Weighted Likelihood Equations With Bootstrap Root Search

Marianthi MARKATOU, Ayanendranath BASU, and Bruce G. LINDSAY

We discuss a method of weighting likelihood equations with the aim of obtaining fully efficient and robust estimators. We discuss the case of continuous probability models using unimodal weighting functions. These weighting functions downweight observations that are inconsistent with the assumed model. At the true model, therefore, the proposed estimating equations behave like the ordinary likelihood equations. We investigate the number of solutions of the estimating equations via a bootstrap root search; the estimators obtained are consistent and asymptotically normal and have desirable robustness properties. An extensive simulation study and real data examples illustrate the operating characteristics of the proposed methodology.

KEY WORDS: Asymptotic efficiency; Density estimation; Influence function; Maximum likelihood; Residual adjustment function; Robustness.

1. INTRODUCTION

We propose a simple method for modifying maximum likelihood that is designed to be used for diagnostics, or possibly inference, in situations where the model is in doubt. The method is based on reweighted likelihood equations. It provides 100% efficient estimators when the model is correct, robust estimators under mild failures of the model as well as diagnostics for identifying data points that are discrepant with the model, and a set of exploratory tools for gross model failure. The method has a 50% breakdown property.

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the density $m_\beta(x)$ corresponding to probability measure M_β. Let $u(x;\beta) = \nabla \ln[m_\beta(x)]$ be the score function, where ∇ denotes differentiation with respect to β. Under regularity, the maximum likelihood estimator (MLE) of β is a solution of the likelihood equation $\sum u(X_i;\beta) = 0$.

Given any point t in the sample space, we construct a weight function $w(t; M_\beta, \hat{F})$ that depends on t, the chosen model distribution M_β, and the sample empirical distribution $\hat{F}$. By our construction, the weight function will take values in [0, 1]. We then consider solutions $\hat{\beta}_w$ to the *weighted likelihood equations* (WLEs),

$$\sum_i w_i u(X_i;\beta) = 0, \qquad (1)$$

where $w_i = w(X_i; M_\beta, \hat{F})$. The weight function will be 1 or nearly so if in a neighborhood of the data point X_i, the data $\hat{F}$ is concordant with the model M_β, and will decline to 0 depending on the degree of their discordance. The solutions are called WLE estimators (WLEEs). The simple form of (1) provides an important motivation for this approach, as it suggests a natural algorithm based on iterative reweighting. In addition, when we are done, the final fitted weights indicate which of those data points were downweighted in the final solution, relative to the MLE.

Such an approach is not new, having been introduced by Green (1984). But the methods that we develop here are based on new weight functions designed to achieve both optimal model efficiency and strong robustness features. In addition, we provide a thorough approach to the problem of multiple solutions, and introduce an algorithm that incorporates a bootstrap root search. Some further discussion of earlier weighted likelihood methods is in the concluding remarks.

The article is organized as follows. Section 2 presents a multivariate normal example that displays the general nature of the methodology. Section 3 describes construction of the weight functions, which are based on ideas that arise in the theory of minimum disparity estimation (Lindsay 1994). Sections 4 and 5 describes the algorithmic methods that we use to find all of the roots to the WLEs. With the description of methods completed, Section 6 presents a simulation study that illustrates how the method works. Section 7 and 8 follow with the mathematical theory that describes the efficiency and robustness properties. The theoretical results are then substantiated by a simulation study.

2. A MULTIVARIATE EXAMPLE

We examined data provided by Lubischew (1962) that concern two species of beetles, *Chaetocnema concinna* and *Chaetocnema heptapotamica.* The data consist of two measurements per beetle: the maximal width and front angle of the aedeagus (male copulative organ). The two species are difficult to distinguish visually, but a careful look at the distribution of these measurements shows a relatively clear pattern of separation. Figure 1 shows the data for 21 measurements from *C. concinna* and 22 measurements from *C. heptapotamica.*

Marianthi Markatou is Associate Professor, Department of Statistics, Columbia University, New York, NY 10027. Ayanendranath Basu is Lecturer, Applied Statistics Unit, Indian Statistical Institute, Calcutta 700 035, India. Bruce G. Lindsay is Distinguished Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802. The work of M. Markatou was partially supported by National Science Foundation (NSF) grant DMS-9008846, RGK Foundation grants RGK CU01583501, and CU01610301, and NSF DMS-9214497 grant to Stanford University (principal investigator Ingram Olkin). The work of A. Basu was partially supported by a Mathematics Department Research Award from the University of Texas at Austin. The work of B. G. Lindsay was partially supported by NSF grant DMS-9403847. The authors thank E. Ronchetti and Ingram Olkin for their comments on an earlier draft of the article.

Journal of the American Statistical Association
June 1998, Vol. 93, No. 442, Theory and Methods

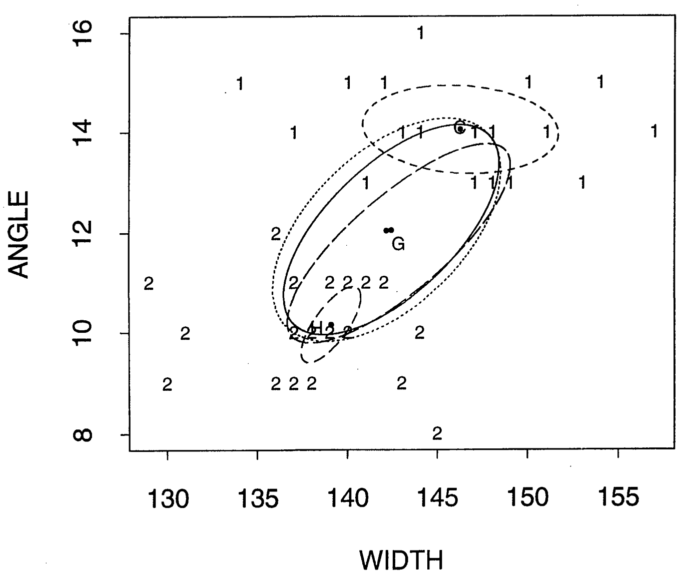

Figure 1. Plot of the Different Roots of the Flea Beetle Dataset (1, Ch. concinna; 2, Ch. heptapotamica). C and H are the sample means of group 1 and 2; G is the MVE location estimate. ——, MLE-like root; · · ·, MLE estimate; - - -, root1; ---, root2; — — —, MVE estimate.

We treat the pooled data as a single sample from a multivariate normal distribution, with parameters μ and Σ, to demonstrate the use of our methodology as a diagnostic of important data structures. We selected tuning parameters so that for true multivariate normal data, the mean downweighting would be 23% of the sample size. When we applied our root search procedure with our WLE method, we found three important roots. One root gave location estimates of (142.395, 12.052), which was very similar to what would be obtained from the MLE, which gave location estimates of (142.139, 12.046). The data points all had weights greater than .5 except for two: the point near (134, 15), with a final weight of .11, and the one near (145, 8), with a weight of .30. The 50% contour ellipsoids of the two parameter fits are also plotted.

Each of the other two roots agreed closely with one of the two species in the sense that the weights were nearly 0 for one or the other species. The corresponding mean estimators are plotted together with the fitted density contours. These location estimates are (139.064, 10.154) and (146.215, 14.057). Additionally, the actual sample means for the two species are (138.272, 10.091) for the *heptapotamica* group and (146.190, 14.095) for the *concinna* group. When *heptapotamica* was used mixed with three data points from *concinna,* so as to introduce a 12.5% contamination, the observation from *concinna* received an approximate weight of 0, and the only root obtained was that corresponding to *heptopotamica.*

As an alternative, one might consider the minimum volume ellipsoid (MVE) methodology, which is designed to determine—robustly—multivariate location and scale (Rousseeuw and Leroy 1987). These estimators are calculated using the S-PLUS code cov.mve. The point estimator of location is plotted, together with the ellipsoid corresponding to 50% normal probability. The MVE estimates of location are (142.805, 11.777). These estimators essentially match those of the MLE-like root. The MVE and MLE-like estimates were qualitatively the same in describing the parameters and the outliers, but we found evidence for data substructures through secondary roots. However, we did not perfectly recapture the original parameters of the two species, as can be seen from the ellipsoid associated with the *concinna* group.

3. THE WEIGHT FUNCTIONS

The particular weighting functions that we propose are based on the existing theory for the robustness and efficiency of minimum disparity estimation, as presented by Basu and Lindsay (1994) and Lindsay (1994). Indeed, if the sample space is discrete, then the methodology offered here has already been proposed and investigated by Lindsay (1994) and Markatou, Basu, and Lindsay (1997). But the method for weight construction for the continuous case that we propose here is new, and it can be motivated by the results for robust minimum disparity estimation provided by Basu and Lindsay (1994). We compare the WLE method to the latter in Section 3.3.

3.1 The Residuals

The downweighting scheme that we propose is based on a special system of residuals. We start with the discrete case. Suppose that the sample space χ is countable and, without loss of generality, let $\chi = \{0, 1, 2, \ldots\}$. Let m_β be a family of probability mass functions on χ. Given $X_1, X_2, \ldots, X_n$, a random sample from m_β, let $d(t)$ be the proportion of observations with value t. Define the Pearson residual $\delta(t)$ to be $\delta(t) = d(t)/m_\beta(t) - 1$, so named because $P^2 = n\sum_t m_\beta(t)\delta^2(t)$ is the Pearson chi-squared statistic for the goodness of fit when the model $m_\beta(t)$ is multinomial.

The Pearson residuals range in the interval $[-1, \infty)$, with $\delta(t) = -1$ only when $d(t) = 0$; that is, cell t is empty. The residual $\delta(t)$ equals 0 when the observed proportion equals the probability of observing t under the model. When the model is correctly specified, the residuals converge to 0 with probability 1.

The foregoing residuals are not suitable in the case where the model is continuous, as the empirical distribution and the model have incompatible densities. To overcome this problem, Basu and Lindsay (1994) proposed the following extension of the discrete methodology. First, construct a nonparametric kernel density estimate f^*, say

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

where k is a smooth family of kernel functions with parameter h, such as the normal densities with mean t and standard deviation h. The window parameter h controls the smoothness of the resulting density, with increasing h corresponding to greater smoothness. Next, apply the same smoothing to the model to get the smoothed density

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

As in the discrete case, one can then construct the Pearson residual $\delta^*(x) = f^*(x)/m^*_\beta(x) - 1$. The advantage of smoothing both data and model, as opposed to the usual tactic of smoothing the data only, is that the residual δ^* at any fixed x will converge with probability 1 to 0, when the model is correctly specified, *even if the window h is kept fixed.*

3.2 The Weight Functions

The weight w_i given to observation X_i has the form $w(\delta(X_i))$ in the discrete case and $w(\delta^*(X_i))$ in the continuous case, where $w(\delta)$ is a prespecified weight function. For the sake of simple interpretation of the final weights, it is natural that the maximal weight be 1, occurring when the residual is 0, and that the $w(\delta)$ decline smoothly to 0 as $\delta \to -1$ or $\delta \to +\infty$, so that larger residuals are downweighted more. For brevity, we call such a weight function *w unimodal.* A unimodal weight function that arises from a minimum chi-squared problem that we discuss later is

$$w(\delta) = 1 - \frac{\delta^2}{(\delta+2)^2}. \tag{2}$$

If a weight is unimodal, then one can perform exploratory data analyses with increasing degrees of outlier downweighting simply by taking powers w^p of the original weight function.

We now turn to our motivation for using this method of weight construction.

3.3 Minimum Disparity Estimation

In the discrete case, following Lindsay (1994), given any convex function $G(\cdot)$, a measure of the disparity between the model m_β and the data d is given by the expression

$$\rho_G(d, m_\beta) = \sum_t m_\beta(t) G(\delta(t)). \tag{3}$$

For strictly convex G, the disparity measure is nonnegative, by Jensen's inequality, equaling 0 only when the densities d and m_β are equal. Through appropriate selection of G, a large class of important divergences and distances can be developed in this manner, including Kullback–Leibler and Hellinger distances. Later in the article, we focus on the disparity generated by $G(\delta) = 2\delta^2/(\delta+2)$; in the discrete case, this corresponds to the *symmetric chi-squared distance,*

$$\sum_t m_\beta(t) G(\delta(t)) = \sum \frac{[m_\beta(t) - d(t)]^2}{\frac{1}{2}m_\beta(t) + \frac{1}{2}d(t)}.$$

It defines a metric as it satisfies the triangle inequality (LeCam 1986) and lies numerically between the total variation distance and the squared Hellinger distance. It is also bounded between 0 and 4.

The value of β that minimizes the disparity is called the *minimum disparity estimator* corresponding to G. Full efficiency of the estimator at the model is automatic in this setting, whereas the robustness properties of the resulting estimator arise from the choice of the G function. Basu and Lindsay (1994) extended this notion to the continuous case via construction of the kernel-smoothed disparity measure

$$\rho^*(\hat{F}, M_\beta) = \int G(\delta^*(x)) m^*_\beta(x)\, dx. \tag{4}$$

This construction preserves many of the desirable robustness features for the corresponding point estimator, but does not in general guarantee full efficiency at the model. The methodology designed here for the continuous case recovers full efficiency and is computationally much simpler.

3.4 The Minimum Disparity Weight Functions

In the discrete case, the minimization of the disparity function leads, under suitable differentiability, to the equation

$$-\nabla\rho = \sum_t A(\delta(t)) \nabla m_\beta(t) = 0, \tag{5}$$

where $A(\delta) = (1+\delta)G'(\delta) - G(\delta)$. It follows that the discrete case minimum disparity equations are the same as the WLE equations in (1), with weight functions defined by

$$w(t; M_\beta, \hat{F}) = w(\delta(t)) = \frac{A(\delta(t)) - A(-1)}{\delta(t) + 1}. \tag{6}$$

Thus by defining the weights as in (6), the WLE estimator is a root of the minimum disparity estimating equation (5). Also, $A(\delta) = \delta$ gives $w(\delta) = 1$, as anticipated. We call weight functions generated in this manner *minimum disparity weights.*

If we apply a minimum disparity weight function in the continuous case, the WLEs do not correspond to minimizing (4) or any other distance measure. This raises the question of root selection when there are multiple roots. Our strategy is to select that root that minimizes the corresponding disparity function; in our case this is the continuous form of symmetric chi-square. We show that the contamination robustness of this selection functional gives breakdown robustness to the point estimator.

We restrict our attention to minimum disparity weights that are also unimodal, twice differentiable at $\delta = 0$, with $w'(0) = 0$ and $w''(0) < 0$. Lindsay (1994) showed that the curvature parameter A_2, defined as $A''(0)$, was a critical determinant of the second-order efficiency and robustness trade-offs for discrete minimum disparity estimation, with larger negative values corresponding to increased robustness and decreased efficiency. For our unimodal weights, we have $A_2 = w''(0)$, which as we show continues to play an important role in the robustness and efficiency properties of WLE estimators.

4. ISSUES IN THE CONSTRUCTION OF RESIDUALS AND WEIGHTS

One unfortunate difference between the discrete and the continuous cases is the need for kernel smoothing. In the discrete case, the second-order statistical properties of the method at the model, and robustness near the model, depend only on the choice of weight function $w(\cdot)$. In the continuous case, these properties depend in a complicated way on the choices of kernel k, smoothing parameter h, and weight function w. We offer some simple analyses that can provide guidance on their selection.

For any fixed method of constructing residuals, the role of the unimodal weight function w is clear intuitively, and substantiated by the discrete case analysis. More severe downweighting leads to more robustness and less efficiency. The exact choice of the kernel function k does not seem critical; thus we recommend choosing it based on convenience, whether to make the weights easy to calculate or to optimize the efficiency of the parallel minimum disparity estimator.

On the other hand, the selection of the smoothing parameter h is difficult, because it has a direct bearing on the efficiency/robustness tradeoff. Our goal of efficiency dictates that the residuals δ^* be close to 0 when the model is correct (so the weights are close to 1), which mandates reasonable bandwidths, so that f^* is stochastically close to m^*_β. On the other hand, the goal of identifying unusual observations suggests using smaller bandwidths for greater sensitivity.

We have found that the sum of the final fitted weights is a useful diagnostic statistic for the comparison of solutions to the WLEE, as it tells us roughly how many observations were deleted from the sample. Correspondingly, we find it insightful to compare weighting methods based on the mean downweighting that occurs.

4.1 The Mean Downweighting Parameter

We consider the mean downweighting that occurs when the model is correct. We let $\tilde{w}^* = n^{-1} \sum w(X_i; M_\beta, \hat{F})$ be the average of the weights at true value β. For an unimodal weight function, and under suitable regularity conditions, the asymptotic distributions will satisfy

$$\begin{aligned} n(1-\tilde{w}^*) &\approx -\frac{A_2}{2} \sum \delta^{*2}(X_i) \\ &\approx -n\, \frac{A_2}{2} \int \left[\frac{\int k(x;t,h)\, d\hat{F}(x)}{m^*_\beta(t)} - 1 \right]^2 dM_\beta(t), \end{aligned}$$

where, as before, $A_2 = w''(0)$. A simple calculation then shows that the asymptotic mean of $n(1-\tilde{w}^*)$ is

$$\Lambda = -\frac{A_2}{2} \left[\int \frac{\int k^2(x;t,h)\, dM_\beta(x)}{m^{*^2}_\beta(t)}\, dM_\beta(t) - 1 \right]. \quad (7)$$

The mean downweighting Λ serves as a simple measure of the interplay of the various parameters in the degree of downweighting that will occur when the model is correct. It can be interpreted simply as the number of observations on average that will be deleted from the sample when the model is correct.

The weight function w appears in the mean downweighting formula (7) separately from the other factors, and only through the curvature parameter $A_2 = w''(0)$. For the symmetric chi-squared distance, $A_2 = -\frac{1}{2}$, and so the leading term in (7) is $+.25$. The term in the brackets shows that the model m_β and the smoothing parameter h interact in a complicated way to determine downweighting. To obtain roughly equal downweighting throughout a parameter space, the smoothing parameter h might well have to depend on β in such a way as to hold the bracketed term constant.

To illustrate this formula, suppose that the model m_β is normal, with mean β and variance σ^2, and the smoothing kernel is normal with variance h^2. A simple calculation shows that the bracketed downweighting factor in (7) is

$$\frac{(\sigma^2+h^2)^{3/2}}{(3\sigma^2+h^2)^{1/2}\cdot h^2} - 1. \quad (8)$$

From this, we can see that increasing h pushes the mean of $(1-\tilde{w}^*)$ toward 0, and hence the weights to 1, which corresponds to becoming more like maximum likelihood (and correspondingly less robust). Also note that the mean downweighting depends on the model parameter σ^2, which implies that if h^2 is held fixed, then the robustness properties will vary with the true value of σ^2.

For this reason, we recommend selecting a constant κ and letting $h^2 = \kappa\sigma^2$, so that expression (8) becomes the constant $[(1+\kappa)^{3/2}]/[\kappa(3+\kappa)^{1/2}] - 1$; the parameter κ can then be selected to determine the degree of downweighting. (Choosing the smoothing parameter for the normal model in this parameter-dependent way has the bonus of making the WLEEs location and scale equivariant.) For $\kappa = 1$, we get the downweighting factor of $\sqrt{2}-1$, which when multiplied by .25 for the chi-squared distance, gives a mean downweighting of only about .1 observations. But for κ near 0, the mean downweighting for the chi-squared distance is approximately $.25[N-1]$, where $N = 1/\kappa\sqrt{3}$.

5. COMPUTATIONAL ISSUES

One of the main advantages of the WLE approach to robustness is that a simple and highly efficient algorithm is automatically available. Given an interim value of the parameter, say b, let $w_i = w(X_i; M_b, \hat{F})$ and solve the equations

$$\sum w_i u(X_i;\beta) = 0 \quad (9)$$

for β, with the weights fixed at the interim value. As we show in the simulations in Section 6, this gives rise to a very quick algorithm when (9) has an explicit solution, typically needing but 5–20 iterations. We show in the Appendix that there is a theoretical justification for its speed, as it has superlinear convergence when the data distribution is the same as the model distribution.

5.1 A Bootstrap Root Search

The proposed estimating equations need not have unique solutions (although our empirical evidence is that they do so in datasets that do not deviate too greatly from the model). Our strategy for tackling the multiple-root problem is to search the parameter space in such a manner that all reasonable solutions are found with some high "probability." Our approach is motivated by the work of Finch, Mendel, and Thode (1989) in which a "prior" was put on the parameter space to generate the start values. It was argued that a bonus of this approach is that one can then construct estimates of the probability that a root not yet found will be found with further searching. Our particular extension here is the use of data-driven starting values, designed to construct automatically a reasonable search region. Hawkins (1993, 1994) and Ruppert (1992) have provided other examples on the use of data-driven starting values.

For our purposes, a reasonable parameter set consists of those values that could plausibly have been used to generate some subset of the data. Let m be the minimum number of observations needed for the MLE of β to exist. For each b from 1 to B, select a bootstrap data set of m distinct elements of the dataset. For each bootstrap data set, let β_b^* be the corresponding MLE. Each value of β_b^* then becomes a starting value for the iterative reweighting algorithm, and we keep track of the roots that are found. In our simulation studies we used $B = 100$. (This rule consistently provided very small estimates of the probability of finding a further root with further searching.)

In a small fraction of such bootstrap searches, always where the bootstrap sample points were close together, we arrived at what we call a *degenerate solution* whose parameters seemed to describe just that small dataset, as was recognizable by the weights being near 1 for those points and near 0 otherwise. But we always found at least one nondegenerate root, with the total of the weights being over half of the sample size, so that the degenerate roots could be safely ignored.

6. A SIMULATION STUDY

The simulation study targeted many of the key issues that surround this method, which is numerically simple but analytically quite complicated. The goal was to try out the method with various weight options, extremely contaminated models, and sample sizes. Within that context, we wished to find out information about the problem of selecting roots, as well as how well the estimator compared with other standard robust methods.

Table 1. MSE of Huber's M Estimate, $\hat{\mu}_1$, With No Iteration on the Scale, Huber's Proposal 2 Estimate, $\hat{\mu}_2$, c = 1.345, and WLEE Estimate, $\hat{\mu}$

Percentage of contamination ε	$MSE(\hat{\mu}_{MLE})$	$MSE(\hat{\mu}_1)$	$MSE(\hat{\mu}_2)$	$h^2 = .015\hat{\sigma}^2$ $MSE(\hat{\mu})$		$h^2 = .005\hat{\sigma}^2$ $MSE(\hat{\mu})$	
				$p = 1$	$p = 1.5$	$p = 1$	$p = 1.5$
0%	.011	.012	.016	.013	.013	.013	.015
5%	.039	.011	.028	.014	.014	.014	.016
10%	.040	.014	.045	.015	.015	.015	.015
20%	.073	.020	.070	.018	.017	.017	.017
30%	.101	.027	.101	.025	.022	.023	.022
40%	.130	.040	.130	.058	.034	.037	.030
50%	.174	.063	.174	.107	.070	.090	.049

NOTE: Data are from $(1 - \varepsilon)N(0, 1) + \varepsilon N(0, 25)$ with sample size 100. The number of Monte Carlo replications is 100; the residual adjustment function used for the WLEE estimates is chi-squared.

In every case, the nominal model was $N(\mu, \sigma^2)$. We also considered 18 different options for construction of the weights. The design was a factorial, with two weight functions corresponding to symmetric chi-squared and a modified Hellinger distance; three powers, $p = .5, 1$, or 1.5; and three levels of κ used in the relationship $h^2 = \kappa\sigma^2$, either .025, .015, or .005. We also considered sample sizes $n = 20$ and 100. At each sample size, we simulated 100 datasets. For each dataset, we carried out a bootstrap root search. Given a sample, we took 100 bootstrap samples of size 2 (no repeats), and used them to construct starting values. We also used the MLE as a starting value, as well as the robust estimators ($\text{med}_i(X_i)$, 1.48 $\text{med}_j|X_j - \text{med}_i(X_i)|$).

6.1 Symmetric Errors

The symmetric sampling models, labelled $\mathbf{S}(\varepsilon)$, were $(1-\varepsilon)N(0,1) + \varepsilon N(0,25)$, with $\varepsilon = 0, .1, .2, .3, .4$, and .5. This being a symmetric error model, we felt that the meaning of robustness would be most clear if we focused on the location problem rather than the scale. When we simulated from the various sampling models, we controlled for the variation in the true contamination by fixing the fractions of observations from the two components at exactly $1 - \varepsilon$ and ε.

When the contaminations were symmetric, there was just one nondegenerate root, regardless of the degree of contamination, sample size, or weighting option. Moreover, all of the weighting options led to roots that were qualitatively the same.

Because there was always one root to the equations in the symmetric case, there was no issue with root selection, and we can easily compare the WLE methodology with other methods defined for the location model under symmetric errors.

Table 1 compares the mean squared error (MSE) of the location estimates under the above described normal gross error model, where the candidates are a standard Huber estimator, the MLE under the normal model, and various WLE candidates. The Huber estimator was calculated using the S-PLUS code rreg with fixed scale selected as 1.48 $\text{med}_j|X_j - \text{med}_i(X_i)|$. The value of the tuning constant is taken to be 1.345 as it guarantees 95% efficiency of the location estimate at the normal model. The initial value for location used was $\text{med}_i(X_i)$. Hampel, Ronchetti, Rousseeuw, and Stahel (1986, p. 105) noted that the scale parameter is often a nuisance parameter. Also, simulations have shown the superiority of the location M estimates with initial scale estimate given as earlier. Thus Hampel et al. (1986, p. 105) recommended using initial median absolute deviation scaling for M estimates. Additionally, the MSE of the Huber proposal 2 location estimate is included (Huber 1981, p. 137). Here we used the same tuning constant and starting values as previously. It is clear that there are only minor differences among the various WLEEs, that they compare favorably with the Huber estimators, and that all perform better under contamination than the MLE.

6.2 Asymmetric Errors

The asymmetric models, labelled $\mathbf{A}(\varepsilon)$, were $(1-\varepsilon)N(0,1)+\varepsilon N(8,1)$, with $\varepsilon = 0, .1, .2, .3, .4$, and $.5$. As in Section 6.1, solutions near (0, 1) will be considered robust, with those near (8, 1) also being informative, as they indicate the presence of data substructure. In the case of asymmetric errors, the equations had multiple nondegenerate roots, with the frequency increasing as the contamination proportion increased. This is exemplified in Table 2, which identifies the frequency of various patterns of roots as found with $B = 100$ bootstrap starting values. Thus, for example, out of 100 data samples of size 100 with contamination level 10%, in 63 the bootstrap search found only a single nondegenerate root, whereas in 37 two roots were found, roughly equaling (0, 1) and (8, 1). The table is exhaustive, in that in no case were more than three nondegenerate roots found. Although classification of roots would generally be a difficult task, the roots usually were quite separated, as we make clear shortly, and classification could be made by ordering the μ values.

Table 3 presents a summary of the properties of the roots organized by type. From the last two columns, it is clear that the individual root types were appropriately estimating the parameters that generated the data subsets, and that the sum of the weights gave a relatively clear indication of the size of the subgroup. On the other hand, the root corresponding to the MLE behaved rather more like Huber estimator, as can be seen from Table 4, where the estimators were calculated as described in Section 6.1.

We note that the Huber estimators were not designed for asymmetric errors, so the comparison is for insight only. We also note that the sum of the weights for this root was quite high relative to the other two roots, substantiating that this sum is not so useful for assessing the quality of the root in fitting the data.

6.3 Other Issues

We examined the properties of the iterative reweighting algorithm across our simulations. Our finding was that the mean (and median) number of iterations increases with the contamination level up to 30%, and then levels out. Thus, for example, at 40% contamination, the mean number of iterations required to find the (0, 1) root was 14.62, with a standard deviation of 9.36, whereas at 50% contamination, the numbers were 14.24 and 11.22. Thus in no case were the calculations onerous. We found that in our large simulation runs, each individual $B = 100$ search averaged about 2 minutes in real time. The programs were written in Fortran and were run on a DEC 5000/50 workstation.

Table 2. Frequency of Identified Roots

Possible roots			% Contamination, ε				
(0, 1)	MLE-like	(8, 1)	10%	20%	30%	40%	50%
+	−	−	63	4	0	0	0
−	+	−	0	0	0	0	0
−	−	+	0	0	0	0	0
+	+	−	0	0	0	0	0
+	−	+	37	86	33	1	0
−	+	+	0	0	0	0	0
+	+	+	0	10	67	99	100
Maximum number of degenerate roots in any single bootstrap search			3	2	4	3	4

NOTE: The model is $(1-\varepsilon)N(0,1)+\varepsilon N(8,1)$, $h^2 = .015\,\hat{\sigma}^2$ and $p = 1$. The chi-squared residual adjustment function was used. The number of Monte Carlo replications is 100; the + sign indicates the presence of the root, − indicates the absence of it. The sample size is 100.

Table 3. WLEE's for the Model $(1-\varepsilon)N(0,1)+\varepsilon N(8,1)$

	$h^2 = 0.015\hat{\sigma}^2$		
% of contamination	MLE-like root	(0, 1)-root	(8, 1)-root
0%		−.0133	
		.9240	
		97.2600	
10%		.0177	7.5558
		.9172	.8347
		86.4083	8.4613
20%	.1429	−.0543	7.9486
	1.5851	.9719	1.0560
	77.4850	77.0251	15.8251
30%	1.1187	−.0024	7.9657
	6.8555	.9129	1.0263
	74.4631	65.1803	24.9576
40%	2.9533	−.0158	7.9874
	14.7835	.9306	1.0063
	79.9801	55.7490	35.3065
50%	3.9628	−.0143	7.9830
	16.2783	.9537	.9728
	81.1892	45.7100	45.8647

NOTE: The first line corresponds to the location estimate, the second to the scale estimate, the third presents the sum of weights. The results are over 100 replications and 100 random starting points. The chi-squared residual adjustment function is used and the power of weights is 1. The sample size is 100.

We also examined using the parallel minimum disparity measure to select roots. At 40% contamination, the measure was not completely reliable at picking the (0, 1) root, doing so only 78% of the time. But, as Table 5 suggests, this might be a function of the separation of the normals. We thus carried out a further simulation from the model $.60N(0,1)+.40N(15,1)$. As anticipated, in this case the measure chose the (0, 1) root 100 times out of 100. Indeed, when the contamination level was pushed to 50%, the measure chose the (0, 1) root or the (15, 1) root in every case out of 100 samples. (It should be noted that an alternative strategy, based on using the robust starting values, fails in this extreme case, as the symmetry of the data leads one to the MLE-like root.)

To give some idea of how the various weighting options affected downweighting at the model as well as robustness characteristics, Table 6 provides some summary statistics for the average values of the WLEEs in the unambiguous cases (i.e., when there is a single root only) at low levels of contamination.

7. INFERENTIAL PROPERTIES

In this section we derive the influence function of the WLEE and show in particular that it equals that of the MLE when the model is correct, so the method is fully efficient. We also present some limit theorems that show that the influence function analysis gives the correct efficiency results.

Table 4. Fixed-Scale and Proposal 2 (c = 1.345) Huber Estimates of Location and Scale for the Model $(1-\varepsilon)N(0, 1) + \varepsilon N(8, 1)$

Percentage of contamination, ε	Fixed-scale estimate of $\hat{\mu}$	Proposal 2 estimates	
		$\hat{\mu}$	$\hat{\sigma}^2$
0%	−.012	−.009	.735
5%	.072	.215	3.043
10%	.185	.748	8.901
20%	.525	1.595	15.570
30%	1.205	2.402	20.074
40%	2.586	3.196	22.858
50%	3.985	3.973	23.736

NOTE: The sample size is 100 and the number of Monte Carlo replications is 100.

7.1 Efficiency

Suppose that the weight function satisfies, for all x and F,

$$w(x; M_\beta, F) \leq 1 \tag{10}$$

with equality for all x when $F = M_\beta$, as is true for our unimodal weight functions. Under some further regularity, this structure will generally suffice to imply that the influence function of the WLE functional β_w, assuming that the model is correct, is exactly that of the MLE, and so the method potentially has full asymptotic efficiency under the model.

We start by writing the WLEE in its functional form. Given a distribution F, the functional $\beta_w(F)$ will be a chosen element of the solution set to the equation

$$\int w(x; M_\beta, F) u(x; \beta)\, dF(x) = 0. \tag{11}$$

We note that if $F = M_{\beta_0}$, then β_0 is among the solutions to this equation, and so the method is Fisher consistent for β if the root is chosen appropriately.

For a fixed distribution F, let $\beta_0 = \beta_w(F)$. Let $F_\varepsilon(x)$ be the ε-contaminated distribution $(1-\varepsilon)F(x) + \varepsilon\Delta_y(x), 0 < \varepsilon < 1$, where $\Delta_y(x)$ is the distribution that assigns mass 1 to the point y. The influence function for the estimator is $\beta_w'(y) = (d/d\varepsilon)\beta_w(F_\varepsilon)$. We can find the influence function at an arbitrary distribution F by taking the derivative of both sides of the equation

$$\int w(x, M_{\beta_w(F_\varepsilon)}, F_\varepsilon) u(x, \beta_w(F_\varepsilon))\, dF_\varepsilon(x) = 0$$

with respect to ε, evaluating at $\varepsilon = 0$, and solving for $\beta_w'(y)$. From this, one obtains $\beta_w'(y) = A(F)^{-1}B(y)$, with

$$A(F) = \left\{ \int w'(\delta(x))(\delta(x)+1) u^*(x; \beta_0) u^T(x; \beta_0)\, dF(x) \right.$$
$$\left. - \int w(\delta(x)) \nabla u(x; \beta_0)\, dF(x) \right\}$$

and

$$\begin{aligned} B(y) &= w(\delta(y)) u(y; \beta_0) \\ &\quad + \int w'(\delta(x)) \, \frac{k(x; y, h)}{m^*_{\beta_0}(x)} \, u(x; \beta_0)\, dF(x) \\ &\quad - \int w'(\delta(x))(\delta(x)+1) u(x; \beta_0)\, dF(x). \end{aligned} \tag{12}$$

Table 5. Chi-squared Distances Between $(1-\varepsilon)N(0, 1) + \varepsilon N(A, 1)$ and $N(\hat{\mu}_1, \hat{\sigma}_1^2)$, $N(\hat{\mu}_2, \hat{\sigma}_2^2)$ and $N(\hat{\mu}_3, \hat{\sigma}_3^2)$ Where $\hat{\mu}_1$ and $\hat{\sigma}_1^2$ are the Roots Close to (0, 1), $\hat{\mu}_2$ and $\hat{\sigma}_2^2$ are Those Close to (A, 1), and $\hat{\mu}_3$ and $\hat{\sigma}_3^2$ are the MLE-Like Roots for the Beran (1997) Example

Percentage of contamination	Distance A	Distinct roots		
		(0, 1)	(A, 1)	MLE-like
40%	A = 6	1.105	2.017	.853
	A = 8	1.106	2.017	1.230
	A = 10	1.106	2.017	1.325
50%	A = 6	1.537	1.481	.915
	A = 8	1.538	1.490	1.251
	A = 10	1.538	1.490	1.492

Here $u^*(x; \beta) = \nabla \ln[m^*_\beta(t)]$.

If the model is correct, then $w(\delta(x)) = 1$ and $w'(\delta(x)) = 0$ (w' now being the derivative with respect to δ), in which case the influence function is the same as that of maximum likelihood, namely

$$\begin{aligned} \beta_w'(y) &= \frac{d}{d\varepsilon} \, [\beta_w(F_\varepsilon)]|_{\varepsilon=0} \\ &= \left[\int -\nabla u(x; \beta_0)\, dM_{\beta_0}(x) \right]^{-1} u(y; \beta_0). \end{aligned} \tag{13}$$

The foregoing analysis also indicates that the asymptotic variance Σ_w of the estimator can be estimated in the "sandwich" fashion as

$$\hat{\Sigma}_w = A(\hat{F}) \left[\frac{1}{n} \sum_{i=1}^{n} \{ B(X_i; \hat{F}) B^T(X_i; \hat{F}) \} \right]^{-1} A^T(\hat{F}).$$

7.2 Limit Theorems

If we proceed from the simple influence analysis to a detailed proof of asymptotic properties, then matters become considerably more difficult but shed little new statistical light. Here we offer two theorems that indicate that when the model is correct, asymptotically the methods work as they should. The conditions are given in the Appendix, and the proofs are available in technical report form from the first author.

Table 6. WLEEs of Location and Scale for the Model $(1-\varepsilon)N(0, 1) + \varepsilon N(8, 1)$ With Starting Values the MLE Estimates of μ and σ^2 and $(\text{med}_i x_i, 1.48\ \text{med}_i|x_i - \text{med}_j x_j|)$

% Contamination	$h^2 = .015\hat{\sigma}^2$		
	p = 1/2	p = 1	p = 1.5
0%	−.014	−.013	−.012
	.959	.924	.889
	98.662	97.260	95.782
5%	−.014	−.013	−.011
	.958	.924	.891
	94.057	92.748	91.343
10%	−.001	−.013	−.011
	1.044	.924	.896
	89.036	87.715	86.452

NOTE: The RAF is chi-squared.

The first result indicates that the WLEs eventually have a root in the neighborhood of the true value β_0.

Theorem 1. Under the assumptions A1–A7, with probability tending to 1 as $n \to \infty$, there exists a root $\hat{\beta}_n$ satisfying $|\hat{\beta}_n - \beta_0| < r$.

As a secondary consequence, we can show the following theorem.

Theorem 2. Under the assumptions A1–A7, there is a root to the WLEs that is consistent and asymptotically normal with mean 0 and asymptotic variance $I^{-1}(\beta_0)$, where $I^{-1}(\beta_0)$ is the Fisher information number.

We return later to the question of root selection.

8. ROBUSTNESS

Establishing the robustness properties of the WLE methodology is considerably more difficult than establishing efficiency. We demonstrate this robustness in a number of different theoretical ways that will give an underpinning to our simulation results.

We start by considering the response of the WLEEs to small amounts of contamination to the model. We note that one cannot use an influence curve analysis to compare the robustness of first-order–efficient estimators, as the influence curve at the model is always just that of the MLE. Nonetheless, we did examine the second-order expansion of the WLE functional at the model, as was done by Lindsay (1994), and found a dampened influence of outliers. Details are given in the technical report.

8.1 Equation Breakdown

When one has an equation with multiple roots, determining the estimator's theoretical breakdown properties is obviously tied to the method of root selection. If, however, the statistician is using the methodology in an exploratory fashion, he might wish to know whether the presence of a root represents some underlying structure. As was seen in the simulation, the WLEs tend to have roots corresponding to any outlying portion of the dataset that is in itself consistent with the model, even when those data subsets contain less than half the observations.

This empirical property corresponds to a stability of the estimating equation and its roots under contamination that can be investigated using an approach similar to that of Lindsay (1994). To do this, consider a fixed model m_{β_0}, and let $\hat{F}$ represent the empirical distribution function. The empirical property is that a large contamination is ignored, provided that it is at sufficient distance from the probabilities specified by β_0. To do this theoretically, we specify exactly what an outlying contamination is via the construction of an outlier sequence.

Let $\{\xi_j\colon j = 1, 2, \ldots\}$ be a sequence of elements of the sample space, let $\hat{F}_j(x) = (1-\varepsilon)\hat{F}(x) + \varepsilon\Delta_{\xi_j}(x)$ be the contaminated distribution, and let $f_j^*(x) = \int k(x; t, h)\, d\hat{F}_j(x)$ be the corresponding kernel-smoothed data. We say that $\{\xi_j\}$ is an *outlier sequence* for the model $m_\beta(x)$ and data $\hat{F}$ provided that the residuals of the contaminated distribution evaluated at the outlying point, namely $\delta^*(\xi_j) = f_j^*(\xi_j)/m_\beta^*(\xi_j) - 1$, converge to infinity as $j \to \infty$ at the same time the smoothed model probabilities at those points, $m_\beta^*(\xi_j)$, go to 0. (If ξ_j is a single observation among $n-1$ other fixed data points, then if the latter limit holds, so will the former. The key is that any observation at an extremely unlikely point causes the delta residual to become large in inverse proportion to the model density at that point.)

The WLE score along this sequence is

$$\int w(x; \hat{F}_j, M_\beta) u(x; \beta)\, d\hat{F}_j.$$

We can make precise the idea that the estimating equations ignore the outliers by showing that the limit of this sequence does not depend on the outliers. To do so, define the subdistribution function $\hat{F}_\varepsilon(x) = (1-\varepsilon)\hat{F}(x)$ that corresponds to discarding the contamination portion from the distribution. Correspondingly, let $f_\varepsilon^*(x) = (1-\varepsilon)f^*(x)$. The corresponding Pearson residual function is denoted by $\delta_\varepsilon^*(x) = f_\varepsilon^*(x)/m_\beta^*(x) - 1$. The score function

$$\int w(x; \hat{F}_\varepsilon, M_\beta) u(x; \beta)\, d\hat{F}_\varepsilon$$

corresponds to "subtracting" the epsilon contamination from the data. As ε converges to 0, note that this estimating function converges to $\int w(x; \hat{F}, M_\beta) u(x; \beta)\, d\hat{F}$, the WLE for the original, uncontaminated data. With this motivation, we say that the WLEEs ignore the contamination sequence if

$$\int w(x; \hat{F}_j, M_\beta) u(x; \beta)\, d\hat{F}_j \to \int w(x; \hat{F}_\varepsilon, M_\beta) u(x; \beta)\, d\hat{F}_\varepsilon. \quad (14)$$

The following theorem indicates that under appropriate assumptions, the WLEEs exhibit this stability property. Let $\tilde{u}(x; \beta) = \nabla \ln[m_\beta^*(x)]$.

Theorem 3. Assume that (a) $E_{m_\beta^*}[|\tilde{u}(X; \beta)|]$ is finite for all β; (b) for some $l > 1, E_{m_\beta^*}[|\tilde{u}(X; \beta)|^l]$ is finite for all β and $A(\delta) = O(\delta^{(l-1)/l})$ as $\delta \to \infty$; (c) $|u(\xi_j; \beta)/\tilde{u}(\xi_j; \beta)|$ remains bounded as $j \to \infty$; (d) the kernel function k is bounded, and (d) $A(-1)$ is finite. Then the WLEEs satisfy relation (14).

The proof of this theorem is a generalization of proposition 14 of Lindsay (1994). Except for (c) and (d), the conditions of Theorem 3 are similar to those of Lindsay. Condition (c) is generally satisfied in the exponential family under bounded kernels. Condition (d) is satisfied by, for example, the normal family of kernels.

With further regularity conditions, the convergence in (14) will be uniform on compact sets of β, and will lead to convergence of the corresponding roots.

8.2 Breakdown Properties

The simulation study demonstrated that if one selects a root based on using the parallel minimum disparity measure, then a 50% breakdown result is entirely plausible. Because Lindsay (1994) gave a detailed argument for the 50% breakdown of the WLEE in, the discrete case, in the Appendix we simply outline the main ideas in the proof to indicate the line of reasoning and the modifications necessary for similar results to apply in the continuous case. We note that a key feature of this type of argument is that, just as in the preceding section, breakdown must be considered in light of the residuals being used, which here are based on local model fit, not distance from the center of the distribution.

9. CONCLUDING REMARKS

We note that there are a number of precedents for using WLE methodology. Our line of descent follows from Lindsay (1994), with an emphasis on distance measures and full efficiency. Green (1984) and Lenth and Green (1987) considered weight functions in the generalized linear model based on deviance residuals (see also Besag 1981 and Pregibon 1982). Field and Smith (1995) investigated two possible downweighting schemes based essentially on downweighting observations that were large or small in magnitude. These methods should be contrasted with our approach, in which the magnitude of the local lack of fit determines downweighting. One particular advantage of our method of construction, in addition to its asymptotic efficiency, is that the WLEs are intrinsically Fisher consistent. Compare this to Field and Smith's methods, which require calculation by numerical integration of an adjustment term to the WLEs.

Although we have derived a general theory, we have focused our simulations on the univariate normal model. Our extensive investigation of this model provides a proper foundation for the use of WLE in such a setting. In this section we discuss some of the issues involved in extending this methodology to other models.

As we noted in the discussion of the weight functions, our primary obstacle in the independent, identically distributed models is in the creation of appropriate kernel smoothing so that the robustness properties are homogeneous throughout the space of model parameters. A sensible global methodology may be based on the mean downweighting function, but properly investigating the operating characteristics of this approach in a wide range of models is necessarily the subject of a separate and substantial investigation.

The development of the WLE methodology in regression-type settings provides further challenges. It is clear that Pearson residuals are not directly applicable, depending as they do on the independent, identically distributed assumption through the empirical distribution function. One could certainly use the empirical distribution of the regression residuals in a normal error model; however, this approach does not apply to the generalized linear model. The regression model also entails several additional statistical issues, including the role of leverage and the distinctions between the lack of fit of the regression and its link versus the lack of fit of the error model.

APPENDIX: CONDITIONS AND BREAKDOWN RESULTS

A.1 Algorithmic Efficiency

For simplicity, we derive our results for a single scalar parameter. Let F be the data distribution. We may write the reweighting algorithm in the following functional form: for fixed v, solve for β in

$$\int w(x; M_v, F)u(x;\beta)\,dF(x) = 0. \tag{A.1}$$

This algorithm will be linearly convergent, and we can determine its rate as follows. Suppose that $\beta_w(F) = \beta_0$, and let the solution to (A.1), given initial value $v = \beta_0 + \tau$, be $\beta_{\text{alg}}(\tau)$. We can differentiate in τ at $\tau = 0$ and obtain

$$\beta'_{\text{alg}}(0) = \frac{\int \frac{d}{d\beta} w(x; M_\beta, F)u(x;\beta)\,dF(x)}{\int w(x; M_\beta, F)\nabla u(x,\beta)\,dF(x)}. \tag{A.2}$$

Because $\beta'_{\text{alg}}(0) \approx (\beta_{\text{alg}}(\tau) - \beta_0)/(v - \beta_0)$, this derivative determines the linear rate of convergence. In particular, if this derivative is 0, then the algorithm is superlinear in convergence. This occurs in our case if the data distribution F exactly equals a model value M_{β_0}, because for an unimodal weight we have $w'(0) = 0$.

A.2 Regularity Conditions

In this section we present and discuss the conditions needed for the existence of solutions and asymptotic normality of the estimators. For simplicity of presentation, we discuss the results in the context of a scalar parameter β, but the results are true for a vector β as well. Assume the following:

A1. The weight function $w(\delta)$ is a nonnegative bounded and differentiable function with respect to δ.

A2. The weight function $w(\delta)$ is regular; that is, $w'(\delta)(\delta + 1)$ is bounded, with $w'(\delta)$ being the derivative of w with respect to δ. Let $\tilde{u}(x;\beta) = \nabla m^*_\beta(x)/m^*_\beta(x)$ and $u(x;\beta) = \nabla m_\beta(x)/m_\beta(x)$, where $m^*_\beta(x)$ is the smoothed version of the model and $m_\beta(x)$ is the true model.

A3. For every $\beta_0 \in \Omega$, there is a neighborhood $N(\beta_0)$ such that for $\beta \in N(\beta_0)$, the quantities $|\tilde{u}(x;\beta)u'(x;\beta)|, |\tilde{u}^2(x;\beta)u(x;\beta)|$, $|\tilde{u}'(x;\beta)u(x;\beta)|$, and $|u''(x;\beta)|$ are bounded by $M_1(x), M_2(x)$, $M_3(x)$, and $M_4(x)$, where $E_{\beta_0}[M_i(X)] < \infty, i = 1, 2, 3, 4$.

A4. $E_{\beta_0}[\tilde{u}^2(X;\beta)u^2(X;\beta)] < \infty$.

A5. $I(\beta) = E_\beta[u^2(X;\beta)] < \infty$; that is, the Fisher information is finite.

A6.

a. $\int |\nabla m_\beta(x)/m^*_\beta(x)|\,dx = \int |u(x;\beta)m_\beta(x)/m^*_\beta(x)|\,dx < \infty$.

b. $\int |\tilde{u}(x;\beta)u(x;\beta)|[m_\beta(x)/m^*_\beta(x)]\,dx < \infty$.

c. $\int |u'(x;\beta)|[m_\beta(x)/m^*_\beta(x)]\,dx < \infty$.

A7. The kernel $k(X;t,h)$ is bounded for all x by a finite constant $M(h)$ that may depend on h but not on t or x.

Note that conditions A1 and A2 are similar to those previously used in the robust literature. Dollinger and Staudte (1991) have used the exact same condition as A1 in the context of linear regression, whereas A2 holds, for example, for weights that use the Hellinger distance RAF, among others. The approach taken here, as in robust estimation, is to increase the restrictions on $w(\delta)$ so as to expand the range of true distributions for which the results hold. Note that A1 and A2 imply that $|w(\delta) - 1| \leq B|(\delta + 1)^{1/2} - 1|$, with B some finite constant.

A.3 Breakdown Argument

We consider the "asymptotic" setting in which the data $\hat{F}$ is replaced by a model value M_{β_0}, so that the contaminated distribution is $M_j = [1-\varepsilon]M_{\beta_0} + \varepsilon\Delta_{\xi_j}$. In this case the right side of (14) has a root at β_0, and so under reasonable assumptions, the terms on the left side have a sequence of roots β_j that approach β_0. Because our goal is to establish that the selection functional will choose a root in the neighborhood of β_0, we have made the first step of establishing that such a root exists, and that it can be made arbitrarily close to β_0.

From this point, the analysis shifts from the WLEs to the properties of the parallel disparity measure $\rho_G^*([1-\varepsilon]M_{\beta_0} + \varepsilon\Delta_{\xi_j}, M_\beta)$. The basic strategy is to show that for j sufficiently large, the value of this disparity measure is larger for all β outside some bounded neighborhood N of β_0 than it is for the sequence of roots β_j. If this is established, then even if we do not select β_j, we cannot select a root outside N, and so the selected roots will eventually be in N as $j \to \infty$. But we can obtain a stronger property, provided that the disparity measures converge uniformly in β in N to a function with global minimum at β_0, as then the selection functional must eventually choose β_j over other roots in N. Because these roots converge to β_0, we then have what might be described as *breakdown consistency.* Suppose that the disparity measure displays *contamination robustness* in the sense that

$$\rho_G^*([1-\varepsilon]M_{\beta_0} + \varepsilon\Delta_{\xi_j}, M_\beta) \to \rho_G^*([1-\varepsilon]M_{\beta_0}, M_\beta)$$

uniformly in a bounded neighborhood N of β_0. Because the right side has the necessary quality of having a global minimum at β_0 (by Jensen's inequality), with value $G(-\varepsilon)$, we know that within N, we will eventually select β_j, and that we next need to show that outside N, the disparity is always greater than $G(-\varepsilon)$. To determine conditions on G that give contamination robustness, we need to modify the results of Lindsay (1994, prop. 12) to account for the change from sample-space summation to kernel-based integration. This can be easily done by modifying the definition of outlier sequence to require that $\Delta_{\xi_j}^*(X)/m_\beta^*(X) \to 0$ a.s. when X is distributed as M_β^* as $j \to \infty$. (This certainly holds in the normal model-normal kernel setup if $\xi_j \to \infty$.) With this adjustment, contamination robustness for G follows from the basic assumptions of Lindsay (1994), the key one being that of $G(\delta)/\delta \to c$ as $\delta \to \infty$, for $c < \infty$, so that G is not too large in the right tail. (This is true of the symmetric chi-squared disparity.) With this assumption, we can modify G to the function $G(\delta) - c\delta$, which generates the same disparity measure and is simpler to work with, as it is a decreasing function. It is also assumed that this modified G is thrice differentiable and strictly convex.

To establish that the values of β outside some neighborhood N fail to fit as well as β_0, we then need to make some assumptions about the relationship between the kernel-smoothed model and the parameter. In particular, we must make an assumption that roughly requires that if $M_{\beta_0}^*$ puts most of its mass on a sample space set A, then for all β sufficiently far from β_0, say outside a bounded neighborhood N', M_β^* puts very little mass on the set A, and so the $M_{\beta_0}^*$ portion of the distribution $[1-\varepsilon]M_{\beta_0}^* + \varepsilon\Delta_{\xi_j}^*$ becomes, in some sense, a *contamination* when viewed from the model M_β^* (see Lindsay (1994, assumption 19). With this assumption and some regularity conditions, we can mimic Lindsay's lemma 20 and show that for every α, there exists a neighborhood N_α of β_0 such that

$$\inf\{\rho_G^*([1-\varepsilon]M_{\beta_0} + \varepsilon\Delta_{\xi_j}, M_\beta) : \beta \notin N_\alpha\} \geq G(\varepsilon - 1) - \alpha$$

for j sufficiently large, which thereby bounds below the disparity values outside the neighborhood N_α.

The argument for 50% breakdown Lindsay (1994, prop. 22) concludes by noting that because G is strictly decreasing, $G(-\varepsilon) < G(\varepsilon-1)$ holds for all $\varepsilon < .5$, and so for any ε arbitrarily close to .5 we can find a sufficiently small α such that $G(-\varepsilon) < G(1-\varepsilon)-\alpha$. Hence the selection method eventually chooses roots in the set N_α as $j \to \infty$. The uniformity of convergence of the disparities on N_α then establishes that we will choose β_j eventually.

[Received June 1995. Revised October 1997.]

REFERENCES

Andrews, D. F., and Herzberg, A. M. (1985), *Data*, New York: Springer-Verlag.

Basu, A., and Lindsay, B. G. (1994), "Minimum Disparity Estimation for Continuous Models: Efficiency, Distribution and Robustness," *Annals of Institute of Statistical Mathematics*, 46, 683–705.

Besag, J. (1981), "On Resistant Techniques and Statistical Analysis," *Biometrika*, 68, 463–469.

Beran, R. (1977), "Minimum Hellinger Distance for Parametric Models," *The Annals of Statistics*, 5, 445–463.

Clarke, B. R. (1983), "Uniqueness and Fréchet Differentiability of Functional Solutions to Maximum Likelihood Type Equations," *The Annals of Statistics*, 11, 1196–1205.

——— (1991), "The Selection Functional," *Probability and Mathematical Statistics*, 11, 149–156.

Crowder, M. (1986), "On Consistency and Inconsistency of Estimating Equations," *Econometric Theory*, 2, 305–330.

Dollinger, M. G., and Staudte, R. G. (1991), "Influence Functions of Iteratively Reweighted Least Squares Estimators," *Journal of the American Statistical Association*, 86, 709–716.

Field, C., and Smith, B. (1995), "Robust Estimation—A Weighted Maximum Likelihood Approach," *International Statistical Review*, 62, 405–424.

Finch, S. J., Mendell, N. R., and Thode, H. C. (1989), "Probabilistic Measures of Adequacy of a Numerical Search for a Global Maximum," *Journal of the American Statistical Association*, 84, 1020–1023.

Green, P. J. (1984), "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives," *Journal of the Royal Statistical Society*, Ser. B, 46, 149–192.

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

Hawkins, D. M. (1993), "The Feasible Set Algorithm for the Least Median of Squares Regression," *Computational Statistics and Data Analysis*, 16, 81–101.

——— (1994), "The Feasible Set Algorithm for the Least Trimmed Squares Regression," *Computational Statistics and Data Analysis*, 17, 185–196.

Heyde, C. C., and Morton, R. (1996), "Multiple Roots and Dimension-Reduction Issues for General Estimating Equations," preprint.

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

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, New York: Springer-Verlag.

Lenth, R. V., and Green, P. J. (1987), "Consistency of Deviance-Based M-Estimators," *Journal of the Royal Statistical Society*, Ser. B, 49, 326–330.

Lindsay, B. G. (1994), "Efficiency Versus Robustness: The Case for Minimum Hellinger Distance and Related Methods," *The Annals of Statistics*, 22, 1018–1114.

Lubischew, A. (1962), "On the Use of Discriminant Functions in Taxonomy," *Biometrics*, 18, 455–477.

Markatou, M. (1996), "Robust Statistical Inference: Weighted Likelihoods or Usual M-Estimation?" *Communications in Statistics, Part A—Theory and Methods*, 25, 2597–2613.

Markatou, M., Basu, A., and Lindsay, B. G. (1996), "Weighted Likelihood Estimating Equations: The Continuous Case," Technical Report 323, Stanford University, Dept. of Statistics.

——— (1997), "Weighted Likelihood Estimating Equations: The Discrete Case With Applications to Logistic Regression," *Journal of Statistical Planning and Inference*, 57, 215–232.

Pregibon, D. (1982), "Resistant Fits for Some Commonly Used Logistic Models With Medical Applications," *Biometrics*, 38, 485–498.

Rousseeuw, P. J., and Leroy, A. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, 1, 253–270.

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

Simpson, D. G. (1987), "Minimum Hellinger Distance Estimation for Analysis of Count Data," *Journal of the American Statistical Association*, 82, 802–807.

——— (1989), "Hellinger Deviance Tests: Efficiency, Breakdown Points and Examples," *Journal of the American Statistical Association*, 84, 107–113.

Tamura, R. N., and Boos, D. D. (1986), "Minimum Hellinger Distance Estimation for Multivariate Location and Covariance," *Journal of the American Statistical Association*, 81, 223–229.