The Annals of Statistics
1994, Vol. 22, No. 2, 1081–1114

EFFICIENCY VERSUS ROBUSTNESS: THE CASE FOR MINIMUM HELLINGER DISTANCE AND RELATED METHODS[1]

BY BRUCE G. LINDSAY

Pennsylvania State University

It is shown how and why the influence curve poorly measures the robustness properties of minimum Hellinger distance estimation. Rather, for this and related forms of estimation, there is another function, the residual adjustment function, that carries the relevant information about the trade-off between efficiency and robustness. It is demonstrated that this function determines various second-order measures of efficiency and robustness through a scalar measure called the estimation curvature. The function is also shown to determine the breakdown properties of the estimators through its tail behavior. A 50% breakdown result is given. It is shown how to create flexible classes of estimation methods in the spirit of M-estimation, but with first-order efficiency (or even second-order efficiency) at the chosen model, 50% breakdown and a minimum distance interpretation.

1. Introduction. There are two fundamental—but potentially competing—ideals in parametric estimation: efficiency when the model has been appropriately chosen and robustness when it has not. The primary goals of this paper are to come to an understanding of how the minimum Hellinger distance estimator and its relatives balance these two ideals, and to develop this understanding into new approaches to efficiency and robustness trade-offs.

In the M-estimation approach to the construction of robust procedures, the robustness is attained at some sacrifice of first-order efficiency [e.g., Hampel, Ronchetti, Rousseeuw and Stahel (1986)]. Mathematically, this trade-off between efficiency and robustness must occur if one adheres to the notion, central to much of robustness theory, that the influence curve carries most of the critical information about the robustness of a procedure. If one insists, for example, that this curve be bounded, then it will generally not equal the influence curve of the fully efficient maximum likelihood estimator.

On the other hand, the fully known properties of minimum Hellinger distance estimation indicate that some of the information about robustness is, in this case, not carried by the influence curve. This method attains first-order efficiency, yet has certain robustness properties [Beran (1977), Tamura and Boos (1986), Simpson (1987, 1989) and Donoho and Liu (1988)]. Note that the efficiency of this method at the model means that it must have the same influence

Received September 1990; revised September 1993.

[1]This research was supported by NSF Grant DMS-91-06795, a Humboldt Foundation Senior Scientist award and the Population Research Institute of Pennsylvania State University.

AMS 1991 *subject classifications.* Primary 62F35; secondary 62F05, 62F10, 62F12.

Key words and phrases. Efficiency, robustness, minimum Hellinger distance, second-order efficiency, breakdown point.

function as the maximum likelihood estimator. One of our basic points is that the influence curve is a very misleading measure of robustness in the case of minimum Hellinger distance; we will illustrate this with an example at the end of this section.

The rest of this paper presents an investigation of this phenomenon, with the following key findings. Within a large class of minimum distance type methods, including both maximum likelihood and minimum Hellinger distance, there exists a key function, here denoted $A(\cdot)$ and called the *residual adjustment function* (RAF), whose shape completely determines the robustness and efficiency behavior of the corresponding estimator. Although similar in interpretation to the ψ-function of M-estimation, the impact of its shape occurs only in second-order calculations of efficiency and robustness.

The analysis here leads to a different paradigm for robustness in which the degree to which an observation is treated as an "outlier" depends on both the sample size and its probability of occurrence under the specified model. This corresponds to the intuition that an observation of $X = 3$ is unusual in a standard normal sample of size 10, but not so surprising in a sample of size 10,000.

We investigate these ideas in the context of the multinomial model because it offers a simple setting with minimal mathematical nuisance. As such the approach taken here is completely free of usual continuous location–scale model ideas and analyses. A companion paper [Basu and Lindsay (1993a)] shows that these ideas can be applied in that setting, with results very similar to conventional robust methods.

First, some notation and basic ideas. Suppose the sample space is a countable set, without loss of generality $\mathbf{X} = \{0, 1, \ldots, K\}$, with K possibly infinite, and that $m_\beta(x)$ is a family of probability densities on $\mathbf{X}$, indexed by $\beta \in \Omega$. To avoid technicalities, it will be assumed that $m_\beta(x) > 0$ for all $x \in \mathbf{X}$. Moreover, suppose that n independent and identically distributed observations $X_1, \ldots, X_n$, are made from $m_\beta(x)$. Let $d(x)$ be the proportion of the n observations which had value x. For any function $h(\cdot)$ on $\mathbf{X}$, we will let $\mathbf{h}$ denote the vector $(h(0), \ldots, h(K))$, with $\mathbf{d}$ thereby being the *data vector* and $\mathbf{m}_\beta$ being the *model vector*. The symbol "Σ" will denote summation on variable x over $\mathbf{X}$ unless otherwise noted. The function $t(x)$ will denote some nominal true density, not necessarily from the model. The minimum Hellinger distance estimator is that value of β that minimizes the squared Hellinger distance, $\Sigma[\sqrt{d(x)} - \sqrt{m_\beta(x)}]^2$.

In Section 2 we will establish that the maximum likelihood estimator and minimum Hellinger distance estimators have the same influence curves at the model $m_\beta(x)$. That is, if $\xi \in \mathbf{X}$, let $\chi_\xi(x)$ be the indicator function for ξ. Let

$$t_\varepsilon(x) := (1 - \varepsilon)t(x) + \varepsilon\chi_\xi(x) \tag{1}$$

be an ε-contaminated version of density $t(x)$. If $T(\mathbf{t})$ is a functional on the space of densities $t(x)$, then it is *Fisher consistent* for our problem if $T(\mathbf{m}_\beta) = \beta$. Its *influence function* is defined to be

$$T'(\xi) := \left.\frac{\partial T\big((1-\varepsilon)\mathbf{t} + \varepsilon\chi_\xi\big)}{\partial\varepsilon}\right|_{\varepsilon=0} = \frac{\partial T(\mathbf{t}_\varepsilon)}{\partial\varepsilon}.$$

If the functional is the maximum likelihood estimator $T_{\mathrm{ML}}(\mathbf{t})$ or the minimum Hellinger distance estimator $T_{\mathrm{HD}}(\mathbf{t})$, then Proposition 1 (Section 2) establishes that, when $t(x)$ corresponds to some $m_\beta(x)$,

$$T'(\xi) := \left.\frac{\partial T\big((1-\varepsilon)\mathbf{m}_\beta + \varepsilon\chi_\xi\big)}{\partial\varepsilon}\right|_{\varepsilon=0} = i(\beta)^{-1}u(\xi;\beta), \tag{2}$$

where $u(\xi;\beta)$ is $\nabla \log(m_\beta(\xi))$, the score function, and $i(\beta)$ is the Fisher information.

Viewed as a function of ε, $\Delta T(\varepsilon) := T(\mathbf{t}_\varepsilon) - T(\mathbf{t})$ represents how the functional changes with contamination. Consider the Taylor series approximation:

$$\Delta T(\varepsilon) := T(\mathbf{t}_\varepsilon) - T(\mathbf{t}) \approx \varepsilon T'(\xi). \tag{3}$$

This approximation is critical to the dual role of the influence function as a measure of efficiency and of robustness. First, the asymptotic variance of the estimator follows from the approximation $T(\mathbf{d}) - T(\mathbf{t}) = \Sigma\, T'(x)\, d(x)$ [von Mises (1947); see also Fernholz (1983)]. In our case, this would imply that, when the model is correctly specified, any estimator with the same influence function as the maximum likelihood estimator, namely, (2), has the same efficiency and so is optimal.

Approximation (3) also plays a role in robustness properties through a shift in perspective. Let $t(x)$ in the above calculations be a model density $m_\beta(x)$, and suppose that the true density is $\mathbf{t}_\varepsilon$, so that the model has been contaminated by observations at ξ in proportion ε. If the goal is to estimate the value $T(\mathbf{t}) = \beta$ and the functional $T(\mathbf{d})$ is used, then $T(\mathbf{d})$ converges to $T(\mathbf{t}_\varepsilon)$. Hence the function $\Delta T(\varepsilon)$ now represents the asymptotic bias in estimating β under contamination. In the case of first-order efficient estimators, this means that

$$\Delta T(\varepsilon) \approx \varepsilon i(\beta)^{-1}u(\xi;\beta) \tag{4}$$

is a first-order approximation to contamination bias. Thus, to this order of approximation, all the first-order efficient estimators have the same sensitivity to contamination as the maximum likelihood estimator and would therefore, in some eyes, be considered nonrobust. However, a key point in our analysis is that the approximation (3) can be very misleading, as we now illustrate.

EXAMPLE 1. To simplify this study, we consider the estimation of the mean value parameter μ in a one-parameter exponential family model. In this case, the influence curves of first-order efficient estimators all equal $T'(\xi) = \xi - \mu$. Moreover, since the maximum likelihood estimator is just the sample mean, the approximation (4) for the maximum likelihood estimator T_{ML} is exact. Thus, for ξ fixed, the plot of the ΔT_{ML} against ε is linear in ε.

The 100% accuracy of this approximation for ΔT_{ML} can be contrasted with its accuracy for ΔT_{HD}, the change in the minimum Hellinger distance functional, through the following example. The model is binomial(12, β), and the true value $\beta = \frac{1}{2}$. The contamination is mass ε at $\xi = 12$. In Figure 1, the diagonal straight

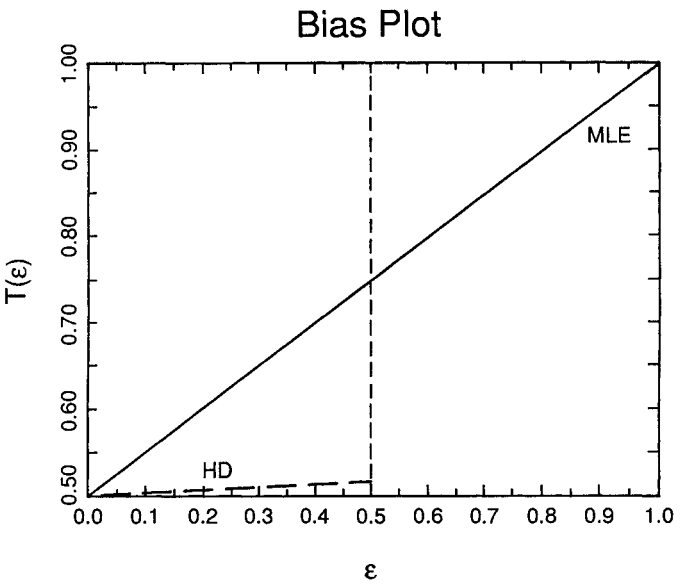

FIG. 1. *Bias plot for the Hellinger MDE in a binomial example.*

line corresponds to $\Delta T_{ML}(\varepsilon)$ for the maximum likelihood functional, where the curve for minimum Hellinger distance $\Delta T_{HD}(\varepsilon)$ is strikingly different. Since both estimators have the same influence curve, the slopes at $\varepsilon = 0$ must be identical; however, this appears to be contradicted by the plot.

The plot has two key features, whose source we will identify in this paper. First, the linear influence curve approximation for Hellinger works very poorly, even for extremely small ε. In Figure 2, we scale the picture to the level of contamination from 0 to 5%. It was found that the actual bias of T was more than 90% of its linear predicted value only for contamination in the neighborhood of 1 in 10,000. At contamination level 5%, the actual bias was only 13.6% of predicted.

We next observe that the estimator then stays nearly constant (not exceeding $\beta = 0.54$) until $\varepsilon \approx 0.50$, at which point the second key feature occurs: suddenly, the global minimum of the distance switches from a local minimum near $\beta = 0.54$ to a local minimum at $\beta = 1$, where it stays for every larger ε. The latter is dramatic visible evidence for the related result of Simpson (1987) that the Hellinger distance functional has an asymptotic breakdown point of 0.5 in the Poisson model.

This example motivates the need for an alternative to the influence curve analysis when we consider the efficiency and robustness of some estimators. The next step is to find the key structural element that links maximum likelihood and minimum Hellinger distance. In the process, we extend to a natural generalization of this type of estimation, all elements having a minimum distance

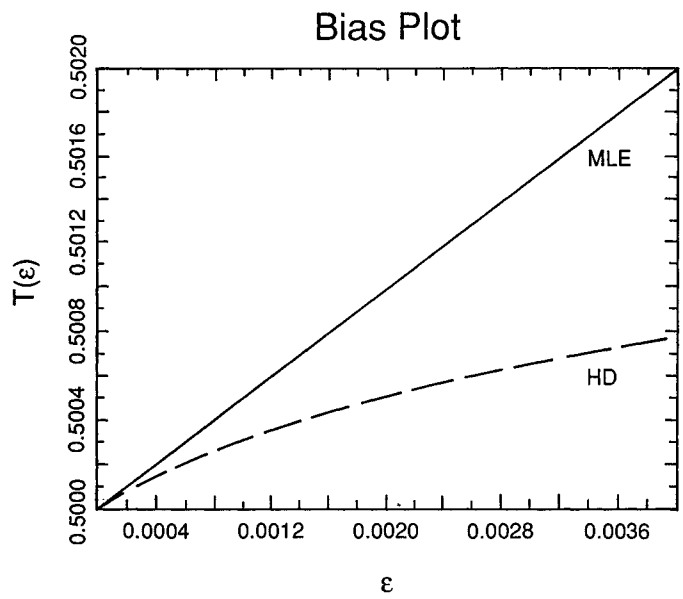

FIG. 2. *Blown-up corner of the plot in Figure* 1.

interpretation, with a wide range of possible efficiency–robustness trade-offs.

2. The estimating functions. Define the Pearson residual function $\delta(x)$ to be

$$\delta(x) = \left[d(x) - m_\beta(x)\right]/m_\beta(x).$$

We use this name because the model-weighted sum of the squared residuals, $\Sigma m_\beta(x)\delta(x)^2$, is Pearson's chi-squared distance. We note that these residuals are not standardized to have identical variances. It is important to note that these residuals have range $[-1, \infty]$.

Let ∇ denote differentiation with respect to β. The focal point of this paper is the class of estimating equations for β of the form

$$\sum A\big(\delta(x)\big)\nabla m_\beta(x) = 0, \tag{5}$$

where $A(\delta)$ satisfies the following assumption.

ASSUMPTION 1. $A(\delta)$ is assumed to be an increasing twice-differentiable function on $[-1, \infty)$, with $A(0) = 0$ and $A'(0) = 1$.

Such a function $A(\delta)$ will be called a *residual adjustment function* (RAF). It will be shown in Section 3 that the estimating equations (5) arise naturally in the minimization of certain distance-type measures. Using the linear RAF $A_{\mathrm{LD}}(\delta) := \delta$ yields the most important special case, the maximum likelihood

equations:

$$0 = \sum \delta(x)\,\nabla m_\beta(x) = \sum \big(d(x) - m(x)\big)u(x;b) = \sum d(x)u(x;\beta).$$

The minimum Hellinger distance equation also has the structural form (5), with

$$A_{\mathrm{HD}}(\delta) := 2\big[\sqrt{\delta+1} - 1\big].$$

Our first point is that such an estimating function automatically gives us first-order efficient estimators. [See also Rao, Sinha and Subramanyan (1983).]

PROPOSITION 1. *For an estimating function of the form* $\Sigma A(\delta(x))\,\nabla m_\beta(x)$, *the influence curve of the estimators* $T(\cdot)$ *has the form* $T'(\xi) = DEN^{-1}\,NUM$, *where, for* $\beta^* = T(\mathbf{t})$ *and* $\delta(x) = [t(x) - m_{\beta^*}(x)]/m_{\beta^*}(x)$,

$$NUM = A'\big(\delta(\xi)\big)u(\xi,\beta^*) - E\Big[A'\big(\delta(X)\big)u(X,\beta^*)\Big],$$

$$DEN = E\Big[u(X,\beta^*)u(X,\beta^*)^t A'\big(\delta(X)\big)\Big] + \sum A\big(\delta(x)\big)\nabla^2 m_{\beta^*}(x).$$

If the density $t(x)$ *is a model point* $m_\beta(x)$, *then* $\beta^* = \beta$, $\delta(x) = 0$ *for all* x, *and the estimators defined by* $\Sigma A(\delta(x))\nabla m_\beta(x) = 0$ *have influence function* $i(\beta)^{-1}u(\xi;\beta)$.

PROOF. The proof is straightforward but tedious differentiation of the estimating equation. □

In Appendix A we have shown that the influence function approximation yields the correct asymptotic distribution for the estimators of interest. This proof is of interest largely because the natural expansions in δ fail to converge when the sample space is infinite. To solve this difficulty, we show how to convert the remainders from Pearson residuals to a Hellinger-type residual.

An important class of residual adjustment functions that contains the above two have the form

$$A_\lambda(\delta) = \frac{(1+\delta)^\lambda - 1}{\lambda + 1}. \tag{6}$$

Arising from minimizing the power-weighted divergence family of distance-type measures [Cressie and Read (1984)], the estimators correspond to maximum likelihood ($\lambda = 0$), Hellinger distance ($\lambda = -\frac{1}{2}$), as well as minimum Pearson's chi-squared ($\lambda = 1$), minimum Neyman's chi-squared ($\lambda = -2$) and minimum Kullback–Leibler divergence ($\lambda = -1$). In Figure 3 we show a variety of shapes available for $A(\delta)$ within this class.

REMARK A. Since $A(\delta)$ is monotone, we could have taken as residual function any invertible increasing function of δ, and we would still generate the same class of estimating functions. The Pearson form of residual was chosen because, on this scale, maximum likelihood occupies the central position of having the linear residual adjustment function, $A_{\mathrm{LD}}(\delta) = \delta$.

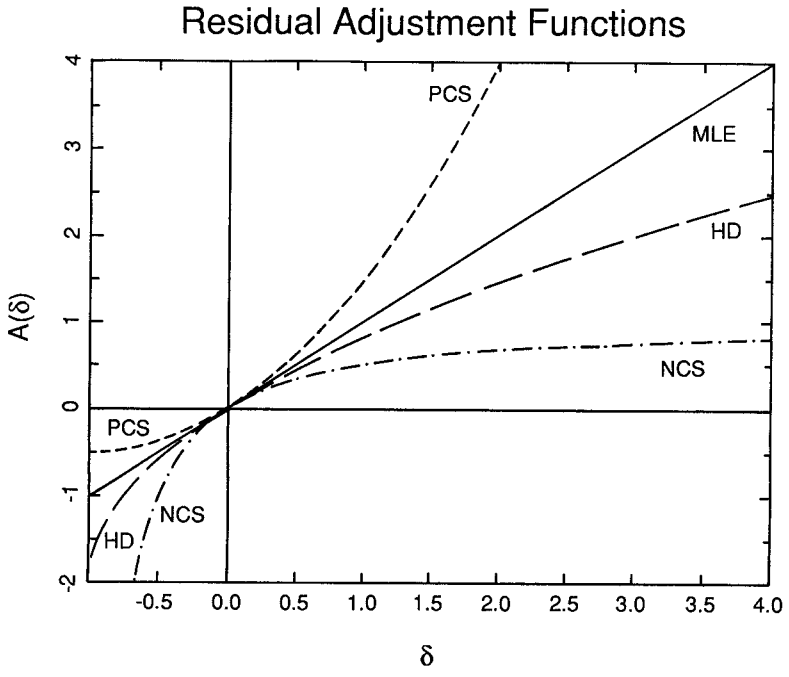

FIG. 3. *The RAFs for NCS, PCS, HD and MLE.*

REMARK B. What is an outlier? If one member of our sample is $X = \xi$, with $d(\xi) = 1/n$, then the Pearson residual is $\delta(\xi) = (nm_\beta(\xi))^{-1} - 1$, so that the magnitude depends on both the sample size and the model probability. As we can see in Figure 3, a large such residual gets quite different weights using different functions.

Exactly as in the case of the ψ-function of M-estimation, the structure of the residual adjustment function $A(\delta)$ has direct bearing on the efficiency and robustness properties of the corresponding estimator. It is clear from (5) that these estimators are completely determined by the form of $A(\delta)$. We can now summarize how the results of this paper give insights into the striking behavior of minimum Hellinger distance in Figure 1.

It is the behavior of $A(\delta)$ for δ near 0 that determines behavior of the bias response curve $\Delta T(\varepsilon)$ near $\varepsilon = 0$. The estimators considered here all have the same influence curve, $\Delta T'(0)$, because they all have the same first-order Taylor approximation $A(\delta) \approx \delta$. The second-order approximation,

$$A(\delta) \approx \delta + A_2\delta^2/2, \tag{7}$$

where $A_2 := A''(0)$, will be shown to determine a number of second-order efficiency and robustness properties of the RAF estimators, including the sharp bend of $\Delta T(\varepsilon)$ near $\varepsilon = 0$ for Hellinger distance in Figure 1. These results will

be presented in Sections 4 and 5.

On the other hand, to understand the behavior of $\Delta T(\varepsilon)$ near $\varepsilon = 0.5$, we must consider the large-δ behavior of the function $A(\delta)$. This outlier analysis will be performed in Section 6. We will show that the breakdown properties of the estimator are determined by the tails of $A(\delta)$.

In addition, in Section 3 we will show that estimating equations of the form (5) are always associated with distance-type measures. In Section 7, we will consider some new measures with a flexible trade-off in efficiency and robustness. Other issues to be considered in this paper included construction of robust tests (Section 5) and computation using iterative reweighting (Section 8).

3. Disparity measures. We now verify a very important property of RAF estimators: that solving the estimating equations (5) corresponds to the minimization of a measure of "distance" between the data $\mathbf{d}$ and the model $\mathbf{m}_\beta$. There are a number of fortunate consequences: for example, this gives a means of selecting between solutions of the estimating equations, it gives a nice conceptual meaning to the estimator and it enables us to create a robust analogue of the likelihood ratio test. In addition, it is *crucial* in the development of the breakdown results.

Suppose that $G(\cdot)$ is a real-valued thrice-differentiable function on $[-1, \infty)$, with $G(0) = 0$. For any pair of densities $m_\beta(x)$ and $d(x)$, define the *disparity measure* determined by G to be

$$\rho(\mathbf{d}, \mathbf{m}_\beta) = \sum m_\beta(x) G\big(\delta(x)\big), \tag{8}$$

where $\delta(x)$ is the Pearson residual defined in Section 2. If G is strictly convex, as will be assumed, then applying Jensen's inequality to the measure ρ shows that it is nonnegative, and it is zero only when $\mathbf{d} = \mathbf{m}_\beta$ [Csiszar (1963)]. As a consequence, if we use as an estimator that value of β, call it $T(\mathbf{d})$, that minimizes ρ, then this *minimum disparity estimator* (MDE) is Fisher consistent. [Such measures are sometimes called f or ϕ-*divergences*; Gelfand and Dey (1991) have used these measures to study Bayesian robustness.]

An important class of such measures is the Cressie–Read [Cressie and Read (1984) and Read and Cressie (1988)] family of *power divergence measures*, defined by

$$\begin{aligned} \mathrm{PWD}(\mathbf{d}, \mathbf{m}_\beta) &= \sum d(x) \frac{\left\{ \left[d(x)/m_\beta(x) \right]^\lambda - 1 \right\}}{\lambda(\lambda+1)} \\ &= \sum m_\beta(x) \frac{\left\{ \left(1 + \delta(x)\right)^{\lambda+1} - 1 \right\}}{\lambda(\lambda+1)}. \end{aligned} \tag{9}$$

For $\lambda = -2, -1, -\frac{1}{2}, 0$ and 1, one obtains the following well-known measures: Neyman's chi-squared divided by 2,

$$\mathrm{NCS}(\mathbf{d}, \mathbf{m}_\beta) = \sum \frac{\left[d(x) - m_\beta(x) \right]^2}{2d(x)}; \tag{10}$$

Kullback–Leibler divergence,

$$\mathrm{KL}(\mathbf{d}, \mathbf{m}_\beta) = \sum m_\beta(x)\Big[\log\big(m_\beta(x)\big) - \log\big(d(x)\big)\Big]; \tag{11}$$

twice-squared Hellinger distance,

$$\mathrm{HD}(\mathbf{d}, \mathbf{m}_\beta) = 2\sum\left[\sqrt{d(x)} - \sqrt{m_\beta(x)}\right]^2; \tag{12}$$

likelihood disparity,

$$\mathrm{LD}(\mathbf{d}, \mathbf{m}_\beta) = \sum d(x)\Big[\log\big(d(x)\big) - \log\big(m_\beta(x)\big)\Big]; \tag{13}$$

Pearson's chi-squared divided by 2,

$$\mathrm{PCS}(\mathbf{d}, \mathbf{m}_\beta) = \sum \frac{\big[d(x) - m_\beta(x)\big]^2}{2m_\beta(x)}. \tag{14}$$

We now show that solving the estimating equation $\Sigma A(\delta(x))\,\nabla m_\beta(x) = 0$ corresponds to minimizing a disparity measure.

Under differentiability of the model, minimization of the disparity measure $\rho(\mathbf{d}, \mathbf{m}_\beta)$ over β corresponds to solving an estimating equation of the form

$$-\nabla\rho(\mathbf{d}, \mathbf{m}_\beta) = \sum\left[G'\big(\delta(x)\big)\frac{d(x)}{m_\beta(x)} - G\big(\delta(x)\big)\right]\nabla m_\beta(x) = 0.$$

[Here we have used the negative gradient of ρ so that in the case of the likelihood disparity (13) the equation, namely, $\Sigma d(x)u(x;\beta) = 0$, agrees with that obtained by maximizing the likelihood rather than minimizing the likelihood disparity.] This minimum disparity equation can then be written in the form of the estimating equation (5) with

$$A(\delta) := (1+\delta)G'(\delta) - G(\delta).$$

Since $A'(\delta) = (1+\delta)G''(\delta)$, second-order differentiability of G, in addition to its strict convexity, implies that $A(\delta)$ is a strictly increasing function of δ on $[-1, \infty)$. In addition, since under regularity conditions $\Sigma\,\nabla m_\beta(x) = 0$, we can redefine $A(\cdot)$ by $A(\delta) := A(\delta) - A(0)$, so that $A(0) = 0$, without changing the solutions to the estimating function above. Finally, since $A'(0) = G''(0) > 0$, we can rescale A so that it satisfies $A'(0) = 1$. Thus we have shown that, by starting with G strictly convex and thrice differentiable, minimizing the disparity measure results in an estimating equation of the form (5), with $A(\delta)$ satisfying the given assumptions.

For the converse, it is easily checked that given a differentiable increasing function $A(\delta)$ [or a nonnegative function $A'(\delta)$], one can construct a disparity measure ρ_G with residual adjustment function $A(\delta)$ by using

$$G(\delta) = \int_0^\delta \int_0^t A'(s)(1+s)^{-1}\,ds\,dt. \tag{15}$$

REMARK C. Suppose for a given $G(\delta)$ we construct $G^*(\delta) := c(G(\delta) - a\delta)$, for scalar a and positive c. Then the disparity measures display the relationship, $\rho_{G*} = c\rho_G$, and so the minimization problem is exactly the same for both measures. By suitable choice of c and a we can arrive at the standardization $A'(0) = 0$ and $A''(0) = 1$.

REMARK D. In a related approach with a different emphasis, Kemp (1986) considers a class of weighted discrepancy estimating equations of the form $\Sigma\, \delta(x)W_\beta(x) = 0$, where W is chosen for simplicity of calculation.

4. Second-order effects: estimation curvature. We now turn to the role of approximation (7), $A(\delta) \approx \delta + A_2\delta^2/2$, and show how the curvature parameter A_2 becomes a measure of the trade-off between efficiency and robustness in a second-order sense. Note that, for the power-weighted divergence family (9), $A_2 = \lambda$, so it equals 0 for maximum likelihood and -0.5 for Hellinger distance. We will show that the structure of the minimum disparity equation is such that, whenever this parameter is nonzero, we can expect the influence curve to give a poor approximation to the actual behavior of the estimator under contamination. Moreover, when it is negative, we expect the actual behavior to be more robust.

We start with a simple mathematical result regarding the second-order deficiency of the MDE's. Since we are in the context of the multinomial model, we may directly apply the concept of second-order efficiency given by Rao (1961, 1962). In this theory, the second-order efficiency E_2 of an estimator T is measured by finding the minimum asymptotic variance of $U(\beta) - a(\beta)[T - \beta] - \lambda(\beta)[T - \beta]^2$ over a and λ, where U is the score function for an iid sample. This quantity is minimized by the MLE; the minimum value is determined by the model and the parameterization, but is zero when estimating the mean value parameter of an exponential family model. Our point here is that the deficiency of an MDE is a simple function of the estimation curvature A_2 and a nonnegative quantity D depending on the model but not on $A(\delta)$.

PROPOSITION 3. *Suppose the sample space is finite ($K < \infty$). The second-order efficiency of a minimum disparity estimator with RAF $A(\delta)$ is*

$$E_2(\text{MDE}) = E_2(\text{MLE}) + A_2^2 D.$$

PROOF. Read and Cressie (1988) give this result for the PWD family (9). However, the calculations depend only on the first and second derivatives of the estimating functions with respect to δ (at $\delta = 0$), so the results are identical for any estimating functions with the same derivatives. However, all the MDE's with the same value of A_2 have the same first- and second-order derivatives at zero. [Proved also in Rao, Sinha and Subramanyan (1982).] □

The statistical significance of Rao's second-order efficiency is subject to some controversy. The interested reader will find a lively discussion in Berkson

(1980). Read and Cressie (1988) show that the MLE is not necessarily second-order optimal in the PWD family when Hodges–Lehmann deficiency is considered instead. In Appendix B we offer a geometric description of the multinomial problem which shows that, in an exponential family model, A_2^2 is a local summary measure of the departure of the estimator's contours from those of the sufficient statistics and hence is a measure of the "lack of sufficiency" of the resulting estimator. However, our main point is not that A_2 is a perfect measure of second-order information loss, but rather that it is a natural single-number summary.

The A_2 curvature is relevant in robustness as well, again at the second order. Recall the inadequacy of the bias approximation (4) that was revealed by Figure 1. Suppose that, as a simple measure of the adequacy of a first-order Taylor approximation, we consider the ratio of the second-order Taylor expansion to the first. If we apply this to $\Delta T(\varepsilon)$, we find that

$$\frac{\text{quadratic approximation}}{\text{linear approximation}} = 1 + \frac{[T''(\xi)/T'(\xi)]\varepsilon}{2}.$$

This indicates that if ε is larger than

$$\varepsilon_{\text{crit}} := \left|\frac{T'(\xi)}{T''(\xi)}\right|,$$

then the two approximations differ by 50% or more. The estimation curvature A_2 plays a pivotal role in the computation of this ratio.

PROPOSITION 4. *Let $i(\beta)$ be the Fisher information about scalar parameter β in model $\mathbf{m}_\beta$. For an estimator defined by an estimating function of the form* (5),

$$\varepsilon_{\text{crit}} = \left|i(\beta)\left[f_1(\xi) + A_2 f_2(\xi)\right]\right|^{-1}, \tag{16}$$

where

$$f_1(\xi) = 2\,\nabla u(\xi;\beta) - 2E\left[\nabla u(X,\beta)\right] + T'(\xi)E\left[\nabla^2 u(X,\beta)\right],$$

and

$$f_2(\xi) = \frac{i(\beta)}{m_\beta(\xi)} + E\left[u(X,\beta)^3\right]\frac{u(\xi;\beta)}{i(\beta)} - 2u(\xi;\beta)^3.$$

COROLLARY 5. *If the model is a one-parameter exponential family and β is the mean value parameter, then $f_1(\xi) = 0$ and so $\varepsilon_{\text{crit}} = |A_2 Q(\xi)|^{-1}$, where*

$$Q(\xi) := \frac{1}{m_\beta(\xi)} + \frac{(\xi-\beta)E(X-\beta)^3}{\left[E(X-\beta)^2\right]^2} - \frac{(\xi-\beta)^2}{\left[E(X-\beta)^2\right]}. \tag{17}$$

PROOF. These are straightforward differentiations and calculations. □

Examining the form of Q, it is clear that the leading term becomes very large for a small cell probability $m_\beta(\xi)$; this is in fact the dominant term in the calculations for Example 1. If we use the approximation $Q(\xi) \approx 1/m_\beta(\xi)$, then we obtain the result that the quadratic and linear approximations to the bias caused by point contamination will differ substantially for ε with order of magnitude of $\varepsilon^*_{\text{crit}} := m_\beta(\xi)/|A_2|$ or larger, and that if A_2 is negative, the quadratic approximation will then predict 50% less bias. However, if A_2 is positive, it predicts that much more bias.

Computation of $Q(\xi)$ for the binomial example gives some idea of the relative dampening effect of the second-derivative term for the different possible contaminations; see Table 1.

The last column of Table 1 gives $1/Q(\xi)$ and so indicates the values of ε for which the Hellinger distance quadratic approximation (using $A_2 = -0.5$) is 75% of the linear prediction. Note that, for $\xi = 12$, this calculation agrees in order of magnitude with the exact calculations shown in Figure 2.

REMARK E. We do note, however, that it is quite possible to have $A_2 = 0$, thus having second-order efficiency, but still have global characteristics that yield bias curves very similar to Figure 1. A second-order-efficient disparity measure of this type is introduced in Section 8.

REMARK F. A consequence of these calculations is that if the influence curve predicts a large influence for an observation ξ in an exponential family, it is in fact likely to have small actual influence on minimum Hellinger distance or its robust relatives. The above calculations indicate that the influence curve becomes unreliable for predicting $\Delta T(\varepsilon)$ when there is a contamination fraction $\varepsilon \geq \varepsilon_{\text{crit}} \approx \varepsilon^*_{\text{crit}} = m_\beta(\xi)/|A_2|$. At $\varepsilon^*_{\text{crit}}$, the linear approximation is

$$\Delta T(\varepsilon^*_{\text{crit}}) \approx \frac{m_\beta(\xi)T'(\xi)}{|A_2|} = m_\beta(\xi)\frac{[\xi - E(X)]}{|A_2|},$$

here using the mean value parameter. If the sample space is infinite, then as ξ goes to ∞ this last expression goes to zero (since the mean exists); that is, the linear predicted bias for contamination at a large ξ is necessarily small unless

TABLE 1
The function Q for Example 1

x or $12 - x$	$Q(x)$	$1/Q(x)$
0	4072.0	0.00025
1	324.7	0.0031
2	51.4	0.019
3	12.6	0.079
4	5.6	0.178
5	4.5	0.222
6	4.4	0.226

ε is significantly larger than $\varepsilon_{\text{crit}}$, in which case the quadratic approximation differs significantly from the linear.

5. Second-order effects: testing. In this section we turn to the role of the approximation $A(\delta) \approx \delta + A_2\delta^2/2$ in measuring the impact of model failure on testing and confidence set inference about the model $\mathbf{m}_\beta$. We will illustrate that the disparity measures can be used to create more robust testing procedures. Our focus here will be on the effect of contamination on the limiting distributions of test statistics. The main point will be that tests based on the robust disparity measures (negative A_2) have a carryover of this robustness to their limiting distribution, and that A_2 continues to serve as a local measure of this robustness. Further results about testing can be found in Basu (1993).

There are some important issues to consider in the creation of robust test procedures. If one assumes that the true density is simply a contaminated model, say,

$$t(x) = m_\varepsilon(x) := (1-\varepsilon)m_\beta(x) + \varepsilon\chi_\xi(x),$$

one could consider the underlying β to be the parameter of interest. However, since there is usually bias in estimating β, it will eventually lie outside any set of model-based confidence intervals that shrink in width as n becomes infinite. Moreover, if the true density $t(x)$ is not simply a contamination, there is no longer a unique well-defined meaning to "true β." For this reason, we suppose rather that we have chosen a functional $T(\mathbf{t})$ that defines the parameter we are interested in—it could well be chosen to have less bias than the MLE so that it more closely represents the underlying β if we have a contaminated model. [For a similar approach, see Kent (1982).] We then ask about the effect of contamination on testing and confidence procedures that concern this parameter. This analysis can be combined with that of the previous section to get an overall picture of bias and misjudged variation.

Testing and confidence methods that are asymptotically correct under any true density can be constructed for the functional $T(\mathbf{t})$ [e.g., Owen; (1988)], provided it is sufficiently regular. Instead, we consider the possible defects in using model-based methods that have optimal power properties when the model $\mathbf{m}_\beta$ is correct. The objective is to create a test procedure that is equivalent to the likelihood ratio test when the model is correct, but is more robust, in the sense of preserving size and hence confidence interval coverage, when the true density is a contaminated model. A local starting point is the likelihood ratio test for testing $H\colon \beta = \beta_0$. We can express it in terms of the likelihood disparity function (13):

$$\text{LRT} = 2n\big[\text{LD}(\mathbf{d}, \mathbf{m}_{\beta 0}) - \text{LD}(\mathbf{d}, \mathbf{m}_T)\big], \quad \text{with } T = T_{\text{ML}}(\mathbf{d}).$$

The analogous form for our class of minimum disparity estimators is then the disparity difference test statistic:

$$\text{DDT} = 2n\big[\rho(\mathbf{d}, \mathbf{m}_{\beta_0}) - \rho(\mathbf{d}, \mathbf{m}_T)\big], \quad \text{with } T = T(\mathbf{d}).$$

This is the test procedure devised by Simpson (1989) in the case of ρ equal to squared Hellinger distance. We consider its behavior as a test statistic for a null hypothesis of the form $H: T(\mathbf{t}) = \beta_0$, where $\mathbf{t}$ may or may not be in the model.

THEOREM 6. *Under the conditions cited in Appendix* A, *the following hold*:

(i) *If* $t(x)$ *is in the model, then, under the null hypothesis,*

$$\text{DDT} \to \chi^2_{\dim(\beta)}. \tag{18}$$

(ii) *Under the null hypothesis, if* $t(x)$ *is the true density and* $\dim(\beta) = 1$, *then*

$$\text{DDT} \to c(\mathbf{t})\chi^2_1,$$

where $c(\mathbf{t}) := \text{Var}_t[T'(t,X)] \cdot \nabla^2\rho(\mathbf{t}, \mathbf{m}_\beta)|_{\beta=\beta 0}$.

PROOF See Appendix A.

This theorem has the following interpretation: suppose we use (18) to construct a confidence statement about $T(\mathbf{t})$. If $\mathbf{t}$ is in the parametric model, then the procedure is asymptotically correct. If not, and we are in the scalar β case, then the interval is conservative or anticonservative depending on whether $c(\mathbf{t})$ is less than 1 or greater than 1.

For the likelihood ratio test we have the simple relationship

$$c_{\text{LD}}(\mathbf{t}) = 1 + \frac{E_t[u(X, \beta_0) + \nabla u(X, \beta_0)]}{E_t[-\nabla u(X, \beta_0)]}.$$

In Table 2 we compare these values for the Hellinger deviance test and the likelihood ratio test, using Example 1 again, with $\mathbf{t} = \mathbf{m}_\varepsilon$, the contamination being at $\xi = 12$. It can be seen that the limiting distribution is considerably more stable for the Hellinger distance.

In order to generalize these results and to find the role of A_2, we examine the effect of letting $\mathbf{t} = \mathbf{m}_\varepsilon$ and taking the derivative of $c(\mathbf{m}_\varepsilon)$ as ε goes to zero. In this manner we discover that those disparity tests whose chi-squared distribution is most stable under small departures from the model are those with negative A_2, with optimal local stability provided by $A_2 = -1$. Let $v(\xi) := u(\xi; \beta_0)^2 + \nabla u(\xi, \beta_0)$ and $\gamma := E[u(X, \beta_0)v(X)]/i(\beta_0)$.

TABLE 2
Effect of contamination on disparity tests

ε	$c(t)$ for HD	$c(t)$ for LD
0.1	1.112	2.000
0.05	1.063	1.523
0.01	1.019	1.109
0.00005	1.00026	1.00055

PROPOSITION 7. *For a minimum disparity estimator with curvature A_2 we have*

$$\left.\frac{dc(\mathbf{m}_\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = i^{-1}\left[v(\xi) - \gamma u(\xi;\beta_0)\right] + A_2 i^{-1}\left[u(\xi;\beta_0)^2 - i - \frac{u(\xi;\beta_0)E\left(u(X)^3\right)}{i}\right].$$

In particular, for a one-parameter exponential family model,

$$\left.\frac{dc(\mathbf{m}_\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = i^{-1}(1+A_2)\left[(\xi-\mu)^2 - i - \gamma(\xi-\mu)\right],$$

where $i = E(Y-\mu)^2$.

PROOF. Again, a straightforward differentiation. Since the result is invariant under reparameterization, the calculation in the exponential family case can be simplified by using the natural parameterization for the calculation. □

For the example in Table 2, we can compute $c'_{\mathrm{L_D}}(\mathbf{m}_\varepsilon) = 12$ and $c'_{\mathrm{HD}}(\mathbf{m}_\varepsilon) = 6$. Thus the first-derivative approximation works well for the bottom entry of Table 2, but does not show the extra stability generated across larger epsilon by the Hellinger distance.

REMARK G. The element of the power divergence family with estimation curvature $A_2 = \lambda = -1$ is the Kullback–Leibler divergence (11), which fails to exist if there are empty cells [$d(x) = 0$]. However, Section 7 contains some new disparity measures with $A_2 = -1$ that do not have this failing.

6. Global robustness features. Despite having an unbounded influence function in the normal model, the minimum Hellinger distance estimator behaves in finite samples as if it were a bounded influence estimator, with a nonzero breakdown point. Beran (1977) and Basu and Lindsay (1993a) have shown in examples that one point ξ in a data set can be moved to infinity without changing the estimator in an unbounded fashion. Simpson (1987) gives a breakdown lower bound in the case of minimum Hellinger distance and shows that the bound equals 50% in the Poisson model. In this section, we uncover the structural features of the residual adjustment function that lead to these robustness features.

6.1. *Outliers.* We need a formal mechanism to identify the effects of a "large" Pearson residual δ. Consider a fixed model $m_\beta(x)$ and contamination level ε. Let $\{\xi_j\colon j = 1, 2, \ldots\}$ be a sequence of elements of the sample space $\mathbf{X}$. Let $\delta_j(\cdot)$ denote the Pearson residual for the ε-contaminated data,

$$d_j(x) := (1-\varepsilon)d(x) + \varepsilon\chi_{\xi_j}(x).$$

DEFINITION 8. We will say that $\{\xi_j\}$ constitutes an outlier sequence for the model $m_\beta(x)$ and data $d(x)$ if $\delta_j(\xi_j) \to \infty$ and $d(\xi_j) \to 0$ as $j \to \infty$.

Suppose we were to add a single new observation to the data. If we were to do this at ξ_j for j sufficiently large, we could make the Pearson residual arbitrarily large for the given model. The following lemma gives a simple way to identify such a sequence, and it shows that the definition does not depend on ε.

LEMMA 9. *The sequence $\{\xi_j\}$ constitutes an outlier sequence if and only if $d(\xi_j) \to 0$ and $m_\beta(\xi_j) \to 0$ as $j \to \infty$.*

PROOF. Since $0 \le d(x) \le 1$, we have, for every j,

$$[\varepsilon/m_\beta(\xi_j)] - 1 \le \delta_j(\xi_j) \le 1/m_\beta(\xi_j). \tag{19}$$

The result now follows in an elementary fashion. □

REMARK H. Although we are here thinking of an infinite sample space, so that we can make the residuals arbitrarily large, the calculations have obvious relevance for large δ in Example 1 because there are points in the sample space with very small model probabilities.

6.2. *Outliers and disparity measures.* In order to evaluate the stability of a disparity measure $\rho(\mathbf{d}, \mathbf{m}_\beta)$ under contamination, we consider its limiting behavior under an outlier sequence ξ_j. Let

$$d_\varepsilon^*(x) = (1 - \varepsilon)d(x).$$

Although $\mathbf{d}_\varepsilon^*$ is no longer a density function, we can still formally calculate $\rho(\mathbf{d}_\varepsilon^*, \mathbf{m}_\beta)$. Note from the definition of ρ we have that

$$\rho(\mathbf{d}_\varepsilon^*, \mathbf{m}_\beta) \to \rho(\mathbf{d}, \mathbf{m}_\beta) \quad \text{as } \varepsilon \to 0, \tag{20}$$

under mild conditions of dominated convergence. Suppose that we also have the following convergence:

$$\rho(\mathbf{d_j}, \mathbf{m}_\beta) \to \rho(\mathbf{d}_\varepsilon^*, \mathbf{m}_\beta) \quad \text{as } j \to \infty. \tag{21}$$

If so, then putting together (20) and (21), for extreme outliers and small contamination fractions ε, the distance from the contaminated data $\mathbf{d_j}$ to $\mathbf{m}_\beta$ is close to that obtained by simply deleting the outlier from the sample, $\rho(\mathbf{d}, \mathbf{m}_\beta)$. There is an elementary sufficient condition on the function G in (8) that gives us convergence (21). First, we let

$$\delta_\varepsilon^*(x) := \left[d_\varepsilon^*(x)/m_\beta(x)\right] - 1. \tag{22}$$

ASSUMPTION 10. *$G(-1)$ is finite, and $G(\delta)/\delta \to 0$ as $\delta \to \infty$.*

Note that if $G(-1)$ is finite, but $G(\delta)/\delta \to c$, for some constant c, then, by Remark C, we can without loss of generality use $G^*(\delta) := G(\delta) - c\delta$, which does satisfy the definition. We restrict attention to G^* because it has the following mathematically useful property.

LEMMA 11. *Under Assumption* 10, $G(\delta)$ *is a decreasing function.*

PROOF. The convexity of G implies G' is increasing. By l'Hôpital's rule and Assumption 10, $G'(\delta) \to 0$ as $\delta \to \infty$, and so G' is negative. Hence G is a decreasing function. □

PROPOSITION 12. *If Assumption* 10 *holds, then convergence* (21) *holds for the disparity measure* ρ_G *determined by* G.

PROOF. We have $\delta_j(x) = \delta_\varepsilon^*(x) + \varepsilon\chi_{\xi_j}(x)/m_\beta(x)$, and so

$$\rho(\mathbf{d_j}, \mathbf{m}_\beta) = \sum G\big(\delta_\varepsilon^*(x)\big)\dot{m}_\beta(x) + A_j - B_j,$$

where $A_j := G(\delta_j(\xi_j))m_\beta(\xi_j)$ and $B_j := G(\delta_\varepsilon^*(\xi_j))m_\beta(\xi_j)$. The claim is proved if both A_j and B_j go to zero as $j \to \infty$. From (19), we have $m_\beta(x) \le |1/\delta(x)|$, so

$$|A_j| \le \left|G\big(\delta_j(\xi_j)\big)/\delta_j(\xi_j)\right| \to 0 \quad \text{as } j \to \infty.$$

Turning to the term B_j, since $-1 \le \delta_\varepsilon^*(\xi_j) \le \delta_j(\xi_j)$, Lemma 11 implies that $|G(\delta_\varepsilon^*(\xi_j))| \le \max[|G(-1)|, |G(\delta_j(\xi_j))|]$, so that a modification of the preceding argument shows $B_j \to 0$ as well. □

As an application, we note that the power divergence family (9) has the outlier stability property (21) provided $\lambda < 0$.

6.3. *Outlier stability of the estimating functions.* We now turn from the convergence of the disparity measure to the convergence of the corresponding estimating functions. Following the ideas of the previous section, it is natural to make the following definition.

DEFINITION 13. We will say that the residual adjustment function $A(\cdot)$ is *outlier stable* for the model $\mathbf{m}_\beta$ if the estimating functions display the following convergence under an outlier sequence:

$$\sum A\big(\delta_j(x)\big)\nabla m_\beta(x) \to \sum A\big(\delta_\varepsilon^*(x)\big)\nabla m_\beta(x), \tag{23}$$

where $\delta_\varepsilon^*(x)$ is given by (22).

The consequences of this convergence will be offered shortly; we first offer easy but statistically important sufficient conditions for outlier stability.

PROPOSITION 14. *If for some* $k > 1$, $E_\beta[|u(X;b)|^k] < \infty$ *for all* β, $A(-1)$ *is finite and* $A(\delta) = O(\delta^{(k-1)/k})$ *as* $\delta \to \infty$, *then* $A(\cdot)$ *is outlier stable for the model* $\mathbf{m}_\beta$.

COROLLARY 15. *If the model* $\mathbf{m}_\beta$ *has finite Fisher information for all* β, *then minimum Hellinger distance, and any other* $A(\delta)$ *with* $A(\delta) = O(\delta^{1/2})$ *and finite* $A(-1)$, *is outlier stable for* $\mathbf{m}_\beta$.

PROOF. We first note that the finiteness of the expectation implies that the terms in its summation representation must converge to zero, and so

$$m_\beta(\xi_j)^{(1-k)/k}|\nabla m_\beta(\xi_j)| \to 0 \quad \text{as } j \to \infty.$$

We then proceed as in the proof of Proposition 12, noting that the difference between the two sides in (23) equals the difference between $C_j := A(\delta_j(\xi_j))\,\nabla m_\beta(\xi_j)$ and $D_j := A(\delta^*(\xi_j))\,\nabla m_\beta(\xi_j)$. Since $A(\delta)$ is increasing, in both C_j and D_j the term involving $A(\cdot)$ is bounded in absolute value by the larger of $|A(-1)|$ and $|A(1/m_\beta(\xi_j))|$. Letting $t_j := 1/m_\beta(\xi_j)$, we then have both C_j and D_j bounded above in absolute value by

$$\underbrace{\{A(t_j) \vee -A(-1)\}t_j^{(1-k)/k}}_{U_j}\ \underbrace{m_\beta(\xi_j)^{(1-k)/k}|\nabla m_\beta(\xi_j)|}_{V_j},$$

where U_j is bounded and V_j converges to zero. □

As an illustration, for minimum Hellinger distance the limiting estimating function in (23) is

$$\sum A\big(\delta_\varepsilon^*(x)\big)\,\nabla m_\beta(x) = \sum 2\sqrt{(1-\varepsilon)\,d(x)/m_\beta(x)}\cdot \nabla m_\beta(x),$$

which is just $\sqrt{1-\varepsilon}$ times the function one would obtain if one ignored the outlier sequence and used the data $d(x)$. Thus the solutions to the limiting equation are exactly the same as one would obtain if the outlier were simply discarded from the data set. Since the estimating equations converge, the solutions will converge as well, provided the convergence is uniform.

In Basu and Lindsay (1993a) it was demonstrated in a numerical example that the minimum Hellinger distance estimator has the outlier stability predicted by this result. That is, if a data point from original data set is moved toward infinity, then the estimator of the mean shifts at first with the data point, but then returns to the value it would have had if the point had been deleted from the data.

6.4. *Robustness of the estimator.* There is a final, and more difficult, issue that must be addressed before a complete picture of outlier behavior can be formed. In particular, we have not addressed the discontinuity in $\Delta T_{\mathrm{HD}}(\varepsilon)$ found in Figure 1 at $\varepsilon = 0.5$. It arises because the disparity measures can have multiple local minima and so it does not suffice to consider the estimating equations in

(23) only. Instead we must return to the convergence of the disparity measures in (21) in order to investigate the conditions under which the *global minimum* $T(\mathbf{d_j})$ of $\rho(\mathbf{d}_j, \mathbf{m}_\beta)$ will converge to $T(\mathbf{d}^*_\varepsilon)$.

DEFINITION 16. The *strong breakdown point* of the minimum disparity estimator $T(\cdot)$ at the density $d(x)$ will be the supremum of those ε such that $T(\mathbf{d_j}) \to T(\mathbf{d}^*_\varepsilon)$ for any outlier sequence.

REMARK I. Since a consequence of the specified convergence is generally boundedness of the estimator sequence, having a strong breakdown point of ε^* will imply that the usual breakdown point [e.g., Hampel, Ronchetti, Rousseeuw and Stahel (1986)] is at least ε^*. Given the functional nature of our estimators, strong breakdown seems a natural mathematical object that is more informative about the actual behavior when large outliers occur.

Recall that under Assumption 10, the following convergence holds:

$$\rho(\mathbf{d}_j, \mathbf{m}_\beta) \to \rho(\mathbf{d}^*_\varepsilon, \mathbf{m}_\beta). \tag{24}$$

Thus it is plausible that, under slightly stronger conditions, the infima of the left-side terms converge to the infimum of the right. On the other hand, it is highly implausible that this convergence could hold in complete generality, as for ε near 1, one is moving over half the data to "infinity." Our result will be as predicted by Figure 1: the strong breakdown point is $\varepsilon = 0.5$ under the right conditions. We start with some basic concepts and assumptions.

ASSUMPTION 17. *It is assumed that* $\rho(\mathbf{d}_j, \mathbf{m}_\beta)$ *and* $\rho(\mathbf{d}^*_\varepsilon, \mathbf{m}_\beta)$ *are continuous in* β, *with the latter having unique minimum at* $T(\mathbf{d}^*_\varepsilon) = b^*$.

ASSUMPTION 18. *It is assumed that the convergence in* (24) *is uniform in* β *for any compact set* $\mathbf{B}$ *of parameter values containing* b^*.

It is easy to see from the proof of Proposition 12 that Assumption 18 holds if

$$\sup_{\beta \in \mathbf{B}} m_\beta(\xi_j) \to 0 \quad \text{as } j \to \infty. \tag{25}$$

This corresponds to an intuitive assumption on the model structure that, as ξ_j gets large, it becomes less and less likely to have arisen from a model distribution with β near b^*. As an example of (25), consider the Poisson model with mean β, or any other exponential family model with infinite sample space.

With the uniformity of convergence it is readily proved that any sequence $\{b^\#_j\}$ of values of β that minimize $\rho(\mathbf{d}_j, \mathbf{m}_\beta)$ over β in $\mathbf{B}$ do converge to b^*. In addition, the minimized values of ρ converge to $\rho_{\min}(\varepsilon, \mathbf{B})$, the infimum of $\rho(\mathbf{d}^*_\varepsilon, \mathbf{m}_\beta)$ over $\mathbf{B}$. In this context, the main task is to determine whether the $b^\#_j$ are eventually the global minima. Our plan of attack is to construct a lower

bound $C(\varepsilon, \mathbf{B}^c)$ on the values of $\rho(\mathbf{d}_j, \mathbf{m}_\beta)$ that is valid for all large j and for $\beta \notin \mathbf{B}$. Suppose we can do so. Then, for any ε^* satisfying

$$\rho_{\min}(\varepsilon^*, \mathbf{B}) < C(\varepsilon^*, \mathbf{B}^c), \tag{26}$$

the global minimum must eventually lie within $\mathbf{B}$ and so be $b_j^\#$, in which case breakdown does not occur for ε^*. For this plan to work, we must choose $\mathbf{B}$ so that we can construct a good lower bound C.

In order to construct this bound, we employ another assumption.

ASSUMPTION 19. *It is assumed that for each $0 < \gamma < 1$ there exists a subset $\mathbf{S}$ of the sample space such that (i) $d(\mathbf{S}) := \Sigma_{x \in \mathbf{S}} d(x) \geq 1 - \gamma$ and (ii) $\mathbf{C} := \{\beta: m_\beta(\mathbf{S}) \geq \gamma\}$ is a compact set.*

We note that such a subset $\mathbf{S}$ will exist for exponential family models: simply choose $\mathbf{S}$ to be a finite set satisfying (i), and the resulting $\mathbf{C}$ will be compact, in the mean value parameter. With this assumption, we can construct a very good lower bound on the disparity measure that has the striking feature of depending only on the function $G(\cdot)$.

LEMMA 20. *Under Assumptions 10, 17, 18 and 19, for every $\alpha > 0$ there exists a compact parameter set $\mathbf{B}_\alpha$ containing b^* such that*

$$\lim_{j \to \infty} \inf_{\beta \in \mathbf{B}_\alpha^c} \rho(\mathbf{d_j}, \mathbf{m}_\beta) \geq G(\varepsilon - 1) - \alpha.$$

PROOF Since the proof is rather long and technical, it is given in Appendix C.

Putting this lemma together with (26), we learn that there can be no breakdown for ε satisfying $G(\varepsilon - 1) - \alpha > \rho_{\min}(\varepsilon, \mathbf{B}_\alpha^c)$. If the density $d(x)$ is a model $m_{\beta_0}(x)$, corresponding to an *asymptotic strong breakdown point* when the contamination model is correct, then we can calculate $\rho_{\min}$ and so be more precise.

LEMMA 21. *For $m_\varepsilon^*(x) := (1-\varepsilon)m_{\beta_0}(x)$, the global minimum $\rho_{\min}$ of $\rho(\mathbf{m}_\varepsilon^*, \mathbf{m}_\beta)$ occurs uniquely at $b^* = \beta_0$, where it equals $G(-\varepsilon)$.*

PROOF. The lemma follows from the strict convexity of G and Jensen's inequality (the uniqueness follows from an assumption of identifiability on the model). □

This leads us to our desired conclusion:

PROPOSITION 22. *Under Assumptions 10, 17, 18 and 19, the asymptotic strong breakdown point of the minimum disparity estimators is $\frac{1}{2}$ or larger.*

PROOF. Since $b^* \in \mathbf{B}_\alpha$, from the preceding lemma $\rho_{\min}(\varepsilon, \mathbf{B}_\alpha^c) = G(-\varepsilon)$ for every α. For those ε such that $G(-\varepsilon) < G(\varepsilon - 1)$, we choose α such that $G(-\varepsilon) <$

$G(\varepsilon - 1) - \alpha$, and we apply Lemmas 20 and 21 to show that eventually the minimum over β in $\mathbf{B}_\alpha$ is smaller than for any β in $\mathbf{B}_\alpha^c$. Since G is strictly decreasing, $G(-\varepsilon) < G(\varepsilon - 1)$ holds true for all ε less than 0.5. □

7. Some new disparity measures. Having completed the robustness analysis, we can turn to the opportunities provided by these ideas.

7.1. *Blended chi-squared measures.* It is sometimes very simple to modify a distance measure to create a family of disparity measures with a wide range of estimation curvatures A_2. In a chi-squared distance, one can modify the "weights" given to the squared discrepancies. To illustrate, for α any fixed number in [0, 1] and $\overline{\alpha} := 1 - \alpha$, define the blended weight chi-squared disparity to be

$$\text{BWCS}(\alpha) = \sum \frac{[d(x) - m_\beta(x)]^2}{2[\alpha d(x) + \overline{\alpha} m_\beta(x)]}. \tag{27}$$

Here Pearson's chi-squared [see (14)] corresponds to $\alpha = 0$ and Neyman's [see (10)] corresponds to $\alpha = 1$. Le Cam [(1986), page 47] considered the case $\alpha = 0.5$, showing that it is a squared distance satisfying the triangle inequality. The corresponding residual adjustment function is

$$A_\alpha(\delta) = \frac{\delta}{1 + \alpha\delta} + \frac{\overline{\alpha}}{2}\left[\frac{\delta}{1 + \alpha\delta}\right]^2.$$

We note that it is an increasing bounded function of δ, with $A_2 = 1 - 3\alpha$.

A second weighting scheme that generalizes Hellinger distance (12) is the following:

$$\text{BWHD}_\alpha = \sum \frac{[d(x) - m_\beta(x)]^2}{2[\alpha\sqrt{d(x)} + \overline{\alpha}\sqrt{m_\beta(x)}\,]^2}.$$

This family includes NCS, $\alpha = 1$, and PCS, $\alpha = 0$, as well as HD, $\alpha = 0.5$. They generate the residual adjustment function family:

$$A_\alpha(\delta) = \delta[w(\delta)]^{-2} + \frac{\overline{\alpha}}{2}\delta^2[w(\delta)]^{-3},$$

where $w(\delta) := \alpha\sqrt{\delta + 1} + \overline{\alpha}$. This gives a family with $\sqrt{\delta}$ tails, $A_2 = 1 - 3\alpha$ and, for $\alpha = \frac{1}{3}$, an estimator for which $A'''(0) = 0$, a form of third-order efficiency. In Figure 4, the RAF $A(\delta)$ for this blended weight Hellinger distance function is shown. Note that the scale is different from Figure 3 and that, since $A'''(0) = 0$, $A(\delta) - \delta$ does not cross zero in a neighborhood of $\delta = 0$; in fact, just as in Hellinger distance, $A(\delta) \leq \delta$, so that this disparity measure has the same qualitative behavior toward residuals as Hellinger distance, even as it provides a form of third-order efficiency. Numerical experience in Basu and Lindsay

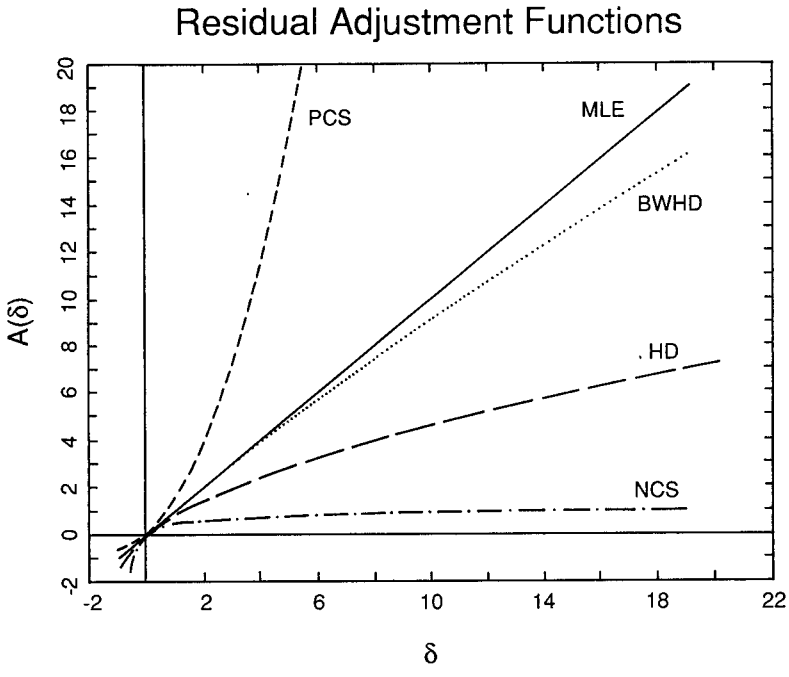

FIG. 4. *The RAF for the second-order efficient blended Hellinger distance.*

(1993a) substantiates the close agreement with maximum likelihood shown in Figure 4.

REMARK J. One of the mathematically nice features of both disparity measures BWCS and BWHD is that, unlike the likelihood disparity (13), they are bounded over all $\mathbf{d}$ and $\mathbf{m}_\beta$, so moments of all orders are guaranteed to exist. To compare these new measures with the power-weighted divergence family, for which $A_2 = \lambda$, we note that we can obtain a range of local curvature A_2, including second-order efficiency, without creating nonrobust tail behavior or having difficulties with existence when cells are empty: for $\lambda \leq -1$ the power-weighted divergences become infinite for any empty cell [$d(x) = 0$]. Further illustration of the geometric behavior of these estimators is given in Appendix B. Basu and Lindsay (1993a) use BWHD to illustrate the trade-off between robustness and efficiency in the normal model.

7.2. *Robustness against inliers.* To this point we have confined attention to the effect on the estimation procedure of shrinking large positive residuals δ, ignoring the behavior of the procedures for "inliers," that is, δ near -1. Such an analysis is relevant if we wish to determine the effect of having missing data in some cells, for instance.

Let us suppose we desire robustness against inliers and proceed by seeking

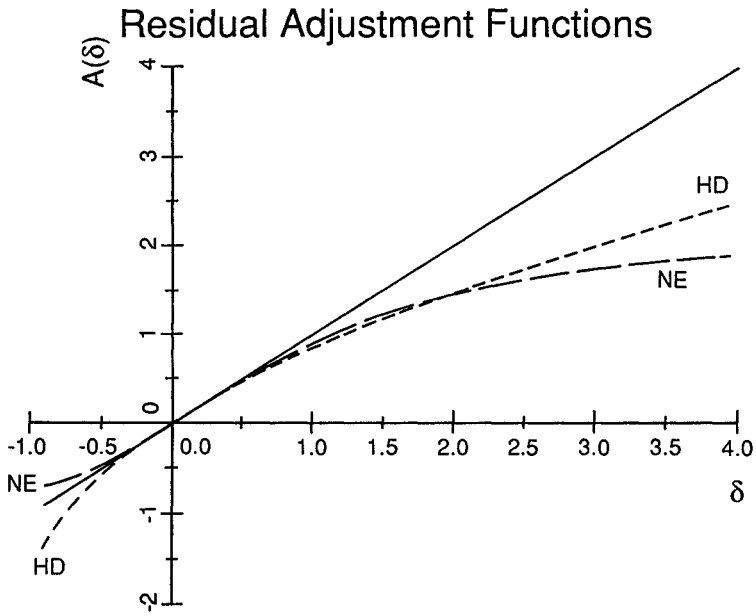

FIG. 5. *The RAF for NE and HD.*

adjustment functions that downweight both positive and negative residuals relative to maximum likelihood, in the sense that $|A(\delta)| \leq |\delta|$. Given that $A(0) = 0$ and $A'(0) = 1$, it is clear that $A(\delta)$ must cross $A_{LD}(\delta) = \delta$ at $\delta = 0$, and since $A'(0) = 1$ we must have $A''(0) = A_2 = 0$, else $A(\delta)$ would stay on one side of $A_{LD}(\delta)$ is some neighborhood of 0; that is, $A(\delta)$ must be second-order efficient. The third derivative $A'''(0)$, if not itself zero, must be negative so that the crossing occurs in the right direction. In order to investigate the implications of this, we develop such a procedure: If we start with the convex function $G(\delta) := e^{-\delta} - 1$, then we obtain the residual adjustment function

$$A_{NE}(\delta) = 2 - (2 + \delta)e^{-\delta}. \tag{28}$$

Thus the *negative exponential disparity measure* (NE) generates a bounded (by 2) RAF. Further, since,

$$A'(\delta) = (1 + \delta)e^{-\delta}, \qquad A''(\delta) = -\delta e^{-\delta} \quad \text{and} \quad A'''(\delta) = (\delta - 1)e^{-\delta},$$

the minimum NE disparity estimator generates a second-order efficient estimator that has the property that it shrinks both positive and negative residuals. These features can be seen in Figure 5, where the Hellinger distance residual adjustment function is shown for comparison.

Returning to Example 1, we find that the comparison of the negative exponential disparity and Hellinger distance follows the implications of Figure 5: the negative exponential has less bias for large positive ε and for negative ε, while

TABLE 3
Bias with winner in boldface

ε	$\Delta T_{HD}(\varepsilon)/\Delta T_{LD}(\varepsilon)$	$\Delta T_{NE}(\varepsilon)/\Delta T_{LD}(\varepsilon)$
0.005	0.47	**0.10**
0.001	0.70	**0.47**
0.0005	0.80	**0.72**
0.0001	**0.94**	0.98
−0.0001	1.08	**0.97**
−0.0002	1.21	**0.83**

Hellinger distance has less bias for very small positive ε and has worse bias than maximum likelihood for negative ε. See Table 3, where we find Hellinger distance has less bias than negative exponential disparity for only one of the selected values of ε.

We note that even though the NE disparity is second order efficient, it has less bias for contaminations above 0.0003, which is roughly the magnitude of the cell probability. Also note that the effect of negative "contaminations" is limited by the lower bound of $-m_\beta(\xi)$ for ε, as smaller values do not yield nonnegative densities.

REMARK K. Understanding the full implications of using second-order efficient residual-shrinking residual adjustment functions will require considerable empirical investigation. However, it is again feasible to construct flexible families of such procedures. For example, we can define an RAF based on the logistic function to be

$$A_\lambda(\delta) = 4\lambda^{-1}\big[\phi(\lambda\delta) - 0.5\big],$$

where $\phi(t) := e^t/(1+e^t)$, and then we get

$$A'(\delta) = 4\phi(1-\phi), \quad A''(\delta) = 4\lambda\big[\phi(1-\phi)(1-2\phi)\big],$$
$$A'''(\delta) = 4\lambda^2\phi(1-\phi)(1-6\phi+6\phi^2).$$

Thus we get a flexible family of second-order efficient estimators, with bounded residual functions and such that they approach the MLE RAF δ as λ goes to zero. The corresponding disparity measure can be found using formula (15).

8. Concluding remarks.

8.1. *Computation.* Basu and Lindsay (1993b) introduce an iteratively reweighted estimating function approach to the calculation of the minimum disparity estimators. Just as in the case of the M-estimator, the weights are a useful diagnostic to the outlier status of the observations. We offer the basic outline of the method:

Suppose that one can explicitly solve the maximum likelihood equation, as in the binomial or Poisson model. The iterative reweighting algorithm arises

from rewriting the estimating equation (5) so that it is a weighted form of the likelihood equation, with weights defined by $w^*(x) := [A(\delta(x)) - A(-1)]/[1+\delta(x)]$:

$$\begin{aligned}\sum A(\delta(x))\nabla m_\beta(x) &= \sum \left[A(\delta(x)) - A(-1)\right] \nabla m_\beta(x) \\ &= \sum w^*(x)(1+\delta(x))\,\nabla m_\beta(x) \\ &= \sum w^*(x)d(x)u(x;\beta).\end{aligned} \tag{29}$$

There is an arbitrary element to the choice of weights in the sense that $A(-1)$ in w^* can be replaced by any other constant without changing the above equalities. Since A is increasing, using $A(-1)$ forces the weights to be nonnegative. [In Basu and Lindsay (1993b), it is shown, however, that allowing negative weights can greatly increase computational efficiency.]

The algorithm is as follows: given current estimate b, create weights $w^*(x)$ and solve for (29) equal to zero with these weights fixed. For example, if $\mathbf{m}_\beta$ is the Poisson model, mean parameter β, then the algorithm gives a reweighted mean as the next value of the parameter:

$$b_{\text{new}} = \frac{\sum w^*(x)d(x)x}{\sum w^*(x)d(x)}.$$

When the algorithm has converged, the final set of weights $w^*(x)$ reflect the relative influence that the observed cells had in the final solution.

8.2. *Outlier philosophy.* It is thus clear that an outlier-stable estimating function [say, with $A(\delta)$ asymptotically $\sqrt{\delta}$] will give a large value of $\delta(x)$ a weight $w^*(x)$ near zero. Thus the philosophical evaluation of the robustness of the procedure hinges on the desirability of downweighting large Pearson residuals. These correspond to observations that are surprising, in the sense that they occur in locations ξ with small probabilities $m_\beta(\xi)$; recall Remark B. This is distinct from downweighting at those observations which are most influential, in the sense that their presence or absence causes the largest numerical change in the maximum likelihood estimator. The latter correspond roughly to large values of the score function $u(\xi;\beta)$.

In some models the concepts of surprising and influential overlap. For example, in the normal location–scale problem an outlying observation is both influential on the maximum likelihood estimates and surprising under the model. On the other hand, in a double-exponential location model, no single observation has a large influence on the maximum likelihood estimator (the median), but it is still possible to have a surprising outlying observation, as would arise, for example, if the true distribution were Cauchy. Since the idea of robustness hinges largely on stability of the parameter estimates under slight departures from the model, an approach based on downweighting data points of dubious authenticity seems quite natural.

In a final note regarding minimum Hellinger distance, we note that, while it has the nice feature of having just the right tail behavior to deal with large

outliers, it has some defects with respect to inliers. In addition to the nonrobust bias properties discussed in Section 7.2, we note that if the true distribution has zero cells, the fact that $A'(-1) = \infty$ means the influence function analysis in Proposition 1 is invalid.

APPENDIX A

Limiting distribution of the estimators. This appendix establishes the asymptotic distribution theory for the minimum disparity estimators, under some mild conditions on the residual adjustment function $A(\cdot)$, the model $m_\beta(x)$ and the true density $t(x)$. This exercise is made more challenging, and therefore worth documenting, by the infinite sample space, as we must therefore be sure convergence occurs in a summable fashion. The main tool is a device for switching the remainders from Pearson residuals to Hellinger residuals. Beyond this, many of the proofs are familiar Taylor expansion arguments in nature and so are merely sketched.

We now subscript with n the Pearson residual,

$$\delta_n(x) := \frac{d_n(x) - m_\beta(x)}{m_\beta(x)},$$

and we let $\delta_t(x) := [t(x) - m(x)]/m(x)$ be its almost-sure limit as $n \to \infty$. The first goal is to find the limiting distribution of

$$F_{1n} := \sqrt{n} \sum \left[A\left(\delta_n(x)\right) - A\left(\delta_t(x)\right)\right] \nabla m_\beta(x),$$

and the method will be to show it is asymptotically equivalent to a Taylor expanded version:

$$F_{2n} := \sqrt{n} \sum \left[\delta_n(x) - \delta_t(x)\right] A'\left(\delta_t(x)\right) \nabla m_\beta(x).$$

Once this is done, we will have proved the following theorem by an elementary application of the central limit theorem to F_{2n}.

THEOREM 23. *Under Assumptions* 24 *and* 28 *(below), and assuming that*

$$V := \operatorname{Var}\left[A'\left(\delta_t(X)\right) u(\beta; X)\right]$$

exists finitely, then

$$F_{1n} \to N(0, V).$$

In order to establish the equivalence of F_{1n} and F_{2n}, the approach will be to bound stochastically the differences in summands:

$$D_n(x) := \left[A\left(\delta_n(x)\right) - A\left(\delta_t(x)\right)\right] - \left[\delta_n(x) - \delta_t(x)\right] A'\left(\delta_t(x)\right).$$

The infinite sums involved force us to exercise some care that bounds are summable over x, and so it is advantageous to convert from Pearson residuals to Hellinger residuals, as then the techniques of Simpson (1987) can be used. Here the Hellinger residuals are

$$r_n(x) := \left[\sqrt{d_n(x)} - \sqrt{m_\beta(x)}\right] \Big/ \sqrt{m_\beta(x)}.$$

Let $r_t(x)$ be the same, but with $t(x)$ replacing $d_n(x)$. Our approach will be to establish a bound of the following type, for some constant B:

$$|D_n(x)| \leq B\left[r_n(x) - r_t(x)\right]^2. \tag{30}$$

For this purpose, the following fairly strong assumption suffices.

ASSUMPTION 24. *It is assumed that $A'(\delta)$ and $A''(\delta)\delta$ are bounded on $[-1, \infty)$.*

REMARK L. This condition is satisfied for the negative exponential disparity (28) and the blended weight chi-squared disparity measures (27), provided α is not 0 or 1. Unfortunately, the boundedness requirements fail at $\delta = -1$ for some of our measures, such as the blended Hellinger disparity measures. For these measures $A'(\delta) \to \infty$ as $\delta \to -1$ and, in particular, the influence function analysis in Proposition 1 fails if the true density satisfies $t(x) = 0$ for some elements of the sample space. We have taken the approach here of increasing the restrictions on $A(\delta)$ so as to expand the range of true distributions $\mathbf{t}$ for which the result holds.

LEMMA 25. *If $A(\delta)$ satisfies Assumption* 24, *then* (30) *holds.*

PROOF. It suffices to show that, for all nonnegative r and s,

$$\left|A(r^2 - 1) - A(s^2 - 1) - (r^2 - s^2)A'(s^2 - 1)\right| \leq B(r - s)^2.$$

We will show that a second-order Taylor series of the function within the absolute values, in r, about s, gives the result. First, the function is zero at $r = s$, and its first derivative is $2rA'(r^2 - 1) - 2rA'(s^2 - 1)$, which is zero at $r = s$. Thus all that is needed is to show that the second derivative in r is bounded in s and r. The second derivative equals $4r^2A''(r^2 - 1) + 2A'(r^2 - 1) - 2A'(s^2 - 1)$, which is bounded by Assumption 24. □

Next we need to establish some technical results which enable us to bound the magnitude of the Hellinger residuals. We here can draw on ideas from Simpson (1987). Let

$$Y_n(x) := \sqrt{n}\left[r_n(x) - r_t(x)\right]^2.$$

LEMMA 26. *For $k \in [0, 2]$, the following hold:*

(i) $E(Y_n(x)^k] \le E|\delta_n(x) - \delta_t(x)|^k n^{k/2} \le [\{t(x)[1 - t(x)]\}^{1/2}/m_\beta(x)]^k$;
(ii) $E[|\delta_n(x) - \delta_t(x)|] \le 2[t(x)(1 - t(x))]/m_\beta(x)$.

PROOF. The first inequality in part (i) follows from the string

$$\left(\sqrt{d} - \sqrt{t}\right)^2 \le |\sqrt{d} - \sqrt{t}|\left(\sqrt{d} + \sqrt{t}\right) = |d - t|.$$

The second inequality in (i) is just the Liapounov inequality for the L_k- and L_2-norms, using the fact that δ_n is an iid average of Bernoulli variates. Part (ii) follows from the triangle inequality for the absolute value function, again writing δ_n as a sum. □

LEMMA 27. $\lim_{n\to\infty} E[Y_n^p(x)] = 0$ *for* $p \in [0, 2]$.

PROOF. First, $Y_n(x)$ converges to zero in probability by elementary Slutsky results. Then, from part (i) of Lemma 26, notice that for each p there exist $k > p$, namely, $k = 2$, such that the L_k-norm of the sequence is bounded. Now use Theorem 4.5.2 of Chung (1974). □

These bounds are sufficient for our purposes, provided we make one tail assumption about the summations involving the true distribution and the model.

ASSUMPTION 28. *It is assumed that*

$$\sum t(x)^{1/2}|u(x;\beta)| < \infty.$$

LEMMA 29. *Under Assumptions* 24 *and* 28,

$$E|F_{1n} - F_{2n}| \to 0 \quad \text{as } n \to \infty,$$

and so Theorem 23 *holds.*

PROOF. From (30), we need only show that L_1-convergence to zero of $\Sigma Y_n(x)\ |\nabla m_\beta(x)|$. It suffices by the dominated convergence theorem to show that the expectation of each summand is dominated by a summable function for all n. However, by Lemma 26, $E[Y_n(x)] \le t(x)^{1/2}/m_\beta(x)$, so Assumption 28 suffices. □

ASSUMPTION 30. *Given that* $t(x)$ *is the true distribution, it is assumed that there exists a unique* β_t *that minimizes the disparity* $\rho(\mathbf{t}, \mathbf{m}_\beta)$, *and that it solves the minimum disparity equation* (5).

For the remaining results, we need to make some assumptions about the family of models. These are just a slight extension of those used for most proofs of the limiting properties of maximum likelihood estimators. We will let u_i, u_{ij}

and u_{ijk} denote the first, second and third derivatives of $\log m_\beta(x)$ with respect to the β's, where the subscripts denote the components of β involved.

ASSUMPTION 31. *It is assumed that the family of models satisfies the regularity conditions in Lehmann* [(1983), *pages* 409 *and* 429] *and, in addition, there exist* $M_{ijk}(x), M_{ij,k}(x)$ *and* $M_{i,j,k}(x)$ *that dominate is absolute value* $u_{ijk}(x;\beta)$, $u_{ij}(x;\beta)u_k(x;\beta)$ *and* $u_i(x;\beta)u_j(x;\beta)u_k(x;\beta)$, *respectively, for all* β *in a neighborhood* ω *of* β_0, *and that are uniformly bounded in expectation* E_β *for all* β *in some, possibly smaller, open neighborhood of* β_0.

ASSUMPTION 32. *It is assumed that the expectations in Assumption* 31 *also exist for the true density* $t(x)$.

THEOREM 33. *Under Assumptions* 24, 28, 30, 31 *and* 32, *there exists a consistent sequence of roots* β_n *to the minimum disparity equation. They satisfy*

$$\sqrt{n}(\beta_n - \beta_t) \to MVN\Big(0, \mathrm{Var}\big(T'(X)\big)\Big),$$

where $\mathrm{Var}(T'(X))$ *is determined by the influence function in Proposition* 1.

PROOF. To prove consistency we follow the arguments of Lehmann [(1983), page 430]. We take a Taylor series expansion of $\rho(\mathbf{d}, \mathbf{m}_\beta)$ about $\beta = \beta_t$ and show that it behaves locally like a positive definite quadratic, so it has a minimum in the neighborhood of β_t. We show in outline how one can deal with the terms of the expansion. Let subscript 0 denote evaluation at $\beta = \beta_t$.

(Linear term.) $\sum A(\delta_{n0}(x))\,\nabla m_0(x)$. We have the corresponding sum with δ_t for δ_{n0} equal to zero by definition of β_t, so it suffices to show that the difference between the sums goes to zero. Since $A'(\delta)$ is bounded (say, by C), the absolute value of this difference is bounded by

$$\sum C|\delta_{n0}(x) - \delta_t(x)|\,|\nabla m_\beta(x)|.$$

Now use the convergence to zero in expectation of the residual difference, from part (i) of Lemma 26, for each x, together with the domination bound generated by part (ii) for a proof like that of Lemma 29.

(Quadratic terms.) A term of the form $\sum\{A'(\delta_{n0})\}u_iu_jm$ has the derivative term bounded by C, so we can take the limit inside. A term of the form $\sum\{A'(\delta_{n0})\ \delta_{n0}\}u_iu_jm$ can have the limit taken inside (and hence converges to zero in L_1) by using

$$|A'(\delta_{n0})\delta_{n0} - A'(\delta_t)\delta_t| < \big|A'(\delta_{n0})(\delta_{n0} - \delta_t)\big| + \big|\delta_t\big(A'(\delta_{n0}) - A'(\delta_t)\big)\big|.$$

Other arguments go similarly.

(Cubic terms.) The arguments are very similar here except that now one must take account of the evaluation at some b^*, not β_t, where b^* depends on the data. Here we need only show that all terms converge, which provides

some simplifications. For example, consider the term $\sum A''(\delta_n^*)\delta_n^{*2}u_i^*u_j^*u_k^*m^*$. The summand will be bounded in absolute value by $|\delta_n^*(x)|M_{i,j,k}(x)m^*(x)$, which is in turn bounded by $[d_n(x)+m^*(x)]M_{i,j,k}(x)$. Hence the expectation is bounded.

This completes the outline of the consistency result. The asymptotic normality result follows from another Taylor series argument without any further difficulty in the convergence of the second-order terms. Here one uses Theorem 23 to obtain the limiting normality of the leading term. □

The limiting chi-squared distributions of Theorem 6 follows from the same standard arguments.

APPENDIX B

The geometry of estimation curvature. Our objective here is to demonstrate geometrically the role of A_2 in estimation. [For a recent review of differential geometry in statistics, see Kass (1989).] For discussing estimation curvature in the multinomial model, it suffices to consider the ordinary Euclidean geometry of the probability simplex of all possible multinomial models, which we write as $\mathsf{S} := \{\mathbf{p}\colon\ \mathbf{p}\cdot\mathbf{1} = 1, \mathbf{p} \geq \mathbf{0}\}$. If we hold β fixed (say, β_0) and consider all potential sample proportions p which have β_0 as a solution to the likelihood equation—the β_0-contour of the maximum likelihood function—we obtain the convex subset of S:

$$\mathsf{A} = \{\mathbf{p}\colon\ \mathbf{p}\cdot\mathbf{1} = 1, \mathbf{p}\cdot\mathbf{u}_{\beta_0} = 0, \mathbf{p} \geq \mathbf{0}\}.$$

Here $\mathbf{u}_{\beta_0}$ is the vectorized score function $u(x;\beta)$.

In order to picture this geometry, we let $K = 2$, so that the vectors $\mathbf{p} = (p(0), p(1), p(2))^t$ are three-dimensional. The probability simplex then lies in a two-dimensional hyperplane. In Figure 6, one-half of the simplex is shown, consisting of all probabilities $\mathbf{p}$ with $p(2) > p(0)$. As our model, we consider the binomial $(2, \beta)$ probabilities

$$\mathbf{m}_\beta = \left((1-\beta)^2, 2\beta(1-\beta), \beta^2\right).$$

These vectors lie along the dashed line in Figure 6 for β in the range 0.5 to 1. For our true value we let $\beta_0 = \frac{2}{3}$ and so the true call probabilities are $\mathbf{m}_0 := \mathbf{m}_{\beta_0} = (\frac{1}{9}, \frac{4}{9}, \frac{4}{9})$.

For this model the set A consists of all points $\mathbf{p}$ which yield $T(\mathbf{p}) = \frac{2}{3}$. In the figure, A is a vertical line segment. In fact, for this exponential family model this is all points $\mathbf{p}$ satisfying $\sum xp(x) = 2\cdot(\frac{2}{3})$.

First-order efficiency has a geometric interpretation. In Figure 6, the $\beta = \frac{2}{3}$ contours of minimum Pearson chi-square [PCS, equation (14)] and minimum Neyman chi-square [NCS, equation (10)] are shown, that is, all points $\mathbf{p}$ on the illustrated curves will yield $\beta = \frac{2}{3}$ as the value of the MDE. One can think of these contours as describing a curved projection operation from the data space to the model. The first-order efficiency of these methods corresponds to

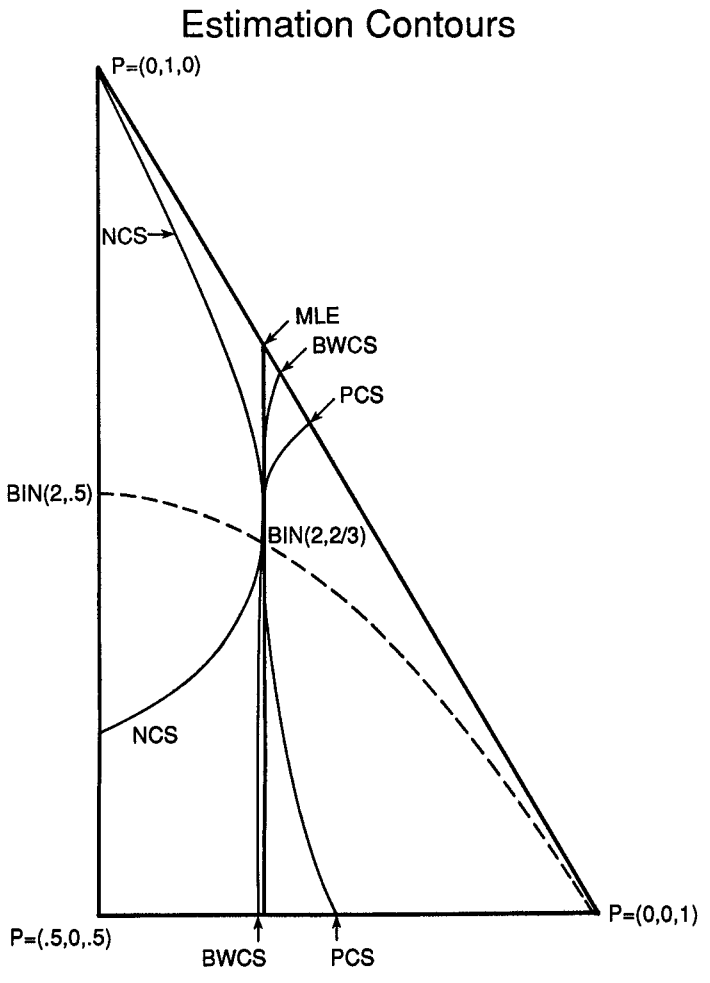

FIG. 6. *The half-simplex, with the binomial model and the contours of the MDE's based on NCS, PCS, MLE and BWCS.*

the tangency of linear set A to their contours at the model point $\mathbf{m}_{\beta_0}$ for each possible true model β_0.

The second-order efficiency of these estimators is related to the curvature of their contours in a neighborhood of $\mathbf{m}_{\beta_0}$. For minimum disparity estimators, this curvature is a function of the second derivative $A''(0) = A_2$ of the RAF: we note that the contour of the estimating function in a neighborhood of $\mathbf{m}_{\beta_0}$ is approximately that of its quadratic approximation (in δ) there. In our case we can write this quadratic approximation as

$$-\nabla\rho = \sum \left[\delta(x) + A_2\delta^2(x)/2\right] \nabla m_\beta(x).$$

Thus for our purposes A_2 represents a simple and direct local measure of the

departure of the estimating function contour from that of the maximum likelihood estimate.

In particular, if $A_2 = 0$, then our MDE contour agrees with that of maximum likelihood to second order. In Figure 6, the contour of the second-order-efficient blended weight chi-squared distance (27) is also shown (with $\alpha = \frac{1}{3}$). Regarding the interpretation of second-order efficiency, in the binomial example it should be noted that for an estimator to be a function of the minimal sufficient statistic it must have the same contours as the maximum likelihood estimator, so departures from those contours represent, in a sense, a departure from dependency on those sufficient statistics.

APPENDIX C

Proof of Lemma 20.

Choose a sample space set $\mathbf{S}$ and parameter set $\mathbf{C}$ as per Assumption 19, corresponding to some yet to be determined γ. We use $\mathbf{S}$ to bound probabilities of the likelihood ratio set $\mathbf{R} := \{x: d(x) \le \gamma m_\beta(x)\}$. For every $\beta \in \mathbf{B}_\alpha^{\mathbf{c}}$, we have

$$\begin{aligned} m_\beta(\mathbf{R}^{\mathbf{c}}) &= m_\beta(\mathbf{R}^{\mathbf{c}} \cap \mathbf{S}) + m_\beta(\mathbf{R}^{\mathbf{c}} \cap \mathbf{S}^{\mathbf{c}}) \\ &\le m_\beta(\mathbf{S}) + d(\mathbf{S}^{\mathbf{c}})/\gamma \le \gamma + \gamma^2/\gamma = 2\gamma. \end{aligned} \tag{31}$$

Another inequality we shall need is

$$d(\mathbf{R}) = d(\mathbf{R} \cap \mathbf{S}) + d(\mathbf{R} \cap \mathbf{S}^{\mathbf{c}}) \le \gamma m_\beta(\mathbf{S}) + d(\mathbf{S}^{\mathbf{c}}) \le 2\gamma. \tag{32}$$

Next, we compute the disparity measure by performing the sample space summation separately over the three disjoint sets:

$$\mathbf{E}_1 := \mathbf{R}\backslash\{\xi_j\}; \quad \mathbf{E}_2 := \{\xi_j\}; \quad \mathbf{E}_3 := \mathbf{R}^{\mathbf{c}}\backslash\{\xi_j\}.$$

The monotonicity of G and the construction of $\mathbf{R}$ gives

$$\mathrm{S1} := \sum_{\mathbf{E}_1} m_\beta(x) G\big(\delta(x)\big) \ge m_\beta(\mathbf{E}_1) G\big((1-\varepsilon)\gamma - 1\big). \tag{33}$$

Let S2 be the single term $m_\beta(\xi_j)G(\delta(\xi_j))$. Finally, for the third summation we get a lower bound using convexity of G and Jensen's inequality:

$$\mathrm{S3} := \sum_{\mathbf{E}_3} m_\beta(x) G\big(\delta(x)\big) \ge m_\beta(\mathbf{E}_3) G\Big(\frac{(1-\varepsilon)d(\mathbf{E}_3)}{m_\beta(\mathbf{E}_3)} - 1\Big).$$

We claim that this last lower bound for S3 can be made as close to zero as desired: since we assume that $G(\delta)/\delta \to 0$ as $\delta \to \infty$, it suffices to show that, for j sufficiently large (say, $j > J$) and for γ sufficiently small (say, $\gamma < \Gamma$), that we can force $m_\beta(\mathbf{E}_3)$ to be as small as we like, without $d(\mathbf{E}_3)$ also getting small. Since $d(\xi_j) \to 0$ and $d(\mathbf{E}_3) = d(\mathbf{R}^{\mathbf{c}}\backslash\{\xi_j\})$, upper bound (32) gives the latter. For the former, we can apply (31). Thus, we can require that $\mathrm{S3} > -\alpha/2$.

Let $M := \mathbf{m}_\beta(E_1 \cup \{\xi_j\})$. From Jensen's inequality and the lower bound (33) for S1, we have

$$\text{S1} + \text{S2} \geq MG\Big(M^{-1}\big[(1-\varepsilon)\gamma m(\mathbf{E_1}) + (1-\varepsilon)d(\xi_j) + \varepsilon\big] - 1\Big). \tag{34}$$

Since with γ small, M can be made arbitrarily close to 1 by (31), γ itself can be made arbitrarily small and $d(\xi_j)$ goes to zero, we can make G's arguments in (34) as close to $\varepsilon - 1$ as we like by choosing γ small and j big. The continuity of G then indicates that we can choose Γ and J such that, for $\gamma < \Gamma$ and $j > J$, $\text{S1} + \text{S2} + \text{S3} \geq G(\varepsilon - 1) - \alpha$.

Finally, we note that we can always include b^* in $\mathbf{B}_\alpha$ without changing the lower bound. □

Acknowledgments. I would like to thank Ayan Basu and Marianthi Markatou for their helpful comments, as well as those involved in the editorial process.

REFERENCES

BASU, A. (1993). Minimum disparity estimation: applications to robust tests of hypotheses. Technical report, Univ. Texas at Austin.

BASU, A. and LINDSAY, B. G. (1993a). Minimum disparity estimation for continuous models: efficiency, distributions, and robustness. Technical report, Univ. Texas at Austin.

BASU, A. and LINDSAY, B. G. (1993b). The iteratively reweighted estimating equation in minimum distance problems. Technical report, Univ. Texas at Austin.

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

BERKSON, J. (1980). Minimum chi-square, not maximum likelihood! (with discussion). *Ann. Statist.* **8** 457–487.

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

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

CSISZÁR, I. (1963). Eine informations theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.* **3** 85–107.

DONOHO, D. L. and LIU, R. C. (1988). The "automatic" robustness of minimum distance functionals. *Ann. Statist.* **16** 552–586.

FERNHOLZ, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Springer, New York.

GELFAND, A. E. and DEY, D. K. (1991). On measuring Bayesian robustness of contaminated classes of priors. *Statist. Decisions* **9** 63–80.

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

KASS, R. E. (1989). The geometry of asymptotic inference. *Statist. Sci.* **4** 188–219.

KEMP, A. W. (1986). Weighted discrepancies and maximum likelihood estimation for discrete distributions. *Comm. Statist. Theory Methods* **15** 783–803.

KENT, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69** 19–27.

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

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

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

RAO, C. R. (1961). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 531–546. Univ. California Press, Berkeley.

RAO, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **24** 46–72.

RAO, C. R., SINHA, B. K. and SUBRAMANYAN, K. (1983). Third order efficiency of the maximum likelihood estimator in the multinomial distribution. *Statist. Decisions* **1** 1–16.

READ, T. R. C. and CRESSIE, N. A. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer, New York.

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.

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

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
422 CLASSROOM BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802