theorem 4.4.4 of Serfling (1980), and equations (3.5) and (3.10) of Basu *et al.* (1997). Now by (21)

$$n(\hat\theta^* - \hat\theta)^\top I(\theta_0)(\hat\theta^* - \hat\theta) = n(\hat\theta^*_{ML} - \hat\theta_{ML})^\top I(\theta_0)(\hat\theta^*_{ML} - \hat\theta_{ML}) + o_p(1). \tag{25}$$

Therefore, by (20), (22), (23) and (25) we have

$$2n(\text{GNED}_\lambda(g_n, f_{\hat\theta^*}) - \text{GNED}_\lambda(g_n, f_{\hat\theta})) = n(\hat\theta^*_{ML} - \hat\theta_{ML})^\top I(\theta_0)(\hat\theta^*_{ML} - \hat\theta_{ML}) + o_p(1).$$

From the proof of Theorem 4.4.4 of Serfling (1980) it follows that the asymptotic distribution of $n(\hat\theta^*_{ML} - \hat\theta_{ML})^\top I(\theta_0)(\hat\theta^*_{ML} - \hat\theta_{ML})$ is $\chi^2(r)$. Thus the proof is complete.

Proof of Theorem 3. Note that

$$\rho_\lambda(h, f_{\theta_H}) \le \rho_\lambda(h, f_{\theta_G}) \le (1-\alpha)\rho_\lambda(g, f_{\theta_G}) + \alpha \le \rho_\lambda(g, f_{\theta_G}) + \alpha(1 - e^{-\lambda}),$$

$$\rho_\lambda(h, f_{\theta_H^*}) \ge \int \exp\left(-\lambda\frac{(1-\alpha)g}{f_{\theta_H^*}}\right) f_{\theta_H^*}\left(1 - \lambda\frac{\alpha k}{f_{\theta_H^*}}\right) \ge \rho_\lambda(g, f_{\theta_H^*}) - \alpha\lambda \ge \rho_\lambda(g, f_{\theta_G^*}) - \alpha\lambda.$$

Therefore, $N_\lambda(H) \ge N_\lambda(G) - \alpha((\lambda+1) - e^{-\lambda})$. Thus, for any $\alpha < N_\lambda(G)/(\lambda + 1 - e^{-\lambda})$,

$$\inf_{K\in\mathcal{G}} N_\lambda(H) \ge N_\lambda(G) - \alpha((\lambda+1) - e^{-\lambda}) > 0 = N_{\lambda,\min}$$

and this establishes the result.

Proof of Lemma 1. If $\rho_\lambda(g, f_{\theta_G^*}) = 1$, then $\rho_\lambda(g, f_{\theta_H^*}) = 1$. Thus, we have

$$\begin{aligned}\rho_\lambda(h, f_{\theta_H^*}) &= \int \exp\left(-(1-\alpha)\lambda\frac{g}{f_{\theta_H^*}}\right)\exp\left(-\alpha\lambda\frac{k}{f_{\theta_H^*}}\right)f_{\theta_H^*}\\ &= \int \exp\left(-\alpha\lambda\frac{k}{f_{\theta_H^*}}\right)f_{\theta_H^*} \geq \exp(-\alpha\lambda)\end{aligned}$$

On the other hand,

$$\rho_\lambda(h, f_{\theta_H}) \leq \rho_\lambda(h, f_{\theta_G}) = \int \exp\left(-(1-\alpha)\lambda\frac{g}{f_{\theta_G}}\right)\exp\left(-\alpha\lambda\frac{k}{f_{\theta_G}}\right)f_{\theta_G} \leq \exp(-(1-\alpha)\lambda).$$

Thus, $N_\lambda(H) \geq \exp(-\alpha\lambda) - \exp(-(1-\alpha)\lambda)$, and this being greater than zero is equivalent to $-\alpha > -(1-\alpha)$, i.e., $\alpha < 1/2$

Proof of Theorem 4. Since (12) holds, we have $N_{\lambda,\max} = 1 - e^{-\lambda}$. Let $H = (1-\alpha)\ G + \alpha K$. Now $\rho_\lambda(h, f_{\theta_H}) \geq e^{-\lambda}$. Also $\rho_\lambda(h, f_{\theta_H^*}) \leq \rho_\lambda(h, f_{\theta_G^*}) \leq (1-\alpha)\rho_\lambda(g, f_{\theta_G^*}) + \alpha\rho_\lambda(k, f_{\theta_G^*})$. Since $\rho_\lambda(k, f_{\theta_G^*}) \leq 1$, it follows that $\rho_\lambda(h, f_{\theta_H^*}) < 1$ unless $\alpha = 1$.

References

AGOSTINELLI, C. & MARKATOU, M. (2001). Test of hypotheses based on the Weighted Likelihood Methodology. *Statistica Sinica*, **11**, 499-514.

AITCHISON, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554.

BASU, A. & LINDSAY, B. G. (1994). Minimum Disparity estimation in the continuous Case: Efficiency, Distributions, Robustness. *Ann. Inst. Statist. Math.*, **46**, 683–705.

BASU, A., HARRIS, I. R. & BASU, S. (1996). Tests of hypothesis in discrete models based on the penalized Hellinger distance. *Statistics and Probability Letters*, **27**, 367–373.

BASU, A., SARKAR, S. & VIDYASHANKAR, A. N. (1997). Minimum negative exponential disparity estimation in parametric models. *J. Statist. Plann. Inf.*, **58**, 349–370.

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

BHANDARI, S. K., BASU, A. & SARKAR, S. (2000). Robust estimation in parametric models using the family of generalized negative exponential disparities. *Technical Report # 7/2000*, Stat-Math Unit, Indian Statistical Institute, Calcutta 700108, India.

BOSE, S. (2003). Multilayer statistical classifiers. *Computational Statistics and Data Analysis*, **42**, 685–701.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. (1986). *Robust Statistics: The Approach Based On Influence Functions*. New York: Wiley.

HARRIS, I. R. & BASU, A. (1994). Hellinger distance as a penalized log likelihood. *Commun. Statist. Simula.* **23**, 1097–1113.

HE, X., SIMPSON, D. & PORTNOY, S. (1990). Breakdown robustness of tests. *J. Amer. Statist. Assoc.*, **85**, 446–452.

JEONG, D. & SARKAR, S. (2000). Negative exponential disparity based family of goodness-of-fit tests for multinomial models. *J. Statist. Comput. Simul.*, **65**, 43–61.

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

MARKATOU, M., BASU, A. & LINDSAY, B. G. (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *J. Statist. Plann. Inf.*, **57**, 215–232.

MARKATOU, M., BASU, A. & LINDSAY, B. G. (1998). Weighted likelihood equations with bootstrap root search. *J. Amer. Statist. Assoc.*, **93**, 740–750.

PETERSON, G. E. & BARNEY, H. L. (1952). Control methods used in a study of vowels. *J. of the Acoustical Society of America*, **24**, 175–185.

SARKAR, S. & BASU, A. (1995). On disparity based robust tests for two discrete populations. *Sankhya, Series B*, **57**, 353-364.

SERFLING, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

SIMPSON, D. G. (1987). Minimum Hellinger distance estimation for 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.

STIGLER, S. M. (1977). Do robust estimators work with real data? (with discussion). *Ann. Statist.*, **7**, 1055–1098.

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