Assessing the Effect of Individual Data Points on Inference From Empirical Likelihood

Nicole A. LAZAR

An oft-cited advantage of empirical likelihood is that the confidence intervals that are produced by this nonparametric technique are not necessarily symmetric. Rather, they reflect the nature of the underlying data and hence give a more representative way of reaching inferences about the functional of interest. However, this advantage can easily become a disadvantage if the resultant intervals are unduly influenced by one of the data points. This article proposes simple methods for evaluating the effect of single points on empirical likelihood confidence intervals. In addition to suggesting diagnostics for detecting important observations, we examine the use of bootstrap and of jackknife influence functions to assess the extremity of suspect points.

**Key Words:** Bootstrap; Jackknife; Likelihood displacement; Sensitivity analysis.

1. INTRODUCTION

Empirical likelihood (Owen 1988; 1990) has been proposed as a nonparametric analog of the familiar likelihood. Over the last decade, as the technique has grown in popularity, its connections to other data-based, computationally intense, inferential methods, such as the bootstrap, have been explored and its advantages touted in the literature. One of the oft-cited features of empirical likelihood is that the shape of the likelihood, and hence of the resultant confidence intervals, does not have symmetry enforced upon it, as is the case with normal-theory intervals. Rather, both likelihood surface and intervals reflect the structure of the data, which are free to “speak for themselves.”

It is evident that this flexibility comes with a price, namely, the possibility that inferences may be unduly influenced by a single, although not necessarily outlying, point in the data. This is an issue that needs to be recognized by users of empirical likelihood. The problem can be explored from a number of different perspectives. This article proposes diagnostic procedures based on these various perspectives. As will be seen throughout, the approach
is a multifaceted one, in which the sensitivity of the empirical likelihood is assessed using multiple diagnostics jointly.

Starting from the profile empirical likelihood itself, consider the *likelihood displacement* as described by Cook (1986). The idea is to evaluate the effects on the value of the likelihood arising from simple changes to the sample (deleting or reweighting points). This allows the user to find observations which have local influence on the likelihood, and also to assess the strength of that influence. An alternative diagnostic for exploring the shape of the likelihood, and how it changes with the deletion of observations, was proposed by DiCiccio and Monti (2001). Their method, which explores the behavior of higher order derivatives of the empirical likelihood and the functionals that define it, measures departures from a normal shape (Kass and Slate 1994; Slate 1999).

Empirical likelihood confidence regions can also be evaluated directly using ideas proposed by Efron (1992) for the bootstrap. Specifically, length and shape measures can be calculated, and influential points detected based on these. Of course, the shapes of the profile empirical likelihood, as measured by DiCiccio and Monti (2001), and of the confidence intervals are closely related; if the likelihood is asymmetric, the confidence intervals are as well.

Tsao and Zhou (2001) studied the robustness of the length of empirical likelihood confidence intervals for two location parameters, the mean and Huber’s $M$-estimator. They found that the length of empirical likelihood confidence intervals for the mean is sensitive to outliers, with a breakdown point of $1/n$. On the other hand, the breakdown point of the length of the confidence interval increases to $.5$, asymptotically, when the robust estimator of location is used. Strictly outlying points therefore can have a considerable effect on empirical likelihood inference, although, not surprisingly, this depends on the functional of interest.

The rest of the article is organized as follows. The next section discusses methods for evaluating the effect of individual points on likelihoods and on the bootstrap. Because these two constructs provide ways of understanding empirical likelihood, it is natural to borrow ideas in our attempt to assess the sensitivity of empirical likelihood inferences to the data at hand. This article shows how likelihood and bootstrap methods can be applied to empirical likelihood. Note that we are not interested in exploring the theoretical properties of empirical likelihood regions for particular functionals, robust or otherwise. Rather, the goal is to provide users of empirical likelihood with a variety of computational tools that will allow them to assess the sensitivity of their analysis. Section 3 takes up the issue of assessing the values of the size and shape diagnostics. Again, two approaches are considered—bootstrap to obtain distributions of the diagnostic measures, and relative jackknife influence functions. These also can be applied, with varying degrees of computational ease, to any empirical likelihood problem. Section 4 looks at three examples that demonstrate the various diagnostics and their usefulness. Although most of our examples deal with the mean, a nonrobust functional, the procedures defined in Sections 2 and 3, and illustrated in Section 4 are completely general. Section 5 concludes with a discussion of the results.
2. LIKELIHOOD AND BOOTSTRAP APPROACHES

2.1 EMPIRICAL LIKELIHOOD DISPLACEMENT

Let \( l(\theta) \) be the empirical likelihood ratio statistic based on the full data. That is, in a general formulation,

\[
l(\theta) = \sum_{i=1}^{n} \log[1 + t^T(\theta)\psi(x_i, \theta)],
\]

where \( \psi(x_i, \theta) \) is an estimating equation that defines \( \theta \), the functional of interest. \( t(\theta) \) is a (vector of) Lagrange multiplier(s) satisfying

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{\psi(x_i, \theta)}{1 + t^T(\theta)\psi(x_i, \theta)} = 0.
\]

Asymptotically, \( 2l(\theta) \) has a \( \chi^2 \) distribution, with degrees of freedom equal to the dimension of \( \theta \); this fact is used to derive confidence regions for the functional of interest. Another approximation for the confidence region (Owen 1990) is

\[
(n - 1)q/(n - q)F(q, n - q),
\]

where \( q \) is the dimension of \( \theta \) and \( n \) is the sample size. For small \( n \), this latter results in somewhat higher confidence thresholds.

In analogy with likelihood displacement (Cook 1986; Cook and Weisberg 1982), define the (jackknife) empirical likelihood displacement by

\[
ELD_i = 2n\{l(\hat{\theta}) - l(\hat{\theta}_{(i)})\} = 2n \left\{ \sum_{j=1}^{n} l(x_j, \hat{\theta}) - \sum_{j=1}^{n} l(x_j, \hat{\theta}_{(i)}) \right\}.
\]

Here, the second equality comes to emphasize that the likelihood calculation is performed on the full sample; in the first term, the likelihood itself is evaluated at the maximum empirical likelihood estimator based on the full data, whereas in the second term, it is evaluated at the maximum empirical likelihood estimator based on the delete-one data. This guarantees that the difference will be positive.

A useful graphical tool related to \( ELD_i \) is to plot empirical likelihood contours, say, those corresponding to 80\%, 90\%, 95\%, and 99\% quantiles for the functional of interest, together with estimates of the functional of interest calculated from the full data and from the data with case \( i \) deleted. Plotting the contours together with each of the estimates shows which points are critical. Delete-one estimates that fall outside the confidence limits from the complete data are suspect and should be examined further.

2.2 SHAPE AND SIZE FOR TWO-DIMENSIONAL INTERVALS

The likelihood approach described in the previous section provides both an analytical and a graphical diagnostic. A user could be satisfied with these alone, for performing at least
an initial screening of points that are likely to be interesting. Note that the emphasis in these
two diagnostics is on the point estimate—how much does the value of the likelihood shift
when it is evaluated at a new point estimate? By contrast, the shape and size diagnostics
proposed here give insight into the behavior of the empirical likelihood confidence intervals
as a whole. Hence they are informative if a more detailed understanding of the specific
changes incurred by deleting observations from the sample, is desired.

In his discussion of using the jackknife after the bootstrap to assess variability and
influence, Efron (1992) defined shape and length measures for confidence intervals. For 90%
intervals, for example, length is given by $s^{*}(0.95) - s^{*}(0.05)$, where $s^{*}(\alpha)$ is the $\alpha$th quantile of
the bootstrap replications. Interval shape is defined as $\log \left( \frac{(s^{*}(0.95) - s^{*}(0.05))}{(s^{*}(0.5) - s^{*}(0.05))} \right)$.

We define analogous summaries for empirical likelihood intervals by length $= ul - ll$
and shape $= \log \left( \frac{|ul - \text{max}|}{(\text{max} - ll)} \right)$, where $ul$ and $ll$ are the upper and lower limits,
respectively, of the confidence interval based on the $\chi^2$ approximation, and max is the value
that maximizes the log-empirical likelihood, that is, the estimate of the functional of interest.
When the interval is perfectly symmetric, the shape is equal to zero. Departures from zero
in either direction indicate skewness.

Efron’s definition of length for one-dimensional confidence intervals is easily extended
to area in two-dimensional regions. This can be further extended to higher dimensions—
volume in three dimensions, and hypervolume in dimensions four and up. In the remainder
of this article, we examine differences in size. Alternatively, as Cook and Weisberg (1982)
suggested for regression, it is possible to look at ratios of size, or logarithms of ratios of
size.

The extension of shape to two dimensions is complicated by the fact that empirical
likelihood intervals can be extremely irregular, reflecting skewness in the data. They are
not simple ellipsoids, as they are under normal theory; they do not even need to be convex
(Hall and LaScala 1990). One way of quantifying the shape is to look along each axis
separately. This involves finding the two most extreme points of the confidence interval on
the $x$-axis direction and the two most extreme points on the $y$-axis direction, as well as
the overall maximum. Along each axis individually, we calculate a modification of Efron’s
one-dimensional shape measure:

\[
\text{shape}_x = \log \left( \frac{(x_{lx} - x_m)^2 + (y_{lx} - y_m)^2}{(x_{ux} - x_m)^2 + (y_{ux} - y_m)^2} \right)^{1/2},
\]

\[
\text{shape}_y = \log \left( \frac{(x_{ly} - x_m)^2 + (y_{ly} - y_m)^2}{(x_{uy} - x_m)^2 + (y_{uy} - y_m)^2} \right)^{1/2}.
\]

Here, $x_m$ and $y_m$ are the coordinates of the maximizing value of the empirical likeli-
hood, giving the empirical likelihood estimates of the functionals of interest; $x_{lx}$, $y_{lx}$ and
$x_{ux}$, $y_{ux}$ are the coordinates of the two extreme points of the interval in the $x$ direction;
$x_{ly}$, $y_{ly}$ and $x_{uy}$, $y_{uy}$ are the coordinates of the two extreme points of the interval in the $y$
direction. An individual shape parameter equals zero when the interval is symmetric in that
direction. For a circular region shape $\text{shape}_x = \text{shape}_y = 0$. Because these two shape measures
look at the Euclidean distances between extreme points and a central point of the interval, it
is straightforward to generalize these as well to higher dimensions. Note that the use of the
standard $x$ and $y$ axes to define shape is completely arbitrary. Other possibilities clearly exist and could be exploited to create other summaries of the shape of the confidence regions. We return to this point in the Discussion.

3. ASSESSING SHAPE AND SIZE

Beyond simply defining measures to diagnose sensitivity to particular data points in the empirical likelihood confidence regions, we also need to know if the jackknifed values are in fact unusual. As we will see in the following discussion of examples, decisions about which points are interesting can be based in part on looking at plots of the diagnostics and forming impressions. It is also desirable to have more objective ways of determining which observations have unusual values of the diagnostic measures. This section considers two approaches to calibrating the shape and size measures: bootstrapping empirical likelihood (Hall and LaScala 1990; Lee and Young 1999) in order to obtain distributions for the diagnostic measures and defining a relative jackknife influence for the shape and size (Efron 1992).

3.1 BOOTSTRAPPING EMPIRICAL LIKELIHOOD

The idea here is simple: draw a large number of bootstrap samples from the original dataset. For each bootstrap sample, calculate the empirical likelihood confidence interval for the functional of interest. From the confidence interval, derive shape and size, resulting finally in bootstrap distributions for the diagnostic measures themselves. By comparing the observed values of shape and size to the bootstrap distribution, it is possible, in principle, to decide which observations are extreme.

3.2 RELATIVE JACKKNIFE INFLUENCE FUNCTIONS

For a given statistic, $s$, Efron (1992) defined the jackknife influence function by $u_i(s) = (n - 1)[s_i - s(i)]$, where $s_i$ is the value of the statistic on the jackknifed sample with the $i$th observation removed, $s(i)$ is the average of the jackknifed values of the statistic, and $n$ is the sample size. We take $s$ to be the shape and size of the empirical likelihood confidence intervals. Efron also defines the relative jackknife influence function,

$$u_i^+(s) = \frac{u_i(s)}{\sqrt{\frac{\sum u_j(s)^2/(n - 1)}{n - 1}^{1/2}}}.$$

Small values of the relative jackknife influence function, $\sup_i(|u_i^+(s)|) < 2$, are a sign of a robust statistic.

The rationale for a cutoff of 2 is as follows. Simple arithmetic manipulation shows that

$$u_i^+(s) = \frac{s_i - s(i)}{\sqrt{\frac{\sum(s_i - s(i))^2/(n - 1)}{n - 1}^{1/2}}}.$$

that is, the relative jackknife influence function is a $t$-like statistic for $s(i)$. This interpretation is useful, because it suggests that we can define a Hotelling $T^2$-like statistic to summarize...
multiple aspects of the jackknifed empirical likelihood confidence intervals: \( U_i^2(s) = (s(i) - \bar{s}_i)^T V^{-1}(s(i) - \bar{s}_i) \), where \( s(i) \) is the \( k \)-vector of means of the jackknifed diagnostics (e.g., the mean of the areas; the mean of the shape_x values; the mean of the shape_y values), \( \bar{s}_i \) is the \( k \)-vector of jackknifed diagnostics for the \( i \)th observation, and \( V \) is the \( k \times k \) variance-covariance matrix of the jackknifed values of the diagnostics.

The question of assessing the values of \( U_i^2(s) \) in order to decide which are extreme enough to cause concern, however, still remains. Following the heuristic underlying Efron’s recommendation for \( u_i^+(s) \), we posit that, roughly,

\[
\frac{n - k}{k(n - 1)} U_i^2(s) \sim F(k, n - k).
\]

This allows us either to set approximate cutoffs, or to calculate approximate significance levels. In either case, adopting a stringent criterion for flagging points as outlying is advisable for a number of reasons. First of all, one should hesitate before declaring any observation to be influential; lenient thresholds increase the number of data points so deemed. Second, there is an issue of multiple testing, because the relative jackknife influence function is calculated for each observation individually. Multiple testing situations induce an adjusted, more conservative, cutoff for significance. Third, the relative jackknife functions are dependent, which again demands a more exacting standard. Finally, since any \( p \) values associated with the \( U_i^2(s) \) are approximate at best, it behooves the user to err on the side of caution in proclaiming points to be influential. In the procedure proposed below, we handle this question by controlling the false discovery rate (FDR), as suggested by Benjamini and Hochberg (1995). These procedures have been offered as an alternative to traditional methods for dealing with multiple testing, such as the Bonferroni correction, as they are generally more powerful; they are also easy to implement.

With this in mind, we propose the following procedure:

1. Calculate the multivariate relative jackknife influence functions, \( U_i^2 \).
2. For each \( U_i^2 \), calculate an approximate \( p \) value based on the \( F(k, n - k) \) distribution.
3. Using the distribution-free version of the false discovery rate (FDR) procedure (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001), flag suspect observations.
4. For the points detected in Step 3, calculate the univariate relative jackknife influence functions, \( u_i^+(s) \), for each shape and size measure.
5. Inspect the univariate influence functions to discern the impact of the observation on the empirical likelihood confidence regions.

Steps 2 and 3 can often be replaced by an informal procedure, because some of the observations will have values of \( U_i^2 \) that are considerably larger than the rest, as is seen in the examples in the next section. Or, following Efron’s heuristic again, we can suggest an approximate threshold of 4, which is an upper bound on the 95th percentile of the \( F(a, b) \) distribution for \( a = 2, 3, 4 \) (corresponding to low-dimensional functionals of interest) and a wide range of values for \( b \) (corresponding to sample size); a more stringent threshold of 8 gives an upper bound on the 99th percentile.
Figure 1. Shape and length of confidence intervals for the mean of the sleep data. The point represented with the + sign is the shape and length of the confidence interval for the complete dataset. The largest observation, with a value of 4.6, is highly influential. When this point is deleted, the confidence interval shifts from being highly skewed to being almost symmetric. The length also decreases dramatically. The smallest observation is also rather influential. When this point is deleted, the confidence interval becomes much more asymmetric, and the length decreases.

4. EXAMPLES

4.1 SLEEP DATA

We first examine the Cushny and Peebles sleep data (1905) analyzed by DiCiccio and Monti (2001). DiCiccio and Monti used this dataset to demonstrate their diagnostic, denoted $F_3$, for the shape of the empirical likelihood. For the scalar mean, $F_3$ has a very simple form, $F_3 = 2 \sum_{i=1}^{n}(X_i - \bar{X})^3 / \{ \sum_{i=1}^{n}(X_i - \bar{X})^2 \}^{3/2}$, a scaled measure of skewness. In general, however, the expression for $F_3$ is complicated, requiring calculation of partial derivatives of third and fourth orders. The shape and size diagnostics we are suggesting, by contrast, can be calculated directly from the empirical likelihood function.

The 10 observations in the dataset are: 0, .8, 1.0, 1.2, 1.3, 1.3, 1.4, 1.8, 2.4, 4.6. As in DiCiccio and Monti (2001), we calculate the empirical likelihood for the mean, jackknifing each data point. The length and shape of the resultant confidence intervals (calculated according to the $\chi^2$ approximation; the only change that results from using the $F$ approximation is somewhat longer intervals and all qualitative conclusions are the same) are plotted jointly in Figure 1. As noted by DiCiccio and Monti, the largest point, with a value of 4.6, is highly influential. Deleting this point renders the confidence interval nearly symmetric, as
Figure 2. The top panel shows the bootstrap distribution of length of the empirical likelihood confidence intervals. The bottom panel shows the bootstrap distribution of shape. This example is based on the sleep data, with 1,000 resamples. Both distributions are bimodal, that for length much more so than that for shape. The lower mode in both cases corresponds to those bootstrap samples that did not include the largest observation. The higher mode corresponds to bootstrap samples that included the largest observation, possibly multiple times.

can be seen here by the value of the jackknifed shape parameter, which is near 0. Furthermore, the length of the interval is considerably shortened, almost cut in half, by deleting this data point. The smallest observation is also somewhat anomalous.

An advantage of the shape and length measures over $F_3$ is that the former relate directly and in an easily interpretable way to the features of the confidence interval itself. Sprott (1980) and Viveros and Sprott (1987) discussed diagnostics $F_3$ and $F_4$, corresponding to skewness and kurtosis, for parametric likelihoods. As discussed there, and by DiCiccio and Monti (2001), the meaning of $F_3$ is straightforward. On the other hand, $F_4$ is interpretable only if the likelihood is symmetric. The length and shape measures have no such limitations or ambiguities. In addition, as mentioned earlier, they are easy to calculate regardless of the dimension of the functional of interest, assuming that the empirical likelihood ratio function can itself be calculated.

$\text{ELD}_i$ values for the observations in this example are: 1.88, .534, .258, .127, .078, .078, .041, .036, .592, 10.789. Again, the last observation is picked out as unusual, compared to the rest.

We next drew 1,000 bootstrap samples from the original sleep dataset. For each, we calculated 95% empirical likelihood confidence intervals for the mean. The bootstrap distributions of the length and shape of the confidence interval are plotted in Figure 2. Interestingly, the bootstrap distribution for length is bimodal, with two clear, separated peaks.
Table 1. Univariate Relative Jackknife Influence Functions for Length and Shape, and Multivariate Relative Jackknife Influence Function, with Associated p Values, Sleep Data Example. The multivariate influence function attributes much larger values to the first and the tenth observations than to any of the others, but they are not significant. Multivariate functions are adjusted for degrees of freedom. The tenth data point is seen to be influential for both length and shape according to the criterion of $|u_i^2| > 2$. The first data point is close to being influential on shape, especially when compared to the other values in the second column, but is not at all influential on length.

| i  | $u_i$(length) | $u_i$(shape) | $U_i^2$ | p value |
|----|---------------|--------------|--------|---------|
| 1  | .487          | -1.461       | 3.305  | .09     |
| 2  | -.385         | .097         | .206   | .82     |
| 3  | -.385         | -.029        | .128   | .88     |
| 4  | -.420         | -.079        | .130   | .88     |
| 5  | -.420         | -.079        | .130   | .88     |
| 6  | -.420         | -.079        | .130   | .88     |
| 7  | -.420         | -.079        | .130   | .88     |
| 8  | -.456         | -.255        | .098   | .91     |
| 9  | -.310         | -.574        | .159   | .86     |
| 10 | 2.730         | 2.537        | 3.585  | .08     |

One of the modes corresponds to those resamples that included the last observation; the confidence interval is strongly pulled in the direction of that measurement, which is much larger than the others. The other mode corresponds to those resamples that did not include this observation. When the largest observation was excluded and the bootstrap procedure repeated, the resulting distribution of length was also bimodal, but much less noticeably so. As seen in Figure 2, the bootstrap distribution of shape is also slightly bimodal, with one peak apparently centered around 0 and the other around 4. The first peak corresponds to the resamples which did not include the largest point. Previous analysis revealed that deleting this point gave nearly symmetrical confidence intervals, that is, intervals with a shape measure close to 0. The second mode comes from those resamples which did include the largest point, and whose confidence intervals were subsequently asymmetric. When the largest observation was excluded from the sample and the bootstrap repeated, the distribution of shape was unimodal.

Table 1 contains the univariate and multivariate relative jackknife influence functions, and approximate p values calculated from the $F(2, 8)$ distribution for the latter, on the sleep data. No points are marked as influential by our procedure, although the p values of the first and tenth observations are much smaller than any of the others. These points would also have obviously been spotted by simply examining the values of $U_i^2$ and using the more informal cutoffs based on Efron’s heuristic. Inspection of the univariate measures reveals that the first observation is somewhat influential on the shape (symmetry) of the confidence interval, but does not appear to have an undue effect on the length. By contrast, the last data point has very large values on both the length and shape summaries.

### 4.2 Labor Costs Data

Table 2 gives hourly labor costs, in 1995, for a number of Western countries (McClave, Benson, and Sincich 1998; originally reported in the *New York Times*). We consider joint
estimation of the mean and the mean absolute deviation from the mean. This is an example of partial $M$-estimation (Stefanski and Boos 2002) and as such can be analyzed under the framework of Qin and Lawless (1994). The estimating equations are

$$\sum_{i=1}^{n} \psi(X_i, \theta_1, \theta_2) = \left( \frac{\sum_{i=1}^{n} (|X_i - \theta_2| - \theta_1)}{\sum_{i=1}^{n} (X_i - \theta_2)} \right).$$

Equating to the zero vector of length two gives the required estimates.

Selected 95% joint confidence regions are given in Figure 3 (again, these were calculated using the usual $\chi^2$ approximation; results using the $F$ approximation were not different.

Figure 3. 95% confidence region for the mean and the mean absolute deviation from the mean. Deletion of most points has little effect on the shape and orientation of the interval, as demonstrated in the panel on the lower left, which is the result of deleting the ninth data point. Deleting the sixth point changes the size of the interval, and somewhat affects the shape and orientation. By way of contrast, deleting the last point changes both size and shape dramatically.
in substance or direction, just in magnitude. The top left panel shows the confidence region based on the complete data. The top right panel is the result of deleting the Netherlands. The bottom panels show the intervals that result from deleting the United States and Portugal, respectively. Deletion of the Netherlands slightly increases the area of the region and causes a moderate shift in orientation. Deleting the U.S. does not have much effect on the overall size or shape of the confidence region. In contrast, deleting Portugal results in a significant change in both size and shape.

We aim to capture these features and changes using the shape and area characteristics that were defined previously. Note that the empirical likelihood here was calculated on a two-dimensional grid of values spanning the range of the functional. At each point in the grid, the value of the empirical likelihood ratio statistic was calculated. This allowed us to calculate the area by simply counting up the number of pairs for which the ELR statistic was less than the $\chi^2$ cutoff value (6 for $\chi^2_{(2)}$ and 95% confidence) and multiplying by the size of a grid point. The shape measures were similarly calculated by finding the extreme points along the 95% contour at 6.

The informal observations are supported by the diagnostics. Pairwise plots of the summary measures reveal two clusters of interesting points, as can be seen in Figure 4.
shape measures show two data points that are close together, but far from the rest of the sample, corresponding to Austria and the Netherlands. The point at the far lower right of this plot might or might not be an unusual case—it is far from the rest, but fits the overall trend relating the two shape measures. This point arises from the deletion of Portugal from the sample. When we look at area and shape \( x \) together, there is only one apparently influential point, and that is Portugal. Deleting this point greatly decreases the area of the confidence region, compared to the others. Removing Portugal from the sample also has a great effect on the shape of the region in the \( x \) direction. Finally, the plot of area against shape \( y \) indicates three influential points, which are Austria, the Netherlands, and Portugal.

From this analysis, there is evidence that three cases influence the size or shape/orientation of the empirical likelihood confidence region calculated jointly for the mean and the mean absolute deviation from the mean.

The values of the empirical likelihood displacement measure for each country in the survey are: 10.822, 6.370, 1.708, 1.050, 0.714, 0.868, 2.856, 3.850, 2.646, 2.394, 1.414, 1.946, 2.688, and 19.418. The last, Portugal, is much greater than any of the others, a further reflection of its status as an influential point. Germany (observation 1) also is interesting by this measure, something that the shape and size diagnostics did not point to.

Figure 5 displays the empirical likelihood contours from the full dataset. The maximum empirical likelihood estimates are marked by the plus sign in the middle, while the dots denote the estimates obtained from deleting each country in turn. Even though the estimates
Table 3. Univariate Relative Jackknife Influence Functions for Area and Shape, and Multivariate Relative Jackknife Influence Function, with Associated $p$ Values, Labor Costs Example. The multivariate influence function attributes much larger values to the first, third, fifth, sixth, and fourteenth observations than to the others, but only the last one is close to being significant. Multivariate functions are adjusted for degrees of freedom. The first data point is mildly influential for area and shape; the fifth and sixth observations are influential for shape in the $y$ direction according to the criterion of $|u_i^y(s)| > 2$. The fourteenth data point is strongly influential on area and shape in the $x$ direction. The third point isn't influential along any of the univariate measures, yet its combined statistic is large.

| $i$ | $u_i$ (area) | $u_i$ (shape$_x$) | $u_i$ (shape$_y$) | $U^2$ | $p$ value |
|-----|-------------|------------------|------------------|------|----------|
| 1   | 1.065       | 1.261            | .938             | 1.744| .22      |
| 2   | .132        | -.063            | .523             | .098 | .96      |
| 3   | -.535       | -.946            | .285             | 1.230| .35      |
| 4   | -.577       | -.284            | -.011            | .354 | .79      |
| 5   | -.568       | -.504            | -2.167           | 1.362| .31      |
| 6   | -.560       | -.504            | -2.030           | 1.191| .36      |
| 7   | -.405       | .158             | .523             | .272 | .84      |
| 8   | -.342       | 1.041            | .938             | .512 | .68      |
| 9   | -.317       | .378             | .701             | .275 | .84      |
| 10  | -.308       | .599             | .404             | .145 | .93      |
| 11  | -.280       | .378             | -.189            | .057 | .98      |
| 12  | -.278       | .378             | -.308            | .082 | .97      |
| 13  | -.177       | .820             | -.486            | .474 | .71      |
| 14  | 3.151       | -.2712           | .879             | 3.204| .07      |

are quite variable, they all fall well within the 95% confidence region defined by the complete data, which is the contour line at 6.

In parallel to the analysis of the first example, we next drew 1,000 bootstrap samples from the labor data, calculated the empirical likelihood confidence regions for the two-dimensional functional of interest, and the appropriate shape and size diagnostics. As expected from the previous discussion, the bootstrap distribution of area was strictly bimodal, with one mode corresponding to samples that included Portugal, and the other mode corresponding to samples that did not. Interestingly, the shape diagnostics were not as clearly bimodal; shape$_x$ did not look bimodal at all, and shape$_y$ was only slightly so.

Relative jackknife influence functions for this example are presented in Table 3. After adjusting for degrees of freedom, only the last observation is close to being significantly flagged as an influential point. Taking a less formal approach, one could instead note that the $U^2_i$ value for Portugal (observation 14) is almost twice as large as the next largest $U^2_i$, and consider that as the only data point warranting possible further investigation.

### 4.3 ETRUSCAN SKULLS DATA

As a final example, we briefly examine a larger dataset, the maximum head breadth (in mm) of Etruscan males. The purpose of this example is two-fold. First, we wanted to look at a large dataset, because empirical likelihood inference is based on asymptotics, and the other two examples were of relatively small datasets. Second, we wanted to look at a dataset that had no apparent outliers. To this end, one observation, that was clearly smaller
Figure 6. Scatterplot of shape and length, confidence intervals for the Etruscan skulls data. The point represented by a + sign (not visible, but present in the middle cluster) gives the measures for the confidence interval based on the complete data.

than the others, was deleted from the dataset. We therefore examined 83 of the original 84 skull measurements, as recorded by Hand et al. (1994).

A scatterplot of the length and shape diagnostics is in Figure 6. Note that the points had to be jittered, since they jointly fell at only five distinct locations. The length and shape of the full data confidence interval fall in the midst of the center cluster, and so are not visible. This scatterplot is interesting, in that there are five clusters, but the two on the left are obviously smaller than the others, comprising only one or two observations. Further inspection revealed that the two small clusters correspond to the smallest and largest observations. Also interesting is to note that the plot of length alone or of shape alone would not have revealed this pattern. Indeed, there is very little variation at all in the length and shape measures. A more detailed analysis, including bootstrapping the diagnostics and calculating the empirical likelihood displacement and the relative jackknife influence functions, showed that the smallest and largest observations in this dataset would get tagged as deserving of further study. The 10 smallest $p$ values in this example ranged from .004 to .0778; the two smallest were .004 and .0246, and would be considered borderline interesting by our informal criteria. Hence the smallest and largest observations in this dataset would be suspect. This is a reflection of the fact that the mean is not robust, and inference for the univariate mean is most affected by the extreme values in the dataset, even if those values themselves are not outlying with respect to the rest of the observations. Again, the multipronged approach to analysis, investigating a variety of diagnostic measures, gives the most complete and accurate assessment of inferential sensitivity.
5. DISCUSSION

Once empirical likelihood confidence regions are obtained from the $\chi^2$ approximation, ELD, shape, and size measures are all easily extracted. Furthermore, the measures, as we have defined them here, capture true features of the intervals, in that they mark points whose deletion changes the characteristics of the confidence regions. Inspection of each measure on its own is informative to identify data points that might be problematic. It is also useful to look at measures jointly, because this reveals observations that are consistently influential. The diagnostics are completely general, in that they can easily be used with any functional of interest that can be analyzed using empirical likelihood methods. Furthermore, they can be supplemented by, or act as a supplement to, the type of breakdown analysis presented by Tsao and Zhou (2001). The breakdown analysis gives a theoretical framework for assessing what a user might expect from empirical likelihood in the worst case, for a given functional; the methods proposed here help the user to identify, for a particular dataset, how (and if) that worst case manifests itself.

It is worth emphasizing again here, that, unlike the size measure and the empirical likelihood displacement, both of which can be defined in only one way, the shape measure, which aims to summarize the entire (possibly very unusual) shape of the confidence region in dimensions higher than one, can be defined in many ways. We made an obvious choice, namely aligning along the standard coordinate axes, but other intuitively pleasing choices are also available. It is not known at this time how important the specific choice of shape summary is. Presumably, any reasonable measure will pick out the cases that have a clear effect on the confidence regions (such as Portugal in the Labor Costs example), and these are the ones that we will usually be most interested in.

Obtaining some type of calibration, even heuristic or informal, for the diagnostics is important, because without this it is hard to know whether values that appear to be extreme really should be taken as such. The concern is that we are simply picking out the largest and smallest data points in the sample. We explored two ways of addressing this concern: bootstrapping the quantities of interest, and evaluating relative jackknife influence functions. These relatively simple diagnostics and evaluations give insight into the sensitivity of the empirical likelihood analysis to vagaries in the data. However, it should be pointed out again that the goal is not to obtain precise significance values for the diagnostics, but rather to provide benchmarks, even rough ones, for distinguishing the “truly interesting” data points from the rest of the sample.

As we have seen, a bimodal bootstrap distribution, with one mode corresponding to samples with the suspect observation and the other to samples without, indicates an observation with an unusual effect; such a data point should be inspected as a possible outlier. However, if the bootstrap distributions are unimodal, or very nearly so, this does not necessarily mean that there are no influential points. In the case of a unimodal distribution, we might want to consider a data point to be influential when the jackknifed shape and size diagnostics are in the far tails of the distribution; this would accord with the standard usage of bootstrap distributions. In either situation we are looking for departures from ordinary
behavior, or “bootstrap pathologies.” These types of abnormalities also motivate the work of Hall and Presnell (1999a, 1999b) on the biased bootstrap.

When using the relative jackknife influence functions, we identify suspect points by calculating approximate $p$ values for a Hotelling $T^2$-like statistic and then using a false discovery rate procedure to winnow out the uninteresting observations. We can also informally pick the points for further investigation by the sheer relative magnitude of the combined $U_i^2$ statistic. Univariate influence functions are calculated for each diagnostic, for the data points flagged in the first stage. The univariate functions should individually be large, in order to conclude that a data point is influential on the empirical likelihood confidence regions in general. The rationale for having several levels of screening is to prevent an observation from being deemed globally influential, when it exerts only a mild amount of influence on all confidence region measures. These “mild but consistent” observations accumulate substantial values of $U_i^2$, but do not necessarily look unusual when compared to the rest of the sample of jackknifed values. Whether or not these points should be considered influential is an interesting question in and of itself. We might define several classes of observations that affect the empirical likelihood confidence regions in different ways: (1) strongly influential on all measures; (2) strongly influential on one or two measures; (3) mildly influential on all measures; or (4) not influential at all. A robust dataset or a robust functional should have most points in category (4) and few (or none) in categories (1) and (2).

The current work draws heavily on ideas from standard likelihood and bootstrap theories, and a direct transfer of results has often been possible. It is worth noting, however, that the closeness of parametric and empirical likelihood cannot always be taken advantage of with such ease. A case in point is Kass, Tierney, and Kadane (1989), who derived diagnostics for case influence based on the relationship $L_{\text{new}}(\theta) = L(\theta) \rho(\theta)$, where $L_{\text{new}}(\theta)$ is the likelihood, or more generally the posterior distribution, with case $i$ deleted and $\rho(\theta) = 1/L_i(\theta)$, the contribution of the $i$th observation to the complete likelihood, $L(\theta)$. Their approach is elegant, and sheds light on the meaning of the likelihood displacement, yet is not applicable to empirical likelihood, since we cannot write $\text{EL}_{\text{new}}(\theta) = \text{EL}(\theta) \rho(\theta)$. The contribution of the $i$th case comes in also through the Lagrange multiplier defining empirical likelihood, which is therefore different for $\text{EL}_{\text{new}}(\theta)$ and $\text{EL}(\theta)$. Hence, unlike standard likelihood, the perturbed empirical likelihood is not proportional to the original one. The implications of this are not at the moment fully understood.

ACKNOWLEDGMENTS

Parts of this work were inspired by conversations with Stephen Lee, while I was visiting the Department of Statistics and Actuarial Science, Hong Kong University. His hospitality, and the hospitality and support of the Department, are gratefully acknowledged. I also thank the anonymous referees and the editor, for their careful reading, comments, and suggestions.

[Received August 2003. Revised October 2004.]
REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), “Controlling the False Discovery Rate—A New and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society*, Series B, 57, 289–300.

Benjamini, Y., and Yekutieli, D. (2001), “The Control of the False Discovery Rate in Multiple Testing Under Dependency,” *The Annals of Statistics*, 29, 1165–1188.

Cook, R. D. (1986), “Assessment of Local Influence” (with discussion), *Journal of the Royal Statistical Society*, Series B, 48, 133–169.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.

Cushny, A. R., and Peebles, A. R. (1905), “The Action of Optical Isomers. II. Hyoscines,” *Journal of Physiology*, 32, 501–510.

DiCiccio, T. J., and Monti, A. C. (2001), “Approximations to the Profile Empirical Likelihood Function for a Scalar Parameter in the Context of $M$-estimation,” *Biometrika*, 88, 337–351.

Efron, B. (1992), “Jackknife-After-Bootstrap Standard Errors and Influence Functions” (with discussion), *Journal of the Royal Statistical Society*, Series B, 54, 83–127.

Hall, P., and La Scala, B. (1990), “Methodology and Algorithms of Empirical Likelihood,” *International Statistical Review*, 58, 109–127.

Hall, P., and Presnell, B. (1999a), “Intentionally Biased Bootstrap Samples,” *Journal of the Royal Statistical Society*, Series B, 61, 143–158.

——— (1999b), “Biased Bootstrap Methods for Reducing the Effects of Contamination,” *Journal of the Royal Statistical Society*, Series B, 61, 661–680.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994), *A Handbook of Small Data Sets*, London: Chapman & Hall.

Kass, R. E., and Slate, E. H. (1994), “Some Diagnostics of Maximum Likelihood and Posterior Nonnormality,” *The Annals of Statistics*, 22, 668–695.

Kass, R. E., Tierney, L., and Kadane, J. B. (1989), “Approximate Methods for Assessing Influence and Sensitivity in Bayesian Analysis,” *Biometrika*, 76, 663–674.

Lee, S. M. S., and Young, G. A. (1999), “Nonparametric Likelihood Ratio Confidence Intervals,” *Biometrika*, 86, 107–118.

McClave, J. T., Benson, P. G., and Sincich, T. (1998), *Statistics for Business and Economics* (7th ed.), New Jersey: Prentice Hall.

Owen, A. B. (1988), “Empirical Likelihood Ratio Confidence Intervals for a Single Functional,” *Biometrika*, 75, 237–249.

——— (1990), “Empirical Likelihood Confidence Regions,” *The Annals of Statistics*, 18, 90–120.

Qin, J., and Lawless, J. (1994), “Empirical Likelihood and General Estimating Equations,” *The Annals of Statistics*, 22, 300–325.

Slate, E. H. (1999), “Assessing Multivariate Nonnormality Using Univariate Distributions,” *Biometrika*, 86, 191–202.

Sprott, D. A. (1980), “Maximum Likelihood in Small Samples: Estimation in the Presence of Nuisance Parameters,” *Biometrika*, 67, 515–523.

Stefanski, L. A., and Boos, D. D. (2002), “The Calculus of $M$-estimation,” *The American Statistician*, 56, 29–38.

Tsao, M., and Zhou, J. (2001), “On the Robustness of Empirical Likelihood Ratio Confidence Intervals for Location,” *Canadian Journal of Statistics*, 29, 129–140.

Viveros, R., and Sprott, D. A. (1987), “Allowance for Skewness in Maximum-Likelihood Estimation with Application to the Location-Scale Model,” *Canadian Journal of Statistics*, 15, 349–361.