# Assessing the Effect of Individual Data Points on Inference from Empirical Likelihood

Nicole A. Lazar
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
nlazar@stat.cmu.edu

*Summary*

An oft–cited advantage of empirical likelihood is that the confidence intervals that are produced by this non–parametric 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 some of the data points. In this paper, we consider the effect of extreme points, not necessarily outliers, on the profile empirical likelihood ratio and 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.

# 1 Introduction

Empirical likelihood has been proposed as a non–parametric analog of the familiar likelihood used in statistics. Over the last decade its connections to other data–based 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 based on empirical likelihood may be unduly influenced by unusual, although not necessarily outlying, points in the data. This problem can be explored from a number of different perspectives.

Starting from the profile empirical likelihood itself, we can consider the *likelihood displacement* as described in Cook (1986). By changing the weight given to a particular data point, while keeping all others fixed, where a weight of zero corresponds to deleting the point outright, and a weight of one corresponds to weighting all points equally, the effects on the value of the likelihood can be evaluated. In this way, it is possible to locate data points 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 in 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. As such, it is in the same family of procedures as Kass and Slate (1994) and Slate (1999), who apply a similar approach in a Bayesian setting for evaluating the shape of posterior distributions. While our focus is not on detecting shifts from normality, but rather on understanding the contributions of individual observations to empirical likelihood inferences, procedures developed for assessing normality are still relevant.

Confidence intervals based on the profile empirical likelihood ratio can also be evaluated using ideas proposed in 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, in the sense that 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 location, namely, the mean and Huber's $M$–estimator of location. 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 0.5, 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 paper is as follows. In the next section, we discuss methods for evaluating

the effect of individual points on likelihoods and on the bootstrap. Since 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. We show how likelihood and bootstrap methods can be applied to empirical likelihood. In section 3, we consider a range of examples to demonstrate the various diagnostics and their usefulness. Section 4 takes up the issue of assessing the values of the size and shape diagnostics. Two approaches are considered – bootstrap to obtain distributions of the diagnostic measures, and relative jackknife influence functions. We conclude in section 5 with a discussion of the results.

# 2 Likelihood and Bootstrap Approaches

## 2.1 Empirical Likelihood Displacement

As in Cook (1986), consider the effect of a small perturbation in the data on the empirical likelihood. This can take the form of either deleting a point altogether, or assigning it a weight somewhere between 0 and 1. We consider the former case here. Let $l(\theta)$ be the log empirical likelihood based on the full data. In analogy with likelihood displacement (Cook, 1986; Cook and Weisberg, 1982), define empirical likelihood displacement by

$$ELD_i = 2\{l(\hat{\theta}) - l(\hat{\theta}_{(i)})\},$$

that is, the difference in the empirical likelihoods evaluated at the maximum empirical likelihood estimator from the full data and from the delete–one data. Values of $ELD_i$ can be

compared to the $\chi^2$ distribution with degrees of freedom equal to the dimension of $\theta$. This gives a way of determining significantly large displacements, that is, displacements in the empirical likelihood that are big enough to affect inference.

A useful graphical tool related to $ELD_i$ is to plot empirical likelihood contours, say, those corresponding to 80%, 90%, 95% and 99% quantiles, 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 crucial to the inference, in the sense that deleting them would lead to different conclusions than those based on the complete data set. Delete–one estimates that fall outside the confidence limits from the complete data are considered influential from this perspective.

## 2.2 Shape and Size for Two–Dimensional Intervals

In his discussion of using the jackknife after the bootstrap to assess variability and influence, Efron (1992) defines 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. This can be normalized by dividing the length by $2\times$ 1.645 $\times$ (bootstrap standard error).

Interval shape is defined as

$$\log|(s^{*(0.95)} - s^{*(0.5)})/(s^{*(0.5)} - s^{*(0.05)})|.$$

We define analogous summaries for empirical likelihood intervals by

$$length = ul - ll$$

and

$$shape = \log|(ul - min)/(min - ll)|,$$

where $ul$ and $ll$ are the upper and lower limits, respectively, of the confidence interval based on the $\chi^2$ approximation, and $min$ is the value that minimizes 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 reveal skewness in the confidence interval.

Efron's definition of length for one–dimensional confidence intervals can be easily extended to a notion of area in two–dimensional regions. This could be further extended to higher dimensions – volume in three dimensions, and hypervolume in dimensions four and up. In the rest of the paper, we will consider differences in size. Alternatively, as Cook and Weisberg (1982) do for regression, it is possible to look at ratios of size, or logarithms of ratios of size.

As for shape, the extension to two dimensions is complicated by the fact that empirical likelihood intervals can be extremely irregular, reflecting skewnesses in the data; they do not even need to be convex (Hall and LaScala, 1990). They are not simple ellipsoids, as they would be under normal theory, for example. One way of quantifying the shape is to look along the left–right axis and along the up–down 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 minimum. Along each axis individually, it is then possible to calculate a modification of Efron's one–dimensional shape measure:

$$shape_x = \log[\{(x_{lx} - x_m)^2 + (y_{lx} - x_m)^2\}^{-1/2} / \{(x_{ux} - x_m)^2 + (y_{ux} - y_m)^2\}^{-1/2}]$$

$$shape_y = \log[\{(x_{ly} - x_m)^2 + (y_{ly} - x_m)^2\}^{-1/2} / \{(x_{uy} - x_m)^2 + (y_{uy} - y_m)^2\}^{-1/2}]$$

Here, $x_m$ and $y_m$ are the coordinates of the minimizing value of the empirical likelihood, 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. As for the one dimensional case, the individual shape parameters will equal zero when the interval is symmetric in that direction. A confidence ellipse that is circular has both $shape_x = 0$ and $shape_y = 0$.

Since these two shape measures look at the Euclidean distances between extreme points and a central point of the interval, it is now straightforward to generalize these as well to higher dimensions. We have therefore defined three measures for exploring the sensitivity of two–dimensional intervals to specific data points: area, $shape_x$ and $shape_y$. The utility of these measures in summarizing the features of the confidence intervals will be explored in more detail in the examples below.

# 3 Examples

## 3.1 Sleep Data

As a first example, we consider the Cushny and Peebles sleep data (1905) analyzed by DiCiccio and Monti (2001). DiCiccio and Monti use this data set to demonstrate the usefulness of 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 ten observations in the data set are: 0.0, 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, deleting each data point in turn. Instead of looking at features of the likelihood itself, however, we here consider the length and shape of the confidence intervals, which 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 likelihood, and the confidence interval, nearly symmetric, as can be seen here by the value of the 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, both by our criteria and that of DiCiccio and Monti.

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

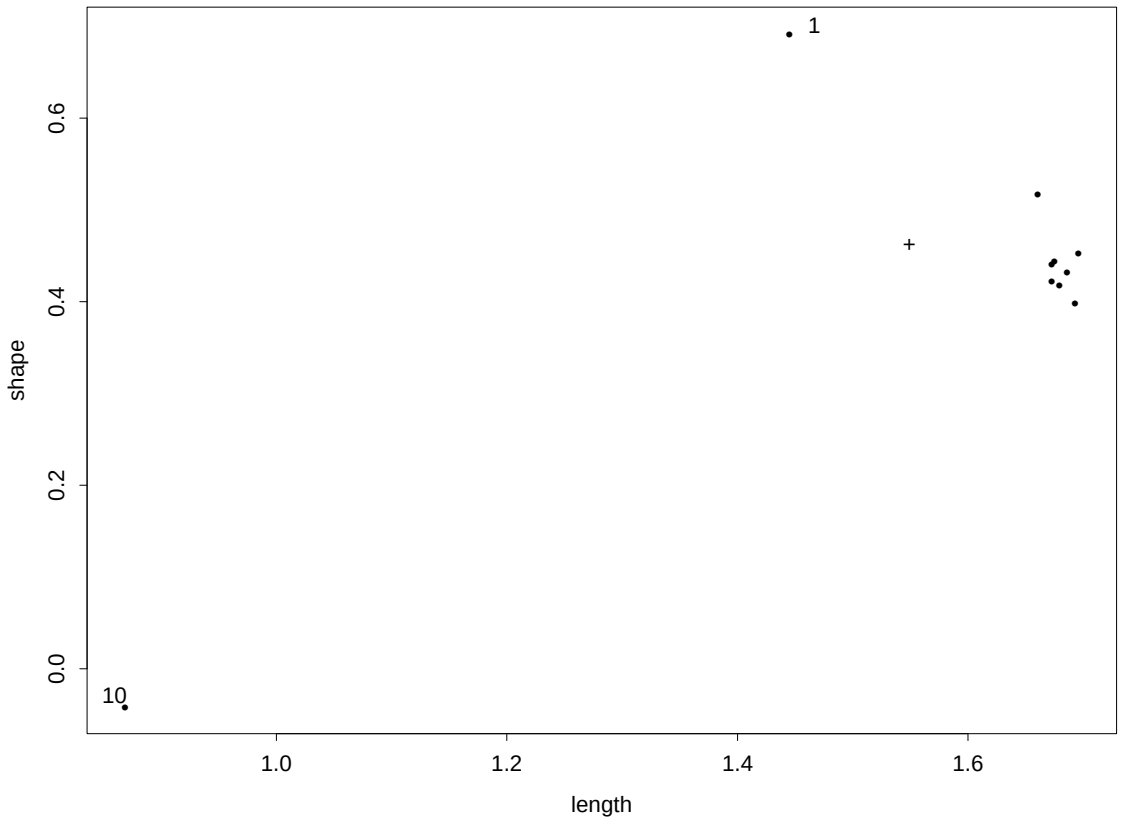


Figure 1: **Shape and length of confidence intervals for the mean of the sleep data. The point represented with the + sign is the shape and length of the confidence interval for the complete data set. The largest observation, with a value of 4.6, is highly influential. When this point is deleted, the confidence interval shifts from being highly skewed to virtually 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.**

(1980) and Viveros and Sprott (1987) discuss diagnostics $F_3$ and $F_4$, corresponding to skewness and kurtosis, for the parametric likelihood case. As discussed there, and in 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.

Empirical likelihood displacements were also calculated for this example. Unlike with the other measures, no cases were picked out as being particularly influential based on their values of $ELD_i$.

## 3.2 Motoring Data

This example looks at male unemployment rates in 1987 in 11 regions of the United Kingdom, and the average percentage of household weekly expenditures on motoring and transportation fares, as reported in Hand *et al.* (1994). The data are given in Table 1.

We performed two empirical likelihood analyses on this data. In the first, we looked at confidence intervals for a ratio estimator for unemployment divided by expenditures. In the second, we considered the two–dimensional confidence intervals for the mean. For both cases, we were interested in seeing whether or not there were data points that exerted undue influence on the shape and size of the intervals.

Since the ratio estimator is unidimensional, we used the empirical likelihood analogies of the shape and length measures defined by Efron (1992) to summarize the confidence intervals. There is some variability in the shape of the intervals, with the measure ranging in value

| Region | Unemployment | Motor Spending |
|---|---|---|
| North England | 14.0 | 12.8 |
| Yorkshire | 11.3 | 14.5 |
| East Midlands | 9.0 | 15.4 |
| East Anglia | 6.8 | 15.0 |
| South East England | 7.1 | 15.0 |
| South West England | 8.2 | 15.3 |
| West Midlands | 11.1 | 14.6 |
| North West England | 12.7 | 14.2 |
| Wales | 12.5 | 14.4 |
| Scotland | 13.0 | 14.1 |
| Northern Ireland | 17.6 | 15.5 |

Table 1: **Data on unemployment rates and transportation expenditures in the United Kingdom.**

from -0.048 to 0.118. In all instances, the confidence interval is close to symmetric. Deleting the first data point, North England, results in the most asymmetrical confidence interval. The orientation of the interval changes when the third, fifth or sixth point, corresponding to East Midlands, South East England and South West England, respectively, is deleted. The length of the interval is not much affected by deletion of any single observation. A joint plot of shape and length (Figure 2) reveals that no data point appears to be particularly influential on the ratio statistic.

Figure 3 shows two dimensional confidence regions for the mean, based on the $\chi^2$ approximation to the empirical likelihood ratio test statistic (Owen, 1990). The top left panel gives the confidence region for the complete data set; the minimizing point, representing the empirical likelihood estimates for the mean vector, is also shown, for reference. This region would be the basis for inference regarding the bivariate mean for this data set. The next three panels show the different patterns of behavior resulting from the deletion of various points. The top right panel is the result of deleting the first data point, North England. As can be seen, the general orientation of the region is the same, although the size is changed considerably. Deleting the seventh case has very little impact on the apparent size and shape of the region; this behavior is typical. The eleventh point is quite influential – deleting this point not only changes the size of the region, but has a serious effect on the shape and orientation as well. Had Northern Ireland not been included in the study, quite different conclusions would have been reached. Somewhat different conclusions might have been reached had North England not been included in the original survey.

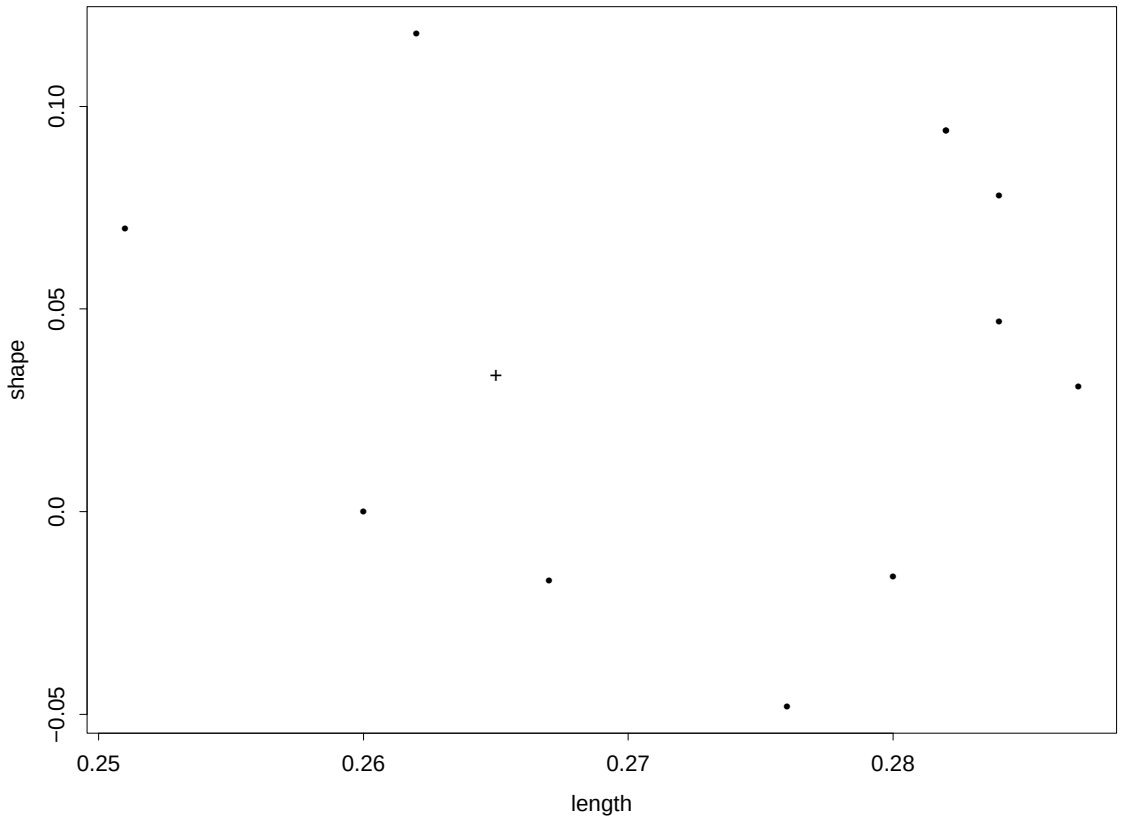


Figure 2: **Shape and length of confidence intervals for the ratio estimator of the motoring data. The + sign represents the confidence interval based on the complete data. No data point appears to be influential. In all cases, the confidence interval is nearly symmetric. There is not much variation in the length of the confidence interval as each point is deleted from the sample.**

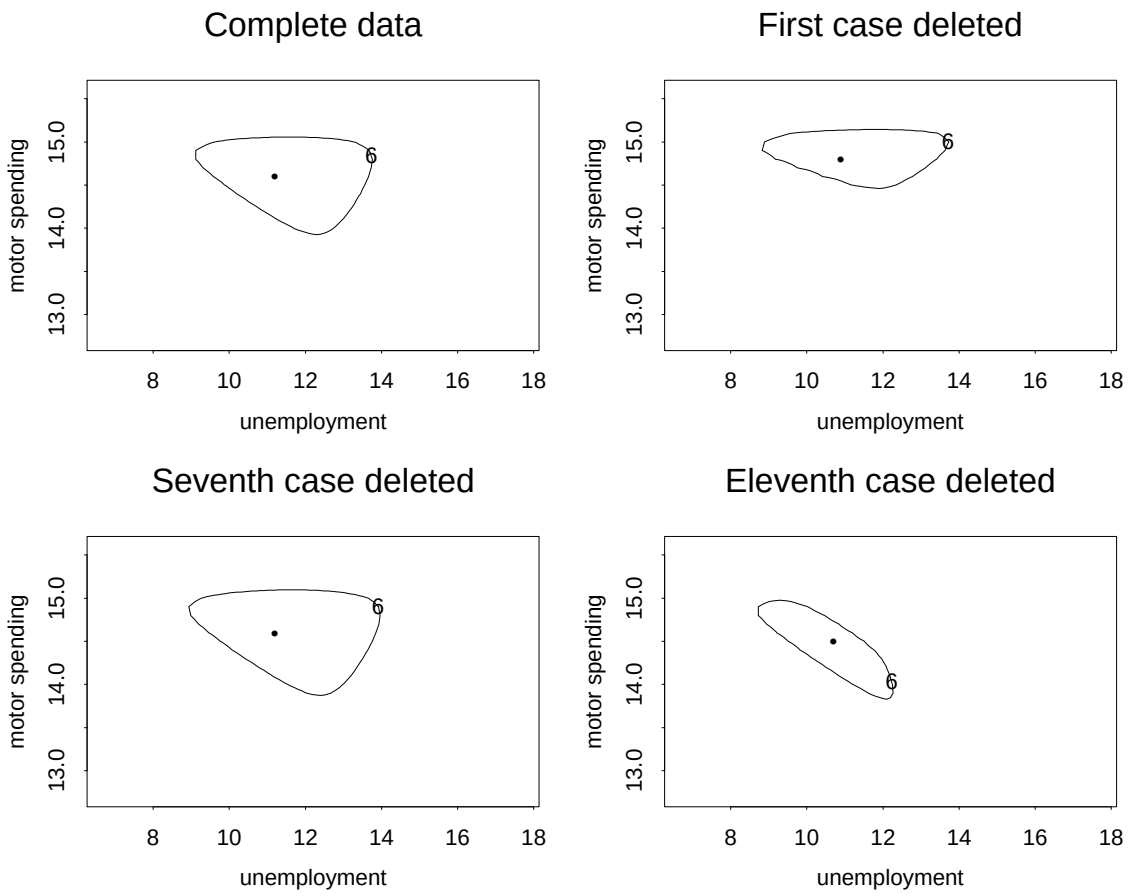


Figure 3: **95% confidence region for the mean. Deletion of most points has little effect on the shape and orientation of the interval, as demonstrated in the panel on the lower left, which is the result of deleting the seventh data point. Deleting the first point changes the size of the interval, although not its general shape and orientation. On the other hand, deleting the last point changes both size and shape.**

These features and changes are what we hope to capture using the shape and area characteristics that were defined in the previous section. Barplots of each of the measures are shown in Figure 4. In each of these panels, the leftmost bar represents the value for the confidence region based on the complete data and is the reference point for the others. Moving from left to right are the values after deleting the first point, the second point, and so on, down to the eleventh point. $shape_x$ is negative for all of the regions save that obtained by deleting Northern Ireland. In magnitude, however, the value of $shape_x$ for Northern Ireland is similar to the rest. Thus both the sign and the magnitude of the shape measure carry relevant information. In terms of $shape_y$, there is a fair amount of variability. Deleting North England or Northern Ireland results in much smaller values of this measure; deleting Yorkshire, East Anglia or Southeast England increases the value of $shape_y$. As for area, smaller confidence regions result from deleting North England and Northern Ireland. Northern Ireland is picked out as unusual along all three measures we have suggested here to summarize two dimensional confidence regions. Northern England is flagged as influential along two of the measures.

Plotting the measures pairwise has some additional diagnostic value. Pairwise plots are in Figure 5. Interestingly, the scatterplot of the two shape measures has very little structure, hence none of the points are flagged as unusual. On the other hand, the plots of both of the shape measures against area highlight the unusual behavior of North England and Northern Ireland relative to the rest.

The analysis has located two points that are influential in the sense that deleting them

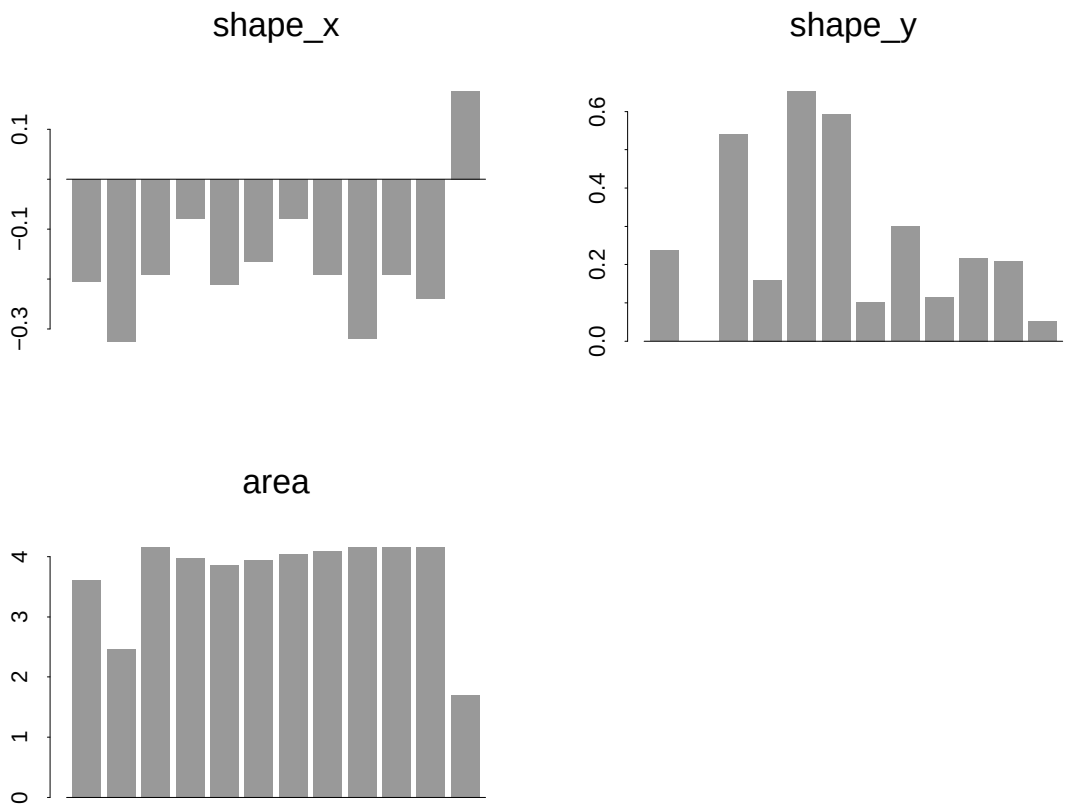


Figure 4: **Barplots for $shape_x$, $shape_y$ and area, confidence regions for the mean of the motoring data. In each barplot, the leftmost bar is the value for the confidence region based on the complete data. Moving from left to right, the first point, second point and so on is deleted in turn. Northern Ireland is clearly aberrant in direction, if not magnitude, of $shape_x$; deleting this point changes the orientation of the confidence region in the x-direction. The distribution of shape in the y-direction is much more variable, although North England and Northern Ireland differ slightly from the rest – deleting either of these points results in a more symmetric interval in the y-direction, motor spending costs. Likewise, deleting either of North England or Northern Ireland substantially reduces the area of the confidence region.**

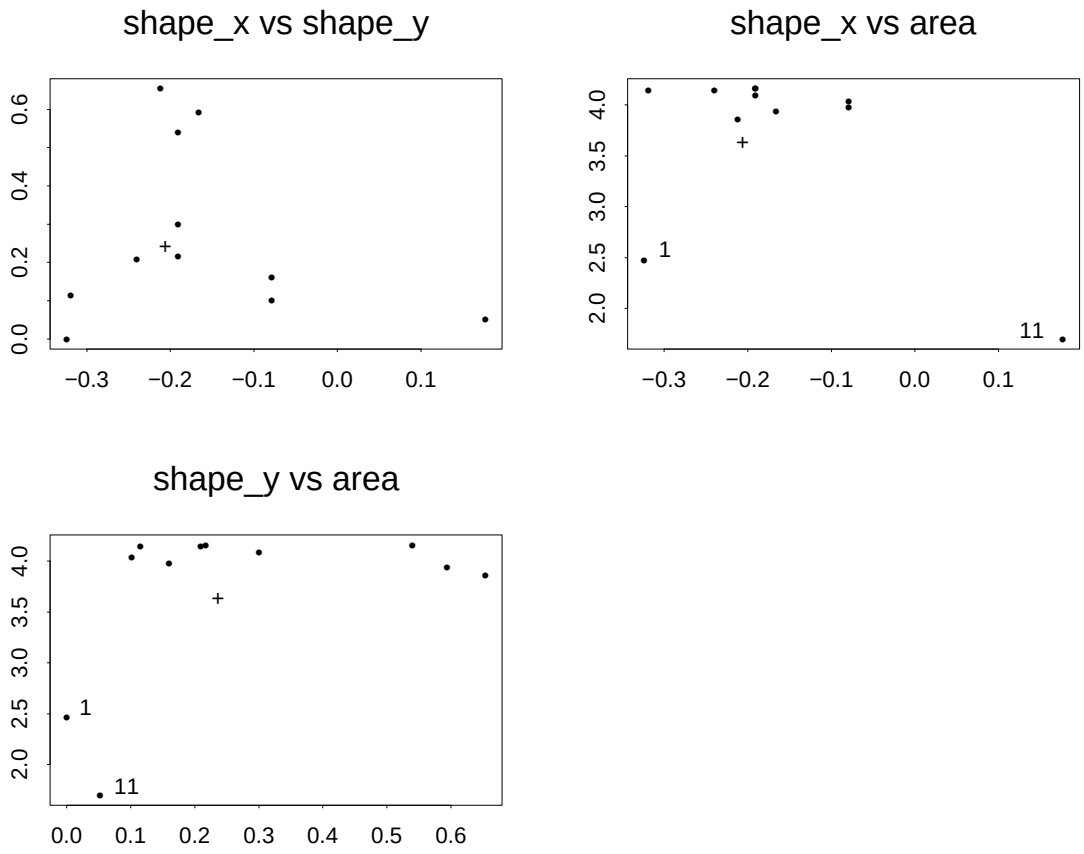


Figure 5: **Pairwise scatterplots of $shape_x$, $shape_y$ and area, confidence regions for the mean of the motoring data. In each scatterplot, the point represented with a + sign is the complete data set. The joint plot for the two shape measures reveals no influential points. However, the first observation (North England) and the eleventh observation (Northern Ireland) have a clear impact on the area of the confidence interval; deleting either of these points results in a much smaller interval.**

would significantly impact empirical likelihood inference for the value of the mean.

Looking at the empirical likelihood displacements reveals a somewhat different picture. Values for each region in turn are: 0.802, 0, 0.266, 0.162, 0.162, 0.317, 0, 0.151, 0.024, 0.151 and 1.216. None of these are big enough, as calibrated against quantiles of the $\chi^2$ distribution, to be considered significant, although the displacements associated with North England and Northern Ireland are relatively large.

## 3.3 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 for this pair are

$$\sum_{i=1}^{n} \psi(X_i, \theta_1, \theta_2) = \begin{pmatrix} \sum_{i=1}^{n}(|X_i - \theta_2| - \theta_1) \\ \\ \sum_{i=1}^{n}(X_i - \theta_2) \end{pmatrix}$$

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

Selected 95% joint confidence regions are given in Figure 6. The top left panel shows the confidence region based on the complete data. The top right panel is the result of deleting the sixth data point, the Netherlands. The bottom panels show the intervals that result from deleting the United States and Portugal respectively. Deletion of the sixth data point slightly increases the area of the region and causes a moderate shift in orientation. Deleting

| Country | Hourly Labor Rates (German Marks) |
|---|---|
| Germany | 43.97 |
| Switzerland | 41.47 |
| Belgium | 37.35 |
| Japan | 36.01 |
| Austria | 35.19 |
| Netherlands | 34.87 |
| Sweden | 31.00 |
| France | 28.92 |
| United States | 27.97 |
| Italy | 27.21 |
| Ireland | 22.17 |
| Britain | 22.06 |
| Spain | 20.25 |
| Portugal | 9.10 |

Table 2: **Data on hourly labor costs, 1995, in German marks.**

the ninth case does not, apparently, have much of an effect on the overall size or shape of the confidence region. In contrast, deleting the fourteenth point results in a significant change in both size and shape. As before, we are interested in seeing whether or not these visual impressions are detected by the three measures.

Looking at pairwise plots of the three summary measures, area, $shape_x$ and $shape_y$ reveals two clusters of possibly influential points. As is evident from Figure 7, the different pairwise combinations indicate different numbers of interesting points. Considering both shape measures, there are two data points that are close together, but far from the rest of the observations. These points correspond to deleting Austria or 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 seems to fit 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 correspond to Austria, the Netherlands, and Portugal.

From this analysis, there is evidence for concluding that three of the cases exert special influence on 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

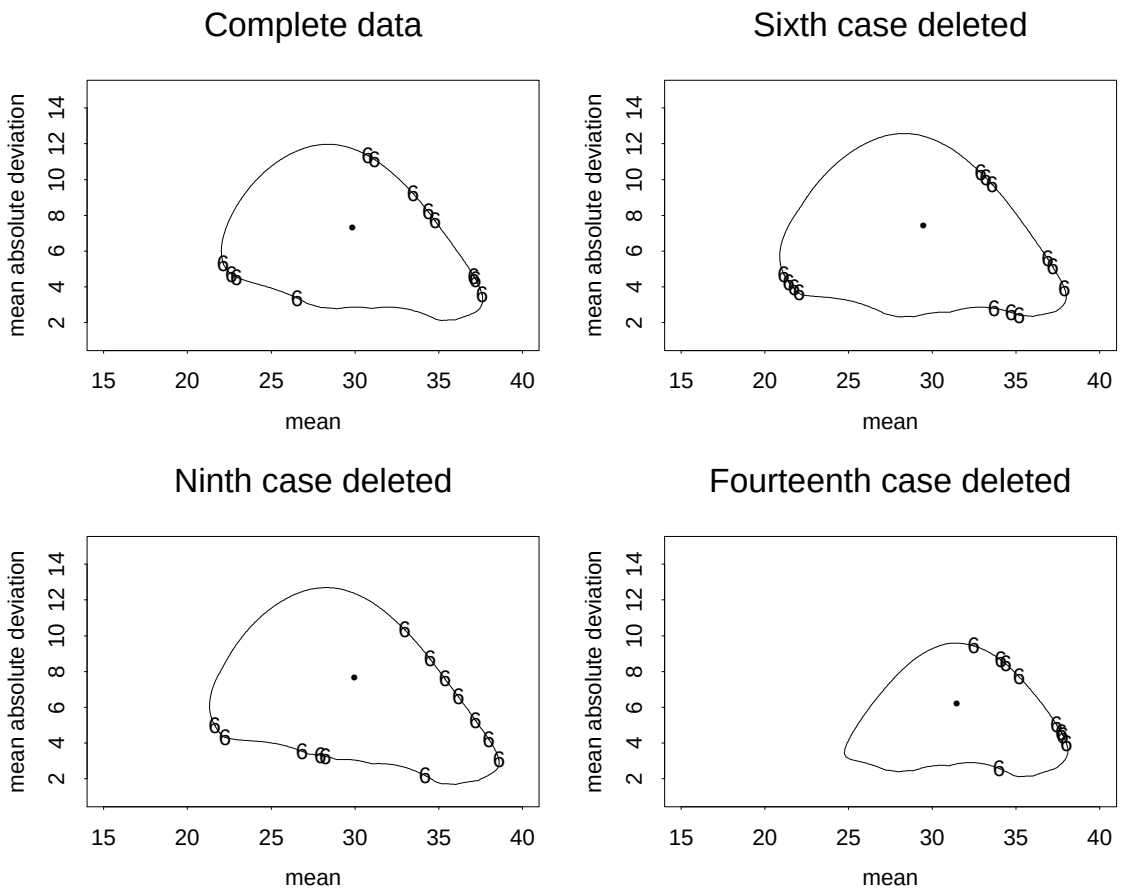


Figure 6: **95% confidence region for the mean and the mean absolute deviation from the mean. Deletion of most points has little effect on the shape and orientation of the interval, as demonstrated in the panel on the lower left, which is the result of deleting the ninth data point. Deleting the sixth 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.**

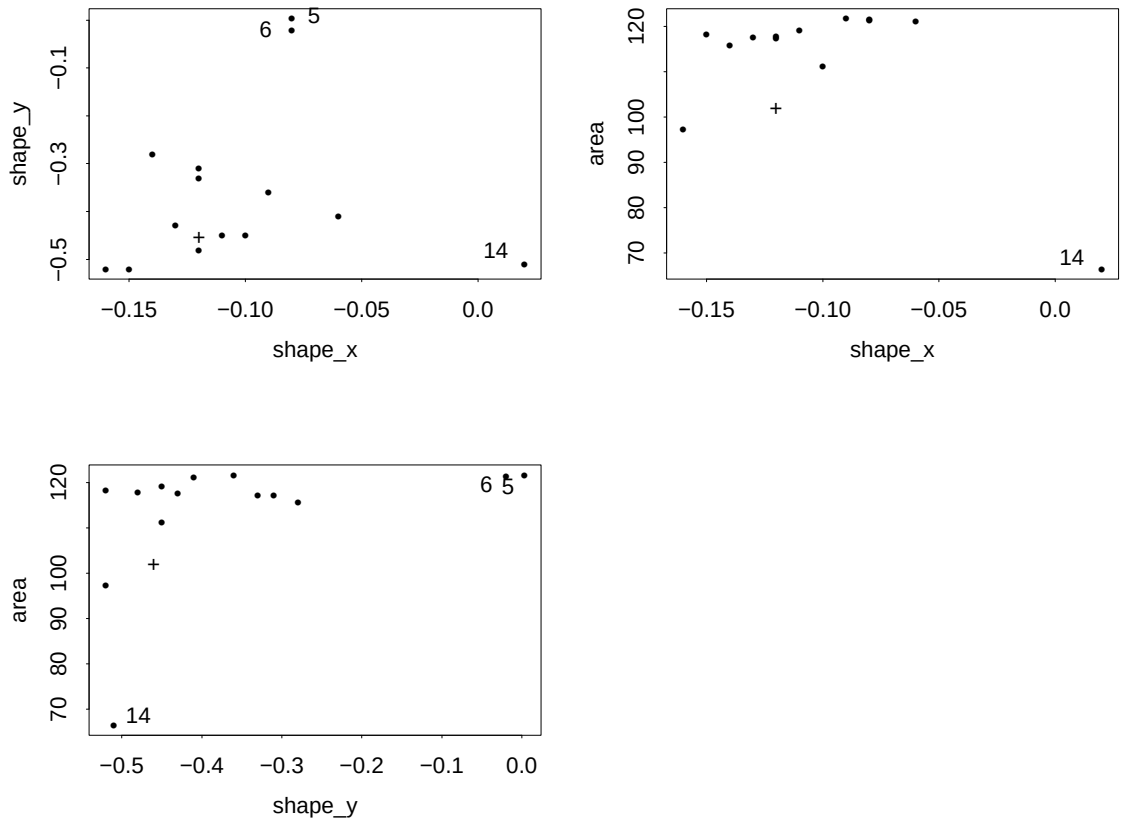


Figure 7: **Pairwise scatterplots of $shape_x$, $shape_y$ and area, confidence regions for the labor costs data. The point represented by a + sign gives the measures for the confidence region based on the complete data. Several points warrant further investigation – observation 5 (Austria), observation 6 (Netherlands) and observation 14 (Portugal). The major effect of deleting observations 5 and 6 is to produce a more symmetric interval in the y-direction, which is the measure of variation in the data. Deleting observation 14 increases symmetry in the x-direction, the mean of the data, and greatly decreases the area of the confidence region.**

were: 0.773, 0.455, 0.122, 0.075, 0.051, 0.062, 0.204, 0.275, 0.189, 0.171, 0.101, 0.139, 0.192 and 1.387. The last, Portugal, is much greater than any of the others, a further reflection of its status as an influential point. However, in comparison to critical points from the $\chi^2_{(2)}$ distribution, none of the values of $ELD_i$ are big enough to be troubling.

Finally, consider Figure 8. This Figure displays the empirical likelihood contours from the full data set. 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 are quite variable, they all fall well within the 95% confidence region defined by the complete data, which is the contour line at 6. The inference obtained from the data doesn't change, although the shape, size and orientation of the confidence regions, on which it is based, might.

# 4 Assessing Shape and Size

Beyond simply defining measures to diagnose sensitivity to particular data points in the empirical likelihood confidence regions, we also wish to know if the values of the shape and size obtained from deleting those points are in fact extreme. In this section, we consider two approaches to this problem. The first approach is to bootstrap empirical likelihood (Hall and LaScala, 1990; Lee and Young, 1999) in order to obtain distributions of the diagnostic measures. The second approach is to define a *relative jackknife influence* for the shape and size, as in Efron (1992). We demonstrate the two approaches on the length and shape measures for the sleep data discussed above.

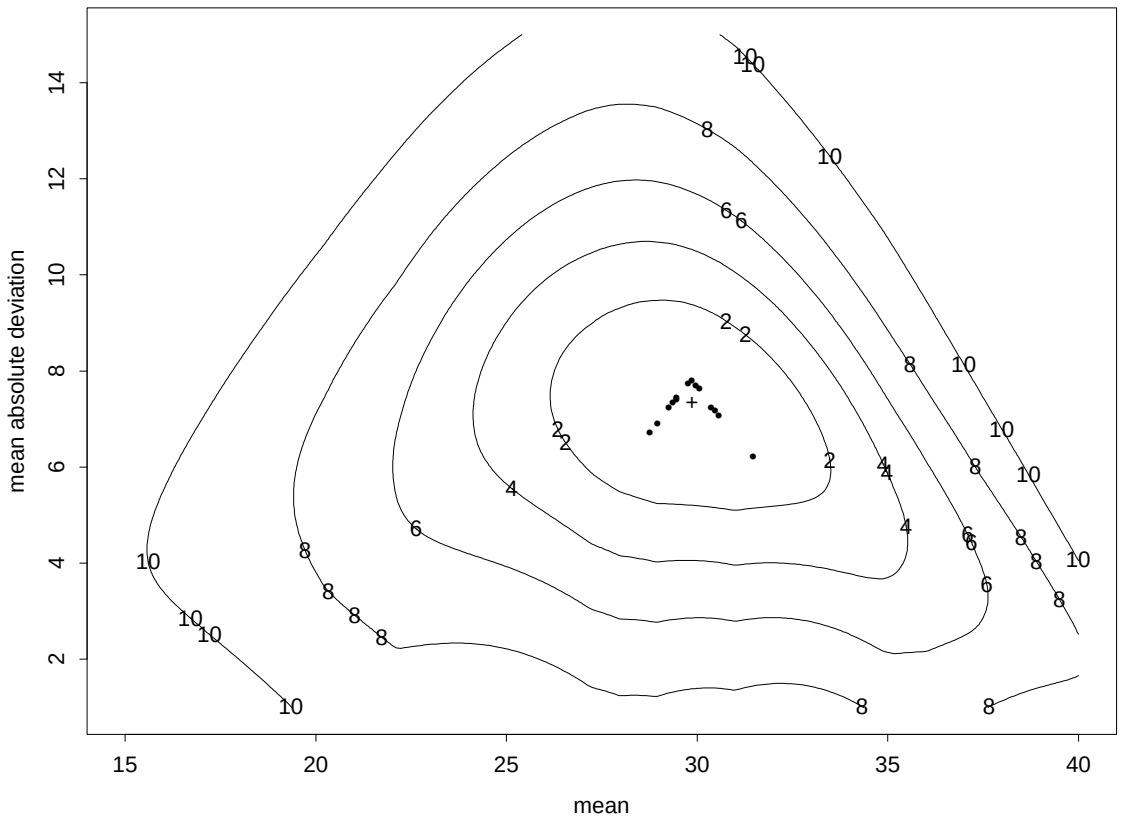


Figure 8: **Empirical likelihood contours based on the complete data. The plus sign denotes the estimates from the full data, while the points represent the estimates from each of the one case deleted data sets. The 95% confidence level for the two dimensional region is approximately 6 for $\chi^2_{(2)}$ and all estimates are well within that limit.**

## 4.1 Bootstrapping Empirical Likelihood

We drew 1000 bootstrap samples from the original data. For each, we then calculated the logarithm of the empirical likelihood ratio statistic for the mean, which in turn gave confidence intervals. On the basis of these, shape and length were found, resulting finally in bootstrap distributions for the two quantities of interest. These are plotted in Figure 9. Interestingly, the bootstrap distribution for length is bimodal, with two clear, separated peaks. One of the modes corresponds to those resamples that included the most anomalous observation; the confidence interval is then strongly pulled in the direction of that measurement, which is much larger than the others. The other mode corresponds to those resamples that didn't include the largest 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 9, the bootstrap distribution of shape is also slightly bimodal, with one peak apparently centered around 0 and the other around 0.4. The first peak corresponds to the resamples which didn't 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. This result is not surprising, in light of the previous analysis which found that deleting this point produced a symmetric empirical likelihood confidence

interval.

Bimodality of the bootstrap distribution can, in and of itself, be taken as an indicator of a point that exerts unusual influence on the empirical likelihood confidence intervals. What, then, is the implication of a unimodal bootstrap distribution? Does it make sense to compare the delete–one diagnostic measures to the tails of the unimodal distribution, as in the case of a standard bootstrap? In other words, exactly how unusual does the bootstrap behavior need to be, for us to conclude that a suspect point should indeed be investigated further, or possibly removed from the sample? Consideration of these types of pathologies is one motivation of the recent work on biased bootstrap (Hall and Presnell, 1999a; Hall and Presnell, 1999b).

## 4.2 Relative Jackknife Influence Functions

For a given statistic, $s$, Efron (1992) defines the *jackknife influence function* by

$$u_i(s) = (n-1)[s_{()} - s_{(i)}],$$

where $s_{(i)}$ is the value of the statistic on the jackknifed sample, with the $i^{th}$ observation removed, $s_{()}$ is the average of the jackknifed values of the statistic, and $n$ is the sample size. In our case, the statistic $s$ is the shape and size measures of the empirical likelihood confidence intervals. From $u_i(s)$, Efron also defines the *relative jackknife influence function*,

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

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

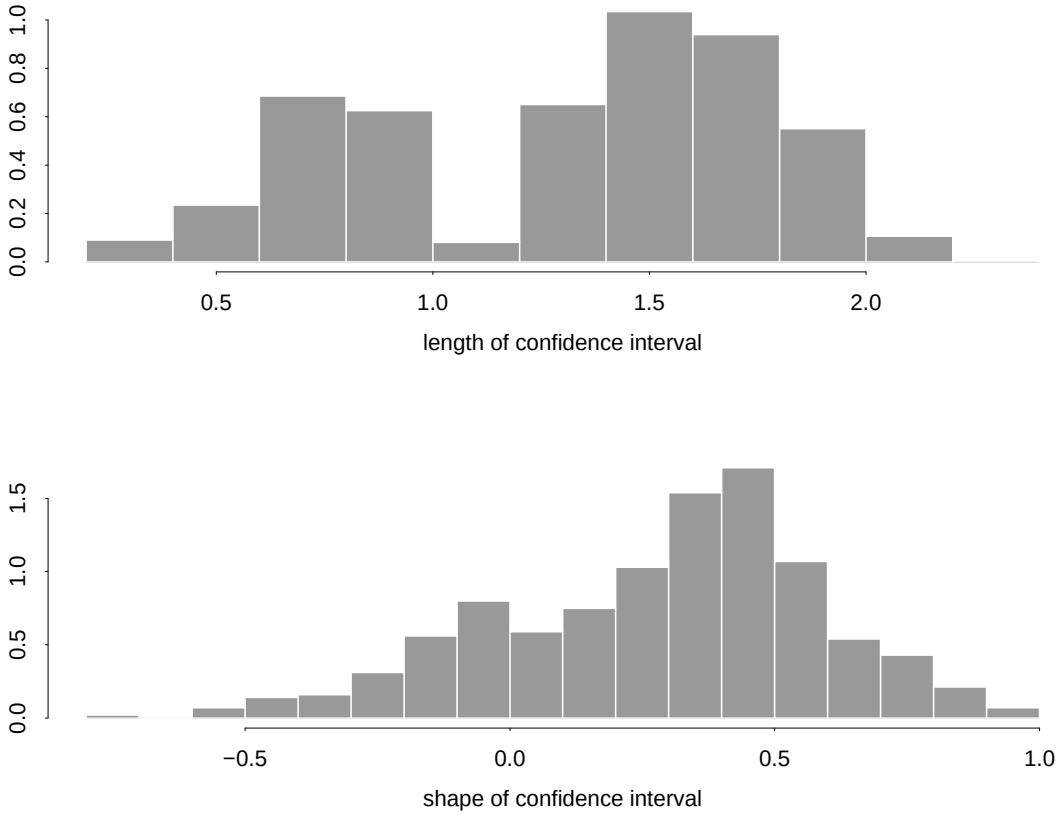


Figure 9: **The top panel shows the bootstrap distribution of length of the empirical likelihood confidence intervals. The bottom panel shows the bootstrap distribution of shape. This example is based on the sleep data, with 1000 resamples. Both distributions are bimodal, that for length much more so 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.**

The rationale for a cutoff of 2 is as follows. Simple arithmetic manipulation of the definition of $u_i^+(s)$ shows that it can be rewritten as

$$u_i^+(s) = \frac{s_{()} - s_{(i)}}{[\sum(s_{()} - s_{(i)})^2/(n-1)]^{1/2}},$$

that is, the relative jackknife influence function is in fact a $t$-like statistic for $s_{(i)}$. This interpretation is useful, since it suggests that we can define a Hotelling $T^2$-like statistic that will summarize multiple aspects of the jackknifed empirical likelihood confidence intervals at once. Therefore define

$$U_i^2(s) = n(\mathbf{s}_{()} - \mathbf{s}_{(\mathbf{i})})^T \mathbf{V}^{-1} (\mathbf{s}_{()} - \mathbf{s}_{(\mathbf{i})}),$$

where $\mathbf{s}_{()}$ is the $k$-vector of means of the jackknifed diagnostics (for example, the mean of the jackknifed area values; the mean of the jackknifed $shape_x$ values; the mean of the jackknifed $shape_y$ values), $\mathbf{s}_{(\mathbf{i})}$ is the $k$-vector of jackknifed diagnostics for the $i^{th}$ observation, and $\mathbf{V}$ is the $k \times k$ variance–covariance matrix of the jackknifed values of the diagnostics under consideration. In one dimensional empirical likelihood problems, using the diagnostics defined in this paper, $k = 2$ (length and shape). For the two dimensional problems, $k = 3$.

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 rough heuristic logic underlying Efron's recommendation for $u_i^+(s)$, we may posit for example 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, since the relative jackknife influence function (in either the univariate or the multivariate version) is calculated for each observation individually. Multiple testing situations often induce an adjusted, more conservative, cutoff for significance. Third, the relative jackknife functions are dependent, which would again seem to demand 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.

Keeping this in mind, we propose the following steps be taken to determine which observations exert undue influence on empirical likelihood confidence regions:

1. Calculate the multivariate relative jackknife influence functions, $U_i^2$.

2. For each $U_i^2$, calculate an approximate $p$-value based on $F(k, n-k)$ distribution.

3. Using the distribution–free version of the False Discovery Rate 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, since some of the observations will have values of $U_i^2$ that are considerably larger than the rest, as is the case in the following example (and several others, which are not included here).

Table 3 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. Using the distribution–free version of the false discovery rate procedure with $q = 0.05$ marks the first and the last observations as influential on the resultant empirical likelihood confidence intervals. These points would also have obviously been spotted by simply examining the values of $U_i^2$. 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.

# 5 Discussion

Once empirical likelihood confidence intervals are obtained, the shape and size measures as defined here are also easily extracted. Furthermore, they capture true features of the intervals, in that they are able to identify points whose deletion has unusual effect on the inference. It is only necessary to have the profile empirical likelihood, since confidence intervals are obtained from the $\chi^2$ approximation, and this can be used to find the edges of the intervals themselves.

Inspection of each measure on its own is informative to identify data points that might be

| $i$ | $u_i$(length) | $u_i$(shape) | $U_i^2$ | $p$-value |
|---|---|---|---|---|
| 1 | 0.4871588 | -1.46089584 | 33.0468489 | 0.00013* |
| 2 | -0.3850126 | 0.09746648 | 2.0585516 | 0.19000 |
| 3 | -0.3850126 | -0.02918488 | 1.2817849 | 0.32894 |
| 4 | -0.4203709 | -0.07874410 | 1.2980351 | 0.32492 |
| 5 | -0.4203709 | -0.07874410 | 1.2980351 | 0.32492 |
| 6 | -0.4203709 | -0.07874410 | 1.2980351 | 0.32492 |
| 7 | -0.4203709 | -0.07874410 | 1.2980351 | 0.32492 |
| 8 | -0.4557292 | -0.25495468 | 0.9850089 | 0.41455 |
| 9 | -0.3103673 | -0.57433636 | 1.5869391 | 0.26275 |
| 10 | 2.7304468 | 2.53688170 | 35.8487289 | 0.00010* |

Table 3: **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. The two data points were flagged by the FDR criterion with $q = 0.05$. The tenth data point is seen to be influential for both length and shape according to the criterion of $|u_i(s)| > 2$. 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.**

problematic. It is also useful to look at more than one measure at once, since this can reveal observations that are consistently influential. We recommend this two–pronged approach. In all three of the examples we considered, the shape and size diagnostics picked out some points that deserved further attention. The shape measure, similarly to DiCiccio and Monti's (2001) $F_3$ diagnostic, also provides a measure of symmetry of the likelihood, which might be important for other purposes. On the other hand, at least in the examples we considered here, there appeared to be less information in the empirical likelihood displacement, which generally had small values far from the critical $\chi^2$ points.

Obtaining a distribution for the diagnostics is thus also important, because without this it is impossible to know whether values that appear to be extreme really should be taken as such. The concern is that the proposed diagnostics 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.

A bimodal bootstrap distribution, with one mode corresponding to samples with the suspected observation and the other to samples without, indicates an observation with an extreme effect, that 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 consider a data point to be influential when the shape and size diagnostics with that observation deleted, are in the far tails of the distribution. In either situation we are looking for departures from ordinary behavior, what might be termed "bootstrap pathologies."

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. Univariate influence functions are then 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 may 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: (i) strongly influential on all measures; (ii) strongly influential on one or two measures; (iii) mildly influential on all measures; (iv) not influential at all. A robust data set should have most points in category (iv) and few (or none) in categories (i) and (ii).

The current work draws heavily on ideas from standard likelihood and bootstrap theories, and in many cases a direct transfer of results has been possible. Examples are the skewness and kurtosis measures of Sprott (1980) and Slate (1999), which have counterparts in the statistics defined by DiCiccio and Monti (2001) for empirical likelihood, and the empirical likelihood displacement measure, defined here, which is completely analogous to its

parametric version. The closeness of parametric and empirical likelihood cannot always be taken advantage of with such ease. Kass, Tierney and Kadane (1989), for instance, derive diagnostics for case influence which are based on the relationship $L_{new}(\theta) = L(\theta)\rho(\theta)$, where $L_{new}(\theta)$ is the likelihood, or more generally the posterior distribution, with case $i$ deleted and $\rho(\theta)$, which can be any perturbation, is, for delete–one data, $\rho(\theta) = 1/L_i(\theta)$, the contribution of the $i^{th}$ observation to the complete likelihood, $L(\theta)$. The approach of Kass *et al.* (1989) is elegant, and sheds light on the meaning of the likelihood displacement, yet is not applicable to empirical likelihood, since we cannot write $EL_{new}(\theta) = EL(\theta)\rho(\theta)$. The contribution of the $i^{th}$ case comes in also through the Lagrange multiplier defining empirical likelihood, which is therefore different for $EL_{new}(\theta)$ and $EL(\theta)$. 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.

# 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. and Weisberg, S. (1982) *Residuals and Influence in Regression.* New York:

Chapman and Hall.

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

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.

Hall, P. and Presnell, B. (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. *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 Edition. New Jersey: Prentice Hall.

Owen, A.B. (1990) Empirical likelihood confidence regions. *Annals of Statistics*, **18**, 90–120.

Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *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.