International Statistical Review (1990), **58**, 2, pp. 109–127. Printed in Great Britain

Methodology and Algorithms of Empirical Likelihood

Peter Hall and Barbara La Scala

Statistics Research Section, School of Mathematical Sciences, Australian National University, G.P.O. Box 4, Canberra, A.C.T. 2601, Australia

Summary

We describe the main features of empirical likelihood and discuss recent developments, including Bartlett correction and location adjustment. Algorithms are provided for implementing empirical likelihood in important cases, for example to means, variances and correlation coefficients. It is shown that empirical likelihood is a serious competitor with contemporary methods such as the bootstrap. Indeed empirical likelihood has several advantages, in that it does not impose prior constraints on region shape, does not require construction of a pivotal statistic, and admits a Bartlett correction which allows very low coverage error. Empirical likelihood deserves a prominent place in the modern statistician's armoury of computer-intensive tools.

Key words: Bartlett correction; Bootstrap; Confidence interval; Confidence region; Coverage error; Empirical likelihood; Likelihood; Location adjustment.

1 Introduction

The empirical likelihood method for constructing confidence regions was introduced by Owen (1988, 1990). It amounts to computing the profile likelihood of a general multinomial distribution which has its atoms at data points. As Owen points out, a version of this technique dates back at least to Thomas & Grunkemeier (1975), in the context of estimating survival probabilities. Owen's remarkable contribution was to show that the idea has very general application. It may be applied to ordinary random sampling models, regression models, autoregressive models etc.

Our purpose in this paper is to describe the main features of the idea and to discuss the developments which have taken place since Owen's early contributions. We argue that empirical likelihood is a serious competitor with contemporary methods such as the bootstrap, and deserves a prominent place in the modern statistician's armoury of computer-intensive tools.

Owen (1990) discussed empirical likelihood as an alternative to likelihood-type bootstrap methods, such as that proposed by Hall (1987). In fact it competes very convincingly with all bootstrap methods, particularly with the more accurate types which are finding increased favour today, such as percentile-t and accelerated bias correction. The advantages of empirical likelihood over classical methods such as normal approximation are rather obvious; for example, empirical likelihood regions are not shaped in a predetermined way which implies a degree of nonexistent symmetry in the sampling distribution. We list below the main advantages of empirical likelihood over the bootstrap.

(i) The shape of empirical likelihood confidence regions automatically reflects emphasis in the observed data set. The regions tend to be concentrated in places where the density of the parameter estimator is greatest. This feature is discussed in § 4.4. Only relatively

complex bootstrap methods can give regions which emulate this property of 'letting the data determine the shape of the region'. Indeed, construction of any multivariate bootstrap region requires a decision on how the region should be shaped, and it can be rather difficult to make that decision in a manner which depends only on objective, data-driven criteria.

(ii) Empirical likelihood regions are Bartlett correctable. That is, a simple correction for the mean of the empirical loglikelihood ratio reduces coverage error from order n^{-1} to order n^{-2}, where n denotes sample size. See § 3.2. The coverage accuracy of bootstrap confidence regions can usually only be improved by bootstrap iteration, e.g. Hall (1986) and Beran (1987), and that is computationally expensive.

(iii) Empirical likelihood regions do not require estimation of scale or skewness. The percentile-t bootstrap method can fail unless a stable estimator of scale is available, and the accelerated bias correction method needs an estimator of skewness. Empirical likelihood does not even require construction of a pivotal statistic. Therefore it is well suited to circumstances where pivotalness is difficult to achieve, for example to interval or region estimation of location for spherical data, and also to interval estimation for the correlation coefficient, whose variance is notoriously difficult to estimate.

(iv) Empirical likelihood regions are range preserving and transformation respecting. For example, the empirical likelihood region for a correlation coefficient always lies between -1 and 1, and the empirical likelihood region for the function $g(\theta)$ of the parameter θ equals g of the empirical likelihood region for θ. Neither of these properties is enjoyed by percentile-t bootstrap confidence regions. The accelerated bias correction method is translation invariant but does not have a multivariate version.

Construction of bootstrap confidence regions usually requires Monte Carlo simulation, whereas empirical likelihood employs numerical methods for constrained maximization. The latter technique needs a somewhat greater degree of numerical sophistication than the former, although as we shall show in § 4, in relatively simple cases a multivariate Newton algorithm may be used to find constrained maxima. The power of the bootstrap resides in the fact that it can be applied to very complex problems, and this feature is not available for empirical likelihood.

In this paper our focus of attention is confidence regions for parameters which are expressible as smooth functions of means. This approach is sufficiently general to stress the wide applicability of empirical likelihood, yet specific enough to enable us to describe theorems under concise regularity conditions, and to discuss corrections (such as Barlett's correction and location adjustment) in some detail. For the most part we assume that the r-variate parameter of interest is $\theta = g(EX)$, where EX denotes the mean of the population from which the s-variate data set $X_1, \ldots, X_n$ was drawn, and where $g : \mathbb{R}^s \to \mathbb{R}^r$ is a smooth function. This does not imply that the data must always be observed in the form $X_1, \ldots, X_n$. For example, when θ is a univariate population variance the observed data set would usually be a scalar sequence $Y_1, \ldots, Y_n$, and we would take $X_i = (Y_i, Y_i^2)^{\mathrm{T}}$, so that $r = 1$, $s = 2$ and

$$\theta = g(EX) = EX^{(2)} - (EX^{(1)})^2 = EY^2 - (EY)^2,$$

bracketed superscripts denoting vector elements.

The method of empirical likelihood may be extended to other contexts, for example to quantiles, U-statistics and related quantities, but as yet we have a unified theory only in the case of a smooth function of a mean.

Section 2 defines empirical likelihood confidence regions and discusses main properties such as Wilks's theorem and region shape. Second-order properties, including coverage error and Bartlett correction, are described in § 3. Section 4 presents algorithms for

computation of empirical likelihood regions in several cases of practical interest, for example univariate and bivariate means, univariate and bivariate variances, and the correlation coefficient. Section 5 illustrates our main points by applying empirical likelihood to a real data set.

The empirical likelihood idea has wider applicability than might at first appear. It may be used to construct confidence intervals and confidence bands in nonparametric density estimation and nonparametric regression (Hall & Owen, 1989), and also to construct confidence intervals in parametric regression (under the so-called 'correlation model'). However, it does not extend easily to problems involving dependent data, or to non-identically distributed data unless the lack of common distribution can be rectified by a simple transformation (e.g. rescaling). It shares these shortcomings with the bootstrap, but as noted earlier, empirical likelihood is not so easily applied as the bootstrap to very complex problems.

2 Main Features of Empirical Likelihood

2.1 Summary

Section 2.2 defines empirical likelihood and the empirical likelihood ratio. Empirical likelihood shares several important properties with ordinary parametric likelihood. One of these, Wilks's theorem, is introduced in § 2.3. Wilks's theorem declares that the logarithm of the empirical likelihood ratio has an asymptotic chi squared distribution. It forms the basis of a general method for constructing empirical likelihood confidence regions. Theorem 2.1 in § 2.3 is a version of Wilks's theorem in the context of empirical likelihood for parameters which are expressible as smooth functions of vector means.

The case where the parameter of interest is the mean itself, possibly a vector, is treated in § 2.4. There we give a simple formula for empirical loglikelihood in the case of means. It is shown that empirical likelihood regions for vector means are necessarily convex, a property also noted by Owen (1990). This implies that empirical likelihood regions for smooth functions of means must be connected and without voids. These results are presented together as Theorem 2.2. Section 2.5 gives a proof of Theorem 2.1.

2.2 Definition of Empirical Likelihood

Let θ denote some characteristic of a population, such as its mean, for which we wish to construct a confidence region. There is no requirement for θ to be a scalar. Indeed, the case where θ is a vector of length 2 is particularly interesting, giving rise to confidence regions in the plane. Let $X_1, \ldots, X_n$ denote a random sample (possibly a sample of vectors) drawn from the population, let $p = (p_1, \ldots, p_n)$ be a vector having each $p_i \geqslant 0$ and $\sum p_i = 1$, and write $\theta(p)$ for the value assumed by the parameter θ when the population is discrete with atom p_i at point X_i, $1 \leqslant i \leqslant n$. For example if θ denotes the population mean then $\theta(p) = \sum p_i X_i$. In this case the vectors θ and X_i would be of the same length, although that would not be so in many other circumstances. Generally, assume that θ is of length r and X_i of length s.

The empirical likelihood for θ, evaluated at $\theta = \theta_1$, is defined to be

$$L(\theta_1) = \max_{p:\theta(p)=\theta_1, \sum p_i=1} \prod_{i=1}^{n} p_i.$$

Thus, empirical likelihood is a multinomial profile likelihood. The number of parameters in the multinomial is only one less than the sample size, and so it is perhaps surprising

that we can ever use $L(\theta)$ to good effect! Empirical likelihood has some, but by no means all, of the properties of parametric likelihood. It admits an integrated version, but a conditional version is in general more difficult to define.

Subject to $\sum p_i = 1$, the product $\prod p_i$ is maximized by taking $p_i = n^{-1}$ for $1 \leq i \leq n$. (The proof of this fact is a simple exercise in Lagrange multipliers.) For this choice of p we have $\theta(p) = \hat{\theta}$, the so-called bootstrap estimator. Therefore the 'maximum empirical likelihood estimator' of θ is none other than the bootstrap estimator. For example when θ is the population mean, $\theta(n^{-1}, \ldots, n^{-1}) = n^{-1} \sum X_i$, the sample mean. This point is not of particular significance in developing the properties of empirical likelihood, although it does help to develop a parallel between classical parametric likelihood and empirical likelihood. It implies that

$$L(\hat{\theta}) = n^{-n},$$

so that the empirical likelihood ratio is

$$L(\theta_1)/L(\hat{\theta}) = \max_{p:\theta(p)=\theta_1, \sum p_i = 1} \prod_{i=1}^{n} (np_i). \tag{2.1}$$

2.3 *Wilks's Theorem*

To appreciate the importance and implications of Wilks's theorem we briefly review the case of parametric likelihood. There $L(\theta)$ denotes the classical likelihood of a sample, $\hat{\theta}$ is the usual maximum likelihood estimator, and

$$l(\theta) = -2 \log \{L(\theta)/L(\hat{\theta})\} \tag{2.2}$$

is the loglikelihood ratio. Let θ_0 denote the true value of θ, assume that there are no nuisance parameters, and let t be the rank of the asymptotic variance matrix of $n^{\frac{1}{2}}\hat{\theta}$. Wilks's theorem states that under appropriate regularity conditions, $l(\theta_0)$ has an asymptotic χ^2_t distribution. This result is the key to constructing parametric likelihood-based confidence regions, as follows. Find from tables the value of c such that

$$P(\chi^2_t \leq c) = 1 - \alpha, \tag{2.3}$$

where $1 - \alpha$ is the desired nominal coverage of the region. Then

$$\mathcal{R}_c = \{\theta : l(\theta) \leq c\} \tag{2.4}$$

is an appropriate region. In view of Wilks's theorem, the asymptotic coverage of $\mathcal{R}_c$ equals $1 - \alpha$:

$$P(\theta_0 \in \mathcal{R}_c) = P\{l(\theta_0) \leq c\} \to 1 - \alpha \tag{2.5}$$

as $n \to \infty$. See for example Wilks (1938) and Chernoff (1954).

A major property of empirical likelihood is that it admits a nonparametric version of Wilks's theorem. Therefore the above prescription for a confidence region may be employed in the case of empirical likelihood. More precisely, define the likelihood ratio and l and c as at (2.1), (2.2) and (2.3), respectively. Construct the confidence region according to (2.4). Then (2.5) holds by the empirical likelihood version of Wilks's theorem; see Theorem 2.1 below.

We shall generalize work of Owen (1990) by establishing Wilks's theorem in the empirical likelihood context where θ is a function of the population mean, say $\theta = g(\mu)$. In this case the bootstrap estimator $\hat{\theta} = g(\bar{X})$ is the same function of the sample mean $\bar{X} = n^{-1} \sum X_i$. Remember that θ is of length r, and X_i and $\mu = (\mu^{(1)}, \ldots, \mu^{(s)})^{\mathrm{T}}$ are of

length s. Let $\mu_0 = E(X)$ denote the true value of the population mean, and let $\theta_0 = g(\mu_0)$ be the true value of θ. If $g = (g^{(1)}, \ldots, g^{(r)})^{\mathrm{T}}$ has a continuous derivative in a neighbourhood of μ_0 then the asymptotic variance matrix of $n^{\frac{1}{2}}\hat{\theta}$ is $V = v_0 \Sigma v_0^{\mathrm{T}}$, where $v_0 = (v_0^{(ij)})$ denotes the $r \times s$ matrix defined by

$$v_0^{(ij)} = \partial g^{(i)}(\mu)/\partial \mu^{(j)}|_{\mu=\mu_0},$$

and $\Sigma = E\{(X - \mu_0)(X - \mu_0)^{\mathrm{T}}\}$ is the population variance matrix.

THEOREM 2.1. *Assume X has finite variance and g has a continuous derivative in a neighbourhood of μ_0. Let $t \leqslant \min(r, s)$ denote the rank of V, and let l be the empirical loglikelihood function. Then $l(\theta_0)$ has an asymptotic χ_t^2 distribution.*

2.4 *The Case $\theta = \mu$*

The case where $\theta = \mu$, the population mean, is of particular interest on several counts. Firstly, empirical likelihood assumes a simple form in that setting. Secondly, empirical likelihood regions for a mean (univariate or multivariate) are always convex. And thirdly, many statistics of practical interest are smooth functions of means, and there it follows from the result in the previous sentence that empirical likelihood regions are connected and without voids. We shall explain these results below.

A little manipulation using Lagrange multipliers shows that the p_i's which maximise $\prod p_i$ subject to $\sum p_i X_i = \mu$ and $\sum p_i = 1$ are given by

$$p_i(\mu) = n^{-1}\{1 + \lambda^{\mathrm{T}}(X_i - \mu)\}^{-1} \quad (1 \leqslant i \leqslant n), \tag{2.6}$$

where the s-vector $\lambda = \lambda(\mu)$ is determined by

$$\sum_{i=1}^{n} \{1 + \lambda^{\mathrm{T}}(X_i - \mu)\}^{-1}(X_i - \mu) = 0. \tag{2.7}$$

Therefore the empirical loglikelihood ratio for the mean is

$$l(\mu) = -2\sum_{i=1}^{n} \log\{np_i(\mu)\} = 2\sum_{i=1}^{n} \log\{1 + \lambda^{\mathrm{T}}(X_i - \mu)\}. \tag{2.8}$$

To see why empirical likelihood regions for a mean are convex, observe that multinomial likelihoods are concave functions of distributions. That is, if $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ are probability distributions satisfying

$$\prod_{i=1}^{n} p_i \geqslant C, \quad \prod_{i=1}^{n} q_i \geqslant C$$

for some $C > 0$, then

$$\prod_{i=1}^{n} \{\beta p_i + (1 - \beta)q_i\} \geqslant C \quad \text{for all } 0 \leqslant \beta \leqslant 1.$$

(Verification of this result is an elementary exercise in calculus). Let $c > 0$ be the index of the empirical likelihood region $\mathcal{R}_c$ defined at (2.4), and put $C = n^{-n}e^{-c/2}$. Then

$$\mathcal{R}_c = \{\mu : l(\mu) \leqslant c\} = \{\mu : L(\mu) \geqslant C\} = \Big\{\mu : \max_{p : \sum p_i X_i = \mu, \sum p_i = 1} \prod p_i \geqslant C\Big\}.$$

If $\mu, \nu \in \mathcal{R}_c$ then there must exist probability distributions p, q such that

$$\sum p_i X_i = \mu, \quad \sum q_i X_i = \nu, \quad \prod p_i \geq C, \quad \prod q_i \geq C.$$

By concavity of multinomial likelihoods, the distribution $r = \beta p + (1-\beta)q$ satisfies $\prod r_i \geq C$, and trivially $\sum r_i X_i = \beta\mu + (1-\beta)\nu$. Therefore $\beta\mu + (1-\beta)\nu \in \mathcal{R}_c$. That is, μ, $\nu \in \mathcal{R}_c$ implies $\beta\mu + (1-\beta)\nu \in \mathcal{R}_c$ for all $0 \leq \beta \leq 1$, and so $\mathcal{R}_c$ is convex.

This result about convexity does not extend to empirical likelihood regions for parameters which are smooth functions of means. However, it does imply that such regions are connected and without voids. To appreciate why, observe from definition (2.4) of an empirical likelihood region that if the parameters θ, ω are related by $\theta = g(\omega)$ for some function g, then the respective regions $\mathcal{R}_{c,\theta}$ and $\mathcal{R}_{c,\omega}$ are related by

$$\mathcal{R}_{c,\theta} = \{g(\omega) : \omega \in \mathcal{R}_{c,\omega}\}. \tag{2.9}$$

Therefore if θ is a continuous function of the population mean then its empirical likelihood region is that same function of a convex region, and so must be connected and without voids.

Suppose ω represents an s-variate mean. A little thought shows that if $s \geq 2$ then for a given sample size n and a given $c > 0$ we may choose a continuous, nondegenerate function $g : \mathbb{R}^s \to \mathbb{R}^s$ such that with positive probability, $\mathcal{R}_{c,\theta}$ (defined at (2.9)) is not convex. However this result is false in the case $s = 1$, as follows from the previous paragraph. Furthermore, in the case $s = 1$ the function $l(\mu)$ defined at (2.8) is convex, as may be checked by using (2.7) to show that $l''(\mu) > 0$.

We conclude this subsection by describing our main results in the form of a theorem, which expands a little on statements in Owen (1990).

THEOREM 2.2 *An empirical likelihood confidence region for a population mean is always convex, and an empirical likelihood confidence region for a continuous function of a population mean is always connected and without voids. However, except in the case where the function is scalar, the region may be non-convex with positive probability.*

2.5 *Proof of Theorem 2.1*

We may assume without loss of generality that X has nonsingular variance matrix Σ. For if var (X) has rank $s' < s$ then we may express $s - s'$ of the elements of X as a linear function in the remaining s' elements, reformulate the problem in terms of a function of the mean of those s' elements, solve the reformulated problem, and thus solve the original problem.

Indicate that we are working with empirical likelihood with respect to θ by appending a subscript θ to L and to l. In this notation we may write

$$L_\mu(\mu_1) = \max_{p : \sum p_i X_i = \mu_1, \sum p_i = 1} \prod p_i$$

for the empirical likelihood defined with respect to the mean μ. Since $\theta = g(\mu)$ then

$$l_\theta(\theta_1) = \min_{\mu_1 : g(\mu_1) = \theta_1} l_\mu(\mu_1). \tag{2.10}$$

We take the minimum rather than the maximum here because the definition of l_θ involves a minus sign.

The probabilities $p_i(\mu_1)$ which maximise $\prod p_i$ subject to $\sum p_i X_i = \mu_1$ and $\sum p_i = 1$ are given by (2.6), in which formula the s-vector λ is determined by (2.7). If μ_1 is distant $n^{-\frac{1}{2}}$

from $\mu_0 = E(X)$ then it may be shown from (2.7) that

$$\lambda = \lambda(\mu_1) = -\Sigma^{-1}(\bar{X} - \mu_1) + O_p(n^{-1}).$$

Therefore, assuming sufficiently many moments of X,

$$\begin{aligned} l_\mu(\mu_1) &= -2\log\{L_\mu(\mu_1)/L_\mu(\bar{X})\} = 2\sum_{i=1}^{n}\log\{1 + \lambda(\mu_1)^{\mathrm{T}}(X_i - \mu_1)\} \\ &= 2\lambda(\mu_1)^{\mathrm{T}}\sum_{i=1}^{n}(X_i - \mu_1) - \lambda(\mu_1)^{\mathrm{T}}\left\{\sum_{i=1}^{n}(X_i - \mu_1)(X_i - \mu_1)^{\mathrm{T}}\right\}\lambda(\mu_1) + O_p(n^{-\frac{1}{2}}) \\ &= n(\bar{X} - \mu_1)^{\mathrm{T}}\Sigma^{-1}(\bar{X} - \mu_1) + O_p(n^{-\frac{1}{2}}). \end{aligned}$$

This argument involves no more than Taylor expansion, and may be used to prove that for each $C > 0$, and under the conditions of the theorem,

$$\max_{\mu_1:\|\mu_1 - \mu_0\| \le Cn^{-\frac{1}{2}}} |l_\mu(\mu_1) - n(\bar{X} - \mu_1)^{\mathrm{T}}\Sigma^{-1}(\bar{X} - \mu_1)| = o_p(1).$$

At this point we see that by changing variable from X_i to $\Sigma^{-\frac{1}{2}}(X_i - \mu_0)$ we may suppose without loss of generality that $\mu_0 = 0$ and $\Sigma = I_{s\times s}$. This assumption is made below. It yields

$$\max_{\mu_1:\|\mu_1\| \le Cn^{-\frac{1}{2}}} |l_\mu(\mu_1) - n(\bar{X} - \mu_1)^{\mathrm{T}}(\bar{X} - \mu_1)| = o_p(1). \tag{2.11}$$

Since $\theta_0 = g(\mu_0) = g(0)$ then by Taylor expansion,

$$g(\mu) = \theta_0 + v_0\mu + o(\|\mu\|)$$

as $\mu \to 0$. Hence by (2.10) and (2.11),

$$\begin{aligned} l_\theta(\theta_0) &= \min_{\mu_1: g(\mu_1) = \theta_0} l_\mu(\mu_1) \\ &= \min_{\mu_1: v_0\mu_1 = 0} n(\bar{X} - \mu_1)^{\mathrm{T}}(\bar{X} - \mu_1) + o_p(1) \\ &= \min_{y: v_0 y = 0} (Y - y)^{\mathrm{T}}(Y - y) + o_p(1), \end{aligned}$$

where $Y = n^{\frac{1}{2}}\bar{X}$. Since Y is asymptotically normal $N(0, I_{s\times s})$ then the limiting distribution of $l_\theta(\theta_0)$ is that of

$$\min_{y: v_0 y = 0} (Z - y)^{\mathrm{T}}(Z - y),$$

where Z is normal $N(0, I_{s\times s})$. According to the Second Fundamental Theorem of Least Squares Theory (Rao, 1965, p. 155), this is the χ^2_t distribution. □

3 Coverage, Bartlett Correction and Location Adjustment

3.1 Summary

Section 2 addressed first order features of empirical likelihood, such as the asymptotic distribution of the empirical loglikelihood ratio. In the present section we describe second order properties, for example the rate of convergence to the asymptotic distribution. Section 3.2 shows that this convergence rate is $O(n^{-1})$, where n denotes sample size. In § 3.3 we demonstrate that this rate may be improved to $O(n^{-2})$ by incorporating an

empirical likelihood version of Bartlett correction, which amounts to correcting for the mean of the loglikelihood ratio. We note that $O(n^{-2})$ coverage accuracy may also be achieved by employing a bootstrap approximation to the distribution of the loglikelihood ratio, rather than a Bartlett correction of the chi squared approximation. Section 3.4 notes that empirical likelihood confidence regions are close to classical likelihood-based regions founded on a certain statistic, except for a centring error of order n^{-1}. We show how to correct for that error by adjusting location. Finally, § 3.5 discusses an example, the case of a univariate mean, which illustrates use of the main ideas in this section: Bartlett correction, location adjustment, and even Bartlett correction of a location adjusted region. Edgeworth expansions are used to describe the effect of all these corrections on the basic empirical likelihood region.

In practice, the necessary Bartlett correction and location adjustment must be estimated from the sample. We give formulae for data-based versions of these corrections. As in the case of Bartlett correction for parametric likelihood, our formulae are based on asymptotic theory, and it is not entirely clear how well they will perform for small samples.

3.2 Coverage Accuracy

We noted in § 2, result (2.5), that an empirical likelihood confidence region is asymptotically of the correct level. That is, if c is chosen according to (2.3) and if $\mathcal{R}_c$ is defined by (2.4), then $P(\theta_0 \in \mathcal{R}_c) \to 1 - \alpha$ as $n \to \infty$. The actual coverage error, $P(\theta_0 \in \mathcal{R}_c) - (1 - \alpha)$, is fortuitously small, being of size n^{-1} rather than $n^{-\frac{1}{2}}$. This is perhaps most easily explained in terms of the signed root empirical loglikelihood ratio statistic, and that is the approach we shall take below.

The empirical loglikelihood ratio $l(\theta_0)$ has an asymptotic χ^2_t distribution, see Theorem 2.1, and so is equal, to first order, to a sum of squares of t random variables which are asymptotically independent and normal $N(0, 1)$. Therefore it should come as no surprise that we may write

$$l(\theta_0) = W^{\mathrm{T}} W,$$

where W is a vector of length t and is asymptotically normal $N(0, I_{t\times t})$. DiCiccio, Hall & Romano (1988b) develop formulae for the distribution of W; see also DiCiccio & Romano (1989). In particular they show that it admits Edgeworth expansions of the common type for statistical functions. For example, the density f of W may be expressed by the formula

$$f(w) = \phi(w) + \sum_{j=1}^{k} n^{-j/2} \pi_j(w) \phi(w) + O(n^{-(k+1)/2}), \tag{3.1}$$

where $k \geq 1$, w is a t-vector, ϕ denotes the $N(0, I_{t\times t})$ density, and π_j is a polynomial of degree $3j$ which is even for even j and odd for odd j. This parity property is crucial to the argument which we shall give below, and it occurs commonly in statistical theory; see for example Petrov (1975, pp. 134–139). Writing $\mathcal{S}$ for the t-dimensional sphere of radius $c^{\frac{1}{2}}$ we find that

$$P(\theta_0 \in \mathcal{R}_c) = P\{l(\theta_0) \leq c\} = P(W^{\mathrm{T}} W \leq c) = P(W \in S) = \int_{\mathcal{S}} f(w)\, dw$$

$$= \int_{\mathcal{S}} \phi(w)\, dw + n^{-\frac{1}{2}} \int_{\mathcal{S}} \pi_1(w) \phi(w)\, dw + O(n^{-1}), \tag{3.2}$$

the last identity following from (3.1). The first integral on the right-hand side of (3.2) is identical to

$$P(\chi_t^2 \leq c) = 1 - \alpha,$$

by choice of c; see (2.3). The second integral vanishes, since $\mathcal{S}$ is a sphere centred at the origin and π_1 is an odd polynomial. Therefore

$$P(\theta_0 \in \mathcal{R}_c) = 1 - \alpha + O(n^{-1}),$$

as had to be shown.

3.3 Bartlett Correction

The source of coverage inaccuracy of the empirical likelihood confidence region defined at (2.4) is the chi squared approximation. As explained in § 3.2, this approximation is in error by order n^{-1}, and in consequence the coverage error is of order n^{-1}. Bartlett correction enhances the accuracy of the chi squared approximation.

The argument behind Bartlett correction is very simple, and runs as follows. Part of the error in a χ_t^2 approximation to the distribution of $l(\theta_0)$ may be explained by the fact that the means of the two distributions do not agree. That is, $E\{l(\theta_0)\} \neq t$. This discrepancy may be eliminated by rescaling $l(\theta_0)$ so that it has the correct mean. In other words, apply the chi squared approximation to $tl(\theta_0)/E\{l(\theta_0)\}$ rather than to $l(\theta_0)$. Now,

$$E\{l(\theta_0)\} = t\{1 + n^{-1}a + O(n^{-2})\},$$

where a is a constant. Therefore, up to an error of order n^{-2}, the mean correction is equivalent to applying the chi squared approximation to $l(\theta_0)/(1 + n^{-1}a)$. If a is unknown, replace a by a $\sqrt{n}$ consistent estimator $\hat{a}$. This is the Bartlett correction: select c from chi squared tables such that

$$P(\chi_t^2 \leq c) = 1 - \alpha, \tag{3.3}$$

and take the confidence region to be

$$\mathcal{R}_c' = \{\theta : l(\theta)/(1 + n^{-1}a) \leq c\} = \{\theta : l(\theta) \leq c(1 + n^{-1}a)\},$$

or $\mathcal{R}_c' = \{\theta : l(\theta) \leq c(1 + n^{-1}\hat{a})\}$ if we are estimating a.

The empirical likelihood version of Bartlett correction is entirely analogous to the ordinary, parametric likelihood Bartlett correction. See Barndorff–Nielsen & Cox (1984) for an account of the latter. In both the parametric and empirical likelihood cases, Bartlett correction reduces coverage error by an order of magnitude, from $O(n^{-1})$ to $O(n^{-2})$.

In essence, Bartlett correction works because the even, 6th degree polynomial π_2 in the Edgeworth expansion (3.1) is actually only of degree 2. That is, all terms of degree 4 and 6 in π_2 vanish. This is a rather deep property of empirical likelihood, and is verified by DiCiccio, Hall & Romano (1988). We shall content ourselves here with proving that if π_2 is in fact of degree 2 then Bartlett correction reduces coverage error to $O(n^{-2})$. More generally, we shall show that if the density f of a t-vector W admits the Edgeworth expansion

$$f(w) = \phi(w) + n^{-\frac{1}{2}}\pi_1(w)\phi(w) + n^{-1}\pi_2(w)\phi(w) + n^{-3/2}\pi_3(w)\phi(w) + O(n^{-2}), \tag{3.4}$$

where π_1, π_3 are odd polynomials and π_2 is even of degree 2; if $c > 0$ is given by (3.3); and if a is defined by

$$E(W^{\mathrm{T}}W) = t\{1 + n^{-1}a + O(n^{-2})\};$$

then

$$P\{W^{\mathrm{T}}W \leqslant c(1+n^{-1}a)\} = 1-\alpha+O(n^{-2}).$$

In the case of empirical likelihood we intend W to be interpreted as the signed root loglikehood ratio.

Integrate (3.4) over the sphere $\{w : \|w\|^2 \leqslant c(1+n^{-1}a)\}$, and note that the polynomials π_1 and π_3 are odd, obtaining

$$P\{W^{\mathrm{T}}W \leqslant c(1+n^{-1}a)\} = P\{\chi_t^2 \leqslant c(1+n^{-1}a)\} + n^{-1}\int_{\|w\|^2\leqslant c} \pi_2(w)\phi(w)\,dw + O(n^{-2}). \tag{3.5}$$

If g denotes the χ_t^2 density then

$$P\{\chi_t^2 \leqslant c(1+n^{-1}a)\} = P(\chi_t^2 \leqslant c) + n^{-1}acg(c) + O(n^{-2}). \tag{3.6}$$

Since π_2 is even and of degree 2, and since $\int \pi_2\phi = 0$ (because f is a density), then for constants a_j and a_{jk},

$$\pi_2(w) = \sum_{j=1}^{t} a_j(w^{(j)2}-1) + \sum\sum_{j\neq k} a_{jk}w^{(j)}w^{(k)},$$

where we have written $w = (w^{(1)}, \ldots, w^{(t)})^{\mathrm{T}}$. Therefore

$$\begin{aligned}\int_{\|w\|^2\leqslant c} \pi_2(w)\phi(w)\,dw &= \sum_{j=1}^{t} a_j \int_{\|w\|^2\leqslant c} (w^{(j)2}-1)\phi(w)\,dw \\ &= \Big(t^{-1}\sum_{j=1}^{t} a_j\Big)\int_0^c (x-t)g(x)\,dx \\ &= -2\Big(t^{-1}\sum_{j=1}^{t} a_j\Big)cg(c). \end{aligned} \tag{3.7}$$

Similarly,

$$ta = \int \|w\|^2\pi_2(w)\phi(w)\,dw = 2\sum_{j=1}^{t} a_j. \tag{3.8}$$

Combining (3.5)–(3.8) we deduce that

$$\begin{aligned}P\{W^{\mathrm{T}}W \leqslant c(1+n^{-1}a)\} &= 1-\alpha+n^{-1}acg(c)-n^{-1}acg(c)+O(n^{-2}) \\ &= 1-\alpha+O(n^{-2}),\end{aligned}$$

as had to be shown.

In the argument above we assumed that Bartlett correction is carried out using the true value of a, not an estimator $\hat{a}$. However, the case of $\hat{a}$ is similar. Since $\hat{a} = a + O_p(n^{-\frac{1}{2}})$ then $1+n^{-1}\hat{a} = 1+n^{-1}a+O_p(n^{-3/2})$, and so it stands to reason that when $\hat{a}$ is used the coverage error after Bartlett correction will be at most $O(n^{-3/2})$. In fact it is $O(n^{-2})$, as follows from parity properties of polynomials in Edgeworth expansions such as (3.1). See Barndorff–Nielsen & Hall (1988) for an account of this phenomenon. Therefore Bartlett correction reduces coverage error from $O(n^{-1})$ to $O(n^{-2})$ when either a mean correction or an estimated mean correction is employed.

The constant a is given by formula (3.12) of DiCiccio, Hall & Romano (1988). In general it is quite complex, and in such cases it is perhaps best estimated via the bootstrap. We shall give details shortly. The formula for a is very much simpler when

$\theta = \mu$ is a population mean. In that context, assume that the population variance matrix Σ is nonsingular, so that $r = s = t$. Put $Y = (Y^{(1)}, \ldots, Y^{(r)})^{\mathrm{T}} = \Sigma^{-\frac{1}{2}}(X - EX)$,

$$\alpha^{jkl} = E(Y^{(j)}Y^{(k)}Y^{(l)}), \quad \alpha^{jklm} = E(Y^{(j)}Y^{(k)}Y^{(l)}Y^{(m)}). \tag{3.9}$$

Then

$$a = \tfrac{5}{3}\sum_j\sum_k\sum_l (\alpha^{jkl})^2 - 2\sum_j\sum_k\sum_l \alpha^{jjl}\alpha^{kkl} + \tfrac{1}{2}\sum_j\sum_k \alpha^{jjkk}.$$

To compute an estimate $\hat{a}$ in this setting, let $\bar{X}$ and $\hat{\Sigma}$ denote the sample mean and variance respectively, put $Y_i = (Y_i^{(1)}, \ldots, Y_i^{(r)})^{\mathrm{T}} = \hat{\Sigma}^{-\frac{1}{2}}(X_i - \bar{X})$, and define

$$\hat{\alpha}^{jkl} = n^{-1}\sum_{i=1}^n Y_i^{(j)}Y_i^{(k)}Y_i^{(l)}, \quad \hat{\alpha}^{jklm} = n^{-1}\sum_{i=1}^n Y_i^{(j)}Y_i^{(k)}Y_i^{(l)}Y_i^{(m)}, \tag{3.10}$$

$$\hat{a} = \tfrac{5}{3}\sum_j\sum_k\sum_l (\hat{\alpha}^{jkl})^2 - 2\sum_j\sum_k\sum_l \hat{\alpha}^{jjl}\hat{\alpha}^{kkl} + \tfrac{1}{2}\sum_j\sum_k \hat{\alpha}^{jjkk}. \tag{3.11}$$

To estimate a via the bootstrap one argues as follows. Let $\chi_b^* = \{X_{b1}^*, \ldots, X_{bn}^*\}$, $1 \leq b \leq B$, denote independent resamples drawn randomly with replacement from the sample $\chi = \{X_1, \ldots, X_n\}$, and let $l_b^*(\theta)$ be the value of $l(\theta)$ computed from χ_b^* instead of χ. Write $l^*(\hat{\theta})$ for a generic version of $l_b^*(\hat{\theta})$. The distribution of $l^*(\hat{\theta})$, conditional on χ, is the bootstrap estimator of the distribution of $l(\theta_0)$. This approximation may be used directly instead of the chi squared approximation, or alternatively a bootstrap estimator of a may be obtained by solving the equation

$$E\{l^*(\hat{\theta}) \mid \chi\} = t(1 + n^{-1}\hat{a})$$

for $\hat{a}$. This estimator is $\sqrt{n}$-consistent. A more readily computable estimator, $\hat{a}_B$, is given by the solution of

$$B^{-1}\sum_{b=1}^B l_b^*(\hat{\theta}) = t(1 + n^{-1}\hat{a}_B).$$

Note that $\hat{a}_B \to \hat{a}$ as $B \to \infty$.

It may be shown that if approximation by the conditional distribution of $l^*(\hat{\theta})$, rather than the χ_t^2 approximation, is used to find the value of $c = \hat{c}$ for constructing the region $\mathcal{R}_c$ at (2.4), then the coverage error will be $O(n^{-2})$ and not $O(n^{-1})$.

3.4 Location Adjustment

Throughout our analysis there has been an implicit suggestion that empirical likelihood is in some sense like 'ordinary' or 'parametric' likelihood. For example we have shown that these two methods share several fundamental properties, such as Wilks's theorem and the availability of Bartlett correction. We might expect that in some sense, empirical likelihood regions should resemble those based on parametric likelihood. In particular the shape and position of the boundary of an empirical likelihood region should reflect important characteristics of the data set.

To the contrary, there is often very little connection between the empirical likelihood of a data set and its parametric likelihood; see DiCiccio, Hall & Romano (1989). However the relationship can be quite close if we focus on parametric likelihood for an appropriate *function* of the data, rather than for the data themselves. If the r-vector θ is the parameter of interest, if $\hat{\theta}$ is its estimator, and if $\hat{\Sigma}$ is an estimator of the asymptotic variance Σ of $n^{\frac{1}{2}}\hat{\theta}$, then the function of greatest interest would usually be $T = n^{\frac{1}{2}}\hat{\Sigma}^{-\frac{1}{2}}(\hat{\theta} - \theta_0)$, where θ_0 denotes the true value of θ. If the density h_T of T were known, and did not

depend on θ_0, then a likelihood based confidence region for θ_0 would be defined by

$$\{\theta : h_T[n^{\frac{1}{2}}\hat{\Sigma}^{-\frac{1}{2}}(\hat{\theta} - \theta)] \leqslant C\}.$$

The coverage probability would be determined by C. See for example Cox & Hinkley (1974, p. 236ff). Empirical likelihood provides a good approximation to this type of region, except that the function of interest should be taken to be

$$U = n^{\frac{1}{2}}(\Sigma^{\frac{1}{2}}\hat{\Sigma}^{-1}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\hat{\theta} - \theta_0)$$

instead of T. Only in the case where θ is a scalar parameter are T and U necessarily identical.

Let h_U denote the density of U. Hall (1990) showed that if the empirical likelihood confidence region $\mathcal{R}_c$ (defined at (2.4)) is translated by an amount $n^{-1}\psi$ to

$$\mathcal{R}_c + n^{-1}\psi = \{\theta + n^{-1}\psi : \theta \in \mathcal{R}_c\},$$

where ψ is a certain fixed r-vector, then the boundary of the translated region is only $O_p(n^{-3/2})$ away from that of the likelihood-based region

$$\{\theta : h_U[n^{\frac{1}{2}}(\Sigma^{\frac{1}{2}}\hat{\Sigma}^{-1}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\hat{\theta} - \theta)] \leqslant C\} \tag{3.12}$$

for an appropriate value of C. We say that $\mathcal{R}_c$ has been 'location adjusted' to $\mathcal{R}_c + n^{-1}\psi$.

In general the vector ψ has a very complicated formula, although it is relatively simple when $\theta = \mu$ is a population mean. There, $\psi = (\psi^{(1)}, \ldots, \psi^{(r)})^{\mathrm{T}}$ is given by

$$\psi^{(j)} = -\tfrac{1}{2}\sum_{k=1}^{r} \alpha^{jkk},$$

where α^{jkk} is defined at (3.9). In practice $\hat{\psi}$ would be estimated by

$$\hat{\psi}^{(j)} = -\tfrac{1}{2}\sum_{k=1}^{r} \hat{\alpha}^{jkk}, \tag{3.13}$$

where $\hat{\alpha}^{jkk}$ is defined at (3.10), and then the location adjusted region would be $\mathcal{R}_c + n^{-1}\hat{\psi}$.

By its very definition, the region defined at (3.12) has as much mass as possible in places where the density h_U is large, and as little as possible in places where h_U is small, subject to the constraint that its coverage probability be equal to a certain value (determined by C). In this respect the shape of the region reflects specific characteristics of the sampling distribution. The shape of the empirical likelihood region, and the shape *and* position of the location adjusted empirical likelihood region, must also reflect those characteristics.

The location adjusted region

$$\mathcal{Q}_c = \mathcal{R}_c + n^{-1}\hat{\psi} = \{\theta + n^{-1}\hat{\psi} : l(\theta) \leqslant c\}$$

has coverage error $O(n^{-1})$. This may be reduced to $O(n^{-2})$ by Bartlett correction. That is, there exists a constant b such that for all $c > 0$, the region

$$\mathcal{Q}_c' = \{\theta + n^{-1}\hat{\psi} : l(\theta) \leqslant c(1 + n^{-1}b)\}$$

satisfies

$$P(\theta_0 \in \mathcal{Q}_c') = P(\chi_t^2 \leqslant c) + O(n^{-2}).$$

The same result is true if b is replaced by a $\sqrt{n}$-consistent estimator $\hat{b}$. However, b is not the same as the a discussed in § 3.3.

3.5 Example: Univariate Mean

Take $r = s = 1$ and assume that $\theta_0 = \mu_0 = E(X)$, the population mean. Put $\sigma^2 = \text{var}(X)$, $\mu_3 = E(X - EX)^3/\sigma^3$, $\mu_4 = E(X - EX)^4/\sigma^4$,

$$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \hat{\mu}_3 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^3/\hat{\sigma}^3,$$

$$\hat{\mu}_4 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^4/\hat{\sigma}^4.$$

The constant a needed for Bartlett correction in § 3.3 is

$$a = \tfrac{1}{2}\mu_4 - \tfrac{1}{3}\mu_3^2,$$

and is estimated by $\hat{a} = \frac{1}{2}\hat{\mu}_4 - \frac{1}{3}\hat{\mu}_3^2$. The constant ψ needed for location adjustment in § 3.4 is

$$\psi = -\tfrac{1}{2}\mu_3,$$

and is estimated by $\hat{\psi} = -\frac{1}{2}\hat{\mu}_3$. The constant b needed for Bartlett correction of the location adjusted region in § 3.4 is

$$b = \tfrac{29}{12}\mu_3^2 - \tfrac{1}{2}\mu_4,$$

and is estimated by $\hat{b} = \frac{29}{12}\hat{\mu}_3 - \frac{1}{2}\hat{\mu}_4$.

To convey an idea of the effect of the various adjustments and corrections, we shall give approximate formulae for the ordinary empirical likelihood interval

$$\mathcal{R}_c = \{\theta : l(\theta) \leqslant c\},$$

the Bartlett corrected interval

$$\mathcal{R}_c' = \{\theta : l(\theta) \leqslant c(1 + n^{-1}\hat{a})\},$$

the location adjusted interval

$$\mathcal{Q}_c = \{\theta + n^{-1}\hat{\psi} : l(\theta) \leqslant c\},$$

and the Bartlett corrected, location adjusted interval

$$\mathcal{Q}_c' = \{\theta + n^{-1}\hat{\psi} : l(\theta) \leqslant c(1 + n^{-1}\hat{b})\}.$$

The formulae are derivable using Edgeworth expansion arguments, much as in Hall (1990), and are accurate up to a remainder of $O_p(n^{-2})$ attached to each interval endpoint.

Let $z = c^{\frac{1}{2}}$. The simplistic normal approximation confidence interval for θ_0 is $(\bar{X} - n^{-\frac{1}{2}}\hat{\sigma}z,\ \bar{X} + n^{-\frac{1}{2}}\hat{\sigma}z)$, and the various empirical likelihood intervals are

$$\begin{aligned}
\mathcal{R}_c &= (\bar{X} + n^{-\frac{1}{2}}\hat{\sigma}\{-z + n^{-\frac{1}{2}}\tfrac{1}{3}\hat{\mu}_3 z^2 + n^{-1}(\tfrac{2}{9}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4 + \tfrac{1}{2})z^3\},\\
&\qquad \bar{X} + n^{-\frac{1}{2}}\hat{\sigma}\{z + n^{-\frac{1}{2}}\tfrac{1}{3}\hat{\mu}_3 z^2 - n^{-1}(\tfrac{2}{9}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4 + \tfrac{1}{2})z^3\}),\\
\mathcal{R}_c' &= (\bar{X} + n^{-\frac{1}{2}}\hat{\sigma}\{-z + n^{-\frac{1}{2}}\tfrac{1}{3}\hat{\mu}_3 z^2 + n^{-1}(\tfrac{2}{9}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4 + \tfrac{1}{2})z^3 - n^{-1}(\tfrac{1}{4}\hat{\mu}_4 - \tfrac{1}{6}\hat{\mu}_3^2)z\},\\
&\qquad \bar{X} + n^{-\frac{1}{2}}\hat{\sigma}\{z + n^{-\frac{1}{2}}\tfrac{1}{3}\hat{\mu}_3 z^2 - n^{-1}(\tfrac{2}{9}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4 + \tfrac{1}{2})z^3 + n^{-1}(\tfrac{1}{4}\hat{\mu}_4 - \tfrac{1}{6}\hat{\mu}_3^2)z\}),\\
\mathcal{Q}_c &= (\bar{X} + n^{-\frac{1}{2}}\hat{\sigma}\{-z + n^{-\frac{1}{2}}\tfrac{1}{6}\hat{\mu}_3(2z^2 - 3) + n^{-1}(\tfrac{2}{3}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4 + \tfrac{1}{2})z^3\},\\
&\qquad \bar{X} - n^{-\frac{1}{2}}\hat{\sigma}\{z + n^{-\frac{1}{2}}\tfrac{1}{6}\hat{\mu}_3(2z^2 - 3) - n^{-1}(\tfrac{2}{3}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4 + \tfrac{1}{2})z^3\}),\\
\mathcal{Q}_c' &= (\bar{X} + n^{-\frac{1}{2}}\hat{\sigma}\{-z + n^{-\frac{1}{2}}\tfrac{1}{6}\hat{\mu}_3(2z^2 - 3) + n^{-1}(\tfrac{2}{3}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4 + \tfrac{1}{2})z^3 - n^{-1}(\tfrac{29}{24}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4)z\},\\
&\qquad \bar{X} - n^{-\frac{1}{2}}\hat{\sigma}\{z + n^{-\frac{1}{2}}\tfrac{1}{6}\hat{\mu}_3(2z^2 - 3) - n^{-1}(\tfrac{2}{3}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4 + \tfrac{1}{2})z^3 + n^{-1}(\tfrac{29}{24}\hat{\mu}_3^2 - \tfrac{1}{4}\hat{\mu}_4)z\}).
\end{aligned}$$

The latter two intervals, $\mathcal{Q}_c$ and $\mathcal{Q}'_c$, agree up to an error of order $n^{-3/2}$ with the 'shortest bootstrap confidence intervals' introduced by Hall (1988).

4 Algorithms for Specific Examples

4.1 *Introduction*

In this section we discuss algorithms which can be used to find empirical likelihood confidence regions in specific cases, including regions for means (both univariate and bivariate), variances and correlation coefficients. Numerical implementation of these methods would usually be based on a multivariate Newton algorithm, which is discussed in § 4.2. Sections 4.3–4.5 treat specific univariate examples, § 4.6 deals with the bivariate mean, and § 4.7 considers the case of a pair of variances.

To describe the algorithms, recall that empirical loglikelihood ratio is defined by

$$l(\theta_1) = -2 \max_{p:\theta(p)=\theta_1, \sum p_i=1} \sum \log (np_i),$$

and that the confidence region is determined by

$$\mathcal{R}_c = \{\theta : l(\theta) \leqslant c\}.$$

Here c might be given directly by chi squared tables, or might incorporate a Bartlett correction, or might be computed from the bootstrap approximation to the distribution of $l(\theta)$; see § 3.3. The boundary or contour of $\mathcal{R}_c$ is the set $\mathcal{C}_c = \{\theta : l(\theta) = c\}$, and equals the collection of all values $\theta(p)$ such that p is a turning point of $\theta(p)$ subject to the constraints

$$-2 \sum \log (np_i) = c, \quad \sum p_i = 1. \tag{4.1}$$

This follows from the definition of $\mathcal{R}_c$, and is the cornerstone of techniques discussed in §§ 4.3–4.7 below. If θ is a scalar (i.e. if $r = 1$) then, provided θ is a function of a (vector) mean, $\mathcal{R}_c$ is always an interval, say $\mathcal{R}_c = (\theta_L, \theta_U)$. In this case the function $\theta(p)$ has just two turning points subject to (4.1), and we may label them p_L and p_U such that $\theta_L = \theta(p_L)$ and $\theta_U = \theta(p_U)$.

The problem of finding a turning point of $\theta(p)$ subject to constraints (4.1) may be solved by the method of Lagrange multipliers, which produces a system of simultaneous equations. A multivariate version of Newton's algorithm may be used to solve such a system, as we now relate.

4.2 *Multivariate Newton Algorithm*

Let $\omega = (\omega^{(1)}, \ldots, \omega^{(m)})^{\mathrm{T}}$ be an m-vector, and suppose we wish to solve the equations

$$f_i(\omega) = 0 \quad (1 \leqslant i \leqslant m),$$

where $f_1, \ldots, f_m$ are known functions. Put $f = (f_1, \ldots, f_m)^{\mathrm{T}}$, $H^{ij} = \partial f_i / \partial \omega^{(j)}$ and $H = (H^{ij})$ (an $m \times m$ matrix). Given an initial approximation $\omega(1)$ to a solution ω_0 of $f(\omega) = 0$, develop successive approximations $\omega(j)$ by iteration:

$$\omega(j+1) = \omega(j) - \{H(\omega(j))\}^{-1} f(\omega(j)) \quad (j \geqslant 1).$$

If the initial solution $\omega(1)$ is sufficiently close to ω_0, and if H is continuous in a neighbourhood of ω_0 and nonsingular at ω_0, then $\omega(j) \to \omega_0$ as $j \to \infty$.

4.3 Univariate Mean

We wish to find a confidence interval for the population mean. Here $\theta(p) = \sum p_i X_i$, and so we seek the turning points of $\sum p_i X_i$ subject to (4.1). That is, we require the turning points of

$$\sum_{i=1}^{n} p_i X_i + \beta\Big\{\sum_{i=1}^{n} \log(np_i) + \tfrac{1}{2}c\Big\} + \gamma\Big(\sum_{i=1}^{n} p_i - 1\Big),$$

where β, γ are Lagrange multipliers. Differentiating with respect to p_i, equating to zero, and changing variable from (β, γ) to new parameters (λ, μ), we see that

$$p_i = n^{-1}\{1 + \lambda(X_i - \mu)\}^{-1} \quad (1 \leqslant i \leqslant n).$$

The constraint $\sum p_i = 1$ is therefore equivalent to $\sum p_i(X_i - \mu) = 0$, and $\sum \log(np_i) = -\frac{1}{2}c$ is equivalent to $\sum \log\{1 + \lambda(X_i - \mu)\} = \frac{1}{2}c$. Therefore, find the two solutions (λ_1, μ_1), (λ_2, μ_2) of the equations

$$f_1(\lambda, \mu) \equiv \sum_{i=1}^{n} \{1 + \lambda(X_i - \mu)\}^{-1}(X_i - \mu) = 0,$$

$$f_2(\lambda, \mu) \equiv \sum_{i=1}^{n} \log\{1 + \lambda(X_i - \mu)\} - \tfrac{1}{2}c = 0,$$

and take μ_L, μ_U to be the smallest, largest respectively of μ_1, μ_2.

4.4 Univariate Variance

We wish to find a confidence interval (τ_L, τ_U) for the population variance $\theta = \tau$, and so we seek the turning points of $\theta(p) = \sum p_i X_i^2 - (\sum p_i X_i)^2$ subject to the constraints (4.1). Arguing as in § 4.3 we see that this amounts to finding the two solutions $(\lambda_1, \mu_1, \tau_1)$ and $(\lambda_2, \mu_2, \tau_2)$ of the three equations

$$f_1(\lambda, \mu, \tau) \equiv \sum_{i=1}^{n} [1 + \lambda\{(X_i - \mu)^2 - \tau\}]^{-1}(X_i - \mu) = 0,$$

$$f_2(\lambda, \mu, \tau) \equiv \sum_{i=1}^{n} [1 + \lambda\{(X_i - \mu)^2 - \tau\}]^{-1}\{(X_i - \mu)^2 - \tau\} = 0,$$

$$f_3(\lambda, \mu, \tau) \equiv \sum_{i=1}^{n} \log[1 + \lambda\{(X_i - \mu)^2 - \tau\}] - \tfrac{1}{2}c = 0,$$

and taking τ_L, τ_U to be the smallest, largest respectively of τ_1, τ_2.

4.5 Correlation Coefficient

Here the data are bivariate, $X_i = (Y_i, Z_i)^{\mathrm{T}}$, $1 \leqslant i \leqslant n$. We seek a confidence interval (ρ_L, ρ_U) for the population correlation coefficient,

$$\rho = \{E(YZ) - E(Y)E(Z)\}\{E(Y^2) - (EY)^2\}^{-\frac{1}{2}}\{E(Z^2) - (EZ)^2\}^{-\frac{1}{2}}.$$

Therefore we require the turning points of

$$\theta(p) = \Big\{\sum p_i Y_i Z_i - \Big(\sum p_i Y_i\Big)\Big(\sum p_i Z_i\Big)\Big\}\Big\{\sum p_i Y_i^2 - \Big(\sum p_i Y_i\Big)^2\Big\}^{-\frac{1}{2}}\Big\{\sum p_i Z_i^2 - \Big(\sum p_i Z_i\Big)^2\Big\}^{-\frac{1}{2}}$$

subject to (4.1). This is equivalent to the following prescription. Define

$$r_i = r_i(\lambda, \mu_Y, \mu_Z, \tau_Y, \tau_Z, \gamma)$$
$$= [1 + \lambda\{(Y_i - \mu_Y)(Z_i - \mu_Z) - (\gamma/2\tau_Y)(Y_i - \mu_Y)^2 - (\gamma/2\tau_Z)(Z_i - \mu_Z)^2\}]^{-1}.$$

Find the two solutions $(\lambda_1, \mu_{Y1}, \mu_{Z1}, \tau_{Y1}, \tau_{Z1}, \gamma_1)$ and $(\lambda_2, \mu_{Y2}, \mu_{Z2}, \tau_{Y2}, \tau_{Z2}, \gamma_2)$ of the six equations

$$\sum r_i(Y_i - \mu_Y) = 0, \quad \sum r_i(Z_i - \mu_Z) = 0,$$

$$\sum r_i\{(Y_i - \mu_Y)^2 - \tau_Y\} = 0, \quad \sum r_i\{(Z_i - \mu_Z)^2 - \tau_Z\} = 0,$$

$$\sum r_i\{(Y_i - \mu_Y)(Z_i - \mu_Z) - \gamma\} = 0, \quad \sum \log r_i^{-1} - \tfrac{1}{2}c = 0.$$

(There are six equations in six unknowns.) Put $\rho_i = \gamma_i/\tau_{Yi}\tau_{Zi}$ for $i = 1, 2$, and take ρ_L, ρ_U to be the smallest, largest respectively of ρ_1, ρ_2.

4.6 Bivariate Mean

Again the data $X_i = (Y_i, Z_i)^T$ are bivariate, and we seek a confidence region (a subset of $\mathbb{R}^2$) for the mean $(\mu_Y, \mu_Z) = (EY, EZ)$. Therefore we require the set of all turning points of $\sum p_i X_i$ subject to (4.1). This may be computed by working through the prescription in the second paragraph of § 4.1, which gives the following algorithm. Define

$$r_i = r_i(\lambda_Y, \lambda_Z, \mu_Y, \mu_Z) = \{1 + \lambda_Y(Y_i - \mu_Y) + \lambda_Z(Z_i - \mu_Z)\}^{-1},$$

and solve the three equations

$$\sum r_i(Y_i - \mu_Y) = 0, \quad \sum r_i(Z_i - \mu_Z) = 0, \quad \sum \log r_i^{-1} = \tfrac{1}{2}c \tag{4.2}$$

for the four unknowns λ_Y, λ_Z, μ_Y, μ_Z. (The three equations imply $\sum r_i = n$). This gives $r_i = r_i(u)$ for $u \in U$, say, where u is a scalar and absorbs the extra degree of freedom. Put

$$\mu_Y(u) = n^{-1} \sum r_i(u) Y_i, \quad \mu_Z(u) = n^{-1} \sum r_i(u) Z_i.$$

Then the contour of the confidence region $\mathcal{R}_c$ is

$$\mathcal{C}_c = \{\theta : l(\theta) = c\} = \{n^{-1} \sum r_i(u) X_i : u \in U\} = \{(\mu_Y(u), \mu_Z(u)) : u \in U\}.$$

In practice the parametrization by u is perhaps most easily accommodated as follows. Let $u \in U = [0, 2\pi)$ and take $\lambda_Y = \lambda \sin u$, $\lambda_Z = \lambda \cos u$. For fixed u, the three equations (4.2) are in only three variables (λ, μ_Y, μ_Z), and may be solved by the multivariate Newton algorithm to give $(\lambda(u), \mu_Y(u), \mu_Z(u))$. Thus we obtain the contour

$$\mathcal{C}_c = \{(\mu_Y(u), \mu_Z(u)) : u \in [0, 2\pi)\}.$$

4.7 Two Variances

The data are again bivariate, $X_i = (Y_i, Z_i)^T$, and we seek a confidence region for the pair of variances $(\tau_Y, \tau_Z) = (\text{var } Y, \text{var } Z)$. This is given by the set of all turning points of

$$\left(\sum p_i Y_i^2 - \left(\sum p_i Y_i\right)^2, \quad \sum p_i Z_i^2 - \left(\sum p_i Z_i\right)^2\right)$$

subject to (4.1), and may be computed as follows. Define

$$r_i = [1 + \lambda_Y\{(Y_i - \mu_Y)^2 - \tau_Y\} + \lambda_Z\{(Z_i - \mu_Z)^2 - \tau_Z\}]^{-1},$$

a function of six variables for which there are only five equations,

$$\sum r_i(Y_i - \mu_Y) = 0, \quad \sum r_i(Z_i - \mu_Z) = 0, \quad \sum r_i\{(Y_i - \mu_Y)^2 - \tau_Y\} = 0,$$
$$\sum r_i\{(Z_i - \mu_Z)^2 - \tau_Z\} = 0, \quad \sum \log r_i^{-1} = \tfrac{1}{2}c. \tag{4.3}$$

Reduce the number of variables to five by putting $\lambda_Y = \lambda \sin u$ and $\lambda_Z = \lambda \cos u$ where $u \in [0, 2\pi)$. For fixed u, solve the five equations (4.3) for the five unknowns (λ, μ_Y, μ_Z, τ_Y, τ_Z), obtaining ($\lambda(u)$, $\mu_Y(u)$, $\mu_Z(u)$, $\tau_Y(u)$, $\tau_Z(u)$). The desired confidence region is bounded by the contour

$$\mathscr{C}_c = \{(\tau_Y(u), \tau_Z(u)) : u \in [0, 2\pi)\}.$$

Others cases may be treated similarly.

5 Numerical Examples

To begin, we illustrate several of the points made in earlier sections by depicting 95% empirical likelihood confidence regions for a bivariate mean. All computations were based on the multivariate Newton algorithm and the method described in § 4.6, and used Owen's (1990) duck data. Sample size was $n = 11$. The unbroken curve in each figure represents the contour of the basic empirical likelihood region

$$\{\theta : l(\theta) = c\},$$

where $c = 5{\cdot}991$ is defined by $P(\chi_2^2 = c) = 0{\cdot}95$. The broken curve in Fig. 1 is the boundary of the Bartlett corrected confidence region, that is the contour defined by

$$\{\theta : l(\theta) = c(1 + n^{-1}\hat{a})\},$$

where c is as for curve (1) and $\hat{a}$ is given by (3.11). The broken curve in Fig. 2 is the

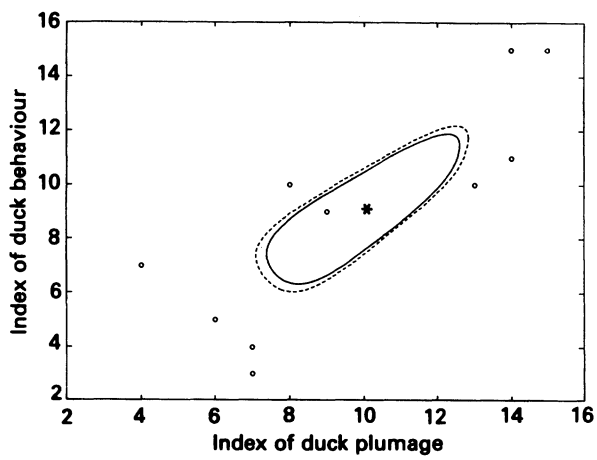

Figure 1. *Contours of ordinary (unbroken) and Barlett corrected (broken) empirical likelihood confidence regions. Circles, data points; asterisk, sample mean; see Owen* (1990) *for further details.*

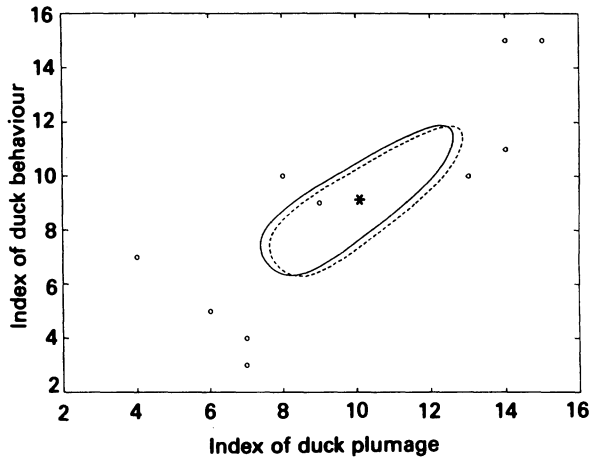

Figure 2. *Contours of ordinary (unbroken) and location adjusted (broken) regions.*
Circles, data points; asterisk, sample mean; see Owen (1990) *for further details.*

boundary of the location adjusted confidence region, that is the contour

$$\{\theta + n^{-1}\hat{\psi} : l(\theta) = c\},$$

where c is as before and $\hat{\psi}$ is given by (3.13).

DiCiccio, Hall & Romano (1989) report a simulation study on the coverage accuracy of empirical likelihood for constructing confidence intervals for an unknown population mean. The performance is quite good, although not spectacular. In the case of samples from normal, chi squared (1 d.f.) and Student's t (5 d.f.) populations, and sample sizes between 20 and 30, the true coverage of a nominal 95% interval varies between 89% and 93%. For the present paper we undertook a simulation study of the more difficult problem of estimating a correlation coefficient ρ, sampling from a bivariate normal distribution. The results are reported in Table 1. There is consistent undercoverage, and

Table 1

Estimated true coverage, mean value of midpoint and mean length of empirical likelihood confidence intervals for the correlation coefficient in a normal population, when the true value, ρ, of the coefficient is either 0 or 0·5. Sample size, n, is the range $7 \leq n \leq 20$. Each value in the table is the average of 1000 simulations; nominal coverage, 0·95.

	$\rho = 0$			$\rho = 0{\cdot}5$		
n	Cov.	Av. midpt.	Av. length	Cov.	Av. midpt.	Av. length
7	75·0	0·012	0·944	76·5	0·361	0·810
8	78·7	0·010	0·918	77·8	0·392	0·778
9	79·0	−0·011	0·901	77·6	0·401	0·731
10	82·6	−0·009	0·889	80·7	0·396	0·724
11	82·6	0·004	0·870	83·8	0·408	0·700
12	84·7	0·010	0·854	83·0	0·410	0·688
13	85·3	0·014	0·827	83·9	0·409	0·680
14	85·4	0·014	0·811	83·1	0·427	0·639
15	86·1	0·010	0·794	84·8	0·424	0·631
16	86·7	0·009	0·779	84·5	0·422	0·614
17	86·0	0·014	0·768	85·3	0·439	0·605
18	87·9	−0·002	0·753	87·5	0·443	0·584
19	88·2	0·005	0·738	88·9	0·441	0·580
20	88·0	0·005	0·719	89·0	0·441	0·576

coverage does not depend appreciably on the value of the coefficient. As expected, given the restricted range of values which the correlation coefficient can assume, the intervals tend to be somewhat shorter in the case $\rho = 0.5$ than they are for $\rho = 0$.

Acknowledgements

The paper has benefited from helpful suggestions by two referees and an Associate Editor.

References

Barndorff–Nielsen, O.E. & Cox, D.R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. R. Statist. Soc.* B **46,** 483–495.

Barndorff–Nielsen, O.E. & Hall, P. (1988). On the level-error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika* **75,** 374–378.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74,** 457–468.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25,** 573–578.

Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

DiCiccio, T.J., Hall, P. & Romano, J.P. (1988). Empirical likelihood is Bartlett-correctable. Unpublished manuscript.

DiCiccio, T.J., Hall, P. & Romano, J.P. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika* **76,** 465–476.

DiCiccio, T.J. & Romano, J.P. (1989). On adjustments based on the signed root of the empirical likelihood ratio statistic. *Biometrika* **76,** 447–456.

Feder, P.I. (1975). The log likelihood ratio in segmented regression. *Ann. Statist.* **3,** 84–97.

Hall, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14,** 1431–1452.

Hall, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika* **74,** 481–494.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16,** 927–985.

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *Ann. Statist.* **18**. To appear.

Hall, P. & Owen, A.B. (1989). Empirical likelihood confidence bands in curve estimation. Unpublished manuscript.

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

Owen, A.B. (1990). Empirical likelihood confidence regions. *Ann. Statist.* **18**. To appear.

Petrov, V.V. (1975). *Sums of Independent Random Variables*. Berlin: Springer.

Rao, C.R. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley.

Thomas, D.R. & Grunkemeier, G.L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Am. Statist. Assoc.* **70,** 865–871.

Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9,** 60–62.

Résumé

Nous décrivons les traits principaux de la méthode de vraisemblance empirique et discutons les développements récents, incluant la correction de Bartlett et l'ajustement de location. Nous donnons des algorithmes pour la méthode de vraisemblance empirique dans certains cas importants, par exemple pour les moyennes, les variances et les coefficients de corrélations. On démontre que la méthode de vraisemblance empirique est une concurrente sérieuse comparée à d'autres méthodes contemporaines comme l'auto-amorcage. D'ailleurs, la méthode de vraisemblance empirique a plusieurs avantages. Elle n'impose pas de constraintes préalables sur la forme de la région. Elle n'exige pas de construction de statistique pivotale et elle permet la correction de Bartlett qui donne une erreur de recouvrement très basse. La méthode de vraisemblance empirique mérite une place importante dans l'arsenal de techniques informatiquement intensive du statisticien moderne.

[*Received February* 1989, *accepted January* 1990]