Sankhyā : The Indian Journal of Statistics
1997, Volume 59, Series B, Pt. 1, pp. 11-27

ROBUST MINIMUM DIVERGENCE PROCEDURES FOR COUNT DATA MODELS

By AYANENDRANATH BASU
SRABASHI BASU
and
GOPAL CHAUDHURI
Indian Statistical Institute, Calcutta

SUMMARY. Simpson (1987) considered minimum Hellinger distance estimation in count data models. Unlike many other robust estimators, the minimum Hellinger distance estimator is simultaneously robust and first order efficient. In particular Simpson provides appealing arguments for the robustness of the minimum Hellinger distance estimator, as well as attractive breakdown results for it. In this paper we show that Simpson's arguments for the Hellinger distance can be extended to a particular subclass of the Cressie-Read (Cressie and Read 1984) family of divergences where the corresponding estimators enjoy similar breakdown properties, and the estimating equations have a very simple weighted likelihood interpretation providing a nice diagnostic tool. The results of Lindsay (1994) provide further justification of the robustness of the estimators. Some numerical results are provided to illustrate the possible improvements in the performance of the estimators and the corresponding test statistics when an empty cell penalty is appropriately applied.

1. Introduction

Minimum Hellinger distance and related methods have been the source of considerable interest in recent statistical literature (Beran 1977; Tamura and Boos 1986; Simpson 1987, 1989a, 1989b; Eslinger and Woodward 1991; Lindsay 1994; Basu and Lindsay 1994; Basu and Harris 1994; Basu and Sarkar 1994a,b; Markatou *et al.* 1997). While such methods require a nonparametric estimate of the true density, it is relatively simple in count data models since one can use the 'empirical' density for this estimate (Simpson 1987).

In this paper we identify a particular family of minimum divergence methods which generalizes the estimating equation of the minimum Hellinger distance estimator in such a way that a single parameter controls the contribution of the

Paper received. March 1996.

AMS (1991) *subject classifications.* Primary 62F10, 62F35; secondary 62F04.

Key words and phrases: Hellinger distance; asymptotic efficiency; influence function; breakdown point; disparity; residual adjustment function; empty cells.

model component and the data component in the estimating equation – and hence the degree of robustness of the corresponding estimator. It turns out that this family is a particular subclass of the Cressie-Read family of divergences. The methods are all first order efficient, and the breakdown results of Simpson (1987) presented in the context of the minimum Hellinger distance estimator can actually be extended to this family of divergences.

One of the reasons for the continued interest in the minimum Hellinger distance and related methods is that many of these procedures combine the property of robustness with asymptotic efficiency – unlike the robust M- estimators. Although asymptotically equivalent to maximum likelihood at the model, some of these more robust minimum divergence procedures may have substantially inferior performance compared to the former in small samples, limiting their practical applicability. In this paper we also discuss a correction for these procedures which has been observed to recover much of their lost efficiency in small samples.

The rest of the paper is organized as follows. The generalized Hellinger divergence is introduced in Section 2, and the weighted likelihood approach is discussed in Section 3. Sections 4 and 5 study the influence function and the breakdown properties of the estimators respectively. An example is considered in Section 6, and the connection of our procedures with other methods is investigated in Section 7. The empty cell penalty is the subject of Section 8, and tests of hypotheses are discussed in Section 9.

2. The Generalized Hellinger Divergence Family

Suppose that we have a random sample $X_1, X_2, \ldots, X_n$ from a parametric model $\{F_\theta; \theta \in \Omega\}$, where Ω is a subset of R^p; assume that the family of distributions $\{F_\theta\}$ is dominated, and that f_θ represents the corresponding density for F_θ. In density-based minimum divergence estimation one minimizes an appropriate measure of discrepancy between a nonparametric density estimate $d(x)$ obtained from the data and the model density $f_\theta(x)$. For count data models, where the sample space $\mathcal{X} = \{0, 1, \ldots\}$, one can take $d(x)$ to be the empirical density function

$$d(x) = \frac{N_x}{n},$$

where N_x is the frequency of x among $X_1, \ldots, X_n$. The minimum Hellinger distance estimator ($MHDE$) of θ minimizes

$$\sum_x (d^{1/2}(x) - f_\theta^{1/2}(x))^2 = 2(1 - \sum_x d^{1/2}(x) f_\theta^{1/2}(x)). \qquad \ldots(2.1)$$

Alternatively therefore the $MHDE$ maximizes $\phi_{n,\theta} = \sum_x d^{1/2}(x) f_\theta^{1/2}(x)$. Let ∇ represent the gradient with respect to θ, and $u_\theta(x) = \nabla f_\theta(x)/f_\theta(x)$ represent the *score function*, the gradient of $log f_\theta(x)$. The corresponding (standardized) estimating equation has the form

$$\phi_{n,\theta}^{-1}\sum_x d^{1/2}(x)f_\theta^{1/2}(x)u_\theta(x) = 0, \qquad \ldots(2.2)$$

as opposed to the maximum likelihood estimating equation

$$\sum_x d(x)u_\theta(x) = 0. \qquad \ldots(2.3)$$

A moments reflection shows that while the estimating equations agree in the limit under the model, the effect of a large deviation from the model at a point x is severely downweighted in (2.2), since the expectation in (2.2) is with respect to the density $\phi_{n,\theta}^{-1}d^{1/2}(x)f_\theta^{1/2}(x)$, unlike (2.3) where the expectation is with respect to $d(x)$. This appealing argument for the robustness of the $MHDE$ relative to the maximum likelihood estimator (MLE) was presented by Simpson (1987).

In this paper we consider generalizations of the above idea in defining a class of robust estimators for count data models. Consider the family of divergences

$$D_\alpha(d, f_\theta) = K_\alpha(1 - \sum_x d^\alpha(x)f_\theta^{1-\alpha}(x)), \quad \alpha \in (0,1), \qquad \ldots(2.4)$$

where K_α is an appropriate nonnegative standardizing constant to be considered later. Note that for $\alpha \in (0,1)$, the minimizer of $D_\alpha(d, f_\theta)$ does not depend on the value of K_α. The minimum Hellinger distance estimator corresponds to $\alpha = 1/2$. We denote the family defined in (2.4) as the generalized Hellinger divergence family, indexed by the parameter α. Let $\phi_{n,\theta,\alpha} = \sum_x d^\alpha(x)f_\theta^{1-\alpha}(x)$; then $\hat{\theta}_{n,\alpha}$, the *minimum generalized Hellinger divergence* estimator with index α ($MGHD_\alpha$ estimator) maximizes $\phi_{n,\theta,\alpha}$ over $\theta \in \Omega$. It can be easily verified that the estimating equation of the $MGHD_\alpha$ estimator has the form

$$\sum_x d^\alpha(x)f_\theta^{1-\alpha}(x)u_\theta(x) = 0, \qquad \ldots(2.5)$$

so that the smaller the value of α, higher is the degree of downweighting applied to an observation inconsistent with the model – for values of α smaller than 1/2 a stronger downweighting effect relative to the $MHDE$ is exerted on such observations. The limiting case $\alpha = 1$ actually corresponds to maximum likelihood.

3. **Weighted Likelihood Equation: Diagnostics**

Consider the generalized Hellinger estimating equation given in (2.5), which we can rewrite as

$$\sum_{x:d(x)\neq 0} d^\alpha(x)f_\theta^{1-\alpha}(x)u_\theta(x) = 0. \qquad \ldots(3.1)$$

By writing $d^\alpha(x)f_\theta^{1-\alpha}(x) = (f_\theta(x)/d(x))^{1-\alpha}d(x) = w(x)d(x)$, and by rewriting the sum over x as a sum over i, (3.1) can be expressed in the equivalent weighted likelihood equation form

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)u_\theta(X_i) = 0, \qquad \ldots(3.2)$$

where $w(X_i) = (f_\theta(X_i)/d(X_i))^{1-\alpha}$. For a cell with an unusually large frequency relative to the model, larger downweighting will be provided for smaller values of α. One can then use the final values of the fitted weights as diagnostics for the aberrant cells. On the other hand, the weights all converge to 1 as $n \to \infty$ when the model is correct. Thus the weighted likelihood estimating equation has the property that it downweights observations discrepant with the model, but asymptotically behaves like the maximum likelihood score equation when the model is correct. Other weighted likelihood estimating equations having the same property have been discussed by Markatou *et al.* (1997), where the weights are based on the residual adjustment functions defined by Lindsay (1994).

The weighted likelihood estimating equation (3.2) can itself be solved by using an iterative reweighting algorithm similar to the iteratively reweighted least squares. One can start with an initial approximation to the unknown θ, and create the weights w; the next approximation to θ can then be obtained by solving the estimating equation (3.2) treating the weights as fixed constants. This process can be continued till convergence.

4. Influence Function and Standard Errors

In this section we determine the influence function of the $MGHD_\alpha$ estimator for the true density given by $g(x)$. For $\xi \in \mathcal{X}$, let $\chi_\xi(x)$ represent the indicator function for ξ. We denote the ϵ contaminated version of the density g as

$$g_\epsilon(x) = (1-\epsilon)g(x) + \epsilon\chi_\xi(x).$$

Let G and G_ϵ represent the distribution functions corresponding to g and g_ϵ. For a functional $T(\cdot)$ defined on the space of distributions, we define its influence function by

$$T'(\xi) = \frac{\partial T(G_\epsilon)}{\partial\epsilon}\Big|_{\epsilon=0}.$$

The functional T is Fisher consistent if $T(F_\theta) = \theta$. From their definition it is clear that the $MGHD_\alpha$ functionals are Fisher consistent.

A straightforward differentiation of the estimating equation of the $MGHD_\alpha$ estimators establishes the following result.

THEOREM 4.1. *For the $MGHD_\alpha$ functional $T_\alpha(\cdot)$, the influence functions defined above has the form $T'_\alpha(\xi) = J_\alpha^{-1} D_\alpha$, where*

$$\begin{aligned} D_\alpha &= \alpha u_{\theta^*}(\xi) g^{\alpha-1}(\xi) f_{\theta^*}^{1-\alpha}(\xi), \\ J_\alpha &= -[(1-\alpha)\sum_x g^\alpha(x) f_{\theta^*}^{1-\alpha}(x) u_{\theta^*}(x) u_{\theta^*}^t(x) + \sum_x g^\alpha(x) f_{\theta^*}^{1-\alpha}(x) u'_{\theta^*}(x)], \end{aligned}$$

and $\theta^* = T_\alpha(G)$ is the $MGHD_\alpha$ functional at G. D_α is a $p \times 1$ vector, and J_α is a $p \times p$ matrix. Also, $u'_{\theta^*}(x)$ is the $p \times p$ matrix of second partial derivatives of $log f_\theta(x)$ evaluated at $\theta = \theta^*$.

As an immediate corollary of the above result we note that if G is a model point F_θ, then $T_\alpha(G) = \theta$ for all $\alpha \in (0,1)$, and the influence function of the $MGHD_\alpha$ functional reduces to $I^{-1}(\theta) u_\theta(\xi)$, $I(\theta)$ being the Fisher information about θ; as this is identical to the influence function of the maximum likelihood estimator, the above result suggests that the $MGHD_\alpha$ estimators are asymptotically fully efficient at the model.

On the other hand, being equal to the influence function of the maximum likelihood estimators, the influence function of the $MGHD_\alpha$ estimators are potentially unbounded. Several authors (Beran 1977; Tamura and Boos 1986; Simpson 1987; Lindsay 1994) have shown that the $MHDE$ and some other density-based minimum divergence estimators have strong robustness properties in spite of this, exhibiting the limitation of the influence function approach in this case. Lindsay has argued that the first order approximation of the bias of the estimator under contamination can be very inaccurate for some minimum distance estimators, and higher order approximations are more appropriate. This in fact turns out to be the case for the $MGHD_\alpha$ estimators as well; this will be further discussed in Section 7.

Although they are not useful for assessing the robustness of the estimators, the influence functions may be used for the estimation of the standard errors of the estimators. Note that by using the above form of the influence function, we have an explicit formula for the asymptotic variance of the estimators, as a function of the true unknown density $g(x)$. This can then be consistently estimated by using the empirical density $d(x)$ in place of $g(x)$. For a random sample $X_1, \ldots, X_n$, let $v_{i,\alpha}$ be the quantity D_α in the above theorem evaluated at $\xi = X_i$ and $g(\cdot) = d(\cdot)$, with θ^* being the $MGHD_\alpha$ estimator; let $\hat{J}_\alpha$ represent the corresponding estimate of the matrix J_α in Theorem 4.1. Also let $\hat{V}_\alpha$ be the $p \times p$ matrix

$$\frac{1}{n-1} \sum_{i=1}^{n} v_{i,\alpha} v_{i,\alpha}^t.$$

Then the standard error of ($\sqrt{n}$ times) the $MGHD_\alpha$ estimators can be estimated as $\hat{J}_\alpha^{-1} \hat{V}_\alpha \hat{J}_\alpha^{-1}$.

5. Breakdown Results

One of the standard measures of the robustness of an estimator is the breakdown point of the estimator when the true distribution is contaminated by some arbitrary distribution (Donoho and Huber 1983). For the $MGHD_\alpha$ estimator, we obtain a general lower bound for the breakdown point at a distribution G. Following Simpson (1987), we consider the contaminated model

$$H_n = (1-\epsilon)G + \epsilon K_n, \qquad \ldots(5.1)$$

where $\{K_n\}$ represents a sequence of contaminating distributions. For any two distributions H and K with densities h and k, we define $\phi_\alpha(H,K) = \sum_x h^\alpha(x)k^{1-\alpha}(x)$. Let $\hat{\phi}_\alpha = max\{\phi_\alpha(G,F_t), t \in \Omega\}$ and suppose that the maximum occurs in the interior of Ω; let $\phi^*_\alpha = lim_{M\to\infty} sup_{|t|>M}\ \phi_\alpha(G,F_t)$.

THEOREM 5.1. *Let*

$$\epsilon < \frac{(\hat{\phi}_\alpha - \phi^*_\alpha)^{1/\alpha}}{1+(\hat{\phi}_\alpha - \phi^*_\alpha)^{1/\alpha}}.$$

Then there exists no sequence of the form (5.1), for which $|T_\alpha(H_n) - T_\alpha(G)| \to \infty$, *as* $n \to \infty$, *where* T_α *represents the* $MGHD_\alpha$ *functional.*

PROOF. Note that

$$\phi_\alpha(H_n, F_\theta) \geq (1-\epsilon)^\alpha \hat{\phi},$$

and

$$\phi_\alpha(H_n, F_{\theta_n}) \leq (1-\epsilon)^\alpha \phi_\alpha(G, F_{\theta_n}) + \epsilon^\alpha,$$

where $\theta = T_\alpha(G)$ and $\theta_n = T_\alpha(H_n)$. The result is then a simple modification of Simpson (1987, Theorem 3). □

When the true distribution is *Poisson* and the model $\{F_\theta : \theta \in \Omega\}$ is the *Poisson* family, then $\hat{\phi}_\alpha = 1$ and $\phi^*_\alpha = 0$, so that the asymptotic breakdown point of the $MGHD_\alpha$ estimator is 1/2 for all $\alpha \in (0,1)$.

6. An Example

In this section we consider a chemical mutagenicity data analyzed previously by Simpson (1987). Details of the experimental protocol are given in Woodruff *et al.* (1984). In the sex-linked recessive lethal test in drosophila (fruit flies), groups of male flies are exposed to different doses of a chemical to be screened. Each male is then mated with unexposed females. Sampling 100 daughters from each male (roughly) one observes the number of daughter flies carrying a recessive lethal mutation on the X chromosome. One then looks at the frequencies of frequencies, of males having 0, 1, 2, ..., recessive lethal daughters. See Simpson and Woodruff *et al.* for more details of the experiment.

TABLE 1. THE DROSOPHILA DATA, AND ESTIMATES OF THE MEAN PARAMETER UNDER A *POISSON* MODEL USING MAXIMUM LIKELIHOOD ESTIMATION (ML), AND MAXIMUM LIKELIHOOD WITH OUTLIER DELETION ($ML + D$).

	Recessive lethal count						$\hat{\theta}$	
Day	0	1	2	3	4	≥ 5	ML	$ML + D$
27	25	4	0	0	0	0	0.138	0.138
28	23	3	0	1	1	0	0.357	0.115
177	23	7	3	0	0	1(91)	3.059	0.394

We consider three particular experimental runs, those on days 27, 28 and the second run of day 177 (for our purpose we will refer to them as the first, second and the third experimental runs). There is one exceptionally large count in the third experimental run, where one male is reported to have produced 91 recessive lethal daughters. For each run *Poisson* models are fit to the data using $MGHD_\alpha$ estimation. The data are presented in Table 1, together with the MLE of the mean, as well as the maximum likelihood estimator from the cleaned data after an outlier detection scheme (Simpson 1987, Section 5) is applied to detect and delete outliers. For the second run the two largest observations (3 and 4), and for the third run the largest observation, 91, are thus deleted.

TABLE 2. ESTIMATES OF THE MEAN PARAMETER UNDER A *POISSON* MODEL USING $MGHD_\alpha$ AND $MPGHD_\alpha$ ESTIMATION FOR THE DROSOPHILA DATA

	Day 27		Day 28		Day 177 (2)	
α	$MGHD_\alpha$	$MPGHD_\alpha$	$MGHD_\alpha$	$MPGHD_\alpha$	$MGHD_\alpha$	$MPGHD_\alpha$
0.10	0.0741	0.1369	0.0677	0.1222	0.2763	0.3737
0.20	0.0964	0.1370	0.0879	0.1245	0.3173	0.3760
0.30	0.1089	0.1372	0.1008	0.1280	0.3390	0.3783
0.40	0.1172	0.1373	0.1121	0.1338	0.3532	0.3806
0.50	0.1230	0.1374	0.1252	0.1437	0.3637	0.3829
0.60	0.1275	0.1375	0.1446	0.1614	0.3719	0.3851
0.70	0.1309	0.1376	0.1767	0.1926	0.3786	0.3874
0.80	0.1337	0.1377	0.2271	0.2403	0.3844	0.3896
0.90	0.1360	0.1378	0.2912	0.2986	0.3894	0.3918
0.99	0.1377	0.1379	0.3507	0.3515	0.4481	0.3937
0.999	0.1379	0.1379	0.3565	0.3566	2.5115	2.5122
0.9999	0.1379	0.1379	0.3571	0.3571	3.0025	3.0026
0.99999	0.1379	0.1379	0.3571	0.3571	3.0532	3.0532
0.999999	0.1379	0.1379	0.3571	0.3571	3.0582	3.0583

In Table 2, the first column for each day represents the $MGHD_\alpha$ estimates. Note that the estimator corresponding to $\alpha = 1/2$ represents the $MHDE$. The following points deserve mention: a) for all the experimental runs, the $MGHD_\alpha$

estimates converge to the MLE as $\alpha \to 1$; b) in the third experimental run, the large count of 91 fails to corrupt the procedure even for an α as high as 0.99, beyond which there is an apparent breakdown in the process (recall the breakdown result in Section 5); c) all the estimates are fairly close to the outlier deleted maximum likelihood estimator in the first experimental run; d) in the second run, where the two largest values are much more difficult to diagnose as outliers there is a more gradual deterioration in the estimate compared to the sharp breakdown observed in the third run; e) the estimates corresponding to very small values of α, while clearly resistant against outliers, are somewhat different from the outlier deleted maximum likelihood estimator, or even the $MHDE$, indicating perhaps a larger variance and/or bias of the estimators in finite samples under the model.

In the context of these data it can therefore be said that the $MGHD_\alpha$ estimates have exhibited strong outlier resistance property for values of α in the range between 0 to approximately 0.7, with smaller α leading to greater robustness. However, it will clearly be helpful if there is a correction possible which makes the estimates corresponding to smaller values of α closer to maximum likelihood when the data roughly follow the model, *without* compromising the robustness properties of these estimators. We will introduce this in Section 8, where we will discuss the rest of Table 1.

7. Connection with other Minimum Distance Methods

Consider the form of the generalized Hellinger divergence defined in equation (2.4). A little algebra shows that one can write this divergence equivalently as

$$-K_\alpha \sum_x d(x) \left[\left(\frac{d(x)}{f_\theta(x)} \right)^{\alpha-1} - 1 \right].$$

Comparing with the power divergence family of Cressie and Read (1984) given by

$$I^\lambda(p,q) = \frac{1}{\lambda(1+\lambda)} \sum_i p_i \left[\left(\frac{p_i}{q_i} \right)^\lambda - 1 \right],$$

where p and q are two specific densities, we see that the generalized Hellinger divergences represent a subclass of the Cressie-Read divergence, with $\alpha = \lambda + 1$, and $k_\alpha = [\alpha(1-\alpha)]^{-1}$; since $\alpha \in (0,1)$, this essentially represents the Cressie-Read family restricted to $\lambda \in (-1,0)$. See also Simpson (1989b).

Harris and Basu (1995) have shown that one can write the Cressie-Read divergence in the equivalent form

$$\sum_i \left\{ \frac{p_i}{\lambda(1+\lambda)} \left[\left(\frac{p_i}{q_i} \right)^\lambda - 1 \right] + \frac{1}{1+\lambda}(q_i - p_i) \right\},$$

the advantage of this form being that each term in this sum is nonnegative, unlike the Cressie-Read form. The second term in the above expression sums to zero, so does not change the divergence. We can therefore write our generalized Hellinger divergence in the form

$$D_\alpha(d, f_\theta) = \sum_x \left\{ \frac{d(x)}{\alpha(\alpha-1)} \left[\left(\frac{d(x)}{f_\theta(x)} \right)^{\alpha-1} - 1 \right] + \frac{(f_\theta(x) - d(x))}{\alpha} \right\}, \ \alpha \in (0,1). \quad \ldots(7.1)$$

For $\alpha = 1/2$, the corresponding divergence is $2\sum_x (d^{1/2}(x) - f_\theta^{1/2}(x))^2$, which is minimized by the *MHDE*. Also, by using L'Hospital's rule, and taking the limit as $\alpha \rightarrow 0$ or 1, the limiting divergences are

$$D_0(d, f_\theta) = \sum_x [f_\theta(x) log(f_\theta(x)/d(x)) + (d(x) - f_\theta(x)],$$

$$D_1(d, f_\theta) = \sum_x [d(x) log(d(x)/f_\theta(x)) + (f_\theta(x) - d(x))].$$

Note that the minimizer of $D_1(d, f_\theta)$ is the maximum likelihood estimator of θ.

Lindsay (1994) has considered a subclass of minimum divergence estimators called *minimum disparity estimators*, where a typical disparity ρ, some measure of discrepancy between the empirical density $d(x)$ and the model density $f_\theta(x)$, has the form

$$\rho(d, f_\theta) = \sum_x G(\delta(x)) f_\theta(x), \quad \delta(x) = \frac{d(x)}{f_\theta(x)} - 1, \quad \ldots(7.2)$$

where the function $G(\delta)$ is convex, thrice differentiable, and satisfies $G(0) = 0$. The class of Cressie-Read divergences belongs to the class of disparities, and consequently so does the class of generalized Hellinger divergences, with the corresponding $G(\delta)$ functions having the form

$$G(\delta) = \frac{(\delta+1)^\alpha - (\delta+1)}{\alpha(\alpha-1)} - \frac{\delta}{\alpha}.$$

For each minimum disparity estimator, there is an associated estimating equation of the form

$$\sum_x A(\delta(x)) \nabla f_\theta(x) = 0, \quad \ldots(7.3)$$

where the function $A(\cdot)$, typical to the disparity in question, satisfies $A(0) = 0$, and $A'(0) = 1$. In particular for the maximum likelihood estimator, $A(\delta) = \delta$, and for the minimum Hellinger distance estimator $A(\delta) = 2[(\delta+1)^{1/2} - 1]$. Since the estimating equations are otherwise identical, the function $A(\cdot)$, called the residual adjustment function of the disparity by Lindsay, controls the theoretical properties of the corresponding estimator. As large outliers correspond to large positive values of δ, one would expect the residual adjustment functions of the more robust minimum disparity estimators to shrink the effect of such residuals closer to zero relative to maximum likelihood. The residual adjustment function of the generalized Hellinger divergence with index α has the form

$$A(\delta) = \frac{(\delta+1)^{\alpha}}{\alpha} - \frac{1}{\alpha}.$$

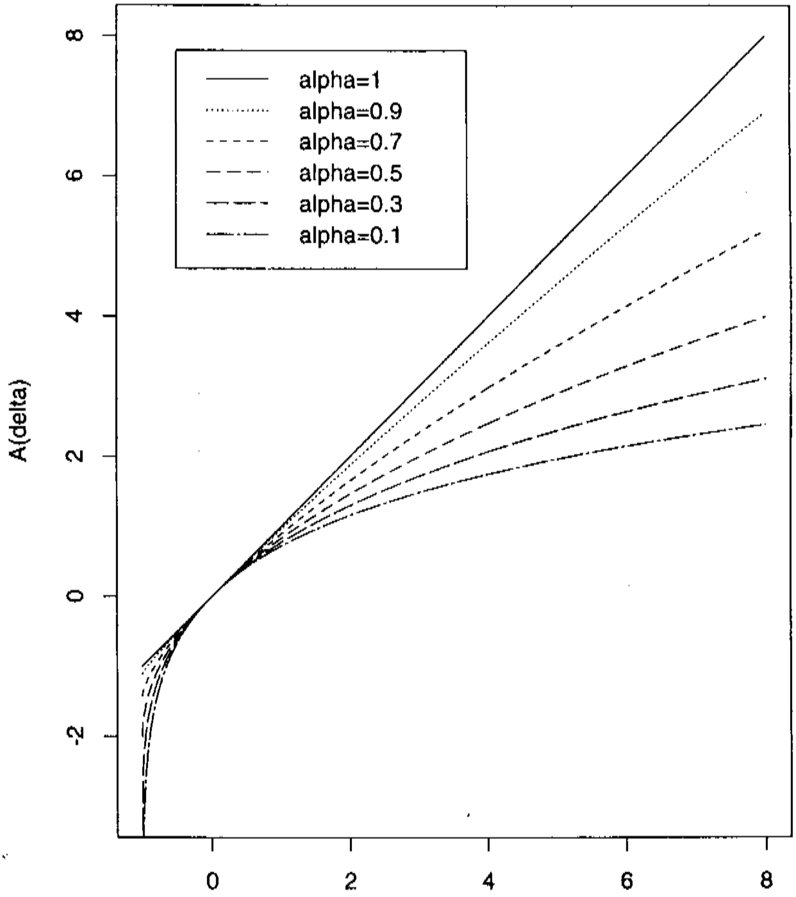

Figure 1. Residual adjustment functions for the generalized Hellinger family

In Figure 1 we represent the residual adjustment functions of several members of the generalized Hellinger family including the boundary cases $\alpha = 0$, and $\alpha = 1$

(maximum likelihood). The graph clearly shows that for smaller values of α, there is a stronger downweighting effect on large δ outliers.

Lindsay also showed that one of the indicators of the robustness of the minimum disparity method is the second derivative of the residual adjustment function evaluated at $\delta = 0$ (denoted by A_2). For maximum likelihood $A_2 = 0$, which in fact is a sufficient condition for the second order efficiency of the method in the sense of Rao (1961), and larger negative values of A_2 correspond to greater robustness. It can be easily seen that for the generalized Hellinger family, $A_2 = \alpha - 1$, as a function of α. Clearly, the smaller the value of α, the larger is the value of A_2 in negative magnitude, indicating greater robustness for smaller values of α. Note that large negative values of A_2 often guarantee that the second order approximation to the bias function of the estimator under contamination can be substantially smaller than the first order approximation (Lindsay 1994; Propositions 3 and 4), illustrating the limitation of the influence function approach in assessing the robustness of the estimators.

In addition, the minimum disparity estimators have certain outlier stability properties against any point mass contamination of the true distribution if $A(\delta) = O(\delta^{1/2})$ (Lindsay 1994; Proposition 14, Corollary 15). From the form of the residual adjustment function of the $MGHD_\alpha$ estimators we see that all estimators within the generalized Hellinger family corresponding to $\alpha \geq 1/2$ satisfy this condition. We have in fact shown a stronger breakdown result in the *Poisson* model for the generalized Hellinger divergence family (Section 5).

8. A Penalized Divergence

It is a simple corollary of Theorem 33 (Lindsay 1994), that all the $MGHD_\alpha, \alpha \in (0,1)$ estimators are asymptotically fully efficient under the model; if $f_\theta(x)$ represents the true density, then in the notation of Section 2, $n^{1/2}(\hat{\theta}_{n,\alpha} - \theta)$ converges to a normal distribution with mean vector zero and covariance matrix given by $I^{-1}(\theta)$ for all $\alpha \in (0,1)$. Thus the $MGHD_\alpha$ estimators have the same asymptotic distribution as the maximum likelihood estimator under the model. This was in fact suggested by the influence function analysis of Section 4.

This asymptotic analysis, however, gives little indication of the possible difference in the performance of the estimators in small samples under the model. It has been noted before (Harris and Basu 1994; Park, *et al.* 1995; Basu *et al.* 1996) that at the model more robust minimum disparity estimators often perform substantially poorly compared to the maximum likelihood estimator if the sample size is small. While a well established theoretical result is lacking, empirical evidence suggests that a part of this inferior small sample performance can be attributed to the disproportionately large weight that these robust disparities attach to the empty cells.

In the context of the generalized Hellinger divergence, one can have an intuitive feeling of such a phenomenon by investigating the form of the divergence given in (7.1). By writing the divergence as a sum of two components, we see that it has the form,

$$\begin{aligned} D_\alpha(d, f_\theta) &= \sum_{x:d(x)\neq 0} \left\{ \frac{d(x)}{\alpha(\alpha-1)} \left[\left(\frac{d(x)}{f_\theta(x)} \right)^{\alpha-1} - 1 \right] + \frac{(f_\theta(x) - d(x))}{\alpha} \right\} \\ &\quad + \frac{1}{\alpha} \sum_{x:d(x)=0} f_\theta(x), \end{aligned} \qquad \ldots (8.1)$$

so that for α close to zero, the weight attached to the empty cells, *i.e.* $\alpha^{-1} \sum_{x:d(x)=0} f_\theta(x)$ becomes extremely large compared to the corresponding weight in $D_1(d, f_\theta)$, which generates the maximum likelihood estimator. In terms of the residual adjustment function $A(\cdot)$ defined in the previous section, this corresponds to a very sharp dip in the left tail of the graph (note that the empty cells correspond to $\delta = -1$). This is unfortunate, since these estimators are otherwise desirable, as they have good robustness properties against outliers.

An alternative therefore is to consider the corresponding penalized version of the generalized Hellinger divergence

$$\begin{aligned} D_{\alpha,\ p}(d, f_\theta) &= \sum_{x:d(x)\neq 0} \left\{ \frac{d(x)}{\alpha(\alpha-1)} \left[\left(\frac{d(x)}{f_\theta(x)} \right)^{\alpha-1} - 1 \right] + \frac{(f_\theta(x) - d(x))}{\alpha} \right\} \\ &\quad + \sum_{x:d(x)=0} f_\theta(x). \end{aligned} \qquad \ldots (8.2)$$

The only difference between equations (8.1) and (8.2) is that the weight of the term $\sum_{x:d(x)=0} f_\theta(x)$ has been changed to 1 from $1/\alpha$, *i.e.* for all α the new penalized divergences put the same weight on the empty cells as $D_1(d, f_\theta)$. We will refer to the minimizers of the divergences $D_{\alpha,\ p}(d, f_\theta)$ by $MPGHD_\alpha$. This penalty does not affect the asymptotic distribution of the estimators; neither does it significantly affect the robustness of the estimators, as it only modifies the weight of the empty cells, not the outlying ones. In addition, $D_{\alpha,\ p}(d, f_\theta)$ is always nonnegative, and is equal to zero if and only if $d \equiv f_\theta$. We will denote the $MPGHD_\alpha$ estimates by $\hat{\theta}_{n,\alpha,p}$.

The following numerical study illustrates the possible gains due to the aforementioned penalty. Data are generated from a *Poisson* distribution with mean 5, sample size being 20. Number of replications used is 1000. For each sample we calculate the $MGHD_\alpha$ and $MPGHD_\alpha$ estimates for different values of α. Then we compute the observed mean square error of the estimates against the true value of 5. The results, presented in Table 3, show that while for small α the $MGHD_\alpha$ estimates may behave quite poorly at this moderately small sample size, the $MPGHD_\alpha$ estimates may be able to recover much of their lost efficiency.

TABLE 3. COMPARISON OF THE $MGHD_\alpha$ AND $MPGHD_\alpha$ ESTIMATORS IN TERMS OF THE ATTAINED MEAN SQUARE ERROR FOR DIFFERENT α

	Mean Square Error	
α	$MGHD_\alpha$	$MPGHD_\alpha$
0.1	0.6892	0.3020
0.2	0.5272	0.2949
0.3	0.4246	0.2908
0.4	0.3668	0.2843
0.5	0.3250	0.2781
0.6	0.2976	0.2722
0.7	0.2792	0.2667
0.8	0.2668	0.2619
0.9	0.2591	0.2580
0.99	0.2556	0.2556

Improvements of roughly similar magnitude were noticed when the experiment was repeated over several underlying distributions within the *Poisson* and the *Geometric* family (which we have not presented here for brevity). It is to be observed, however, that while applying the penalty makes intuitive sense, a completely general statement of the possible improvements due to the penalty is difficult to make since a rigorous theoretical result is unavailable. Also, not unexpectedly, the effect of the penalty becomes less and less noticeable as the sample size increases.

Application of this penalty to the example of Section 6 produces extremely nice results (see Table 2). In this case, the estimates corresponding to small values of α do not differ by quite as much from the outlier deleted maximum likelihood estimator. For the first experimental run, the estimates are practically uniform over the whole range of α; it is the same in the third run, except in the range where it breaks down; in the second run, the estimates are slightly more discrepant, but clearly much more uniform compared to the non penalized version. The results indicate that the penalty is not compromising the robustness of the estimators, but is perhaps making the estimates closer to maximum likelihood when the data fit the model.

9. Tests of Hypotheses

The minimum divergence ideas considered in this paper can be utilized to construct robust tests of hypotheses and confidence intervals in discrete parametric models. Simpson (1989a) considered the Hellinger deviance test. A generalization to the class of disparities was given in Lindsay (1994). An extension in the case of multiple sample problems was provided by Sarkar and Basu (1995).

Under the notation of Section 2, consider the simple null hypothesis $H_0 : \theta = \theta_0$. It can be checked easily that $TS_1 = 2n[D_1(d, f_{\theta_0}) - D_1(d, f_{\hat{\theta}})]$ equals the negative of twice log likelihood ratio, where $\hat{\theta}$ represents the MLE; it is well known that this converges to a $\chi^2(p)$ distribution under the null hypothesis (*e.g.* Serfling 1980). Simpson showed that the Hellinger deviance test statistic, $TS_{1/2} = 2n[D_{1/2}(d, f_{\theta_0}) - D_{1/2}(d, f_{\hat{\theta}_{n,1/2}})]$ is asymptotically equivalent to TS_1 under the null hypothesis. It follows from the results of Lindsay (1994) that $TS_\alpha = 2n[D_\alpha(d, f_{\theta_0}) - D_\alpha(d, f_{\hat{\theta}_{n,\alpha}})]$ have the same asymptotic chi-square distribution under the null hypothesis for all $\alpha \in (0, 1)$.

TABLE 4. COMPARISON OF TS_α AND $TS_{\alpha,p}$ TESTS FOR DIFFERENT α

		Nominal Level		
α		10%	5%	1%
0.1	TS_α	0.561	0.510	0.371
	$TS_{\alpha,\ p}$	0.078	0.041	0.006
0.2	TS_α	0.394	0.306	0.179
	$TS_{\alpha,\ p}$	0.082	0.040	0.006
0.3	TS_α	0.294	0.198	0.100
	$TS_{\alpha,\ p}$	0.083	0.044	0.007
0.4	TS_α	0.212	0.138	0.055
	$TS_{\alpha,\ p}$	0.085	0.044	0.008
0.5	TS_α	0.160	0.108	0.030
	$TS_{\alpha,\ p}$	0.086	0.044	0.008
0.6	TS_α	0.139	0.085	0.022
	$TS_{\alpha,\ p}$	0.088	0.048	0.008
0.7	TS_α	0.126	0.074	0.016
	$TS_{\alpha,\ p}$	0.093	0.049	0.008
0.8	TS_α	0.109	0.067	0.012
	$TS_{\alpha,\ p}$	0.094	0.050	0.008
0.9	TS_α	0.106	0.060	0.009
	$TS_{\alpha,\ p}$	0.097	0.052	0.008
0.99	TS_α	0.110	0.061	0.009
	$TS_{\alpha,\ p}$	0.104	0.061	0.009

The TS_α test statistics corresponding to smaller values of α can generally perform much better than the likelihood ratio test TS_1 in keeping the level and the power of the tests stable under contamination. In particular the Hellinger deviance test $TS_{1/2}$ have been studied by several authors (Simpson 1989a; Lindsay 1994; Basu and Sarkar 1994b; Basu *et al.* 1996) which demonstrate the desirable robustness properties of this test. For small samples, the chi-square approximation for this test statistic under the null hypothesis, however, can be quite inaccurate, with the observed levels being considerably inflated compared to the nominal levels; consequently, the confidence intervals obtained by inverting the test statistic has a true confidence coefficient lower than the nominal one. Basu *et al.* (1996) considered the use of the penalized Hellinger distance $D_{\alpha,\ p}$ in constructing robust tests of hypothesis, and showed that the chi-square approximation works substantially better for the penalized test statistic $TS_{1/2,\ p} = 2n[D_{1/2,\ p}(d, f_{\theta_0}) - D_{1/2,\ p}(d, f_{\hat{\theta}_{n,1/2,p}})]$, without compromising the robustness properties of the test.

Here we introduce the family of test statistics obtained by applying the empty cell penalty to the family of generalized Hellinger divergences. Accordingly we define the penalized family of test statistics

$$TS_{\alpha,\ p} = 2n[D_{\alpha,\ p}(d, f_{\theta_0}) - D_{\alpha,\ p}(d, f_{\hat{\theta}_{n,\alpha,p}})].$$

Note that the families TS_α and $TS_{\alpha,\ p}$ have the same asymptotic distribution under the null hypothesis since they differ between themselves only in the empty cells. Consequently, it follows from Lindsay (1994, Theorem 6) that the $TS_{\alpha,\ p}$ statistics have an asymptotic $\chi^2(p)$ distribution under the null hypothesis.

We perform a modest simulation study to demonstrate the effects of the penalty on the levels of these test statistics. Data are generated from a *Poisson* distribution with mean 5; here we test the hypothesis $\theta = 5$, assuming a $Poisson(\theta)$ model. The sample size was 20, and the number of replications was 1000. The results, obtained by using the chi-square critical values, are presented in Table 4. The numbers clearly show that compared to TS_α, the tails of the $TS_{\alpha,\ p}$ statistic are much better approximated by the chi-square distribution. For smaller α values the chi-square approximation is totally unsuitable for the former. For these values of α, the observed levels of the $TS_{\alpha,\ p}$ statistic slightly underestimate the nominal levels, but the difference is negligible compared to the performance of the TS_α statistic. Further calculations show (numbers are not presented here) that as the sample size increases, the levels of penalized test statistic quickly approach the nominal values, whereas the ordinary test statistic needs an extremely large sample size for the chi-square approximation to be reasonable.

The testing procedures described in this section extend straightforwardly to the case when the null hypothesis is composite, using the techniques of Serfling (1980). The tests based on $TS_{\alpha,\ p}$ again have the same asymptotic distribution as the likelihood ratio test under the null.

Acknowledgements. The research of the third author was supported by a Grant from CSIR, India.

References

BASU, A. and HARRIS, I.R. (1994). Robust predictive distributions for exponential families. *Biometrika* **81**, 790-794.

BASU, A., HARRIS, I.R. and BASU, S. (1996). Tests of hypotheses in discrete models based on the penalized Hellinger distance. *Statist. Prob. Letters*, **27**, 367-373.

BASU, A. and LINDSAY, B.G. (1994). Minimum disparity estimation for continuous models. *Ann. Inst. Stat. Math.* **46**, 683-705.

BASU, A. and SARKAR, S. (1994a). Minimum disparity estimation in the errors-in-variables model. *Statist. Prob. Letters* **20**, 69-73.

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

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

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

DONOHO, D.L. and HUBER, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich. L. Lehmann*, eds. P. Bickel, K. Doksum, and J. L. Hodges, Jr., Belmont CA: Wadsworth.

ESLINGER, P.W. and WOODWARD, W.A. (1991). Minimum Hellinger distance estimation for normal models. *J. Statist. Comput. Simul.* **39**, 95-114.

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

– – –– (1995). A generalized divergence measure. *Technical Report, Center for Statistical Sciences, University of Texas at Austin, Austin, TX 78712, USA.*

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

MARKATOU, M., BASU, A. and LINDSAY, B.G. (1997). Weighted likelihood estimating equations: the discrete case with applications to logistic regression. *J. Statist. Planning. Inf.*, **57**, 215-232.

PARK, C., BASU, A. and BASU, S. (1995). Robust minimum distance inference based on combined distances. *Commun. Statist.: Simula.*, **24**, 653-673.

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

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

SERFLING, R. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

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

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

– – –– (1989b). Choosing a discrepancy for minimum distance estimation: multinomial models with infinitely many cells. *Technical Report, Department of Statistics, University of Illinois, Champaign, IL 61820, U. S. A.*

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

WOODRUFF, R.C., MASON, J.M., VALENCIA, R. and ZIMMERING, S. (1984). Chemical mutagenesis testing in drosophila–I: Comparison of positive and negative control data for sex-linked recessive lethal mutations and reciprocal translocations in three laboratories. *Environmental Mutagenicity* **6**, 189-202.

APPLIED STATISTICS UNIT
INDIAN STATISTICAL INSTITUTE
203 B. T. ROAD
CALCUTTA 700 035
INDIA

STAT-MATH UNIT
INDIAN STATISTICAL INSTITUTE
203 B. T. ROAD
CALCUTTA 700 035
INDIA

STAT-MATH UNIT
INDIAN STATISTICAL INSTITUTE
203 B. T. ROAD
CALCUTTA 700 035
INDIA