Biometrika (1998), **85**, 3, *pp.* 549–559
Printed in Great Britain

Robust and efficient estimation by minimising a density power divergence

BY AYANENDRANATH BASU
Applied Statistics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India
ayanbasu@isical.ernet.in

IAN R. HARRIS
Department of Mathematics, Northern Arizona University, Flagstaff, Arizona 86011, U.S.A.
irh@odin.math.nau.edu

NILS L. HJORT
Department of Mathematics and Statistics, University of Oslo, P.B. 1053 Blindern, N-0316 Oslo, Norway
nils@math.uio.no

AND M. C. JONES
Department of Statistics, The Open University, Milton Keynes, MK7 6AA, U.K.
m.c.jones@open.ac.uk

SUMMARY

A minimum divergence estimation method is developed for robust parameter estimation. The proposed approach uses new density-based divergences which, unlike existing methods of this type such as minimum Hellinger distance estimation, avoid the use of nonparametric density estimation and associated complications such as bandwidth selection. The proposed class of 'density power divergences' is indexed by a single parameter α which controls the trade-off between robustness and efficiency. The methodology affords a robust extension of maximum likelihood estimation for which $\alpha = 0$. Choices of α near zero afford considerable robustness while retaining efficiency close to that of maximum likelihood.

Some key words: Asymptotic efficiency: Influence function; *M*-estimation; Maximum likelihood; Minimum distance estimation; Robustness.

1. INTRODUCTION

In parametric estimation, density-based minimum divergence methods, i.e. methods which estimate the parameter through minimising a data-based estimate of some appropriate divergence between the assumed model density and the true density underlying the data, have a long history. These procedures include the classical maximum likelihood method as well as minimum chi-squared methods based on families of chi-squared distances (Neyman, 1949; Rao, 1963; Cressie & Read, 1984; Lindsay, 1994; Victoria-Feser & Ronchetti, 1997). Beran (1977), using Hellinger distance, was the first to use density-based minimum divergence estimation in continuous models to develop parameter estimators

with good robustness properties relative to maximum likelihood. Among others, Tamura & Boos (1986) and Simpson (1987) have followed up on this line of research. Under some regularity conditions, these methods have full asymptotic efficiency at the model. However, in continuous models the methods suffer from the drawback that it is necessary to use some nonparametric smoothing technique such as kernel density estimation to produce a continuous estimate of the true density. They therefore involve all the associated complications such as bandwidth selection. See also Cao, Cuevas & Fraiman (1995). Basu & Lindsay (1994) considered another modification of this approach where the model is smoothed with the same kernel as the data to reduce the dependence of the procedure on the smoothing method.

The present paper introduces a new family of density-based divergence measures, to be called density power divergences. Note that these measures are not closely related to the 'power divergences' of Cressie & Read (1984). The family is indexed by a single parameter α which controls the trade-off between robustness and asymptotic efficiency of the parameter estimators which are the minimisers of this family of divergences. When $\alpha = 0$, the density power divergence is the Kullback–Leibler divergence (Kullback & Leibler, 1951) and the method is maximum likelihood estimation; when $\alpha = 1$, the divergence is the mean squared error, and a robust but inefficient minimum mean squared error estimator ensues. For any α, the estimation procedure has the considerable advantage of not requiring any nonparametric smoothing. Various examples are explored to investigate the interplay between robustness and efficiency. It is found that the estimators with small α have strong robustness properties with little loss in asymptotic efficiency relative to maximum likelihood under model conditions. It should be noted that the minimum density power divergence estimators are a particular case of M-estimators, but they are a novel case motivated by other attractive considerations.

The rest of the paper is organised as follows. In § 2 we develop the new class of estimation procedures. Some of their properties are described in § 3 by appealing to the M-estimation interpretation of the methodology. In § 4, we investigate the performance of the estimators in several common parametric families, look at the breakdown of the methods in the normal model, and illustrate the performance of the method in two examples. Brief remarks on selecting α are presented in § 5.

2. THE DENSITY POWER DIVERGENCE

Consider a parametric family of models $\{F_t\}$, indexed by the unknown parameter $t \in \Omega \subset R^s$, possessing densities $\{f_t\}$ with respect to Lebesgue measure, and let $\mathscr{G}$ be the class of all distributions G having densities g with respect to Lebesgue measure. The latter is for the sake of keeping a clear focus in our presentation, but results hold for discrete models as well.

Define the divergence $d_\alpha(g, f)$ between density functions g and f to be

$$d_\alpha(g, f) = \int \left\{ f^{1+\alpha}(z) - \left(1 + \frac{1}{\alpha}\right) g(z) f^\alpha(z) + \frac{1}{\alpha} g^{1+\alpha}(z) \right\} dz \quad (\alpha > 0). \tag{2·1}$$

When $\alpha = 0$, the integrand in expression (2·1) is undefined, and we define the divergence $d_0(g, f)$ as

$$d_0(g, f) = \lim_{\alpha \to 0} d_\alpha(g, f) = \int g(z) \log\{g(z)/f(z)\}\, dz.$$

Note that $d_0(g, f)$ is the Kullback–Leibler divergence. The estimation procedure that we discuss in this paper consists of choosing parameter values to minimise an estimate of $d_\alpha(g, f_t)$.

We are most interested in smaller values of $\alpha \geqslant 0$, say between zero and one, although values greater than one can be considered too. The procedure typically becomes less and less efficient as α increases as we will see later.

THEOREM 1. *The quantity $d_\alpha(g, f)$ is a divergence in that it is nonnegative for all $g, f \in \mathscr{G}$ and is equal to zero if and only if $f \equiv g$ almost everywhere.*

Proof. For $\alpha > 0$, it is straightforward to show that the integrand is nonnegative for any fixed value of z, and takes the value zero if and only if $g = f$ identically. When $\alpha = 0$, the result is well known. □

The family of divergences d_α, as a function of α, will be called the class of density power divergences. The following is a simple consequence of Theorem 1: for any given α the minimum density power divergence functional at G, defined by the requirement $d_\alpha(g, f_{T_\alpha(G)}) = \min_{t \in \Omega} d_\alpha(g, f_t)$, is Fisher consistent (Rao, 1965, § 5c.1). In addition, given a random sample $X_1, \ldots, X_n$ from G, the minimum density power divergence estimator $\hat{\theta}$, generated by minimising

$$\int f_t^{1+\alpha}(z)\, dz - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} f_t^\alpha(X_i) \tag{2·2}$$

with respect to t, is weakly consistent for $\theta = T_\alpha(G)$; see Theorem 2. We assume here that $T_\alpha(G)$ exists and is unique, as will normally be the case. Verifying this is perhaps most easily done on a case by case basis, and would depend on the parameter space and the complexity of the $\{f_t\}$ family as well as on the true density g.

Given the data, $T_0(G_n)$ maximises $\int \log f_t(z)\, dG_n(z)$, where G_n is the empirical distribution function, and is therefore the maximum likelihood estimate of the parameter if it exists. On the other hand, for the value $\alpha = 1$, $d_1(g, f) = \int \{f(z) - g(z)\}^2 dz$, and the estimator minimises the L_2 distance between the densities. Notice that this L_2 minimisation does not require a smooth nonparametric estimate of g, in contrast to work of Cao et al. (1995). Thus, for $0 < \alpha < 1$, the class of density power divergences provides a smooth bridge between the L_2 distance and the Kullback–Leibler divergence. For general families it can be checked easily that the estimating equations have the form

$$U_n(t) \equiv n^{-1} \sum_{i=1}^{n} u_t(X_i) f_t^\alpha(X_i) - \int u_t(z) f_t^{1+\alpha}(z)\, dz = 0, \tag{2·3}$$

where $u_t(z) = \partial \log f_t(z)/\partial t$ is the maximum likelihood score function. Note that this estimating equation is unbiased when $g = f_t$.

Some motivation for the form of the divergence (2·1) can be obtained by looking at the location model, where $\int f_t^{1+\alpha}(z)\, dz$ is independent of t. In this case, the proposed estimators maximise $\sum f_t^\alpha(X_i)$, with the corresponding estimating equations having the form

$$\sum_{i=1}^{n} u_t(X_i) f_t^\alpha(X_i) = 0. \tag{2·4}$$

This can be viewed as a weighted version of the efficient maximum likelihood score equation. When $\alpha > 0$, (2·4) provides a relative-to-the-model downweighting for outlying observations; observations that are wildly discrepant with respect to the model will get

nearly zero weights. In the fully efficient case $\alpha = 0$, all observations, including very severe outliers, get weights equal to one.

A few examples of the robustness of some variants of the minimum L_2 distance estimator, $\alpha = 1$, in the normal model have been presented by Brown & Hwang (1993), while minimising the L_2 distance between a normal density and a histogram estimating g. Consideration of the small contribution of outliers to L_2 distance based on histograms or kernel density estimates makes this robustness intuitively apparent; see also Terrell (1993), Hjort (1994) and Jones & Hjort (1994). Unfortunately, the robustness of the minimum L_2 distance estimator is achieved at a fairly stiff price in asymptotic efficiency, as we will see later. By choosing a value of α close to zero, one makes all the weights closer to 1 compared to the minimum L_2 method, improving the asymptotic efficiency of the procedure. The proposed estimators, therefore, represent compromises between efficiency and robustness, with the degree of compromise controlled by the tuning parameter α.

The idea of downweighting with respect to the model rather than the data is also the motivating principle of Field & Smith (1994) and Windham (1995). Field & Smith's weights involve the model distribution function. Windham describes a fixed point algorithm that also uses density power weighting. A version of Windham's procedure is equivalent to choosing t such that

$$\frac{\sum_i u_t(X_i) f_t^{\alpha}(X_i)}{\sum_i f_t^{\alpha}(X_i)} = \frac{\int u_t(z) f_t^{1+\alpha}(z)\, dz}{\int f_t^{1+\alpha}(z)\, dz}. \tag{2·5}$$

If f_t is a location family then (2·5) reduces to (2·4), but in general (2·5) does not reduce to (2·3). The relationship between the methods is the subject of a paper currently in preparation. Estimating equations that are weighted sums of $u_t(X_i)$ also appear elsewhere, for instance in the general theory of M-estimation (Hampel et al., 1986).

Note that the divergence given by (2·1) is close to a weighted L_2 distance (Hjort, 1994) in the sense that, for fixed α, and f close to g, $d_\alpha(g, f)$ becomes close to

$$\frac{1}{2}(1+\alpha) \int g^{\alpha-1}(z)\{f(z) - g(z)\}^2\, dz. \tag{2·6}$$

Observe how minimum L_2 corresponds exactly to a unit weighting, maximum likelihood corresponds to a $1/g$ weighting, and minimum density power divergence for $0 < \alpha < 1$ corresponds to an intermediate $1/g^{\gamma}$ for $0 < \gamma < 1$ weighting. Unlike (2·6), however, the beauty of (2·1) is that, if we ignore the last term because it does not depend on f, g appears only as a multiplier of terms in f. Thus, while f will be replaced by f_t, g can appropriately be replaced by its empirical version rather than a smooth estimate of g, which is necessary in other robust density-based minimum divergence approaches (Beran, 1977; Cao et al., 1995). The same holds, of course, for maximum likelihood estimation. Formula (2·6) also gives insight into the properties of minimum divergence estimation described in § 3.

3. PROPERTIES

3·1. *Link with M-estimation*

Properties of minimum divergence estimators follow immediately from existing theory once it is recognised that minimum divergence estimators are M-estimators, i.e. they solve an equation of the form $\sum_i \psi(X_i, t) = 0$ (Huber, 1981; Hampel et al., 1986). Our ψ function

is

$$\psi(x, t) = u_t(x) f_t^{\alpha}(x) - \int u_t(z) f_t^{1+\alpha}(x)\, dx.$$

As a result of this, further theoretical development can be presented in outline form only. A technical report with the same title and authors as this paper, Statistical Research Report No. 7, Department of Mathematics, University of Oslo, henceforth referred to as 'our report', fills in many of the details.

3·2. *Asymptotic properties*

Suppose the data are generated from the true distribution G, not necessarily in the model. In the following, θ represents the best fitting value of the parameter, in the sense of minimising the discrepancy $d_\alpha(g, f_t)$, whereas t denotes a generic element of Ω. Let $X_1, \ldots, X_n$ be independent and identically distributed with distribution G and density g, and let $\hat{\theta}$ be the minimiser of (2·2). Define $i_t(x) = -\partial\{u_t(x)\}/\partial t$, the so-called information function of the model, which is positive definite for all required t.

THEOREM 2. *Under certain regularity conditions, given in our report, there exists $\hat{\theta}$ such that, as $n \to \infty$,*

(i) *$\hat{\theta}$ is consistent for θ, and*

(ii) *$n^{\frac{1}{2}}(\hat{\theta} - \theta)$ is asymptotically multivariate normal with vector mean zero and covariance matrix $J^{-1}KJ^{-1}$, where $J = J(\theta)$ and $K = K(\theta)$ are given by*

$$J = \int u_\theta(z) u_\theta^{\mathrm{T}}(z) f_\theta^{1+\alpha}(z)\, dz + \int \{i_\theta(z) - \alpha u_\theta(z) u_\theta^{\mathrm{T}}(z)\}\{g(z) - f_\theta(z)\} f_\theta^{\alpha}(z)\, dz, \quad (3·1)$$

$$K = \int u_\theta(z) u_\theta^{\mathrm{T}}(z) f_\theta^{2\alpha}(z) g(z)\, dz - \xi\xi^{\mathrm{T}} \quad (3·2)$$

with $\xi = \int u_\theta(z) f_\theta^{\alpha}(z) g(z)\, dz$.

The essential M-estimation formulae underlying the above are in Hampel et al. (1986, § 4.2c).

3·3. *Influence function and standard error*

From M-estimation theory, the influence function of the density power divergence functional is immediately

$$\mathrm{IF}(y;\, T_\alpha, G) = J^{-1}\{u_\theta(y) f_\theta^{\alpha}(y) - \xi\},$$

where $\theta = T_\alpha(G)$ and J and ξ are as in (3·1) and (3·2). If we assume that J and ξ are finite, this is a bounded function of y whenever $u_\theta(y) f_\theta^{\alpha}(y)$ is bounded. This is true, for example, for any $\alpha > 0$ in the normal location-scale problem, unlike other density based minimum divergence procedures such as those based on the Hellinger distance. The influence functions for the estimation of the normal mean when $\sigma = 1$ are plotted in Fig. 1 for several values of α; note their redescending nature for all $\alpha > 0$.

The asymptotic variance of $\sqrt{n}$ times the minimum density power divergence estimator can be consistently estimated in a sandwich fashion by using the above influence function, e.g. Huber (1981). Let $K_i = u_\theta(X_i) f_\theta^{\alpha}(X_i) - \xi$, and let $\hat{K}_i$ be the corresponding quantity

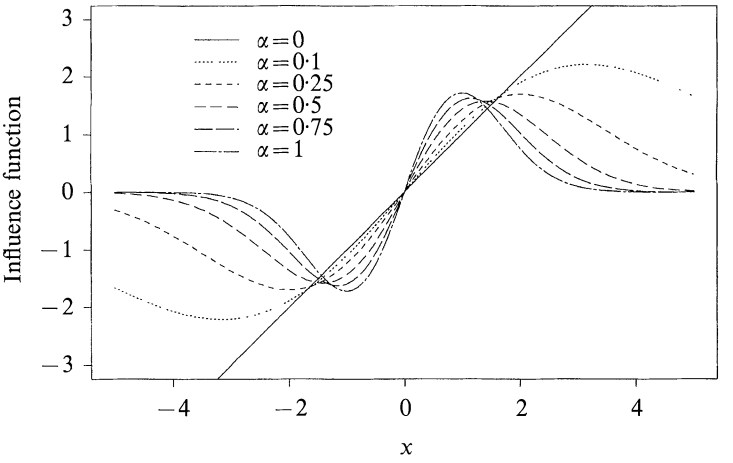

Fig. 1. Influence functions for estimation of a normal mean, with known variance, for various choices of α.

evaluated at $\hat{\theta}$, with G_n in place of G. Let $\hat{K} = (n-1)^{-1}\sum_i (\hat{K}_i \hat{K}_i^{\mathrm{T}})$. Then the asymptotic variance of $\sqrt{n}$ times the parameter estimators can be consistently estimated by $\hat{J}^{-1}\hat{K}\hat{J}^{-1}$, where $\hat{J}$ is obtained from J by replacing θ with $\hat{\theta}$, with G_n in place of G. Consistent estimators of the asymptotic variance of the method can also be obtained by the jackknife and bootstrap techniques.

3·4. *Equivariance*

The maximum likelihood method has two important equivariance properties; estimates are equivariant with respect to both reparameterisations and transformation of the data. Our minimum density power divergence method shares the first general property: if the model is reparameterised to $\psi = \psi(\theta)$ with a one-one transformation, then the density power divergence estimate of ψ is simply $\hat{\psi} = \psi(\hat{\theta})$, in terms of the density power divergence estimate $\hat{\theta}$ of θ, using the same α.

The second maximum likelihood property does not generally hold for the new estimation method, however. If data are transformed from X_i to $Y_i = h(X_i)$, then the minimum density power divergence estimator, θ^*, say, is defined as the minimiser of

$$\int [f_t\{\eta(y)\}\,|\eta'(y)|]^{1+\alpha}\,dy - \left(1+\frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} [f_t(X_i)|\eta'\{h(X_i)\}\,|]^{\alpha},$$

where $X_i = \eta(Y_i)$ is the inverse transformation. We see by comparison with (2·3) that θ^* is equal to $\hat{\theta}$ if $\eta'(y)$ is a nonzero constant. Thus the estimation method is equivariant under a transformation of the form $Y_i = aX_i + b$, but not in general under other transformations, unless $\alpha = 0$.

3·5. *Hypothesis testing, model choice and regression*

Further aspects of ordinary statistical likelihood inference can also be robustified by the use of minimum divergence estimation, paralleling similar work carried out in the literature for other M-estimators. One can develop robust hypothesis testing as in Heritier & Ronchetti (1994) and robust model choice criteria as in Ronchetti (1997); versions of

these, including a robustified Akaike information criterion for model choice, are detailed in our report. Finally, the important generalisation to robust inference in general regression models can be carried out; see Hampel et al. (1986, Ch. 6) for some general suggestions. In our report, in the context of a model $f_\beta(y|x)$ for data pairs (x_i, Y_i), we analyse the behaviour of estimators formed by minimising

$$n^{-1}\sum_{i=1}^{n}\int f_\beta^{1+\alpha}(y|x_i)\,dy-\left(1+\frac{1}{\alpha}\right)n^{-1}\sum_{i=1}^{n}f_\beta^{\alpha}(Y_i|x_i);$$

again the limiting case $\alpha\to 0$ corresponds to ordinary likelihood analysis.

4. Special parametric families: Efficiency, breakdown and examples

4·1. *Introduction to* § 4·2

Suppose that the true distribution g belongs to the parametric family $\{f_t\}$, θ being the true value of the parameter. Then the formulae for J, K and ξ simplify to

$$J=\int u_\theta(z)u_\theta^{\mathrm{T}}(z)f_\theta^{1+\alpha}(z)\,dz,$$

$$K=\int u_\theta(z)u_\theta^{\mathrm{T}}(z)f_\theta^{1+2\alpha}(z)\,dz-\xi\xi^{\mathrm{T}},\quad \xi=\int u_\theta(z)f_\theta^{1+\alpha}(z)\,dz.$$

Note that, in the limit $\alpha\to 0$, J and K both become equal to the Fisher information. These formulae can be used to investigate the asymptotic efficiency of the estimators, and in particular to judge how much is lost relative to the maximum likelihood estimator under model conditions. In § 4·2, some examples for particular parametric families are considered. We will define the asymptotic relative efficiency of an estimator to be the ratio of the asymptotic variance of the maximum likelihood estimator to that of the estimator in question.

4·2. *Efficiencies for particular families*

(a) *Mean of univariate normal.* For a location family $\xi=0$. If we let f_θ be the $N(\mu,\sigma^2)$ density with known σ^2 and u_θ the score function with respect to the mean parameter μ, elementary integration gives

$$K=(2\pi)^{-\alpha}\sigma^{-(2+2\alpha)}(1+2\alpha)^{-3/2},\quad J=(2\pi)^{-\alpha/2}\sigma^{-(2+\alpha)}(1+\alpha)^{-3/2}.$$

The asymptotic variance of $n^{\frac{1}{2}}$ times the estimator of μ is then given by

$$\left(1+\frac{\alpha^2}{1+2\alpha}\right)^{3/2}\sigma^2.$$

Since the asymptotic variance of $n^{\frac{1}{2}}$ times the maximum likelihood estimator is σ^2, the asymptotic relative efficiency of the density divergence estimator is immediate. For $\alpha=0{\cdot}25$ it is 0·941, for example, already quite close to one. Results for different values of α are given in the first row of Table 1.

(b) *Standard deviation of univariate normal.* Again let f_θ be the $N(\mu,\sigma^2)$ density but treat both parameters as unknown. Calculations for the two 2×2 matrices J and K show that both have zeros off the diagonals, that is, the estimators $\hat\mu$ and $\hat\sigma$ are asymptotically

Table 1. *Asymptotic relative efficiencies of the density power divergence estimators*

	α						
Model	0·00	0·02	0·05	0·10	0·25	0·50	1·00
Normal μ	1·000	0·999	0·997	0·988	0·941	0·838	0·650
Normal σ	1·000	0·999	0·993	0·976	0·888	0·731	0·541
Exponential (θ)	1·000	0·998	0·991	0·968	0·858	0·684	0·509
Poisson ($\lambda = 3$)	1·000	0·999	0·997	0·988	0·944	0·850	0·679
Poisson ($\lambda = 10$)	1·000	0·999	0·997	0·988	0·941	0·840	0·656

independent. The limiting distribution for $n^{\frac{1}{2}}(\hat{\mu} - \mu)$ is therefore as found in case (a) even when σ is unknown.

Here, we concentrate on estimation of σ. Lengthy calculations show that the asymptotic variance of $n^{\frac{1}{2}}$ times the estimator is

$$\frac{(1+\alpha)^2}{(2+\alpha^2)^2}\left\{\frac{2(1+\alpha)^3(1+2\alpha^2)}{(1+2\alpha)^{5/2}} - \alpha^2\right\}\sigma^2.$$

Efficiency calculations, in comparison with $\sigma^2/2$, are presented in the second row of Table 1. Small α density power divergence estimation continues to retain high efficiency. The values in Table 1 clearly show that the minimum L_2 distance estimators of μ and σ are quite inefficient; see also Hjort (1994).

(c) *Exponential distribution.* For the density $f_\theta(x) = \theta^{-1}\exp(-x/\theta)\,(x > 0)$, the quantities K and J in the asymptotic variance of $n^{\frac{1}{2}}$ times the minimum density power divergence estimator of θ are given by

$$K = \left\{\frac{1+4\alpha^2}{(1+2\alpha)^3} - \frac{\alpha^2}{(1+\alpha)^4}\right\}\theta^{-(2+2\alpha)}, \quad J = \left\{\frac{1+\alpha^2}{(1+\alpha)^3}\right\}\theta^{-(2+\alpha)}.$$

The asymptotic variance is then given by

$$\frac{(1+\alpha)^2}{(1+\alpha^2)^2}\left\{\frac{(1+\alpha)^4(1+4\alpha^2)}{(1+2\alpha)^3} - \alpha^2\right\}\theta^2.$$

The asymptotic relative efficiencies are given for certain α in the third row of Table 1. Again, efficiencies remain high for small α.

(d) *Mean of multivariate normal.* The family is $N_p(\mu, \Sigma)$. The limiting covariance matrix of $n^{\frac{1}{2}}$ times the minimum density power divergence estimator of μ, whether or not Σ is known, can be shown to be

$$\left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{p/2+1}\Sigma.$$

Thus one loses efficiency for increasing p if α is kept fixed.

(e) *Poisson distribution.* Calculation of the asymptotic variance of the estimator can be carried out numerically, although not via a closed-form formula. It involves an infinite but rapidly convergent sum. In Table 1 we also provide the asymptotic relative efficiencies of the estimators for two different values of the mean parameter λ and several choices of α. Note that the Poisson results are very similar to those for normal μ for $\lambda = 10$.

4·3. *Breakdown in the normal distribution*

The breakdown point of an estimator, crudely described as the proportion of bad observations that an estimator can tolerate before it becomes completely uninformative, is one of the descriptors of the robustness of the method. The gross-error breakdown point (Hampel et al., 1986, p. 97) of the minimum density power divergence estimator of the parameters of the normal distribution can be determined when the data come from the contaminated model $q(z) = (1 - \varepsilon)g(z) + \varepsilon\delta_x(z)$, where δ is the Dirac delta function and $x \to \infty$. This is done in our report, where the derivation proceeds from first principles because we have not found a general result which covers this situation.

Simultaneous location and scale breakdown, in the sense that location 'explodes' and scale 'implodes' (Hampel et al., 1986, p. 98), is found to occur if

$$\varepsilon > \alpha/(1 + \alpha)^{3/2}.$$

The breakdown point increases monotonically from zero when $\alpha \doteq 0$, in line with the zero breakdown of the maximum likelihood estimator which can easily be shown separately, to $1/(2\sqrt{2}) = 0{\cdot}354$ when $\alpha = 1$. In fact, the breakdown continues to increase until its maximal value of $2/(3\sqrt{3}) = 0{\cdot}385$ at $\alpha = 2$, but by then the efficiency of the estimator is unacceptably low.

4·4. *Examples*

In our first example we consider Newcomb's light speed data, e.g. Moore & McCabe (1993). The data were also analysed by Brown & Hwang (1993), who were trying to fit the 'best approximating normal distribution' to the corresponding histogram. The limiting case of their approach generates the normal distribution whose mean and standard deviation are the minimum L_2 distance estimates of μ and σ under a normal model. This estimator, it was observed, quite successfully downweighted the extreme outliers in the Newcomb data.

For the dataset, Table 2 gives the values of the minimum density power divergence estimates of μ and σ for various values of α under the normal model. These estimators exhibit strong outlier resistance properties even for quite small values of α. When α is as small as 0·1, for which the minimum density power divergence estimator of σ has an efficiency loss of only 2·4% under the model, the estimate of σ is 5·39, fairly close to the estimate obtained for $\alpha = 1$. A visual representation of this is provided in Fig. 2, where the normal densities $N(\hat{\mu}, \hat{\sigma}^2)$, for $\alpha = 0$, 0·1, 0·25, 0·5 and 1, are superimposed on a histogram of the Newcomb data. Except when the maximum likelihood estimate is used, all the normal densities fit the main body of the histogram quite well, even the one with $\alpha = 0{\cdot}1$.

Table 2. *Estimated parameters for the Newcomb data under the normal model*

	α						
	0·00	0·02	0·05	0·10	0·25	0·50	1·00
$\hat{\mu}$	26·21	26·74	27·44	27·60	27·64	27·52	27·29
$\hat{\sigma}$	10·66	8·92	5·99	5·39	5·04	4·90	4·67

In the second example our estimation method is applied to chemical mutagenicity data previously analysed by Simpson (1987) in the context of minimum Hellinger distance estimation. In the sex-linked recessive lethal test in drosophila, male flies are exposed to

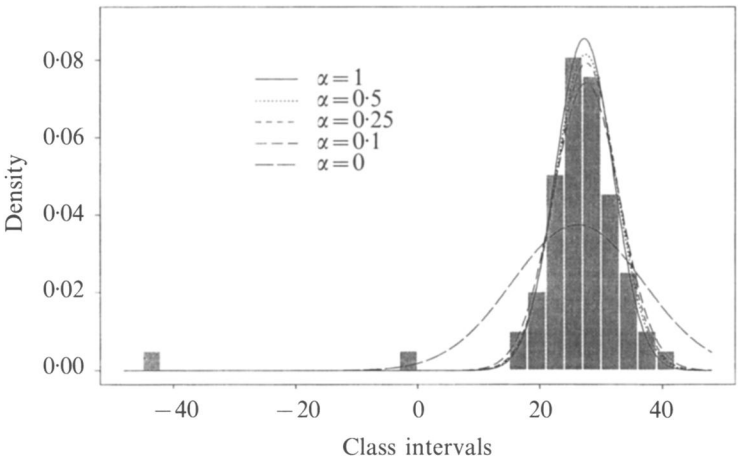

Fig. 2. A histogram of the Newcomb data with superimposed normal densities fitted using density power divergence parameter estimation with various values of α.

different doses of a chemical to be screened. They are then mated with unexposed females and for each male the number of daughter flies carrying a recessive lethal mutation on the X chromosome is noted. One such experiment with 34 males resulted in 23, 7, 3 and 1 males having 0, 1, 2 and 91 such daughters respectively. Note that the last value is a very large outlier. Simpson considered a Poisson fit for these data, and found that the minimum Hellinger distance estimate of the mean parameter λ successfully downweights the large outlier, unlike the maximum likelihood method.

Here we compute the minimum density power divergence estimates for these data under the Poisson(λ) model. The results are presented in Table 3. As expected the more robust members of the family downweight the large outlier successfully. However, what is more interesting is that this downweighting can be observed even for very small values of α. The procedure apparently loses robustness for some α between 0·01 and 0·001. For comparison, the maximum likelihood estimate of λ after deleting this outlier is 0·394, and the minimum Hellinger distance estimate of λ for these data, with and without the outlier, is 0·364.

Table 3. *Estimated parameters for the drosophila data under Poisson model*

	α								
	0·00	0·001	0·01	0·02	0·05	0·10	0·25	0·50	1·00
$\hat{\lambda}$ (all observations)	3·059	2·056	0·447	0·394	0·393	0·392	0·386	0·374	0·365
$\hat{\lambda}$ (outlier deleted)	0·394	0·394	0·394	0·393	0·392	0·390	0·382	0·366	0·349

In the above examples, we successfully used a simple bisection method for the one-parameter case and Newton–Raphson in the two-parameter case, with fast results. Computational questions for larger and more difficult problems are left for future research.

5. REMARKS ON SELECTING α

There can be no universal way of selecting an appropriate α parameter when applying our estimation methods. It specifies the underlying distance measure and typically dictates

to what extent the resulting methods become statistically more robust than the maximum likelihood methods, and should be thought of as an algorithmic parameter. One way of selecting it is to fix the efficiency loss, at the ideal parametric model employed, at some low level, like five or ten percent. A related idea is to fix the maximum level of the influence curve at some acceptable level. Other ways could in some practical applications involve prior motions of the extent of contamination of the model.

Acknowledgement

The authors would like to thank Professor Probal Chaudhuri for helpful comments and the referees for persuading us to take advantage of the theory of M-estimation.

References

Basu, A. & Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **48**, 683–705.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5**, 445–63.

Brown, L. D. & Hwang, J. T. G. (1993). How to approximate a histogram by a normal density. *Am. Statistician* **47**, 251–5.

Cao, R., Cuevas, A. & Fraiman, R. (1995). Minimum distance density-based estimation. *Comp. Statist. Data Anal.* **20**, 611–31.

Cressie, N. & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. R. Statist. Soc.* B **46**, 440–64.

Field, C. & Smith, B. (1994). Robust estimation—a weighted maximum likelihood approach. *Int. Statist. Rev.* **62**, 405–24.

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

Heritier, S. & Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Am. Statist. Assoc.* **89**, 897–904.

Hjort, N. L. (1994). Minimum L_2 and robust Kullback–Leibler estimation. In *Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Ed. P. Lachout and J. Á. Víšek, pp. 102–5. Prague: Academy of Sciences of the Czech Republic.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Jones, M. C. & Hjort, N. L. (1994). Comment on Brown & Hwang (1993). *Am. Statistician* **48**, 353–4.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.* **22**, 1081–114.

Moore, D. S. & McCabe, G. P. (1993). *Introduction to the Practice of Statistics*, 2nd ed. New York: W.H. Freeman.

Neyman, J. (1949). Contribution to the theory of χ^2 tests. In *Proceedings of the First Berkeley Symposium in Mathematics and Statistics*, pp. 239–73. Berkeley, CA: University of California Press.

Rao, C. R. (1963). Criteria of estimation in large samples. *Sankhyā* A **25**, 189–206.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications.* New York: Wiley.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statist. Sinica* **7**, 327–38.

Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Am. Statist. Assoc.* **82**, 802–7.

Tamura, R. N. & Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Am. Statist. Assoc.* **81**, 223–9.

Terrell, G. R. (1993). Spline density estimates. In *Proc. Statist. Comp. Sect.*, pp. 255–60. Washington, DC: Am. Statist. Assoc.

Victoria-Feser, M. P. & Ronchetti, E. (1997). Robust estimation for grouped data. *J. Am. Statist. Assoc.* **92**, 333–40.

Windham, M. P. (1995). Robustifying model fitting. *J. R. Statist. Soc.* B **57**, 599–609.

[*Received March* 1997. *Revised October* 1997]