The Iteratively Reweighted Estimating Equation in Minimum Distance Problems

BY

Ayanendranath Basu
Applied Statistics Unit
Indian Statistical Institute
203 B.T.Road
Calcutta 700 108
INDIA

AND

Bruce G. Lindsay
Department of Statistics
Penn State University
326 Thomas Building
University Park, PA 16802
USA

Abstract: The class of density based minimum distance estimators provide attractive alternatives to the maximum likelihood estimator because several members of this class have nice robustness properties while being first order efficient under the assumed model. A helpful computational technique – similar to the iteratively reweighted least squares used in robust regression – is introduced which makes these estimators computationally much more feasible. This technique is much simpler than the Newton-Raphson (NR) method to implement. The loss suffered in the rate of convergence compared to the NR method can be made to vanish in some exponential family situations by a little modification in the weight function – in which case the performance is comparable to the NR method. For a large number of parameters the performance of this modified version is actually expected to be better than the NR method. In view of the widespread interest in density based robust procedures, this modification appears to be of great practical value.

Key words and phrases: Disparity, robustness, Hellinger distance, iteratively reweighted least squares, iteratively reweighted estimating equation, Newton-Raphson algorithm, Fixed point algorithm.

The second author's research was partially supported by the National Science Foundation under grant DMS 0104443.

1 Introduction

Minimum distance estimation forms an important subclass of statistical methodology. Originally, minimum distance functions were developed for goodness of fit purposes. The popular distances in the early literature were the Kolmogorov-Smirnov distance, the Cramér-von Mises distance, as well as weighted versions and other variants of these. The basic ingredient in this approach is the measurement of a distance between the data, summarized by the empirical distribution function, and the hypothesized probability distribution. During the last few decades statisticians have become increasingly aware of the potential of this approach in robust estimation. Much of the work in the minimum distance area was pioneered by Wolfowitz (1953, 1954, 1957) in the mid fifties. There was a revival of this line of research in the early eighties as evidenced by the works of Parr and Schucany (1980), Boos (1981), Parr and DeWet (1981), Heathcote and Silvapulle (1981), and others. Parr (1981) also provides an extensive bibliography of minimum distance estimation up to that point of time. Works of Wiens (1987), Hettmansperger et al. (1994), Özturk (1994), Özturk and Hettmansperger (1997), and Özturk et al. (1999) have further extended this line of research.

Many of the estimators proposed in the above papers have strong robustness properties under model misspecification. However their robustness is usually achieved at the cost of first order efficiency at the model. On the other hand, a second and a relatively more modern branch of minimum distance estimation, that based on density based distances (or divergences in general) has been shown to produce a large class of estimators which combine attractive robustness properties will full asymptotic efficiency. Beran (1977) appears to be the first to use a density based distance for the purpose of robust inference. He demonstrated that the minimum Hellinger distance estimators are simultaneously robust and first order efficient. Other authors, such as Stather (1981), Tamura and Boos (1986), Simpson (1987, 1989) Donoho and Liu (1988a,b), Eslinger and Woodward (1991), Basu and Harris (1994), Cao, Cuevas and Fraiman (1995), Markatou (1996), Basu, Sarkar and Vidyashankar (1997), and Basu and Basu (1998) have further investigated

related estimators. Lindsay (1994) introduced a class of minimum disparity estimators and illustrated the geometry behind their robustness. These ideas were extended to continuous models by Basu and Lindsay (1994). Also see Basu, Harris and Basu (1997) for a comprehensive review of minimum disparity inference.

The density based minimum distance estimators (or minimum disparity estimators in particular) provide attractive alternatives to the maximum likelihood estimator. However, the defining equations of the minimum disparity estimators are usually nonlinear and numerical methods have to be applied to solve them. The numerical difficulty increases greatly with the number of parameters. For example, to carry out the estimation of (μ, Σ) in a multidimensional normal model in d dimensions, there are $p = d + d(d+1)/2 = d(d+3)/2$ unknown parameters. Each step of Newton-Raphson requires $(p+1)(p+2)/2$ numerical integrations and the inversion of a p dimensional Hessian matrix.

In this paper we consider a method closely related to iteratively reweighted least squares with the aim to reduce the numerical difficulty described in the previous paragraph. Our initial motivation came from the fact that the new method is *vastly* simpler to program and, in the d dimensional normal, requires $(p+2)$ numerical integrations and no matrix inversion per step. Even for the case $d = 3$, the number of parameters is 9 and so each Newton-Raphson step requires 55 numerical integrations and the inversion of a 9×9 matrix, while the new method requires only 11 numerical integrations per step. While the price one might expect to pay for this is a decrease from quadratic to linear convergence, our most striking finding was that by a careful (but very simple) selection of weights, we could make the method competitive in speed with Newton-Raphson even in the univariate model, where $p = 2$. (A simple scalar adjustment makes the method quadratically convergent when the data exactly fit the model.) Our theoretical calculations are substantiated by several numerical investigations. We believe this paper demonstrates generally applicable techniques for applying iterative reweighting algorithms in new problems and for making them more efficient. In particular we expect the algorithm described here to be of great practical use in view of the widespread interest in the minimum Hellinger distance and related methods.

The rest of the paper is organized as follows: In Section 2, we provide a brief review of minimum disparity estimation. The main contributions of the paper are presented in Section 3, where we first develop the iteratively reweighted estimating equation ($IREE$) algorithm in the spirit of iteratively reweighted least squares ($IRLS$) used in robust regression, and then demonstrate that a simple refinement can make the method comparable in performance to the Newton-Raphson algorithm, while keeping the implementation substantially simpler. Some further issues including a second order analysis, some discussion on the range of applicability of the method in small samples, and a weighted likelihood modification resulting from the $IREE$ idea are discussed in Section 4. A small appendix presents a step by step implementation of the algorithm.

2. Minimum Disparity Estimation

Let us briefly review minimum disparity estimation leading up to the estimating equation that we will be concerned with. We start with the discrete model. Let $m_\beta(x)$ represent the model density function indexed by an unknown $\beta \in \Omega$; without loss of generality let the sample space be $\mathcal{X} = \{0, 1, \ldots, \}$. Let $d(x)$ represent the proportion of observations in a sample of size n that have the value x. Define $\delta(x) = d(x)/m_\beta(x) - 1$ to be the *Pearson* residual at x. Then a general disparity measure ρ can be expressed in the form

$$\rho_G(d, m_\beta) = \sum_x G(\delta(x)) m_\beta(x) \tag{2.1}$$

where G is a strictly convex, thrice differentiable function, with $G(0) = 0$. Minimization of a disparity measure over the parameter space Ω generates the corresponding minimum disparity estimator (MDE). For example $G(\delta) = 2(\sqrt{\delta + 1} - 1)^2$ generates twice squared Hellinger distance $HD(d, m_\beta) = 2\sum(d^{1/2}(x) - m_\beta^{1/2}(x))^2$, $G(\delta) = (e^{-\delta} + \delta - 1)$ generates the negative exponential disparity (Lindsay 1994; Basu, Sarkar and Vidyashankar 1997) $NED(d, m_\beta) = \sum(e^{-\delta(x)} + \delta(x) - 1)m_\beta(x)$, and $G(\delta) = (\delta + 1)\log(\delta + 1)$ generates the likelihood disparity

$$LD(d, m_\beta) = \sum d(x) \log[d(x)/m_\beta(x)]. \tag{2.2}$$

The last disparity is minimized by the maximum likelihood estimator of β.

For continuous models, one constructs a nonparametric density estimator and minimizes its distance from the model density. Beran (1977) used a kernel density estimator. Given the empirical distribution function $\hat{F}(t)$ and a smooth kernel function $k(x;t,h)$, the kernel density estimate f^* is given by

$$f^*(x) = \int k(x;t,h)d\hat{F}(t).$$

The parameter h controls the smoothness of the corresponding density estimate. One can then minimize an appropriate disparity $\int G(\delta^*(x))m_\beta(x)dx$ between f^* and m_β; the smoothing parameter must tend to zero at the appropriate rate for the density estimator f^* to converge to m_β in the limit. Beran (1977) used this approach to find the minimum Hellinger distance estimate of β in continuous models. In this case $\delta^*(x) = f^*(x)/m_\beta(x) - 1$, and we will refer to this approach as the "Beran approach"; a similar approach was used by Basu, Sarkar and Vidyashankar (1997) for the negative exponential disparity.

In a departure from this approach, Basu and Lindsay (1994) integrated the model with the same kernel to obtain a smoothed version of $m_\beta(x)$. Thus,

$$m^*_\beta(x) = \int k(x;t,h)dM_\beta(t),$$

where M_β represents the distribution function corresponding to m_β. In this case the Pearson residual is defined as $\delta^*(x) = f^*(x)/m^*_\beta(x) - 1$. The minimum disparity estimators are then obtained by minimizing $\rho_G(f^*, m^*_\beta) = \int G(\delta^*(x))m^*_\beta(x)dx$. We will refer to this approach as the "Basu-Lindsay" approach. In this case, one gets consistent estimates even when the smoothing parameter h is kept *fixed*. (A more detailed discussion on the role of the smoothing parameter is provided at the end of Section 3). In addition, Basu and Lindsay have shown that in some cases, the kernel can be appropriately chosen so that the minimum disparity estimators are asymptotically fully efficient, the most prominent example being the normal kernel in the normal model.

To keep a clear focus here, we will concentrate on the Basu-Lindsay approach in this paper, and present our development and illustrations with this approach in mind. Our specific motivation in doing so is to demonstrate that the implementation of the

robust minimum disparity estimation scheme can be carried out eliminating the rate of convergence concerns related to bandwidth selection. The results apply equally to the Beran approach also; in the latter case, however, the experimenter has to take on the additional issue of choosing the bandwidth properly from the point of view of the convergence of densities, as discussed later in Section 3.

Under differentiability of the model, let ∇ represent the gradient with respect to β. Then, minimizing the disparity (2.1) is equivalent to solving the equation

$$\sum A(\delta(x))\nabla m_\beta(x) = 0 \tag{2.3}$$

where the function $A(\delta)$ equals $(1+\delta)G'(\delta) - G(\delta)$. (For continuous models (2.3) has a similar form involving integrals.) The properties of the function G allow us to center and scale A so that $A(0) = 0$ and $A'(0) = 1$, without changing the estimating properties of the disparity. This centered and scaled function A is called the residual adjustment function (RAF) of the disparity ρ_G. Lindsay (1994) has shown how the theoretical properties of the minimum disparity estimators are determined by the form of the RAF. The strict convexity of $G(\cdot)$ guarantees that $A(\cdot)$ is strictly increasing on $[-1, \infty)$.

For the likelihood disparity, in particular, $A(\delta) = \delta$ and the estimating equation becomes

$$\sum d(x)\frac{\nabla m_\beta(x)}{m_\beta(x)} = \sum \delta(x)\nabla m_\beta(x) = 0. \tag{2.4}$$

An analogous analysis is possible in continuous models for either of the two approaches.

3 A New Algorithm: Iteratively Reweighted Estimating Equation ($IREE$)

3.1 The Iteratively Reweighted Least Squares ($IRLS$)

We first discuss the $IRLS$ and then develop the $IREE$ along those lines. The $IRLS$ is an algorithm often used in determining the parameter estimates in robust regression. It is generally attributed to Beaton and Tukey (1974), and is far simpler to apply than the Newton-Raphson method. Holland and Welsch (1977), McCullagh and Nelder (1989)

and Green (1984) are good general references. Byrd and Pyne (1979) and Birch (1980) discuss convergence results and Del Pino (1989) provides an extensive bibliography.

Consider the standard regression model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

A robust estimate $\hat{\beta}$ of β is found by minimizing

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i\beta}{\sigma}\right),$$

where σ is a known or previously estimated scale parameter and X_i is the i-th row of X. Let ψ represent the first derivative of ρ. Then $\hat{\beta}$ satisfies the estimating equation

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{Y_i - X_i\hat{\beta}}{\sigma}\right) = 0, \tag{3.1}$$

for $j = 1, 2, \ldots, p$. Here x_{ij} is the j-th component of X_i. Solving this set of equations directly typically requires application of numerical methods.

Define the weight function $w(r)$ as $\psi(r)/r$; the equation (3.1) can be written as

$$\sum_{i=1}^{n} \left(\frac{Y_i - X_i\hat{\beta}}{\sigma}\right) w\left(\frac{Y_i - X_i\hat{\beta}}{\sigma}\right) x_{ij} = 0. \tag{3.2}$$

One can solve (3.2) iteratively using a weighted least squares algorithm. Let W_β be the $n \times n$ diagonal matrix whose i-th diagonal element is $w\left(\frac{Y_i - X_i\beta}{\sigma}\right)$. Then for a given starting value β_0, the first iteration yields

$$\beta_1 = (X^T W_{\beta_0} X)^{-1} X^T W_{\beta_0} Y. \tag{3.3}$$

This iteration scheme is continued till convergence is achieved. Note that if Y exactly fits the model, in the sense $Y = X\hat{\beta}$, then (3.2) converges in one step.

In numerical analysis terms, the $IRLS$ is using a fixed point method to solve (3.2). This is a simple algorithm to determine the roots of the equation $f(x) = 0$ in some interval (a, b). The conditions leading to its convergence are well known results of numerical analysis; among others, Ralston and Rabinowitz (1978), and Ortega (1990) are sources of detailed discussions on this subject.

However, since our development of the method described in this paper depends critically on the convergence mechanism of the fixed point algorithm, we briefly describe the same in the following. Consider a target value α, a starting value $x^{(0)}$, and a sequence $\{x^{(i)} : i = 1, 2, \ldots\}$ in $\mathbb{R}^p$. We will say that $x^{(i)}$ converges linearly to α if for a starting value sufficiently close to the target value there exists a constant $c \in (0, 1)$ such that

$$||x^{(i+1)} - \alpha|| \leq c||x^{(i)} - \alpha||,$$

where $||\cdot||$ denotes the Euclidean norm. The sequence converges quadratically if for a sufficiently close starting value there exists a constant c such that

$$||x^{(i+1)} - \alpha|| \leq c||x^{(i)} - \alpha||^2.$$

The fixed point iteration method can be used to determine the root of an equation $f(x) = 0$, when the equation has been written in the alternative form $x = F(x)$. First consider the univariate case, i.e. f is a real valued function of a single real variable x. Let $x = F(x)$ be the fixed point formulation of the equation $f(x) = 0$. In this case we start with an initial approximation $x^{(0)}$ and at the i-th stage perform the next iteration as $x^{(i+1)} = F(x^{(i)})$. Given that $x^{(i)}$ is the value at the i-th stage, we assume that $F(x)$ has a continuous derivative in the closed interval bounded by $x^{(i)}$ and the true solution α. Since $\alpha = F(\alpha)$, it follows that

$$x^{(i+1)} - \alpha = F(x^{(i)}) - F(\alpha) = (x^{(i)} - \alpha)F'(\zeta^{(i)}),$$

where $\zeta^{(i)}$ lies between $x^{(i)}$ and α. When the iteration converges $x^{(i)} \to \alpha$ and $F'(\zeta^{(i)}) \to F'(\alpha)$. Thus we get $x^{(i+1)} - \alpha \sim (x^{(i)} - \alpha)F'(\alpha)$, and hence also

$$x^{(i)} - \alpha \sim A[F'(\alpha)]^i,$$

for a constant A. Thus $|F'(\alpha)| < 1$ is a necessary condition for the iteration to be asymptotically stable. When $|F'(\alpha)| < 1$, and the initial value is sufficiently close to α, the sequence $x^{(i)}$ will converge to α. Since $x^{(i+1)} - \alpha = (x^{(i)} - \alpha)F'(\zeta^{(i)})$, sufficient closeness of $x^{(i)}$ to α and the continuity of $F'(x)$ at α will mean that

$$|x^{(i+1)} - \alpha| \sim |F'(\alpha)|\ |x^{(i)} - \alpha|,$$

and so $x^{(i)}$ converges to α at a linear rate.

It is well known that if $F'(\alpha) = 0$, then it leads to quadratic convergence for the fixed point method; in this case

$$\begin{aligned}(x^{(i+1)} - \alpha) &= F(x^{(i)}) - F(\alpha) \\ &= (x^{(i)} - \alpha)F'(\alpha) + \tfrac{1}{2}(x^{(i)} - \alpha)^2 F''(\zeta^{(i)}) \\ &= \tfrac{1}{2}(x^{(i)} - \alpha)^2 F''(\zeta^{(i)})\end{aligned}$$

where $\zeta^{(i)}$ is between $x^{(i+1)}$ and α. So if the method converges, the error in $x^{(i+1)}$ tends to be proportional to the square of the error in $x^{(i)}$.

The fixed point formulation $x_i = F_i(x_1, x_2, \ldots, x_p), \quad i = 1, \ldots, p$, has to be solved in the case where there are p unknowns. Let α be the true solution and let $x^{(j)} = (x_1^{(j)}, \ldots, x_p^{(j)})$ be the value at the j-th stage. For a suitably close starting value $x^{(0)}$, $x^{(j)} - \alpha \sim D^j z$ where $D = F'(\alpha)$, F' being the Jacobian matrix, and z a constant vector. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of D. The necessary condition for convergence now is that the spectral radius of D given by $\rho(D) = \max_i |\lambda_i|$ is less than 1. When the method converges, the rate of convergence is linear. However, as in the scalar case, the rate of convergence becomes quadratic when the matrix D is a null matrix or a nilpotent matrix (a matrix is nilpotent if some power of it is the null matrix).

3.2 The Standard *IREE*

Let us now examine how ideas similar to those in Section 3.1 can be used to solve for the roots of the minimum disparity estimating equations. Assume that $\Omega \in I\!R^p$. We are solving (in the Basu-Lindsay approach, for instance) the estimating equation

$$\int A(\delta^*(x)) \nabla m_\beta^*(x) dx = 0.$$

Assuming that $\int m_\beta^*(x)dx$ can be differentiated under the integral sign, we can write

$$\int (A(\delta^*(x)) - \lambda) m_\beta^*(x) \frac{\nabla m_\beta^*(x)}{m_\beta^*(x)} dx = 0$$

for any constant λ, or

$$\int w(x) \frac{\nabla m_\beta^*(x)}{m_\beta^*(x)} dx = 0, \tag{3.4}$$

where

$$w(x) = (A(\delta^*(x)) - \lambda)m^*_\beta(x). \tag{3.5}$$

This is a weighted version of the estimating equation of the likelihood disparity (see equation 2.4), just as (3.2) is a weighted version of the ordinary least squares.

Let $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, and ∇_i be the gradient with respect to β_i. If $m^*_\beta(x)$ is in the exponential family, a relationship of the form

$$\frac{\nabla_i m^*_\beta(x)}{m^*_\beta(x)} = K(\beta)[S_i(x,\beta) - \beta_i], \tag{3.6}$$

is often found to be true. It is always true if β represents the set of natural parameters. The function S_i may depend on β. Assuming that we have a relationship of the form (3.6), we can write the i-th equation of (3.4) as

$$\int w(x)[S_i(x,\beta) - \beta_i]dx = 0,$$

or

$$\beta_i = \frac{\int w(x)S_i(x,\beta)dx}{\int w(x)dx}, \tag{3.7}$$

and hence we arrive at the fixed point equation $\beta = F(\beta)$, where F is a function from $\mathbb{R}^p$ to $\mathbb{R}^p$. The iteration will be carried on till convergence to a specific level of tolerance is reached. We will refer to this algorithm as the iteratively reweighted estimating equation ($IREE$) algorithm.

This $IREE$ method does not require the evaluation of the second partial derivatives, and the inversion of the Hessian matrix. The discrete case is similar with integrals replaced with summations. In Theorems 3.1 and 3.2, we will show how to improve the rate of convergence of the algorithm when the functions S_i are independent of β.

Since the RAF $A(\delta)$ is increasing on $[-1, \infty)$, the weights $w(x)$ in (3.5) will be non-negative if we use $\lambda = A(-1)$. We will refer to this case as the standard $IREE$ (or $IREE$ with standard weights).

For the purpose of illustration, if $m^*_\beta(x)$ is the $N(\mu, \sigma^2)$ density, then letting $\beta = (\beta_1, \beta_2) = (\mu, \sigma^2)$, the fixed point equations for solving β_1 and β_2 are

$$\beta_i = \frac{\int w(x)S_i(x,\beta)dx}{\int w(x)dx}, \ i = 1, 2 \tag{3.8}$$

where $S_1(x,\beta) = x$, $S_2(x,\beta) = (x-\beta_1)^2$; this is easily verified by differentiating $\log(m^*_\beta(x)) = constant - (\log\beta_2)/2 - 0.5(x-\beta_1)^2/\beta_2$. Clearly, this simple technique can also be used to solve for the parameter estimates in the Beran approach, in which case $\delta^*(x)$ (and hence $w(x)$) is now a function of $m_\beta(x)$.

In the discrete case, the simplicity of the method is even more apparent. To illustrate this point let us look at the one parameter exponential family. Let $\mu = \beta$ be the mean and V be the variance for the model m_β. Then

$$\frac{\nabla m_\beta(x)}{m_\beta(x)} = \frac{(x-\mu)}{V}$$

where ∇ represents the gradient with respect to the mean value parameter. The $IREE$ will solve the equation

$$\sum w(x)\left\{\frac{(x-\mu)}{V}\right\} = 0$$

for μ. This gives us the fixed point equation for μ as

$$\mu = F(\mu) = \frac{\sum xw(x)}{\sum w(x)}. \tag{3.9}$$

In general, for a univariate parameter β,

$$F(\beta) = \frac{\sum S(x,\beta)w(x)}{\sum w(x)}, \tag{3.10}$$

if $\nabla m_\beta/m_\beta = K(\beta)[S(x,\beta) - \beta]$ for some functions K and S.

3.3 Comparison of the Standard $IREE$ and the Newton-Raphson Method

We now apply the $IREE$ (with standard weights) to the data set introduced in Beran (1977) to find the minimum Hellinger distance estimates ($MHDE$s) of the parameters for the Basu-Lindsay approach and repeat the calculations with the Newton-Raphson (NR) method to compare the rate of convergence of the methods. All the computations in this paper are performed using Splus on a SUN ULTRA 5 workstation at the Department of Statistics, Pennsylvania State University; for all the examples convergence was declared when the decrease in the disparity was less than 10^{-7}.

The data set in Beran is a pseudo random sample of size 40 generated from $N(0,1)$. Assuming the $N(\mu,\sigma^2)$ model, we compute the $MHDE$s of μ and σ^2 using the two methods. We use a normal kernel with bandwidth $h = 0.5$. We do this for the original sample as well as after replacing the 22nd observation (the value closest to 0) by several contaminating values. The initial estimates were $\hat{\mu}^{(0)} = median\{X_i\}$ and $\hat{\sigma}^{(0)} = (0.674)^{-1} median\{|X_i - \hat{\mu}^{(0)}|\}$, and the final results are given in Table 1. (Table 1 also has the corresponding numbers for another version of the $IREE$ which we will introduce later.)

As expected, the quadratically convergent NR algorithm converges substantially faster than the $IREE$. However it is fair to say that the convergence of the standard $IREE$ is moderately quick. Note that while the $IREE$ requires about 2.5 to 3.5 times the number of steps needed for the Newton-Raphson method to converge, overall it only requires just about double the number of numerical integrations or less compared to what is necessary for the NR. This is because at each step the $IREE$ requires only four numerical integrations whereas the NR method requires six integrations involving much more complex functions. In terms of convergence, each iteration of an NR step took approximately 2.5 times the amount of time necessary for each iteration of the $IREE$ algorithm considered here (as well as the optimal $IREE$ discussed in the next section) as determined by the user time component of the proc.time() function of Splus.

As noted earlier, as the number of parameters grow, the number of extra numerical integrations necessary for the NR method will grow rapidly. For the bivariate normal with $p = 5$ parameters, the NR will require three times as many numerical integrations as the $IREE$ at every step. Even discounting programming difficulty and matrix inversion, this makes the standard $IREE$ method competitive even at three times the number of steps. Note also that the second derivative functions being integrated in the Newton-Raphson case are far more complicated. In our bivariate normal example (reported later in Section 3.4), we observed that each Newton-Raphson iteration took about six times as much time as taken by a single iteration of the optimal $IREE$.

3.4 Optimally Weighted *IREE*

It may be possible to have some amount of control over the convergence of the *IREE*. For instance, being a linearly convergent algorithm, one could improve its rate of convergence using the Aitken acceleration (eg. Ralston and Rabinowitz, 1978). However a simpler and more effective modification to the *IREE* is often possible in exponential family models.

Consider a scalar parameter β, and let the weight function w of the *IREE* be as defined in equation (3.5). The standard *IREE* is obtained by replacing λ with $A(-1)$. This keeps the weights nonnegative. A nice improvement is possible if we allow negative weights. As discussed in Section 3.1 the convergence of the fixed point algorithm applied to the fixed point formulation (3.10) depends on the derivative of $F(\beta)$ at the solution, and the rate of convergence is quadratic if this derivative is zero. If $S(x,\beta) = S(x)$ is independent of β, direct differentiation of (3.10), combined with the result that at the solution $\beta = F(\beta) = \sum w(x)S(x)/\sum w(x)$, gives

$$F'(\beta) = \frac{\sum w'(x)(S(x)-\beta)}{\sum w(x)} \tag{3.11}$$

at the solution, where $w'(x) = \partial w(x)/\partial\beta$. Thus the form of the data will determine convergence properties. An important special case occurs when the data fit the model well.

Theorem 3.1 *Suppose that* $d(x) = m_\beta(x)$, *and* $\nabla m_\beta(x)/m_\beta(x) = K(\beta)[S(x)-\beta]$ *where* $S(x)$ *is independent of* β. *Then for* $\lambda = -1$, *we get* $F'(\beta) = 0$ *at the solution and thus the IREE converges at a quadratic rate.*

Proof: In this case the derivative $F'(\beta)$ at the solution is as in equation (3.11). By direct differentiation,

$$w'(x) = \frac{\partial A(\delta(x))}{\partial\delta(x)}\frac{\partial\delta(x)}{\partial\beta}m_\beta(x) + (A(\delta(x)) - \lambda)\nabla m_\beta(x),$$

and when $d(x) = m_\beta(x)$ we get $\delta(x) = 0$, $A(\delta(x)) = 0$, $\partial A(\delta(x))/\partial\delta(x) = 1$, and $\partial\delta(x)/\partial\beta = -\nabla m_\beta/m_\beta$, so that $w'(x) = -(1+\lambda)\nabla m_\beta$, which vanishes for $\lambda = -1$.

As a result, the right hand side of equation (3.11) vanishes as well, implying $F'(\beta) = 0$ at the solution if $\lambda = -1$. Thus, under the conditions of the theorem, the $IREE$ converges quadratically. □

In particular for the mean value parameter $\beta = \mu$ we get, at the solution,

$$F'(\mu) = \frac{\sum w'(x)(x-\mu)}{\sum w(x)}. \tag{3.12}$$

For the conditions of the above theorem we get $w(x) = -\lambda m_\beta(x)$ and

$$w'(x) = -\{\frac{(x-\mu)}{V}\}m_\beta(x) - \lambda\{\frac{(x-\mu)}{V}\}m_\beta(x),$$

and replacing these values in (3.12) we obtain $F'(\mu) = 1 + 1/\lambda$. At $\lambda = -1$, $F'(\mu) = 0$.

We will refer to the case where $\lambda = -1$ is used in the weight function $w(x)$ as the optimal $IREE$ (or the $IREE$ with optimal weights). We illustrate the performance of the optimal $IREE$ with an example. Taking m_β to be the $Poisson$ model with mean parameter β and letting $d(x)$ be the model vector for the $Poisson(2)$ distribution we minimize $HD(d, m_\beta)$ over β. By Fisher consistency, the final solution is $\beta = 2$, but we are interested in checking how many steps the methods require to converge to the true value when the iteration starts at some other starting value; in particular we used initial $\beta = 3$. Table 2 gives the performance of the NR and the $IREE$ (with standard as well as optimal weights). The optimally weighted $IREE$ is clearly far superior than the standard $IREE$ and comparable to the NR method (in fact it converges in 4 steps compared to 5 for the NR). Using the user time component of the proc.time() function of Splus, we observed that the Newton-Raphson Algorithm and the optimal $IREE$ algorithm took 50% and 36% of the time taken by the ordinary $IREE$ to converge.

The same modification to the $IREE$ can be made in the continuous case in the Basu-Lindsay approach if the smoothed model m^*_β is a one parameter exponential family model. We will now replace d by f^*, m_β by m^*_β, and the summations by integrals. It can be easily seen that it works for the Beran approach as well.

It will be most helpful if we can use the optimally weighted $IREE$ in multiparameter situations, as it is really more useful in such cases. The error in the i-th stage tends to

be described by $D^i z$, where D is the Jacobian matrix at the solution and z is a fixed vector. As described in Section 3.1, an enhancement in the rate of convergence of the $IREE$ similar to that in Theorem 3.1 can be obtained in this case if the weights can be chosen so that the Jacobian matrix at the solution is a null matrix or a nilpotent matrix. The following result is proved in the context of the Basu-Lindsay approach in continuous models, but holds for discrete models and the Beran approach as well.

Theorem 3.2 *Suppose that β is p-dimensional. Assume that the quantities $S_i(x, \beta)$ used in equation* (3.6) *are independent of β for each i, $i = 1, \ldots, p$. In such cases the $IREE$ will converge quadratically at the model ($f^*(x) = m^*_\beta(x)$) if we use $\lambda = -1$.*

Proof: Equation (3.7) can now be represented as

$$\beta_i = \frac{\int w(x) S_i(x) dx}{\int w(x) dx},$$

where $S_i(x)$ depends on depend on x only. The ij-th element of the Jacobian matrix D at the solution is

$$\frac{\int \nabla_j w(x)(S_i(x) - \beta_i) dx}{\int w(x) dx},$$

where ∇_j represents the gradient with respect to β_j. As in the unidimensional case, the above expression is 0 at the model when $\lambda = -1$, and this is true for all i and j, making the Jacobian matrix at the solution a null matrix. In this case, therefore, the optimally weighted $IREE$ will converge quadratically. □

In the univariate normal model, for example, if we use the parameterization $\beta_1 = E(X)$ and $\beta_2 = E(X^2)$ (instead of the (μ, σ^2) parameterization), we get $S_1(x) = x$ and $S_2(x) = x^2$, so that the above Jacobian matrix is a null matrix at the model. However we note that actual calculation shows that the Jacobian matrix is a null matrix at the model for $\lambda = -1$ in the (μ, σ^2) parameterization also – in fact this is true for the multivariate normal density for any dimension, showing that the condition $S_i(x, \beta)$ be independent of β is not necessary, although sufficient. We then employ the optimal $IREE$ to determine the $MHDE$s of the parameters (using the Basu-Lindsay approach) for Beran's data set, already analyzed by the NR method and the standard $IREE$ in Section 3.3. The results

are available in Table 1, which now gives a comprehensive picture of the comparison of the three methods. The performance of the optimal $IREE$ is superior or comparable to the NR in terms of the number of numerical integrations necessary for each of these cases.

Some comments are necessary here about the small decrease in efficiency of the optimal $IREE$ when X_{22} is in the range 3 to 5. Notice that the optimal $IREE$ is a quadratically convergent algorithm only *at the model*. In terms of real data examples this means that the algorithm will perform best when the data roughly follow the pattern dictated by the model. Thus for small positive values of X_{22} the algorithm performs well as this observation, together with the rest of the data, is not inconsistent with a normal model. As X_{22} starts getting larger the observation looks more and more like an outlier inconsistent with the rest of the data and the normal model, and the optimal $IREE$ requires more steps to converge. In fact the actual minimum Hellinger distance estimators (not reported here) also are affected most by the mid sized outlier like 3, 4 and 5. However when the outlier becomes unacceptably large (say 6 or larger in this case) most robust minimum disparity estimators would be able to clearly distinguish it as such and downweight it almost entirely, and the performance of the estimator (as well as the $IREE$ algorithm) would now be governed primarily by the majority of the data (excluding the outlier) which follow the model closely. For large outliers and robust initial estimates, the weights for values of X around the outlier are practically equal to zero (either for optimal or standard $IREE$), so that in extreme cases the algorithm works as if the outlier was simply not there, and the algorithm converges quickly. Basu and Lindsay (1994, Figure 4) provide an example which demonstrates that the effect of an outlier on the $MHDE$ quickly dissipates as the outlier becomes unusually large.

Next we present an example where the data were generated from a bivariate normal, and compare the rates of convergence of the optimally weighted $IREE$ and the Newton-Raphson method. The pseudo random sample generated from the $BVN(0, 0, 1, 1, 0)$ distribution is presented in Basu (1991). The $MHDE$s of the five parameters of the bivariate normal (two means, two variances and the covariance) were calculated using the Basu-Lindsay approach, and the bivariate normal kernel.

The results are presented in Table 3. Two sets of starting values were used, the true parameter values 0, 0, 1, 1, 0, and the $UMVUE$s of the parameters. It is quite apparent that the optimal $IREE$ is very competitive in terms of the number of steps necessary and far superior in terms of the number of numerical integrations required. Roughly speaking, the amount of code that had to be generated for the NR method was more than double the amount necessary for the $IREE$.

However, the authors feel that even that does not accurately quantify the amount of simplicity the $IREE$ brings in to this optimization problem. It is not easy for the reader to get a full idea of the comparison of the two methods without actually programming the two methods for the same problem, but it is the view of the authors that in the bivariate normal problem the NR is far worse than "twice as difficult" to program as the $IREE$, given the coding, debugging and convergence obstacles. In particular the use of the user time component of the proc.time() function in Splus showed that one iteration for the Newton-Raphson algorithm in the bivariate normal example requires approximately six times the computer processing time necessary for one iteration of the the optimal $IREE$ algorithm.

At this stage we must address the very important issue of selecting the bandwidth h. The choice of the bandwidth is, by itself, an important problem in kernel density estimation. This is because the smoothing introduces a bias in the density estimate, and the bandwidth h must go to zero at the appropriate rate as a function of the sample size so that this bias is asymptotically zero. In addition, there are other efficiency considerations which have to be addressed to generate an optimal rate for choosing the bandwidth.

However in the Basu-Lindsay approach – on which we have focused in this paper – the proper rate of bandwidth selection for the convergence of the density estimate to the true density is not a critical issue for the following reason: since here the model is also smoothed, the bias that is introduced in the data due to smoothing, is also introduced in the model through the same smoothing. Thus it is no longer necessary to let the bandwidth go to zero. For any *fixed* bandwidth, the density estimate converges to a biased version of the true density. However, because the model has been smoothed, it is

this biased version of the true density (assuming it is in the model) that is our target, and not the true density itself. Thus, instead of adjusting the bandwidth to make the density estimate converge to the true model density, we shift the model density to the biased version to which the kernel density estimate converges for that fixed value of h.

This of course would lead to a problem if our aim was density estimation per se. But our real aim is the estimation of the unknown parameter, and density estimation is just an intermediate tool that we have to use. The smoothed version of the model is the function of the same set of parameters as was the original model density. Thus parameter estimates obtained by minimizing distances between the kernel density estimate and smoothed versions of the model are consistent for *fixed* values of the smoothing parameter. Basu and Lindsay (1994) provide details of this method of estimation. Avoiding the problem of bandwidth selection was, in fact, one of the main motivations of their work.

Therefore, the choice of the smoothing parameter in the Basu-Lindsay approach does not have to be dictated by the consideration that the density estimate must converge to the true data generating density. The density estimate converges to the biased version of the model density in any case. Instead, the choice of the smoothing parameter is governed by the considerations of robustness and numerical stability of the algorithm. In this connection note that as the value of h increases, the smoothing begins to have a bigger impact over the resulting densities, and for very big values of the smoothing parameter the resulting smoothed empirical and the smoothed model density begin to look alike. As a result the $A(\delta)$ values tend to get closer to 0, and the estimating equation in (3.4) begins to look more and more like that of the likelihood disparity. Notice that if $\delta^*(x) = 0$ for all x, the estimating equation coincides with that of the likelihood disparity and the $IREE$ algorithm converges in one step.

Thus our expectation is that the estimators will get closer to the MLEs for larger h, and the convergence will also become faster (which is also our observation in numerical studies). In Table 3 one can see that the methods converge faster for larger h. What has not been reported (but is true) is that the estimators also tend towards the MLE as h increases. However with increasing h the robustness will get weaker. We take the view

that the choice of h should primarily be guided by robustness considerations. Algorithmic considerations such as faster convergence are also important but should not be achieved by compromising the robustness aspect. Choosing small values of h, however, should also be done with caution. Extremely small values of h can make the smoothed empirical very spiky unless the sample size is very large, possibly making the objective function a badly behaved one. In general the choice of h should be related to the scale of the data, and choosing h to be a constant multiple of an equivariant estimate of scale makes the method location scale equivariant (Basu and Lindsay, 1994, Section 7.1). Typically one should choose a robust initial estimate of scale when applying the above idea.

For the Beran approach, there is no equivalent smoothing in the model to compensate for the extra smoothing in the data when the smoothing parameter increases. As a result the estimates of the scale get inflated when there is an increase in the smoothing parameter (see Beran 1977, Table 1). In this case one could still choose the smoothing parameter as $c_n s_n$, where s_n is a robust estimate of scale, and c_n is a sequence of real numbers satisfying condition (v), Theorem 4 (Beran 1977). From the point of view of convergence of densities, the optimal bandwith for univariate data is proportional to $n^{-1/5}$ in the sense of minimizing the mean integrated square error (see, for example, Silverman, 1986). The books by Devroye and Gyorfi (1985) and Devroye (1987) also provide many details of these methods. Also see Härdle et al. (1988), Marron (1989) and Hall and Marron (1991), as well as Schimek (2000) for the most recent developments in this field. Cao and Devroye (1996) provides another interesting approach. However, the guidelines provided by the above authors must be combined with the robustness issue in these problems, and precise recommendations of bandwidth selections in this respect will require an extensive study beyond the scope of this paper.

4. Some Additional Issues

4.1 Second Order Analysis for the Optimally Weighted $IREE$

Let $A_2 = A''(0)$ represent the second derivative of the residual adjustment function

of the disparity evaluated at zero. Lindsay (1994) and Basu and Lindsay (1994) have shown that this plays an important role in determining the theoretical properties of the estimator. In this section we will show that the right hand side of equation (3.12) can be expressed as a function of A_2 when the residuals are small.

Direct differentiation of $w(x)$ gives

$$w'(x) = -[A'(\delta(x))d(x) - m(x)(A(\delta(x)) - \lambda)]u(x)$$

where $u(x) = \nabla m(x)/m(x)$ is the score function. Replacing this in the numerator of the right hand side of (3.12) gives us

$$\begin{aligned}\sum w'(x)(x-\mu) = -\{&\sum A'(\delta(x))d(x)u(x)(x-\mu) \\ &- \sum m(x)u(x)(x-\mu)[A(\delta(x)) - \lambda]\}.\end{aligned}$$

If the δs are small, so that we can write $A(\delta) \cong \delta + A_2\delta^2/2$ and $A'(\delta) \cong 1 + A_2\delta$, the above equation reduces to

$$\begin{aligned}\sum w'(x)(x-\mu) = -\{&A_2[\sum \delta^2(x)u(x)(x-\mu)m(x)/2 \\ &+ \sum \delta(x)u(x)(x-\mu)m(x)]\} - (1+\lambda).\end{aligned}$$

Similarly,

$$\begin{aligned}\sum w(x) &= \sum m(x)A(\delta(x)) - \lambda \\ &= \sum m(x)[\delta(x) + A_2\delta^2(x)/2] - \lambda \\ &= A_2[\sum m(x)\delta^2(x)/2] - \lambda.\end{aligned}$$

Thus at $\lambda = -1$, we have, by replacing the above expressions in (3.12)

$$F'(\mu) = \frac{A_2 \sum m(x)[\delta^2(x)/2 + \delta(x)](x-\mu)^2/V}{A_2 \sum m(x)\delta^2(x)/2 + 1}. \tag{4.1}$$

For disparities which have $A_2 = 0$, the optimally weighted $IREE$ will behave like a quadratically convergent algorithm for small residuals δ. Even when $A_2 \neq 0$, as $n \to \infty$, the numerator of (4.1) converges to 0 and the denominator converges to 1, so that for large n, the rate $F'(\mu)$ converges to the optimal value 0.

4.2 A Note of Caution in Using the $IREE$ when the Sample Size is Small

In this section we will investigate the convergence of the $IREE$ (as a function of λ) when the sample size is 1, and show that small samples may require conservative choice of λ. Letting X_{obs} represent the single observation, the discrete model version of equation (3.4) equals

$$\sum_{x \neq X_{obs}} [A(-1) - \lambda] \nabla m(x) + [A(1/m(X_{obs}) - 1) - \lambda] \nabla m(X_{obs}) = 0. \tag{4.2}$$

If $\lambda = A(-1)$, the above equation simplifies to $\nabla m(X_{obs}) = 0$ and convergence is obtained in one step. Thus in this case, the standard weights $\lambda = A(-1)$ are optimal. Note that the choice of the standard weights reduces the estimating equation in (4.2) to a sum of over a single point.

Equation (3.9) can now be rewritten as

$$\mu = F(\mu) = \frac{\sum x[A(\delta(x)) - \lambda]m(x)}{\sum [A(\delta(x)) - \lambda]m(x)} = \frac{N}{D} \tag{4.3}$$

where

$$N = A(-1)[\mu - X_{obs} m(X_{obs})] - \lambda\mu + X_{obs} m(X_{obs}) A(1/m(X_{obs}) - 1)$$

and

$$D = A(-1)[1 - m(X_{obs})] - \lambda + m(X_{obs}) A(1/m(X_{obs}) - 1).$$

Direct differentiation of (4.3) gives us, at the true solution,

$$F'(\mu) = \frac{A(-1) - \lambda}{A(-1) - \lambda + m(X_{obs})[A(1/m(X_{obs}) - 1) - A(-1)]}.$$

Since the quantity within square brackets in the denominator is necessarily positive, λ must satisfy

$$\lambda < A(-1) + \{m(X(_{obs}))[A(1/m(X_{obs}) - 1) - A(-1)]\}/2$$

to achieve convergence. The upper bound of the acceptable values of λ may be smaller than -1, and in such cases the $IREE$ will not converge for $\lambda = -1$. For small samples,

therefore, it may be safer to choose values of λ closer to $A(-1)$ to guarantee convergence. This presents no difficulty for disparity measures such as the Pearson's chi-square and the negative exponential disparity where $A(-1) > -1$, so that $\lambda = -1$ is in the safe range.

4.3 Weighted Likelihood Estimation

An investigation of this reweighting scheme leads to the development of an attractive estimation procedure in continuous models. The method has been studied in detail by Markatou et al. (1998). Here we briefly describe how it follows from the idea of the iterative reweighting algorithm. In continuous models, solving the estimating equation (3.4) requires numerical evaluation of integrals. Assume that the density estimate $f^*(x) > 0$ over the whole sample space, as is the case when using a kernel like the normal. Equation (3.4) then looks like $\int v(x)(\nabla m^*_\beta(x)/m^*_\beta(x))f^*(x)d(x) = 0$ where $v(x) = (A(\delta(x)) - \lambda)m^*_\beta(x)/f^*(x)$. If we keep the $v(\cdot)$ part intact in the above equation and replace the smoothed quantities elsewhere with their unsmoothed versions, we get

$$\int v(x)\frac{\nabla m_\beta(x)}{m_\beta(x)}dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} v(X_i)u_\beta(X_i) = 0,$$

where F_n is the empirical distribution function, and u_β is the likelihood score function. Thus we have a *sum* over the observed data, rather than an *integral* over the entire support. At the model the $v(X_i)$'s all converge to a constant as the sample size increases; asymptotically the estimating equation behaves like the likelihood equation. However, when the $v(X_i)$'s arise from robust disparities like the Hellinger distance, $A(\delta) \ll \delta$ for large positive δ, and the $v(\cdot)$ function can severely downweight large outliers.

The same technique works for the Beran approach as well, although in this case one has to let the bandwidth go to zero as $n \to \infty$ to get consistency.

Appendix

Here we provide a step by guideline for the implementation of the algorithm in the continuous case with the Basu-Lindsay approach. The discrete case can be handled similarly. The Beran approach will require additional considerations.

1. Decide on the parametric model. Although the method will work in other cases as well, the biggest benefit of the method will come when the likelihood score function of the smoothed model admits of relationships of the form (3.6).

2. Create the smoothed empirical density by choosing an appropriate kernel function. For the multivariate normal model, choose the multivariate normal kernel. Choose the bandwidth to be a multiple of a robust equivariant scale estimator.

3. Choose a robust starting value $\beta^{(0)}$. For the univariate normal model one can choose $\hat{\mu}^{(0)} = median\{X_i\}$ and $\hat{\sigma}^{(0)} = (0.674)^{-1} median\{|X_i - \hat{\mu}^{(0)}|\}$ as the starting values.

4. Create the smoothed model density and construct the Pearson residuals δ. Use the same kernel applied to item 2 above to construct the smoothed model density.

5. Choose an appropriate RAF to create the weight functions. In this paper we have based all our calculations on the Hellinger distance, but in practice any of several robust disparities may be reasonable choices (see Lindsay 1994).

6. Choose the tuning parameter λ. Notice that $\lambda = -1$ is algorithmically optimal, but conservative choices of λ closer to $A(-1)$ may be preferable for small sample sizes.

7. Create weights $w(x)$ as in (3.5) and solve the corresponding weighted likelihood estimating equation assuming the weights to be fixed constants to get the next iterate.

8. Repeat steps 4-7 with the current iterate until an appropriate convergence criterion has been satisfied. The form of the RAF used and the tuning parameter λ does not change from iteration to iteration.

Bibliography

Basu, A. (1991). *Minimum disparity estimation in the continuous case: efficiency, distributions, robustness and algorithms.* Ph.D dissertation, The Pennsylvania State University, University Park, PA 16802, USA.

Basu, A., and Basu, S. (1998). "Penalized minimum disparity methods for multinomial models". *Statistica Sinica*, **8**, 1998, 841–860.

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. (1997). "Minimum distance estimation: the approach using density based distances". In *Handbook of Statistics*, Vol. **15**, G. S. Maddala and C. R. Rao, (Ed.), North Holland, Amsterdam, 21–48.

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

Basu, A., Sarkar, S., and Vidyashankar, A. N. (1997). "Minimum negative exponential disparity estimation in parametric models". *J. Statist. Plan. Inf.*, **58**, 349–370.

Beaton, A. E., and Tukey, J. W. (1974). "The fitting of power series, meaning polynomials, illustrated on band spectroscopic data". *Technometrics* **16**, 147–185.

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

Birch, J. B. (1980). "Some convergence properties of iterated least squares in the location model". *Commun. Statist.* **B 9**, 359–369.

Boos, D. D. (1981). "Minimum distance estimators for location and goodness-of-fit". *J. Amer. Statist. Assoc.* **76**, 663-670.

Byrd, R. H., and Pyne, D. A. (1979). "Some results on the convergence of the iteratively reweighted least squares". *ASA Proc. Statist. Comput. Sec.*, 87–90.

Cao, R., Cuevas, A., and Fraiman, R. (1995). "Minimum distance density-based estimation". *Comput. Statist. Data Analy.* **20**, 611–631.

Cao, R. and Devroye, L. (1996). "The consistency of a smoothed minimum distance

estimate." *Scand. J. Statist.* **23** , 405–418.

Del Pino, G. E. (1989). "The unifying role of the iterative generalized least squares in statistical algorithms (with discussions)". *Statist. Science*, **4**, 394–408.

Devroye, L. (1987). *A course in density estimation*. Birkhauser, Boston.

Devroye, L., and Gyorfi (1985). *Nonparametric density estimation: The L_1 view*. John Wiley, New York.

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

Donoho, D. L., and Liu, R. C. (1988b). "Pathologies of some minimum distance estimators". *Ann. Statist.* **16**, 587–608.

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

Green, P. J. (1984). "Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussions)". *J. Roy. Statist. Soc.* B. 46, 149–192.

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

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

Heathcote, C. R., and Silvapulle, M. J. (1981). "Minimum mean squared estimation of location and scale parameters under misspecifications of the model". *Biometrika* **68**, 501–514.

Hettmansperger, T. P., Hueter, I. and Hüsler, J. (1994). "Minimum distance estimators". *J. Statist. Plan. Inf.* **41**, 291–302.

Holland, P. W., and Welsch, R. E. (1977). "Robust Regression using Iteratively Reweighted Least Squares". *Commun. Statist.* **A 6**, 813–827.

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

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models*, 2nd Ed. Chapman and Hall, New York.

Markatou, M. (1996). "Robust statistical inference: Weighted likelihood or usual *M*-estimation?". *Communications in Statistics: Theory and Methods*, **25**, 2597–2613.

Markatou, M., Basu, A., Lindsay, B. G. (1998). "Weighted likelihood equations with bootstrap root search". *J. Amer. Statist. Assoc.* **93**, 740–750.

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

Ortega, J. M. (1990). *Numerical Analysis–a Second Course.* Society for Industrial and Applied Mathematics, Philadelphia.

Özturk, Ö. (1994). *Minimum distance estimation.* Ph.D. dissertation, The Pennsylvania State University, University Park, PA 16802, USA.

Özturk, Ö., and Hettmansperger, T. P. (1997). "Generalized Cramér-von Mises distance estimators". *Biometrika* **84**, 283–294.

Özturk, Ö., Hettmansperger, T. P., and Hüsler, J. (1999). "Minimum distance and non-parametric dispersion functions". In *Asymptotics, Nonparametrics, and Time Series*, Subir Ghosh Ed., Marcel Dekker, New York, 511-531.

Parr, W. C. (1981)."Minimum distance method: a bibliography". *Commun. Statist. Theory. Meth.* **A10**, 1205–1224.

Parr, W. C., and De Wet, T. (1981)."On minimum Cramér-von Mises-norm parameter estimation". *Commun. Statist. Theory. Meth.* **A10**, 1149–1166.

Parr, W. C., and Schucany, W. R. (1980).“Minimum distance and robust estimation”. *J. Amer. Statist. Assoc.* **75**, 616–624.

Ralston, A., and Rabinowitz, P. (1978). *A First Course in Numerical Analysis.* McGraw-Hill, New York.

Schimek, M. G., Editor, (2000). *Smoothing and Regression: Approaches, Computation and Application.* John Wiley and Sons, New York.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

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

Simpson, D. G. (1989). “Hellinger deviance tests: efficiency, breakdown points and examples”. *J. Amer. Statist. Assoc.* **84**, 107-113.

Stather, C. R. (1981). *Robust Statistical Inference using Hellinger Distance Methods.* Ph.D dissertation, LaTrobe University, Melbourne, Australia.

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

Wiens, D. P. (1987).“Robust weighted Cramér-von Mises estimators of location with minimax varinace in ϵ-contamination neighborhoods”. *Canad. J. Statist.* **15**, 269–278.

Wolfowitz, J. (1953). “Estimation by the minimum distance method”. *Ann. Inst. Statist. Math.* **5**, 9–23.

Wolfowitz, J. (1954). “Estimation by the minimum distance method in nonparametric difference equations”. *Ann. Math. Statist.* **25**, 203–217.

Wolfowitz, J. (1957). “The minimum distance method”. *Ann. Math. Statist.* **28**, 75–88.

Table 1: Comparison of the NR method and the $IREE$ for Beran's data. The $MHDE$s have been obtained by replacing X_{22} with several contaminating values

	Original sample	$X_{22}=2$	$X_{22}=3$	$X_{22}=4$	$X_{22}=5$	$X_{22}=10$
No. of steps in which the NR converged	3	4	4	4	4	3
No. of steps in which the standard $IREE$ converged	10	10	12	14	13	9
No. of steps in which the opt. $IREE$ converged	4	4	5	6	5	3
Total no. of num. int. for the NR	19	25	25	25	25	19
Total no. of num. int. for the standard $IREE$	41	41	49	57	53	37
Total no. of num. int. for the opt. $IREE$	17	17	21	25	21	13

Table 2: Comparison of the NR method and the $IREE$ at the $Poisson$ model

Iteration #	β: NR method	β: standard $IREE$	β: optimal $IREE$
0	3.000000	3.000000	3.000000
1	1.461136	2.449490	1.838822
2	1.901533	2.213364	1.996883
3	1.996449	2.103979	1.999999
4	1.999995	2.051331	2.000000
5	2.000000	2.025503	.
.	.	.	.
10	.	2.000792	.
11	.	2.000396	.
12	.	2.000179	.

Table 3: Comparison of the NR method and the $IREE$ for bivariate normal data.

Starting values		$h = 0.5$	$h = 0.6$	$h = 0.7$	$h = 0.8$
True Parameters	No. of steps in which the NR converged	6	5	5	5
True Parameters	No. of steps in which the opt. $IREE$ converged	6	5	5	5
True Parameters	Total no. of num. int. for the NR	127	106	106	106
True Parameters	Total no. of num. int. for the opt. $IREE$	43	36	36	36
$UMVUE$s	No. of steps in which the NR converged	3	3	3	3
$UMVUE$s	No. of steps in which the opt. $IREE$ converged	4	4	3	3
$UMVUE$s	Total no. of num. int. for the NR	64	64	64	64
$UMVUE$s	Total no. of num. int. for the opt. $IREE$	29	29	22	22