Ann. Inst. Statist. Math.
Vol. 50, No. 4, 697–714 (1998)

COMBINED AND LEAST SQUARES EMPIRICAL LIKELIHOOD

BRUCE M. BROWN[1] AND SONG XI CHEN[2]

[1] *Department of Mathematics, University of Tasmania, GPO Box 252L, Hobart, TAS 7001, Australia*
[2] *School of Statistical Science, La Trobe University, VIC 3083, Australia*

(Received February 13, 1997; revised October 30, 1997)

Abstract. In conventional empirical likelihood, there is exactly one structural constraint for every parameter. In some circumstances, additional constraints are imposed to reflect additional and sought-after features of statistical analysis. Such an augmented scheme uses the implicit power of empirical likelihood to produce very natural adaptive statistical methods, free of arbitrary tuning parameter choices, and does have good asymptotic properties. The price to be paid for such good properties is in extra computational difficulty. To overcome the computational difficulty, we propose a 'least-squares' version of the empirical likelihood. The method is illustrated by application to the case of combined empirical likelihood for the mean and the median in one sample location inference.

Key words and phrases: Empirical likelihood, least squares empirical likelihood, maximum likelihood estimate, mean, median.

1. Introduction

Empirical likelihood, introduced by Owen (1988, 1990), is a remarkable computation device which enables the development of parameter tests, free of any distributional assumptions, but with identical first-order asymptotics to those of classical likelihood ratio tests. Empirical likelihood (EL) is defined by maximizing a multinomial likelihood subject to various structural constraints. To illustrate, an empirical likelihood of an unknown mean θ allocates notional weight p_i to X_i, the i-th of n observations, then defines the empirical likelihood function of θ as

$$L(\theta) = \max\left(\prod_{i=1}^{n} p_i\right) \tag{1.1}$$

subject to $\sum_i p_i = 1$, $p_i \geq 0$, and the structural constraint

$$\sum_{i=1}^{n} p_i(X_i - \theta) = 0. \tag{1.2}$$

The structural constraint reflects the fact that θ is a mean. Different types of parameters yield different structural constraints; for instance, if θ had been a population median, then (1.2) would be replaced by

$$\sum_i p_i \operatorname{sgn}(X_i - \theta) = 0 \tag{1.3}$$

where sgn() is the sign function.

It has been shown in a wide range of situations that an inference based on empirical likelihood has the same order of accuracy as the bootstrap. At the same time it has some good properties not shared by the bootstrap, such as respecting the range of parameters and determining the shape and orientation of a confidence region naturally from the data; see Hall and La Scala (1990) and Chen (1993, 1994). Qin and Lawless (1994) studied constructing empirical likelihood for parameters determined by estimation equations. They showed that when the number of structural constraints exceeds the number of parameters the maximum empirical likelihood estimate is no longer equivalent to the bootstrap estimate. The log empirical likelihood ratio is still asymptotically chi-square distributed which enables construction of confidence regions and tests for the parameters. Recently, Wood *et al.* (1996) study the computation of empirical likelihood in situations where the constraints are not linear, like U-statistics, and the computation is difficult. They propose a method of sequential linearization of the constraints in the computation of the empirical likelihood.

The character of empirical likelihood inference is determined by the constraints, and it is of great interest to see how empirical likelihood behaves when additional structural constraints are imposed. When the number of structural constraints exceeds the number of parameters, it is natural to hope that the resulting empirical likelihood might have 'the best of all worlds', in adapting naturally to whichever of the structural constraints that puts the most pressure on the data. For instance, in estimating a location parameter θ, the structural constraint (1.2) for the mean might be combined with (1.3) for the median to provide robustness. If the data were 'well-behaved', an estimate like the sample mean might result, but if severe outliers were present, something closer to the sample median could occur. This combined empirical likelihood can be viewed as a way of using the implicit power of empirical likelihood to produce natural adaptive statistical methods, entirely free of subjective choices such as tuning parameters.

The present paper studies the use of combined empirical likelihood as a way of constructing natural adaptive statistical inference. It turns out that the asymptotic properties of combined empirical likelihood are very good, but that there is a price to be paid for such good properties, namely, in extra computational difficulties. It becomes necessary to construct instead a smoothed version, making combined empirical likelihood schemes become workable practical propositions.

Our proposal for smoothing centers around the observation that in regular empirical likelihood, maximizing $\sum \log(np_i)$ can be replaced by maximizing $\sum g(np_i)$, where g is any concave function, smooth near $np_i = 1$. The reason is that the constraint $\sum p_i = 1$ enables an arbitrary linear term in np_i to be extracted from g, until $g \leq 0$ with maximum 0 at $np_i = 1$, whence concavity makes this g locally equivalent to $\log(np_i) - np_i + 1$, with consequent identical asymptotic behaviour. The simplest version of g is the least-squares choice $g(np) = (np-1)^2$. This 'least-squares' version of empirical likelihood is actually the Euclidean likelihood discussed briefly in Owen (1991). However, to emphasize its link with empirical likelihood, we call it least-squares empirical likelihood.

The rest of the paper is organised as follows. In Section 2 we introduce the least-squares empirical likelihood and compare it with the empirical likelihood in a general setting. Sections 3 and 4 contain results from applying least-squares empirical likelihood to combined mean and median location parameter inference. Results from a simulation study are presented in Section 5. Section 6 is a short discussion section, and proofs are deferred to the Appendices.

2. Least-squares empirical likelihood

In this section we introduce the least-squares empirical likelihood in a general setting and study its asymptotic behaviour.

Let $Z_1(\theta), Z_2(\theta), \ldots, Z_n(\theta)$ be k dimensional independent but not necessarily identically distributed random vectors, relying on an unknown parameter θ of dimension p. We assume θ has a true value θ_0 and $E\{Z_i(\theta_0)\} = 0$. In this paper, we concentrate on situations when k, the number of structural constraints, is larger than p, the dimension of the parameter θ. The components of the $\{Z_i(\theta)\}$ provide the structural components for the constraints in empirical likelihood.

The log empirical likelihood ratio for θ is

$$\ell(\theta) = \min\left[-2\sum \log(np_i)\right]$$

where p_i satisfy

$$\sum_{i=1}^{n} p_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} p_i Z_i(\theta) = 0. \tag{2.1}$$

Using Lagrange multipliers, it can be shown that the optimal p_i are

$$p_i = n^{-1}\{1 + \lambda^T Z_i(\theta)\}^{-1},$$

where $\lambda \in R^k$ satisfies

$$\sum \frac{Z_i(\theta)}{1 + \lambda^T Z_i(\theta)} = 0. \tag{2.2}$$

Thus

$$\ell(\theta) = 2\sum \log\{1 + \lambda^T Z_i(\theta)\}.$$

The computation of $\ell(\theta)$ involves either solving λ as a root of equation (2.2) by Newton's method (Hall and La Scala (1990)) or minimizing a dual function $f(\lambda) = -\sum \log\{1 + \lambda^T Z_i(\theta)\}$ within a compact set $D = \{\lambda \mid 1 + \lambda^T Z_i(\theta) \geq 0\}$ (Owen (1990)). However, when $k > p$, the number of structural constraints exceeds the number of parameters, the competition among the constraints may cause the empirical likelihood not to exist at some θ values, and make the computation of empirical likelihood impossible at those parameter values.

In order to make the idea of the combined empirical likelihood computationally possible, we consider the least-squares empirical likelihood, which is defined as

$$lsl(\theta) = \min \sum (np_i - 1)^2 \tag{2.3}$$

where p_i are subject to (2.1). It is easy to show that $lsl(\theta)$ is just the dominant term in the Taylor expansion of $\ell(\theta)$ at $p_i = 1/n$. Because $lsl(\theta) = n^2 Q(\theta) - n$ where $Q(\theta) = \min \sum p_i^2$ subject to (2.1), we shall find $Q(\theta)$ directly.

Using Lagrange multipliers $\alpha = (\alpha_1, \ldots, \alpha_p)^T$, choose p_i to minimize

$$\sum_i p_i^2 + \alpha_0 \sum_i p_i + \alpha^T \sum_i p_i Z_i(\theta).$$

This is quadratic in $\{p_i\}$, minimized by

$$p_i = -\frac{1}{2}\left\{\alpha_0 + \sum_j \alpha_j Z_{ij}(\theta)\right\}. \tag{2.4}$$

Let $\alpha^T = (\alpha_0, \alpha_1, \ldots, \alpha_k)$, $V^T = (V_1, \ldots, V_k)$ where $V_j = \sum_i Z_{ij}(\theta)$, and $R = (R_{jj'})_{k\times k}$ where $R_{jj'} = \sum_i Z_{ij}(\theta) Z_{ij'}(\theta)$. From (2.4) the constraints in (2.1) yield

$$(1\ 0\ \cdots\ 0)^T = -\frac{1}{2}\begin{pmatrix} n & V^T \\ V & R \end{pmatrix}\alpha. \tag{2.5}$$

Define $e_1 = (1, 0, \ldots, 0)^T$ and

$$B = \begin{pmatrix} n & V^T \\ V & R \end{pmatrix},$$

then (2.5) can be expressed in matrix form as $e_1 = -\frac{1}{2}B\alpha$. This and (2.4) imply that the optimal p_i are

$$p_i = (1, Z_i^T(\theta))B^{-1}e_1 = n^{-1} + n^{-1}(n^{-1}V - Z_i(\theta))^T H^{-1} V$$

where $H = R - n^{-1}VV^T$. Like the empirical likelihood the least-squares empirical likelihood also assigns unequal weights p_i. However, the least-squares weights can be negative, for instance when the location parameter θ is outside the convex hull of the data. However, as pointed out by Owen (1991), this can be advantageous for allowing a small sample based confidence interval to be extended outside the convex hull of the data.

Some simple algebra reveals that

$$Q(\theta) = \frac{1}{4}\alpha^T B \alpha = u_1^T B^{-1} u_1 = (B^{-1})_{11}.$$

Therefore,

$$lsl(\theta) = V^T H^{-1} V. \tag{2.6}$$

A referee has pointed out that (2.6) is equivalent to a generalized method of moment estimator considered in Hansen (1982).

In more conventional adaptive schemes, $Z_{ij}(\theta)$ could be the derivative with respective to θ_j of some convex function $\Phi(r_i)$ of the i-th residual, in some linear or other model. Here, because $k > p$, there are additional components in each Z_i and the particular function Φ could also depend on j. The null means of $\{Z_i\}$ are zero. The natural test statistic is $n^{-1}V$, and its estimated covariance matrix is $n^{-2}H$. Therefore, an asymptotically χ^2_k test statistic is $(n^{-1}V)^T(n^{-2}H)^{-1}(n^{-1}V) = V^T H^{-1} V$, which is the same form of least-squares empirical likelihood given in (2.6). But, the least-squares empirical likelihood can avoid tuning parameter choice or nuisance parameter estimation by the optimization procedure in (2.3). Another advantage of combined empirical likelihood is through the resulting internal studentization. In contrast, the sample covariance matrix H has to be used in conventional inference. This could be a problem when the sample variance estimates are erratic.

Next we compare $lsl(\theta)$ and $\ell(\theta)$ for θ within an $n^{-1/2}$-neighbourhood of θ_0. According to the expansions (A.1) and (A.2) developed in Appendix A.1, we see that both $\ell(\theta)$ and $lsl(\theta)$ have the same first order term. The second order term and one of the third order terms of $lsl(\theta)$ also appear in the expansion for $\ell(\theta)$. Therefore, $lsl(\theta)$ approximates $\ell(\theta)$ at least to the first order for θ within an $n^{-1/2}$-neighbourhood of θ_0.

It is clear from (A.1) and (A.2) that at $\theta = \theta_0$, both $\ell(\theta)$ and $lsl(\theta)$ have a limiting chi-square distribution under some mild conditions, for instance the ones given in Chen (1994, p. 284). These allow us to construct confidence intervals for θ by looking at chi-square tables. Similar calculations to those of Chen (1993) reveal that the third order cumulant of the sign root of $lsl(\theta_0)$ is not of an order smaller than n^{-1}. This implies that the least-squares empirical likelihood based confidence intervals are not Bartlett-correctable. However, the coverage accuracy can be improved by bootstrap calibration. A description of the bootstrap calibrations of empirical likelihood is available in Hall and Owen (1993) and Chen and Hall (1993), and can be directly applied for calibrating least-squares empirical likelihood.

3. Combining mean and median

We now concentrate on inference for a location parameter θ defined by (1.2) and (1.3). Suppose that $X_1, \ldots, X_n$ is a random sample from a distribution which is symmetric about θ. Then, inference on θ can be made by treating θ both as the population mean and the population median. In a related work, Jing (1995) studied empirical likelihood inference for the mean by assuming a symmetric distribution, but without the median constraint.

3.1 *Maximum likelihood estimates*

The log empirical likelihood and least-squares empirical likelihood ratios for θ are

$$(3.1) \qquad \ell(\theta) = \min_{\sum p_i = 1} -2 \sum_{i=1}^{n} \log(np_i)$$

$$(3.2) \qquad lsl(\theta) = \min_{\sum p_i = 1} \sum_{i=1}^{n} (np_i - 1)^2$$

subject to the mean constraint (1.2) and the median constraint (1.3).

Due to discontinuity of the sign function, both $\ell(\theta)$ and $lsl(\theta)$ jump at each data point X_j and are continuous between two successive data points (X_j, X_{j+1}) for $j = 1, \ldots, n-1$, assuming the data have been ranked in ascending order. The maximum empirical likelihood estimate or the maximum least-squares empirical likelihood estimate is either at a stationary point inside (X_j, X_{j+1}) or at a data point X_j for some j.

Define $\bar{X}_j = j^{-1}\sum_{i=1}^{j} X_i$ and $\bar{X}^{n-j} = (n-j)^{-1}\sum_{i=j+1}^{n} X_i$ as the lower-j and the upper-$(n-j)$ sample means respectively, and let

$$\hat{\theta}_j = \frac{1}{2}(\bar{X}_j + \bar{X}^{n-j}) \qquad \text{if} \quad \hat{\theta}_j \in (X_j, X_{j+1}).$$

The following theorem, whose proof is deferred to the Appendix, reveals a close relationship between $\ell(\theta)$ and $lsl(\theta)$.

THEOREM 1. *$lsl(\theta)$ and $\ell(\theta)$ have the same stationary points, at some $\hat{\theta}_j$.*

A good gauge for the behaviour of the two likelihood curves is $\hat{\theta}_j = \frac{1}{2}(\bar{X}_j + \bar{X}^{n-j})$. If $\hat{\theta}_j$ falls outside (X_j, X_{j+1}), both $\ell(\theta)$ and $lsl(\theta)$ are monotone in (X_j, X_{j+1}). If $\hat{\theta}_j < X_j (> X_j)$, both are monotonically increasing (decreasing). Therefore, the two likelihood functions increase or decrease at the same time, which indicates that the least-squares empirical likelihood basically captures most of information in the empirical likelihood. Note in passing that $\frac{1}{2}(\bar{X}_j + \bar{X}^{n-j})$ is something which has characteristics of both the mean and the median.

As there are three possible definitions for sgn(0), there are three values for both $\ell(\theta)$ and $lsl(\theta)$ at any data point X_j for $2 \le j \le n-1$. They are $\{\ell(X_j^+), \ell(X_j^0), \ell(X_j^-)\}$ and $\{lsl(X_j^+), lsl(X_j^0), lsl(X_j^-)\}$ respectively. Neither likelihood function has finite left limits at X_1 nor right limits at X_n. However, the probability that the maximum likelihood estimates are at X_1 or X_n is zero. So, we need to consider only X_j for $j = 2, \ldots, n-1$. We define

$$\ell(X_j) = \min\{\ell(X_j^-), \ell(X_j^0), \ell(X_j^+)\} \qquad \text{and}$$
$$lsl(X_j) = \min\{lsl(X_j^-), lsl(X_j^0), lsl(X_j^+)\}.$$

Let $\hat{\theta}_{mel}$ and $\hat{\theta}_{mlsl}$ be the maximum empirical likelihood and maximum least-squares empirical likelihood estimates for θ respectively. Then, each of them is either a $\hat{\theta}_j$ satisfying $X_j < \hat{\theta}_j < X_{j+1}$, or a data point X_j, whichever has the smallest likelihood value.

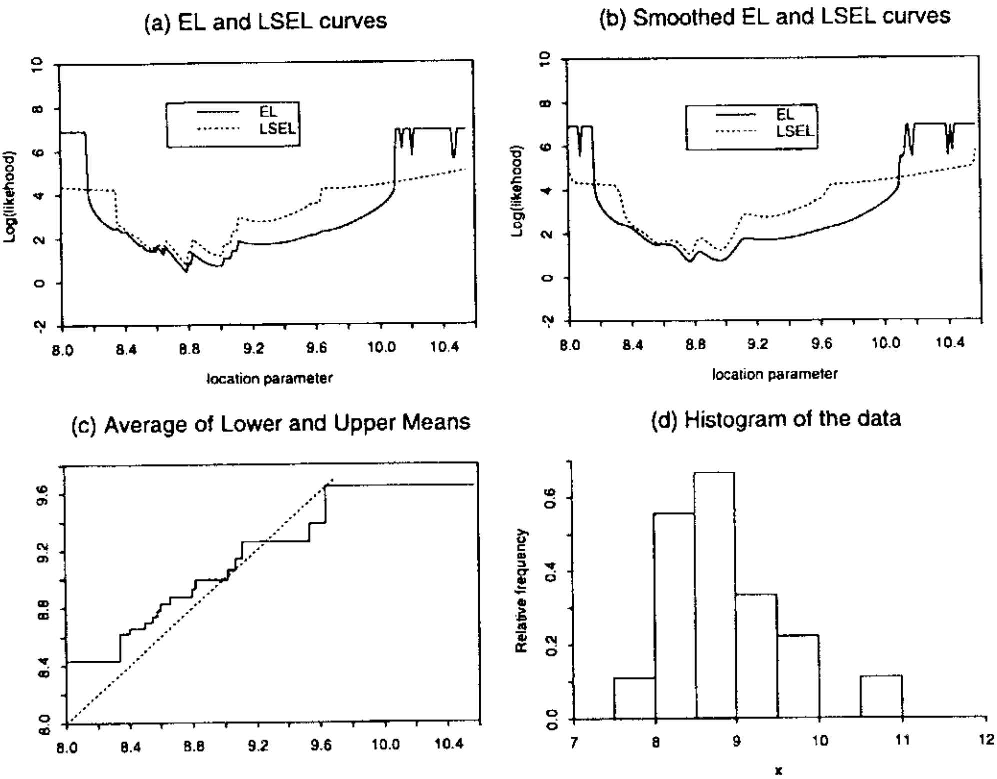

Fig. 1. Empirical likelihood and least-squares empirical likelihood curves for the location parameter of the Short's data set.

3.2 *An example*

To illustrate our findings in Subsection 4.1, the two likelihood curves are plotted in (a) of Fig. 1 for a data set of Stigler (1977), which is Short's 1763 measurements of the parallax of the sun. There are 18 observations ranging from $X_1 = 7.99$ to $X_n = 10.57$. In the figure we also plot kernel smoothed empirical likelihood and its least-squares empirical likelihood curves by smoothing the sgn() function in the median constraint. The median constraint becomes

$$\sum p_i G\left(\frac{X_i - \theta}{h}\right) = 0$$

where $G(x) = \int_{y<x} K(y)dy$, K is the Gaussian kernel function, and h is the smoothing bandwidth chosen by a standard method in kernel smoothing. To identify the stationary points $\hat{\theta}_j$, we plot in (c) $\hat{\theta}_j = \frac{1}{2}(\bar{X}_j + \bar{X}^{n-j})$ as a step function of X_j and the identity function. Clearly, $\hat{\theta}_j$ are the crossing points of the two functions. Also a histogram of the data is given in (d).

We observe from (a) and (b) of Fig. 1 that both likelihood curves have similar shapes but are not convex at all. When θ is in the middle of the data range, the likelihood curves go up and down representing a tense competition between the mean and the median constraints. When θ is in the two extreme data ranges, there are plateaus associated with the empirical likelihood curves. These plateaus represent computational difficulties for empirical likelihood as an optimal weight p_i does not exist and an artificial value of 1000 has to be assigned to rescue the computer program from an infinite number of iterations. Note that the height of the plateaus is $\log(1000) = 6.91$. We observe that when the empirical likelihood curves have finite values, both empirical likelihood and least-squares empirical likelihood curves in Fig. 1(a) increase or decrease at the same time. We also see that both likelihood curves have the same three stationary points, which correspond to the three crossing points in Fig. 1(c) as predicted by Theorem 1. By replacing the sign function sgn() with the Gaussian kernel function, we see some of the ups-and-downs in the likelihood curves are removed. However, there are still plateaus with the kernel smoothed empirical likelihood curve, which means that kernel smoothing does not solve the computation difficulty.

4. Asymptotic results for LSEL

In this section we show first that the maximum least-squares empirical likelihood estimate for the location parameter θ is asymptotically normally distributed. Then we present a result which has implications for least-squares empirical likelihood confidence intervals.

4.1 *Asymptotic normality*

The asymptotic normality will be derived heuristically, as it is a complicated exercise to do rigorously. However, a rigorous proof can be pursued alone the lines outline below. Without loss of generality we take the true parameter value $\theta_0 = 0$. Define

$$\hat\sigma_x^2 = n^{-1}\sum(X_i-\bar X)^2, \qquad \rho(\theta) = n^{-1}\sum(X_i-\bar X)\,\mathrm{sgn}(X_i-\theta) \qquad \text{and}$$
$$\hat\sigma_{\mathrm{sgn}}^2(\theta) = n^{-1}\sum \mathrm{sgn}(X_i-\theta)^2 - \Big\{n^{-1}\sum \mathrm{sgn}(X_i-\theta)\Big\}^2.$$

According to (2.6), $\hat\theta_{mlsl}$ maximizes $V^T H^{-1} V$ where $V = n(\bar X-\theta, n^{-1}\sum \mathrm{sgn}(X_i-\theta))$ and

$$H = n\begin{pmatrix} \hat\sigma_x^2 & \rho(\theta) \\ \rho(\theta) & \hat\sigma_{\mathrm{sgn}}^2(\theta)\end{pmatrix}.$$

For the simplicity of notation, we write $\hat\theta_{mlsl}$ as θ. It may be shown that $\theta = \hat\theta_{mlsl} = O_p(n^{-1/2})$. For small θ write V and H as

$$V = V_0 + \theta V_1 + \theta^2 V_2 + \cdots \qquad \text{and}$$
$$H = H_0 + \theta H_1 + \theta^2 H_2 + \cdots. \tag{4.1}$$

Denote $\bar X = Z_1/\sqrt{n}$ by S_1, $n^{-1}\sum \mathrm{sgn}(X_i) = Z_2/\sqrt{n}$ by S_2 and $f = f(0)$, the value of the probability density function of X_1 at the true center. Note that

$$(Z_1, Z_2) \to N\left((0,0), \begin{pmatrix}\sigma^2 & \beta \\ \beta & 1\end{pmatrix}\right) \qquad \text{in distribution}$$

as $n \to \infty$, where $\sigma^2 = \mathrm{Var}(X_1)$ and $\beta = E(|X_1|)$. Checking details carefully gives

$$
\begin{aligned}
&H_0 = n\begin{pmatrix} \sigma^2 & \beta \\ \beta & 1 \end{pmatrix}, \qquad H_1 = O_p(n^{1/2}), \qquad H_2 = O_p(n), \\
&V_0 = n^{1/2}(Z_1, Z_2)^T, \qquad V_1 = -n(1, 2f)^T \quad \text{and} \quad V_2 = O_p(n^{1/2}).
\end{aligned} \tag{4.2}
$$

Use (4.1) and the expansion

$$(H_0 + \Delta)^{-1} = H_0^{-1} + H_0^{-1}\Delta H_0^{-1} + H_0^{-1}\Delta H_0^{-1}\Delta H_0^{-1} + \cdots .$$

We can approximate to $V^T H^{-1} V$ in small powers of θ and then choose θ to minimize. Assembling the terms which are linear and quadratic in θ, and using the orders of magnitude given in (4.2), eventually yields

$$\hat{\theta}_{mlsl} = -\frac{V_1^T H_0^{-1} V_0}{V_1^T H_0^{-1} V_1}\{1 + o_p(1)\}.$$

In turn, from the expressions for V_0, V_1 and H_0 in (4.2),

$$n^{1/2}\hat{\theta}_{mlsl} = \frac{Z_1(1 - 2\beta f) + Z_2(2\sigma^2 f - \beta)}{(1 - 4\beta f + 4\sigma^2 f^2)}. \tag{4.3}$$

Noting that $\mathrm{Var}(Z_1) = \sigma^2$, $\mathrm{Var}(Z_2) = 1$ and $\mathrm{Cov}(Z_1, Z_2) = \beta$ gives the asymptotic variance for $\hat{\theta}_{mlsl}$ as

$$n\,\mathrm{Var}(\hat{\theta}_{mlsl}) = \frac{\sigma^2 - \beta^2}{(1 - 2\beta f)^2 + 4f^2(\sigma^2 - \beta^2)}. \tag{4.4}$$

As checks, the special cases $X \sim N(0, \sigma^2)$ and $X \sim \mathrm{Laplace}(\sigma)$ give in (4.3)

$$\hat{\theta}_{mlsl} = S_1 \quad \text{and} \quad \hat{\theta}_{mlsl} = S_2$$

respectively. This means that in two classical cases where either the mean or the median is efficient for θ, the combined least-squares empirical likelihood estimator reduces asymptotically to each of those two cases.

It may be shown from (4.4) that the asymptotic variance of $\hat{\theta}_{mlsl}$ is less than $\mathrm{Var}(\bar{X})$ and $\mathrm{Var}(\hat{m})$ respectively. In fact, the asymptotic variance in (4.4) coincides with that of an arbitrary linear combination of sample mean and sample median, with weights chosen to minimize asymptotic variance. Consequently, least squares combined empirical likelihood estimation is at least as efficient as both sample mean and sample median. This extends a result in Qin and Lawless (1994) which maintains that the variance of the empirical likelihood estimate cannot decrease if a constraint is dropped, provided that all the constraints are differentiable with respect to θ. Here the median constraint is not differentiable.

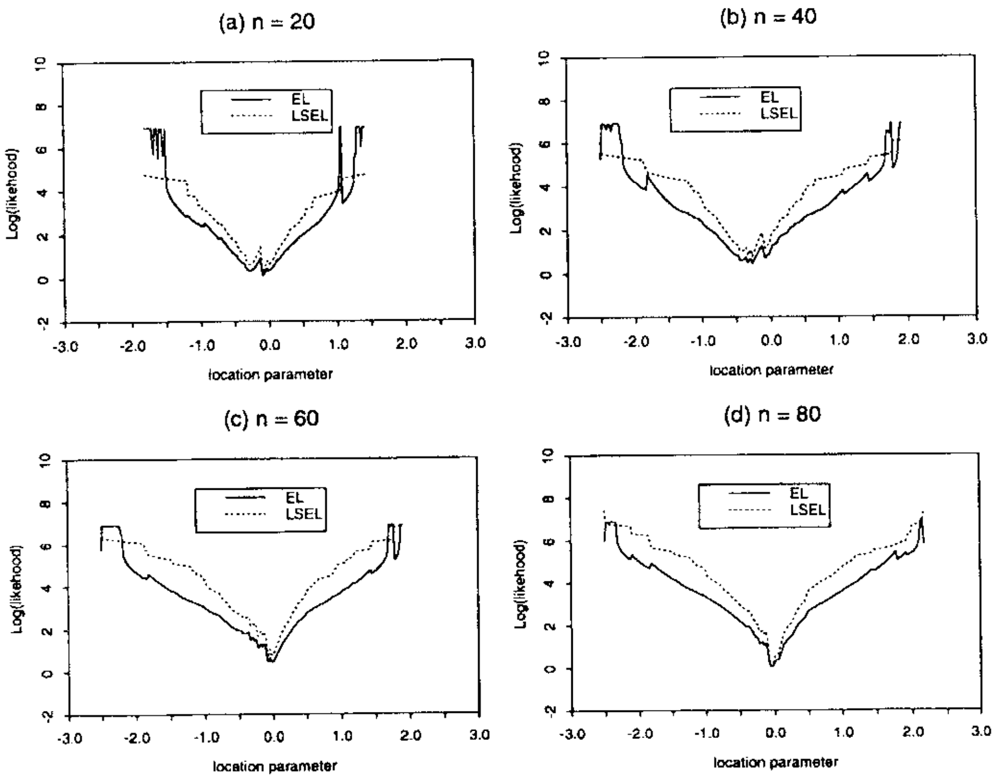

Fig. 2. EL and LSEL curves for the location parameter for simulated $N(0,1)$ data sets.

4.2 *An issue on confidence intervals*

For a regular empirical likelihood where $k = p$, Hall and La Scala (1990) showed that if θ is a smooth function of means, the empirical likelihood confidence interval for θ (region) is connected without voids. In case of $k > p$, the Wilks' theorem was established by Qin and Lawless (1994) for estimating functions derived from an independent and identically distributed sample. We have shown in Section 2 and Appendix A.1 that Wilks' theorem holds for more general cases. However, there is no guarantee that the empirical likelihood confidence intervals (regions) are connected without voids when $k > p$. Indeed, we have observed in Fig. 1 that both $\ell(\theta)$ and $lsl(\theta)$ are neither continuous nor convex for the location parameter.

However, the following theorem implies that in the combined mean and the median case least-squares empirical likelihood confidence intervals for θ are asymptotically connected without voids.

THEOREM 2. *Suppose $j = [np] + 1$ for a fixed $p \in (0,1)$ where $[np]$ is the largest integer smaller than np. Then, as $n \to \infty$,*

$$\text{(i)} \quad P\left\{\frac{1}{2}(\bar{X}_j + \bar{X}^{n-j}) \geq X_{j+1}\right\} \to 1 \quad \text{if} \quad p < 1/2,$$

(ii) $P\left\{\frac{1}{2}(\bar{X}_j + \bar{X}^{n-j}) \leq X_j\right\} \to 1 \quad \textit{if} \quad p > 1/2 \quad \textit{and}$

(iii) $lsl(X_j^-)/lsl(X_j^+) = 1 + O_p(n^{-1}) \quad \textit{and}$

$lsl(X_j^0)/lsl(X_j^+) = 1 + O_p(n^{-1}).$

The proof is deferred to Appendix A.3. The theorem means that: (a) asymptotically $lsl(\theta)$ is monotonically non-increasing (non-decreasing) within each (X_j, X_{j+1}) if $j < n/2$ ($j > n/2$); (b) asymptotically the three least-squares empirical likelihood values at each X_j become closer and closer. Therefore, asymptotically $lsl(\theta)$ is a continuous convex function, and the least-squares empirical likelihood confidence intervals are connected without voids.

To illustrate the asymptotic results, in Fig. 2 we plot the empirical likelihood and least-squares empirical likelihood curves based on the standard normal samples of size $n = 20, 40, 60$ and 80. We observe that as the sample size increases, both curves become smoother. We also find that both likelihood curves are very close to each other around $\theta_0 = 0$. This is reassuring as we have shown in Section 2 that $lsl(\theta)$ approximates $\ell(\theta)$ up to at least the first order when θ is within an $n^{-1/2}$ neighbourhood of θ_0.

5. Empirical results

We have indicated in Section 1 that the objective for the combined empirical likelihood is to seek 'the best of all worlds', hoping that the empirical likelihood would adapt naturally to whichever of the structural constraints put the most pressure on the data. For combining the mean and the median, if the data are 'well-behaved', either the maximum empirical likelihood or least-squares empirical likelihood estimates will be close to $\bar{X}$, the maximum likelihood estimates under the single mean constraint, but if severe outliers are present they will be closer to the sample median $\hat{m}$.

To see how maximum likelihood estimates behave, we deliberately corrupt the Short's data set used in Fig. 1 by increasing the largest data value x_n by 5 each time for 7 times. Table 1 contains the maximum likelihood estimates $\hat{\theta}_{el}$ and $\hat{\theta}_{mlsl}$, the sample mean $\bar{X}$ and the sample median $\hat{m}$. We see that (i) both $\hat{\theta}_{el}$ and $\hat{\theta}_{lsel}$ are closer to the sample mean when no corruption is made and $x_n = 10.57$; (ii) as x_n is corrupted, both $\hat{\theta}_{el}$ and $\hat{\theta}_{lsel}$ are very robust and become closer to the sample median. It seems that both empirical likelihood and least-squares empirical likelihood have achieved 'the best of all worlds'.

We also conducted a simulation study designed to investigate the bias and the standard error of the least-squares empirical likelihood estimate $\hat{\theta}_{lsel}$ in comparison with $\bar{X}$ and $\hat{m}$, where the latter two are the maximum empirical likelihood estimates under the single mean and the single median constraint respectively. We generate random samples from the standard normal distribution $N(0,1)$ and a Laplace(1.0) distribution with a density function

$$f(x) = \frac{1}{2}\exp(-|x|),$$

Table 1. Robustness of θ_{el} and θ_{lsel} on Short's 1763 sun data.

x_n	10.57	15.57	20.57	25.57	30.57	35.57	40.57	45.57
$\hat{\theta}_{lsel}$	8.80	8.80	8.66	8.66	8.66	8.60	8.60	8.60
$\hat{\theta}_{el}$	8.80	8.80	8.80	8.50	8.50	8.50	8.50	8.50
$\bar{X}$	8.83	9.10	9.38	9.65	9.93	10.21	10.49	10.77
$\hat{m}$	8.63	8.63	8.63	8.63	8.63	8.63	8.63	8.63

Table 2. Simulated bias and standard error (S.E.) for $\hat{\theta}_{lsel}$, $\bar{X}$ and $\hat{m}$.

n	$\bar{X}$		$\hat{\theta}_{lsel}$		$\hat{m}$	
	bias	S.E.	bias	S.E.	bias	S.E.
(1) standard normal data						
10	0.0005	0.3139	−0.0003	0.3311	0.0005	0.3711
20	0.0008	0.2196	−0.0012	0.2365	0.0009	0.2671
30	0.0008	0.1804	−0.0005	0.1926	−0.0017	0.2216
40	−0.0017	0.1550	−0.0005	0.1646	−0.0002	0.1910
50	−0.0012	0.1401	−0.0007	0.1510	0.0008	0.1716
60	−0.0001	0.1275	0.0000	0.1359	0.0016	0.1572
70	0.0002	0.1199	0.0008	0.1275	0.0005	0.1462
80	−0.0006	0.1111	−0.0014	0.1195	−0.0008	0.1393
90	−0.0009	0.1050	−0.0016	0.1126	−0.0008	0.1297
100	0.0003	0.1020	0.0007	0.1082	0.0006	0.1247
(2) double exponential data						
10	0.0089	0.4505	0.0027	0.4631	0.0025	0.3786
20	−0.0015	0.3250	−0.0071	0.3240	−0.0030	0.2578
30	−0.0043	0.2583	−0.0067	0.2462	−0.0042	0.2029
40	−0.0042	0.2229	−0.0055	0.2032	−0.0051	0.1683
50	−0.0027	0.2019	−0.0012	0.1802	−0.0024	0.1549
60	−0.0004	0.1853	−0.0031	0.1667	−0.0002	0.1419
70	0.0019	0.1708	0.0009	0.1499	0.0008	0.1298
80	0.0007	0.1583	0.0008	0.1380	−0.0026	0.1201
90	−0.0009	0.1503	−0.0013	0.1316	−0.0023	0.1147
100	−0.0022	0.1411	−0.0001	0.1226	−0.0018	0.1063

using the routines given by Press *et al.* (1992). The sample sizes considered range from 10 to 100.

Table 2 contains the simulated bias and standard errors of $\bar{X}$, $\hat{\theta}_{lsel}$ and $\hat{m}$ for data generated from the normal and the Laplace distributions. Each entry in Table 2 is based on 2000 simulations.

We observe that the bias is negligible when compared with the size of the standard error for each of the three estimators. The standard error of $\bar{X}$ is less than that of $\hat{m}$ for the standard normal data, and vice versa for the Laplace data. These just reflect the well known theory about sample mean and median estimators. It is interesting to see that the standard error of the maximum least-squares empirical likelihood estimator $\hat{\theta}_{lsel}$ is between those of the $\bar{X}$ and $\hat{m}$ for both cases. The standard errors were closer to those of $\bar{X}$ for the standard normal data, and were closer to those of the sample median for the Laplace data were generated. These are not surprising, having be anticipated by the asymptotic study in Subsection 4.1.

6. Discussions

We introduced the notion of combined empirical likelihood for situations when the number of structure constraints exceeds the number of parameters, hoping that the resulting empirical likelihood will have "the best of all worlds" in adapting naturally to whichever of the structure constraints that puts the most pressure on the data. To facilitate the computation of the combined empirical likelihood, a least-squares version of the empirical likelihood is introduced. We demonstrated in Section 2 that this empirical likelihood has a high level of accuracy in approximation to full empirical likelihood, in quite general settings. The computation of least squares empirical likelihood is straightforward.

To illustrate the power of least squares empirical likelihood, a comprehensive study was carried out in the one sample location problem by combining the mean and median constraints. The least-squares empirical likelihood curves follow those of the full empirical likelihood closely. The least-squares empirical likelihood estimate is asymptotically normally distributed, has an asymptotic variance not larger than those of sample mean and sample median, and at the same time is more robust than the sample mean as demonstrated in the analysis of Short's data.

The least-squares empirical likelihood can be extended to other situations. For instance in linear regression the constraints on the mean residuals can be combined with those on the median residuals to make regression inference more robust. Also, additional constraints can be added to protect any particular parameter against outliers in the residual errors or wrong entries in the design matrix.

Acknowledgements

The authors would like to thank two referees for comments which have improved the presentation of the paper.

Appendix: Expansions and proofs

A.1 *Expansions for EL and LSEL*

We present Taylor expansions for $\ell(\theta)$ and $lsl(\theta)$ in this appendix, which are used to compare the two likelihood functions in Section 2. To be specific, we consider $\theta = \theta_0 + n^{-1/2}\Sigma^{1/2}\tau$, where τ is a k-dimensional constant vector and $\Sigma = n^{-1}\sum \mathrm{Cov}\{Z_i(\theta_0)\}$, the average covariance matrix of $Z_i(\theta_0)$, and is assumed to be non-singular. We define $W_i = \Sigma^{-1/2} Z_i(\theta_0)$ with the j-th component W_{ij},

$$\bar{\alpha}^{j_1\cdots j_k} = n^{-1}\sum E(W_{ij_1}\cdots W_{ij_k}) \quad \text{and}$$
$$A^{j_1\cdots j_k} = n^{-1}\sum (W_{ij_1}\cdots W_{ij_k} - \bar{\alpha}^{j_1\cdots j_k}).$$

From a Taylor expansion given in (3.1) of Chen (1994),

$$\begin{aligned}
\text{(A.1)}\quad n^{-1}\ell(\theta) &= (A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^j \\
&\quad - \{A^{jk} + n^{-1/2}\tau^j(A+n^{-1/2}\tau)^k[2] + n^{-1}\tau^j\tau^k\} \\
&\quad \cdot (A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^k \\
&\quad + \frac{2}{3}(\bar{\alpha}^{jkl} + A^{jkl} + n^{-1/2}\tau^j\delta^{kl}[3] - 2\bar{\alpha}^{jkm}A^{lm}) \\
&\quad \cdot (A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^k(A+n^{-1/2}\tau)^l \\
&\quad + \left(\bar{\alpha}^{jkn}\bar{\alpha}^{lmn} - \frac{1}{2}\bar{\alpha}^{jklm}\right)(A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^k \\
&\quad \cdot (A+n^{-1/2}\tau)^l(A+n^{-1/2}\tau)^m \\
&\quad + A^{jl}A^{kl}(A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^k + O_p(n^{-5/2}),
\end{aligned}$$

where δ^{jk} is the Kronecker delta. We use here a convention that terms with repeated indices are to be summed over, and a rule that $\tau^j A^k[2] = \tau^j A^k + \tau^k A^j$ and the same rule applies for $\tau^j\delta^{kl}[3]$.

Using the same method in Chen (1994), it may be shown that $lsl(\theta)$ admits the following Taylor expansion:

$$\begin{aligned}
\text{(A.2)}\qquad n^{-1}lsl(\theta) &= (A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^j \\
&\quad - \{A^{jk} + n^{-1/2}\tau^j(A+n^{-1/2}\tau)^k[2] + n^{-1}\tau^j\tau^k\} \\
&\quad \cdot (A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^k \\
&\quad + (A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^j \\
&\quad \cdot (A+n^{-1/2}\tau)^l(A+n^{-1/2}\tau)^l \\
&\quad + \{A^{jl} + n^{-1/2}\tau^j(A+n^{-1/2}\tau)^l[2] + n^{-1}\tau^j\tau^l\} \\
&\quad \cdot (A+n^{-1/2}\tau)^j(A+n^{-1/2}\tau)^k \\
&\quad \times \{A^{kl} + n^{-1/2}\tau^k(A+n^{-1/2}\tau)^l[2] + n^{-1}\tau^l\tau^k\} \\
&\quad + O_p(n^{-5/2}).
\end{aligned}$$

A.2 *Proof of Theorem* 1

In the notation of Section 2, we have $k = 2$, $p = 1$, $Z_i(\theta) = (X_i - \theta, \operatorname{sgn}(X_i - \theta))^T$ and

$$H = n\begin{pmatrix} \hat{\sigma}_x^2 & \rho(\theta) \\ \rho(\theta) & \hat{\sigma}_{\text{sgn}}^2(\theta) \end{pmatrix},$$

where

$$\hat{\sigma}_x^2 = n^{-1}\sum(X_i - \bar{X})^2, \qquad \rho(\theta) = n^{-1}\sum(X_i - \bar{X})\operatorname{sgn}(X_i - \theta) \qquad \text{and}$$
$$\hat{\sigma}_{\text{sgn}}^2(\theta) = n^{-1}\sum \operatorname{sgn}(X_i - \theta)^2 - \left\{n^{-1}\sum \operatorname{sgn}(X_i - \theta)\right\}^2.$$

According to (2.6),

$$n^{-1}lsl(\theta) = \left(\bar{X} - \theta, n^{-1}\sum \operatorname{sgn}(X_i - \theta)\right) H^{-1} \cdot \left(\bar{X} - \theta, n^{-1}\sum \operatorname{sgn}(X_i - \theta)\right)^T. \tag{A.3}$$

As both $\rho(\theta)$ and $\hat{\sigma}^2_{\rm sgn}(\theta)$ are step functions, we write them as $\rho(j)$ and $\hat{\sigma}^2_{\rm sgn}(j)$ for $\theta \in (X_j, X_{j+1})$ respectively. It turns out that

$$\rho(j) = (1 - j/n)\bar{X}^{n-j} - j\bar{X}_j/n + (1 - 2j/n)\bar{X} \quad \text{and}$$
$$\hat{\sigma}^2_{\rm sgn}(j) = 1 - (1 - 2j/n)^2.$$

Put $c_j = \{\hat{\sigma}^2_x \hat{\sigma}^2_{\rm sgn}(j) - \rho(j)^2\}^{-1}$. Then for $\theta \in (X_j, X_{j+1})$,

$$lsl(\theta) = c_j\{(\bar{X} - \theta)^2 \hat{\sigma}^2_{\rm sgn}(j) + (1 - 2j/n)^2 \hat{\sigma}^2_x - 2(\bar{X} - \theta)(1 - 2j/n)\rho(j)\}.$$

A stationary point in (X_j, X_{j+1}) at which $lsl(\theta)$ has zero derivative value is

$$\hat{\theta}_j = \bar{X} - (1 - 2j/n)\rho(j)/\hat{\sigma}^2_{\rm sgn}(j) = \frac{1}{2}(\bar{X}_j + \bar{X}^{n-j}),$$

provided $\hat{\theta}_j$ is inside (X_j, X_{j+1}).

So to finish the proof of the theorem we only need to show that any stationary point of $\ell(\theta)$ is equal to $\frac{1}{2}(\bar{X}_j + \bar{X}^{n-j})$ for some j. The optimal weights p_i determined by empirical likelihood at θ are

$$p_i = n^{-1}\{1 + \lambda_1(X_i - \theta) + \lambda_2 \operatorname{sgn}(X_i - \theta)\}^{-1}, \tag{A.4}$$

where λ_1 and λ_2 are the Lagrange multipliers. As $\ell(\theta) = -2\sum \log(np_i)$,

$$\frac{\partial \ell(\theta)}{\partial \theta} = -2\sum p_i^{-1}\frac{\partial p_i}{\partial \theta} = -2\lambda_1 \sum p_i + 2\lambda_2 \sum p_i \frac{\partial \operatorname{sgn}(X_i - \theta)}{\partial \theta}.$$

So, if $\theta \in (X_j, X_{j+1})$ and $\frac{\partial \ell(\theta)}{\partial \theta} = 0$, λ_1 must be zero and

$$p_i = 1/(n + \lambda_2) \quad \text{if } i \le j \qquad \text{and} \qquad p_i = 1/(n - \lambda_2) \quad \text{if } i > j.$$

Substituting the above p_i back to the median constraint (1.3), we have $\lambda_2 = 2j - n$. Therefore,

$$p_i = 1/(2j) \quad \text{if } i \le j \qquad \text{and} \qquad p_i = 1/(2n - 2j) \quad \text{if } i > j.$$

The mean constraint gives $\theta = \frac{1}{2}(\bar{X}_j + \bar{X}^{n-j})$.

A.3 *Proof of Theorem* 2

We start with proofs for (i) and (ii). As $\bar{X}^{n-j} = -\frac{j}{n-j}\bar{X}_j + \frac{n}{n-j}\bar{X}$, then

$$\bar{X}_j + \bar{X}^{n-j} = \frac{n-2j}{n-j}\bar{X}_j + \frac{n}{n-j}\bar{X} = \frac{2j-n}{j}\bar{X}^{n-j} + \frac{n}{j}\bar{X}. \tag{A.5}$$

For a continuous distribution and any $1 < j < n-1$

$$P(\bar{X}_j < X_j) = 1 \quad \text{and} \quad P(\bar{X}^{n-j} > X_{j+1}) = 1.$$

Also for $j = [np]+1$, the probabilities of $\bar{X} \geq X_{j+1}$ if $p < 1/2$ and $\bar{X} \leq X_j$ if $p > 1/2$ approach to 1 as $n \to \infty$. Thus, if $p < 1/2$, using the last expression of (A.5),

$$P\left(\bar{X}_j + \bar{X}^{n-j} \geq \frac{2j-n}{j}X_{j+1} + \frac{n}{j}X_{j+1}\right) \to 1 \quad \text{as} \quad n \to \infty.$$

Because $j = [np]+1$, $\frac{2j-n}{j} \to \frac{2p-1}{p}$ and $\frac{n}{j} \to \frac{1}{p}$ as $n \to \infty$. So

$$P(\bar{X}_j + \bar{X}^{n-j} \geq 2X_{j+1}) \to 1 \quad \text{as} \quad n \to \infty.$$

So (i) of Theorem 2 is obtained.

Similarly, if $p > 1/2$ using the middle expression of (A.5), we have

$$P(\bar{X}_j + \bar{X}^{n-j} \leq 2X_j) \to 1 \quad \text{as} \quad n \to \infty,$$

which leads us to (ii) of Theorem 2.

To prove (iii) of Theorem 2, we define for $1 \leq j \leq n$

$$\Sigma_j = \begin{pmatrix} \hat{\sigma}_x^2 & \rho(j) \\ \rho(j) & 1-(1-2j/n)^2 \end{pmatrix} \quad \text{and}$$
$$\Sigma_{j0} = \begin{pmatrix} \hat{\sigma}_x^2 & \rho(j)+n^{-1}(X_j-\bar{X}) \\ \rho(j)+n^{-1}(X_j-\bar{X}) & 1-(1-(2j-1)/n)^2-1/n \end{pmatrix}.$$

Then from (A.3),

$$lsl(X_j^-) = n(\bar{X}-X_j, 1-2(j+1)/n)\Sigma_{j-1}^{-1}(\bar{X}-X_j, 1-2(j+1)/n)^T,$$
$$lsl(X_j^0) = n(\bar{X}-X_j, 1-(2j-1)/n)\Sigma_{j0}^{-1}(\bar{X}-X_j, 1-(2j-1)/n)^T \quad \text{and}$$
$$lsl(X_j^+) = n(\bar{X}-X_j, 1-2j/n)\Sigma_j^{-1}(\bar{X}-X_j, 1-2j/n)^T,$$

where $lsl(X_j^-)$ is defined for $j = 2, \ldots, n$, $lsl(X_j^0)$ for $j = 1, \ldots, n$ and $lsl(X_j^+)$ for $j = 1, \ldots, n-1$.

We define furthermore for $j = 1, \ldots, n$

$$\Gamma_j = \begin{pmatrix} 0 & X_j-\bar{X} \\ X_j-\bar{X} & (4j+2)/n-2 \end{pmatrix} \quad \text{and}$$
$$\Gamma_{j0} = \begin{pmatrix} 0 & X_j-\bar{X} \\ X_j-\bar{X} & -1-(2n-4j+1)/n \end{pmatrix}.$$

As $\rho(j-1) = \rho(j) + 2n^{-1}(X_j - \bar{X})$, we have $\Sigma_{j-1} = \Sigma_j + 2n^{-1}\Gamma_j$ and $\Sigma_{j0} = \Sigma_j + n^{-1}\Gamma_{j0}$. Thus

$$\Sigma_{j-1}^{-1} = \Sigma_j^{-1} + 2n^{-1}\Sigma_j^{-1}\Gamma_j\Sigma_j^{-1} + O_p(n^{-2}) \quad \text{and}$$
$$\Sigma_{j0}^{-1} = \Sigma_j^{-1} + n^{-1}\Sigma_j^{-1}\Gamma_{j0}\Sigma_j^{-1} + O_p(n^{-2}).$$

Substituting the above expressions into $lsl(X_j^-)$, we have

$$\begin{aligned} lsl(X_j^-) = lsl(X_j^+) &+ 2(\bar{X} - X_j, 1 - 2j/n)\Sigma_j^{-1}\Gamma_j\Sigma_j^{-1}(\bar{X} - X_j, 1 - 2j/n)^T \\ &+ (0, -4)\Sigma_j^{-1}(\bar{X} - X_j, 1 - (2j+2)/n)^T + n^{-1}(0, -2)\Sigma_j^{-1}(0, -2)^T \\ &+ O_p(n^{-1}). \end{aligned}$$

Let $j = [np] + 1$ for some $p \in (0, 1)$, ξ_p be the p-th population quantile and

$$\theta^p = \int_{\xi_p}^{\infty} x f(x) dx \Big/ \int_{\xi_p}^{\infty} f(x) dx.$$

It may be shown that as $n \to \infty$, $1 - 2j/n \to 1 - 2p$,

$$\bar{X} - X_j \to \theta - \xi_p \quad \text{in probability},$$

and Σ_j, Σ_{j-1} and Σ_{j0} all converge in probability to

$$\begin{pmatrix} \mathrm{Var}(X) & 2(1-p)(\theta^p - \theta) \\ 2(1-p)(\theta^p - \theta) & 1 - (1-2p)^2 \end{pmatrix}.$$

Thus,

$$lsl(X_j^-)/lsl(X_j^+) = 1 + O_p(n^{-1}).$$

Similarly, we can show that

$$lsl(X_j^0)/lsl(X_j^+) = 1 + O_p(n^{-1}).$$

References

Chen, S. X. (1993). On the coverage accuracy of empirical likelihood confidence regions for linear regression model, *Ann. Inst. Statist. Math.*, **45**, 621–637.

Chen, S. X. (1994). Comparing empirical likelihood and bootstrap hypothesis tests, *J. Multivariate Anal.*, **51**, 277–293.

Chen, S. X. and Hall, P. (1993). On the calculation of standard error for quotation in confidence statements, *Statist. Probab. Lett.*, **19**, 147–151.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood, *Internat. Statist. Rev.*, **58**, 109–127.

Hall, P. and Owen, A. (1993). Empirical likelihood confidence bands in density estimation, *Journal of Computational and Graphical Statistics*, **2**, 273–289.

Hansen, L. P. (1982). Large sample properties of generalized method of moment estimators, *Econometrica*, **50**, 1029–1054.

Jing, B. Y. (1995). Some resampling procedures under symmetry, *Austral. J. Statist.*, **37**, 337–344.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, **75**, 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions, *Ann. Statist.*, **18**, 90–120.

Owen, A (1991). Empirical likelihood for linear model, *Ann. Statist.*, **19**, 1725–1747.

Press, W. H., Flannery, B. F., Teukolsky, S. A. and Vetterling, W. T. (1992). *Numerical Recipes in C*, Cambridge University Press, Cambridge.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating functions, *Ann. Statist.*, **22**, 300–325.

Stigler, S. M. (1977). Do robust estimators work with real data? (with discussion), *Ann. Statist.*, **5**, 1055–1098.

Wood, A. T. A., Do, K.-A. and Broom, B. M. (1996). Sequential linearization of empirical likelihood constraints with application to U-statistics, *Journal of Computational and Graphical Statistics*, **5**, 365–385.