On the small sample distributions of the empirical loglikelihood ratio

Key words and phrases: Confidence region; bounds on coverage level; $E$ distributions; empirical loglikelihood ratio; estimating equations; Hotelling’s $T^2$; multivariate normal distribution; parametric likelihood ratio.


Abstract: The author shows that the empirical loglikelihood ratio is also pivotal for the multivariate normal mean and derives $E$ distributions, a family of distributions equivalent to Hotelling’s $T^2$ in the context of empirical likelihood. For empirical likelihood ratio confidence regions based on estimating equations, the author discusses bounds on their coverage levels derived from the atoms of the $E$ distributions. The author also introduces $E$ calibration which calibrates a general empirical loglikelihood ratio using the $E$ distributions. Discussions on the atoms and bounds provide further understanding on the undercoverage problem of the asymptotic Chi-square calibrated empirical likelihood ratio confidence regions. The $E$ calibration respects bounds on coverage levels and is a simple, effective method for correcting the undercoverage problem.

1 Introduction

Although it is entirely based on data, the empirical likelihood (Owen, 2001) has some of the important asymptotic properties of the parametric likelihood. These include the fundamental properties that the limiting distribution of the empirical loglikelihood ratio is a Chi-square distribution (Owen, 1988, 1990) and that the empirical likelihood is Bartlett-correctable (DiCiccio, Hall and Romano, 1991).

While many asymptotic similarities between the empirical and parametric likelihoods have been discovered, there are fewer results on similarities between the two likelihoods for finite sample sizes. The fact that the empirical likelihood does not have an analytic expression certainly makes it difficult to study its finite sample properties. This paper is concerned with its finite sample properties for the normal case and the applications of these properties. In the normal case, we can bypass the problem of lack of analytic expres-
sion through a combination of affine transformation and simulation. It is well-known that for a multivariate normal mean, the parametric loglikelihood is pivotal, leading to Wilk’s lambda statistic and Hotelling’s $T^2$ distribution. We show that the empirical loglikelihood ratio is also pivotal for the multivariate normal mean. That is, the distribution of the empirical loglikelihood ratio at the true mean is independent of the mean and covariance matrix of the underlying normal distribution. For convenience, we will refer to this distribution as the $E$ distribution. The $E$ distribution depends on only the dimension of the normal distribution and the sample size just like Hotelling’s $T^2$.

There is, however, an important difference between the $E$ and the $T^2$ distributions; the $E$ distributions each has an atom at the positive infinity and the $T^2$ does not. The significance of the atom is two fold. Firstly it imposes a bound on the coverage level of the empirical likelihood ratio confidence region, and secondly it affects the accuracy of the asymptotic Chi-square approximation to the finite sample distribution of the empirical loglikelihood ratio; the Chi-square approximation cannot accommodate an atom at the infinity. The asymptotic Chi-square calibrated empirical likelihood ratio confidence region suffers from an undercoverage problem in that the actual coverage level (probability) tend to be lower than the nominal level. Numerical evidence of the undercoverage problem can be found throughout the literature. Qin and Lawless (1994), for example, noted the undercoverage problem in empirical likelihood inference for estimating equations. Research on the asymptotics of the empirical likelihood have led to more accurate methods of calibration. The most noted of these is the estimated Bartlett correction of DiCiccio, Hall and Romano (1991). In real applications, the coverage error of an asymptotically exact calibration method is always the consequence of the finite sample sizes. This asymptotic explanation may be satisfactory if the calibration method results in undercoverage sometimes but not nearly always. In light of the persistent undercoverage problem of the asymptotic Chi-square calibration, we seek understanding of its coverage error in addition to the asymptotic explanation. Owen (2001) was the first to observe the constraint on the coverage level placed by some convex halls in various problems. Through discussions on the atom and the bound, we build on this observation and provide a more detailed finite sample perspective on the undercoverage problem. We also introduce an $E$ distribution based calibration method, the $E$ calibration, which respects the bound and consistently corrects the undercoverage problem of the Chi-square calibration.
The rest of this paper is organized as follows: in Section 2 we show that the empirical likelihood is pivotal for the normal mean and present the $E$ distributions. An affine equivariance property of the empirical loglikelihood ratio is also given. In Section 3, we discuss the bounds on the coverage levels of empirical likelihood ratio confidence regions based on estimating equations. In Section 4, we provide a finite sample perspective on the undercoverage problem of the Chi-square calibration through the atoms of the $E$ distributions and the bounds on the coverage levels. We introduce the $E$ calibration and discuss its advantages relative to other methods of calibration.

2 The distribution of the empirical loglikelihood ratio for the multivariate normal mean

Let $X_1, X_2, \cdots, X_n$ be independent copies of a non-singular multivariate normal random vector $X \sim \text{MVN}(\mu_T, \Sigma)$ in $\mathbb{R}^k$ and $n > k$. The empirical loglikelihood ratio for the mean is

$$l(\mu) = 2 \sum_{i=1}^{n} \log\{1 + \lambda^T(X_i - \mu)\}, \quad (2.1)$$

where $\mu$ and $\lambda = \lambda(\mu)$ are in $\mathbb{R}^k$, and $\lambda$ is determined by

$$\sum_{i=1}^{n}\{1 + \lambda^T(X_i - \mu)\}^{-1}(X_i - \mu) = 0, \quad (2.2)$$

where 0 is the origin in $\mathbb{R}^k$. See Owen (1990). When $\mu$ is not in the convex hall of $X_i$, (2.2) does not have a solution and $l(\mu) = +\infty$.

Let $Y$ be a standard multivariate normal random vector, $Y \sim \text{MVN}(0, I)$, where $I$ is the $k \times k$ identity matrix. Let $Y_1, Y_2, \cdots, Y_n$ be independent copies of $Y$. We define the $E$ distribution, $E_{k,n}$, as the distribution of the $E$ random variable

$$E = l(0) = 2 \sum_{i=1}^{n} \log\{1 + \lambda^TY_i\}, \quad (2.3)$$

where $\lambda = \lambda(0)$ is determined by

$$\sum_{i=1}^{n}(1 + \lambda^TY_i)^{-1}Y_i = 0. \quad (2.4)$$
For fixed $n$ and $k$, there is a positive probability $a(k, n)$ that the convex hall of $Y_i$ does not contain the mean $0$. That is, the $E_{k,n}$ distribution has an atom of value $a(k, n)$ at the positive infinity. The following lemma shows that the distribution of $l(\mu_T)$ also has an atom of value $a(k, n)$ at the positive infinity.

**Lemma 1.** For a non-singular multivariate normal distribution $MVN(\mu_T, \Sigma)$, the probability that the convex hall of $X_i$ does not contain $\mu_T$ is $a(k, n)$.

**Proof of Lemma 1:** Since $X$ is non-singular, $B = \Sigma^{1/2}$ is symmetric, positive-definite and there is a corresponding $Y \sim MVN(0, I)$ such that $X = BY + \mu_T$. So it suffices to show that the convex hall of $X_i$ contains $\mu_T$ if and only if the convex hall of $Y_i$ contains $0$. If the convex hall of $X_i$ contains $\mu_T$, then there is a set of weights $w_1, w_2, \ldots, w_n$ such that $w_i \geq 0, \sum w_i = 1$ and $\sum w_i X_i = \mu_T$. This implies $B \sum w_i Y_i = 0$. Since $B$ is non-singular, $\sum w_i Y_i = 0$. Thus $0$ is in the convex hall of $Y_i$. It is easy to see that the converse is also true. 

We now show that the empirical loglikelihood ratio is pivotal for $\mu_T$.

**Theorem 1.** If $X \sim MVN(\mu_T, \Sigma)$ is non-singular, then the distribution of $l(\mu_T)$ is $E_{k,n}$, regardless of the values of $\mu_T$ and $\Sigma$.

**Proof of Theorem 1:** Suppose $\mu_T$ is in the convex hall of $X_i$. Then $0$ is in the convex hall of the corresponding $Y_i$. From the proof of Lemma 1, $X = BY + \mu_T$. By (2.1) and (2.2),

$$l(\mu_T) = 2 \sum_{i=1}^{n} \log\{1 + \lambda^T (X_i - \mu_T)\} = 2 \sum_{i=1}^{n} \log\{1 + (B\lambda)^T Y_i\},$$

(2.5)

where $\lambda = \lambda(\mu_T)$ is determined by

$$\sum_{i=1}^{n} \{1 + \lambda^T (X_i - \mu_T)\}^{-1} (X_i - \mu_T) = B \sum_{i=1}^{n} \{1 + (B\lambda)^T Y_i\}^{-1} Y_i = 0.$$  

(2.6)

Since $B$ is non-singular, (2.6) implies that

$$\sum_{i=1}^{n} \{1 + (B\lambda)^T Y_i\}^{-1} Y_i = 0.$$  

(2.7)

Letting $\beta = B\lambda$, (2.5) and (2.7) may be written as

$$l(\mu_T) = 2 \sum_{i=1}^{n} \log\{1 + \beta^T Y_i\} \quad \text{and} \quad \sum_{i=1}^{n} \{1 + \beta^T Y_i\}^{-1} Y_i = 0,$$

4
respectively, which are exactly the same equations as (2.3) and (2.4). It follows that the
distribution of $l(\mu_T)$ in (2.5), conditional on that $\mu_T$ is in the convex hall of $X_i$, is the
same as that of the conditional distribution of $l(0)$ in (2.3).

Furthermore, by Lemma 1, $P[l(\mu_T) = +\infty] = a(k, n)$. Hence $l(\mu_T) \sim E_{k,n}$. #

If $X \sim MVN(\mu_T, \Sigma)$ is singular where rank($\Sigma$) = $d < k$, then there is a $k \times d$ matrix
$B$ with rank($B$) = $d$ and a $Y \sim MVN(0, I_{d \times d})$ such that $X = BY + \mu_T$. With minor
modifications, the above proof can be used to show that the distribution of $l(\mu_T)$ is $E_{d,n}$, regardless of the values of $\mu_T$ and $\Sigma$.

A more concise proof may be obtained by using the profile empirical likelihood function. See, e.g., (3.1) in Owen (2001, p30). Nevertheless, we adopted the above proof here because it made clear the dependence of the $E$ random variables on $Y_i$ and the atoms of the $E$ distributions. Note that the proof did not use any information regarding the multivariate normal distribution except for the fact that any multivariate normal random vector $X$ may be expressed as an affine transformation of the standard multivariate normal vector $Y$. Thus it implies the following affine equivariance property of the empirical loglikelihood ratio.

**Theorem 2.** Let $\{U_1, U_2, \cdots, U_n\}$ and $\{V_1, V_2, \cdots, V_n\}$ be two sets of iid random vectors
where $U_i \in \mathcal{R}^p$, $V_i \in \mathcal{R}^q$, $E(U_i) = \mu_U$, $E(V_i) = \mu_V$ and $p \geq q$. Suppose there exists a $B_p \times q$
with rank($B$) = $q$ and a $b \in \mathcal{R}^p$ such that $U_i = BV_i + b$. Then the distribution of $l(\mu_U)$ is
the same as that of $l(\mu_V)$.

The empirical likelihood is asymptotically pivotal regardless of the parameter of interest and the underlying distribution. Theorem 1 indicates that it is also pivotal for the mean for finite sample sizes when the underlying distribution is multivariate normal. Since $l(\mu_T)$ has an asymptotic $\chi^2_k$ distribution, for large sample sizes, the $E_{k,n}$ distribution is essentially a $\chi^2_k$ distribution. For small to moderately large sample sizes, however, their differences can be substantial. The atom of $E_{k,n}$ can be quite large for small sample sizes, contributing to the difference.

The value of the atom equals the probability that a random sample of size $n$ from an
$MVN(0, I) \in \mathcal{R}^k$ are all on one side of a hyperplane through the origin. Wendel (1962)
proved the following formula for this probability. For any $n > k$,

$$a(k, n) = \left\{ \binom{n-1}{0} + \binom{n-1}{1} + \ldots + \binom{n-1}{k-1} \right\} 2^{-(n-1)}. \tag{2.8}$$

It is interesting to note that by (2.8), when the sample size is twice the dimension the value of the atom $a(k, 2k) = 0.5$. The asymptotic behaviour of the atom when $k = \varepsilon n$ for some $\varepsilon \in (0, 1)$ will be considered later in the proof of Theorem 4.

Table 1: Some estimated critical values of the $E_{k,n}$ distributions.

<table>
<thead>
<tr>
<th>$n \backslash \alpha$</th>
<th>Dimension $k = 1$</th>
<th>Dimension $k = 2$</th>
<th>Dimension $k = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.20</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>3.439</td>
<td>9.501</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>6</td>
<td>2.812</td>
<td>6.130</td>
<td>14.181</td>
</tr>
<tr>
<td>7</td>
<td>2.594</td>
<td>5.018</td>
<td>9.416</td>
</tr>
<tr>
<td>8</td>
<td>2.342</td>
<td>4.733</td>
<td>7.794</td>
</tr>
<tr>
<td>9</td>
<td>2.279</td>
<td>4.211</td>
<td>6.690</td>
</tr>
<tr>
<td>10</td>
<td>2.166</td>
<td>4.009</td>
<td>6.141</td>
</tr>
<tr>
<td>11</td>
<td>2.035</td>
<td>3.726</td>
<td>5.712</td>
</tr>
<tr>
<td>$\chi^2_k$</td>
<td>1.642</td>
<td>2.705</td>
<td>3.841</td>
</tr>
</tbody>
</table>

Except the values of the atom, the exact $E_{k,n}$ distribution is not available. To estimate this distribution through simulation, we may generate $n$ random observations from $MVN(0, I)$ and compute $l(0)$ using the $n$ observations. The $l(0)$ is then a random observation from $E_{k,n}$. Table 1 contains estimates of critical values of some $E_{k,n}$ distributions. Each estimate is the $(1 - \alpha)100\text{th}$ quantile of a random sample of 10,000 from the corresponding $E_{k,n}$ distribution adjusted slightly to meet an order constraint (????, 2001). For computing $l(0)$ with the $n$ observations, we used a program $el.S$ by Professor A. B. Owen. Later we will refer to the $\alpha \times 100\%$ critical value of $E_{k,n}$ as $e_{\alpha,k,n}$. Also, a $+\infty$ at $(k, n, \alpha) = (1, 5, 0.05)$, for example, indicates that the atom of $E_{1,5}$ is more than 5%. More complete tables of the $E$ distributions may be found in ???? (2001).

Table 1 indicates that the continuous Chi-square approximation to the $E$ distribution is poor when the sample size is not large. For $k = 2$, for example, the atom is more
than 5% for sample sizes as large as 8. Comparing the critical values of the asymptotic Chi-square distributions (last row) with the corresponding values of the $E$ distributions, we see that the latter are consistently larger. When $n$ is not large and $k$ is high, they are substantially larger.

3 Bounds on the coverage level

Consider a parameter of interest $\theta$ of a continuous random vector $Y$, which may or may not have a multivariate normal distribution. Let $Y_1, Y_2, \cdots, Y_n$ be $n$ iid copies of a $Y$ whose $\theta = \theta_0$. Let $m(Y, \theta) \in \mathcal{R}^k$ be an estimating function for $\theta_0$ that is continuous in $Y$. The empirical likelihood ratio function for $\theta$ is

$$R(\theta) = \sup \left\{ \prod_{i=1}^{n} w_i \left\{ \sum_{i=1}^{n} w_i m(Y_i, \theta) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\} \right\}. \quad (3.9)$$

See Qin and Lawless (1994) and Owen (2001). The loglikelihood ratio $l(\theta)$ is given by $l(\theta) = -2 \log R(\theta)$. The empirical likelihood ratio confidence region for $\theta_0$ is given by

$$\{\theta | l(\theta) < r\}, \quad (3.10)$$

where $r$ is a finite quantity determined by the desired nominal level and the method of calibration that one uses. We impose no conditions on $m(Y, \theta)$ other than that it be continuous in $Y$, and asymptotically (3.10) may or may not be a meaningful confidence region for $\theta_0$. The latter does not affect our discussion on the bounds.

Let $X_i = m(Y_i, \theta_0)$. Then $X_i$ are iid copies of the continuous random vector $X = m(Y, \theta_0) \in \mathcal{R}^k$. Denote by $H(X_1, X_2, \cdots, X_n)$ the convex hall of $X_i$. Consider the coverage level of (3.10), i.e., the probability that it covers the true value $\theta_0$. Because $l(\theta_0)$ equals positive infinity when $0 \in H(X_1, X_2, \cdots, X_n)$, for all finite $r$ the coverage level of (3.10) is bounded above by

$$B = P\{0 \in H(X_1, X_2, \cdots, X_n)\}. \quad (3.11)$$

Let $B^c = 1 - B$. Then there is a probability of at least $B^c$ that confidence region (3.10) does not cover the true value $\theta_0$. 

7
The distribution of $X$ and thus $B$ depend on the estimating function $m(Y,\theta)$, the true value $\theta_0$ and the distribution of $Y$. In the nonparametric setting of the empirical likelihood, the distribution of $Y$ is unknown. So is the value of $\theta_0$. Thus the distribution of $X$ is unknown. The bound $B$ cannot be determined. Nevertheless, if we put aside $m(Y,\theta)$, $\theta_0$ and the distribution of $Y$ for a moment and allow $X_i$ in (3.11) to be an arbitrary set of continuous iid random vectors in $\mathcal{R}^k$, then because $B$ is bounded above by 1, there exists a least upper bound $b(k,n)$ for $B$:

$$b(k,n) = \sup\{B : \forall X_1, X_2, \cdots, X_n \text{ iid and continuous in } \mathcal{R}^k\}.$$  

(3.12)

And if we can find a combination of $Y$, $\theta_0$ and $m(Y,\theta)$ for which the $B$ equals $b(k,n)$, then $b(k,n)$ is also the least upper bound on the coverage level of the empirical likelihood ratio confidence region based on estimating equation (3.10). For the case of the multivariate normal mean, the bound $B = 1 - a(k,n)$. Theorem 3 below shows that for $k = 1, 2$, this $B$ is the least upper bound $b(k,n)$. Theorem 4 gives two monotone properties and an asymptotic property of $b(k,n)$. To prove Theorem 3, we need Lemma 2.

Let $v_i = \|X_i\|_2$ and without loss of generality assume $v_i > 0$ for $i = 1, 2, \cdots, n$. Then $U_i = v_i^{-1}X_i$ are the projections of $X_i$ on the unit sphere. We have

**Lemma 2:** $\theta \notin H(X_1, \cdots, X_n)$ if and only if $\theta \notin H(U_1, \cdots, U_n)$.

**Proof of Lemma 2:** The convex hall $H(X_1, \cdots, X_n)$ does not contain $\theta$ if and only if all $X_i$ are on one side of a hyperplane through $\theta$. All $X_i$ are on one side of a hyperplane through $\theta$ if and only if their projections $U_i$ are on one side of a hyperplane. All $U_i$ are on one side of a hyperplane if and only if their convex hall $H(U_1, \cdots, U_n)$ does not contain $\theta$. Thus the lemma. #

**Theorem 3:** Let $X_1, X_2, \cdots, X_n$ be iid copies of a continuous random vector $X$ in $\mathcal{R}^k$. For $k = 1, 2$ and any $n > k$,

$$a(k,n) \leq B_c,$$

(3.13)

where $B_c = P\{\theta \notin H(X_1, X_2, \cdots, X_n)\}$. Further, the equality holds if and only if the distribution of $X$ is symmetric with respect to $\theta$.

**Proof of Theorem 3:** The projection of a standard multivariate normal random vector on the unit sphere has a uniform distribution. To prove Theorem 3, by Lemma 2, we only need to show that for continuous distributions on the unit sphere, the probability that
the convex hall of a random sample of size $n$ does not contain $0$ is the smallest when the underlying distribution is uniform.

For $k = 1$, the uniform distribution on the sphere degenerates into a discrete uniform distribution on $\{-1, 1\}$. Let $p = P\{X_i = 1\}$. Theorem 3 amounts to the simple observation that function $p^n + (1-p)^n$ has its unique minimum at $p = 1/2$ and that this minimum is $a(1, n) = (1/2)^{n-1}$. For $k = 2$, let $X$ be a random variable on the unit circle ($0 \leq X < 2\pi$) and $f(x)$ be its density function. Define

$$G(x) = \int_{x}^{2\pi} f(x) dx,$$

where $f(x) = f(x + 2\pi)$. For $X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n$, denote the event that they are in the half-circle $(X_j, X_j + \pi)$ by $A_j$. If $X_j > \pi$, this half-circle represents the union of $(X_j, 2\pi)$ and $[0, X_j - \pi]$. Since $X_i$ are iid, we have for $j = 1, 2, \cdots, n$

$$P\{A_j\} = \int_{0}^{2\pi} f(x)[G(x)]^{n-1} dx.$$

Further, $A_i \cap A_j = \emptyset$ for $i \neq j$, where $\emptyset$ denotes the empty set, and

$$\{0 \notin H(X_1, X_2, \cdots, X_n)\} = \bigcup_{i=1}^{n} A_i.$$

It follows that for any $n \geq 1$,

$$B^c = \sum_{i=1}^{n} P\{A_i\} = n \int_{0}^{2\pi} f(x)[G(x)]^{n-1} dx. \quad (3.14)$$

An equivalent expression for $P\{A_j\}$ is

$$P\{A_j\} = \int_{0}^{2\pi} f(x + \pi)[G(x)]^{n-1} dx.$$

This leads to another expression for $B^c$:

$$B^c = \sum_{i=1}^{n} P\{A_i\} = n \int_{0}^{2\pi} f(x + \pi)[G(x)]^{n-1} dx. \quad (3.15)$$

Adding up (3.14) and (3.15) gives yet another expression for $B^c$:

$$B^c = n \int_{0}^{\pi} \frac{1}{2} \{f(x) + f(x + \pi)\}\{[G(x)]^{n-1} + [G(x + \pi)]^{n-1}\} dx. \quad (3.16)$$

To show that the equality in (3.13) holds if the distribution of $X$ is symmetric with respect to $0$, note that the distribution is symmetric if and only if $G(x) = 1/2$ for all $x \in [0, 2\pi)$. Thus (3.14) implies $B^c = n(1/2)^{n-1} = a(2, n)$ if $X$ is symmetric.
To show that the inequality in (3.13) holds strictly if the distribution of $X$ is not symmetric and thus it also must be symmetric if the equality holds, first note that $G(x), G(x + \pi) \geq 0$ and $G(x) + G(x + \pi) = 1$. Thus for $n \geq 1$

$$(1/2)^{n-2} \leq [G(x)]^{n-1} + [G(x + \pi)]^{n-1}. \quad (3.17)$$

If the distribution of $X$ is not symmetric, $G(x)$ cannot be $1/2$ for all $x \in [0, \pi)$. Further, $G(x)$ is continuously differentiable. There exists an open subinterval of $[0, \pi)$ in which $G(x) \neq 1/2$ and $G'(x) \neq 0$. Over this subinterval, $f(x) + f(x + \pi)$ is strictly positive and the inequality in (3.17) holds strictly. Multiply both sides of (3.17) by $f(x) + f(x + \pi)$ and then integrate from 0 to $\pi$. We have

$$(1/2)^{n-1} < \int_0^\pi \frac{1}{2} \{f(x) + f(x + \pi)\} \{[G(x)]^{n-1} + [G(x + \pi)]^{n-1}\} dx, \quad (3.18)$$

where the left-hand side is strictly smaller than the right-hand side because of the subinterval. It follows from (3.16) and (3.18) that $B^c > n(1/2)^{n-1} = a(2, n)$. #

Theorem 3 implies that for $k = 1, 2$, $b(n, k) = 1 - a(n, k)$ and the value of the atom $a(k, n)$ is the greatest lower bound on the atom of the distribution for $l(\theta_0)$. For $k \geq 3$, a proof of (3.13) has eluded us so far due to difficulties in finding an analytic expression for $B^c$ in higher dimensions. We conjecture that (3.13) holds for any $k$. The rest of our discussion assumes this conjecture holds. Expression (3.16) had previously appeared in Efron (1965).

**Theorem 4:** Let $[x]$ be the largest integer smaller than $x$. For any $n > k$,

(a) $b(k, n + 1) > b(k, n)$ and $b(k, n) > b(k + 1, n)$, and

(b) for any $\varepsilon \in (0, 0.5)$, $b([\varepsilon n], n) \to 1$ and $b([n - \varepsilon n], n) \to 0$ as $n \to \infty$.

**Proof of Theorem 4:** The inequalities in (a) are equivalent to $a(k, n + 1) < a(k, n)$ and $a(k, n) < a(k + 1, n)$, respectively. Their proof follows easily from (2.8).

To see (b) is true, consider the binomial random variable $X \sim \text{Bin}(1/2, n - 1)$. Denote by $Z$ the standard normal random variable. By (2.8), we have

$$a([\varepsilon n], n) = P\{X \leq [\varepsilon n] - 1\} \sim P\left\{Z \leq \frac{[\varepsilon n] - 1 - (n - 1)/2}{\sqrt{n/2}}\right\}.$$

It is easy to see that the probability on the right-hand side goes to zero when $n$ goes to infinity. Similarly, $a([n - \varepsilon n], n)$ goes to 1. These imply (b). #
4 Applications

[4.1] Discussion on the bounds

The bound on the coverage level provides a finite sample viewpoint for looking at the coverage property of the empirical likelihood ratio confidence region. It does not depend on the dimension of the data $Y_i$ and that of the parameter $\theta_0$. The asymptotic distribution of $l(\theta_0)$ is also not a factor. In this sense, the bound reflects a constraint on the coverage level inherent in the way empirical likelihood ratio confidence regions are constructed.

The bound $b(k,n)$ is a strictly decreasing function in $k$. This presents a dilemma for empirical likelihood inference; in the case where extra information on $\theta_0$ is available in the form of some unbiased estimating functions, should we use them regardless of $n$? Asymptotically, empirical likelihood makes efficient use of that extra information (Qin and Lawless, 1994). Since $b(k,\infty) = 1$ for any finite $k$, the bound is also of no concern. Nevertheless, at a fixed $n$, the atom of the $E$ distribution increases as $k$ increases, causing the bound to decrease and making high coverage levels impossible. Table 2 gives the atoms for small sample sizes at various dimensions. Looking across each row, we see the atom increases rapidly as $k$ increases. If we look at the number of free parameters in the empirical likelihood alone, up to $(n-2)$ constraints on $\theta_0$ can be handled. But at $k = n/2$, the bound is already 0.5. Owen (2001, p54) had suggested that $k$ should be much smaller than $n$. Theorem 4 allows us to make this statement more precise; asymptotically in order to avoid the restriction on the coverage level imposed by the bound, $k$ must be less than $\varepsilon n$ for some $\varepsilon \in (0, 0.5)$. Further research is needed to determine the benefit of incorporating more constraints at different finite sample sizes.

Apart from the estimating equations (3.9) considered in this paper, the empirical loglikelihood ratio $l(\theta_0)$ for other problems may also be positive infinity when the origin is not in the convex hall of a set of $n$ random vectors $X_i \in \mathcal{R}^k$, which may or may not be iid (Owen, 2001). Whenever $X_i$ are iid, the bound $b(k,n) = 1 - a(k,n)$ applies to the coverage level of the corresponding confidence region. When $X_i$ are not iid, $b(k,n) = 1 - a(k,n)$ is not the least upper bound if $X_i$ are allowed to be a set of arbitrary random vectors. Examples may be found where $B$ exceeds $b(k,n)$. Nevertheless, whether or not $b(k,n)$ is still an upper bound on the coverage level is not clear; in empirical likelihood, $X_i$ may not be iid but they are not arbitrary. In such cases, one may instead look for the least
upper bound on the coverage level for the class of non-iid problems of interest.

In two sample problems where there are two sets of iid random vectors, denote the dimension and sample size of each set by \((k_i, n_i)\) where \(i = 1, 2\). Suppose the loglikelihood ratio of the parameter of interest is infinity or undefined if the convex halls of the two sets of iid random vectors do not contain their respective origins at the same time. Then an upper bound on the coverage level is \(\min\{b(k_1, n_1), b(k_2, n_2)\}\). If the two sets of random vectors are also independent between themselves, then the upper bound may be reduced to \(b(k_1, n_1) \times b(k_2, n_2)\).

[4.2] The \(E\) calibration

The existence of an atom for the distribution of \(l(\theta_0)\) suggests that the accuracy of the continuous Chi-square approximation to the finite sample distribution of \(l(\theta_0)\) needs to be carefully examined, even when the atom is small. When the ratio \(n/k\) is not large, the upper percentiles of the distribution of \(l(\theta_0)\) tend to be substantially larger than the corresponding percentiles of the asymptotic Chi-square distribution, causing serious undercoverage for the Chi-square calibration. To illustrate this point, consider the distribution of \(l(\theta_0)\) where \(\theta_0\) is the parameter vector of a bivariate normal random vector \((X, Y)\); \(\theta_0 = (\mu_x, \mu_y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})\). Let \(\theta_0 = (0, 0, 10, 17, 7)\) and formulate \(R(\theta_0)\) through an estimating function which may be found on page 42 in Owen (2001). Table 1 gives the simulated critical values of the distribution of \(l(\theta_0)\) and the corresponding critical values of the \(E\) and asymptotic Chi-square distributions. Each critical value for the distribution of \(l(\theta_0)\) is estimated by using the corresponding percentile of a sample

Table 2: Values of the atoms of the \(E_{k,n}\) distribution.

<table>
<thead>
<tr>
<th>(n)</th>
<th>(k = 1)</th>
<th>(k = 2)</th>
<th>(k = 3)</th>
<th>(k = 4)</th>
<th>(k = 5)</th>
<th>(k = 6)</th>
<th>(k = 7)</th>
<th>(k = 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.031</td>
<td>0.188</td>
<td>0.500</td>
<td>0.813</td>
<td>0.969</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>7</td>
<td>0.016</td>
<td>0.109</td>
<td>0.344</td>
<td>0.656</td>
<td>0.891</td>
<td>0.984</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>0.008</td>
<td>0.063</td>
<td>0.227</td>
<td>0.500</td>
<td>0.773</td>
<td>0.938</td>
<td>0.992</td>
<td>N/A</td>
</tr>
<tr>
<td>9</td>
<td>0.004</td>
<td>0.035</td>
<td>0.145</td>
<td>0.363</td>
<td>0.637</td>
<td>0.855</td>
<td>0.965</td>
<td>0.996</td>
</tr>
<tr>
<td>10</td>
<td>0.002</td>
<td>0.020</td>
<td>0.090</td>
<td>0.254</td>
<td>0.500</td>
<td>0.746</td>
<td>0.910</td>
<td>0.980</td>
</tr>
<tr>
<td>11</td>
<td>0.001</td>
<td>0.011</td>
<td>0.055</td>
<td>0.172</td>
<td>0.377</td>
<td>0.623</td>
<td>0.828</td>
<td>0.945</td>
</tr>
<tr>
<td>12</td>
<td>0.000</td>
<td>0.006</td>
<td>0.033</td>
<td>0.113</td>
<td>0.274</td>
<td>0.500</td>
<td>0.726</td>
<td>0.887</td>
</tr>
</tbody>
</table>
Table 3: Comparison of critical values of the finite sample distribution of \( l(\theta_0) \), that of the corresponding \( E_{k,n} \) distributions and asymptotic \( \chi^2 \) distribution.

<table>
<thead>
<tr>
<th>n</th>
<th>Distribution</th>
<th>( \alpha = 0.50 )</th>
<th>0.40</th>
<th>0.30</th>
<th>0.25</th>
<th>0.20</th>
<th>0.15</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( l(\theta_0) )</td>
<td>16.21</td>
<td>31.38</td>
<td>+\infty</td>
<td>+\infty</td>
<td>+\infty</td>
<td>+\infty</td>
<td>+\infty</td>
<td>+\infty</td>
<td>+\infty</td>
</tr>
<tr>
<td>15</td>
<td>( E_{5,15} )</td>
<td>8.32</td>
<td>10.69</td>
<td>14.31</td>
<td>17.45</td>
<td>22.33</td>
<td>31.48</td>
<td>62.31</td>
<td>+\infty</td>
<td>+\infty</td>
</tr>
<tr>
<td>20</td>
<td>( l(\theta_0) )</td>
<td>9.50</td>
<td>12.89</td>
<td>18.59</td>
<td>23.78</td>
<td>31.82</td>
<td>50.70</td>
<td>+\infty</td>
<td>+\infty</td>
<td>+\infty</td>
</tr>
<tr>
<td></td>
<td>( E_{5,20} )</td>
<td>6.52</td>
<td>7.77</td>
<td>9.64</td>
<td>10.90</td>
<td>12.68</td>
<td>15.23</td>
<td>19.20</td>
<td>28.20</td>
<td>+\infty</td>
</tr>
<tr>
<td>30</td>
<td>( l(\theta_0) )</td>
<td>6.30</td>
<td>7.86</td>
<td>9.66</td>
<td>11.01</td>
<td>12.43</td>
<td>15.50</td>
<td>19.52</td>
<td>30.23</td>
<td>88.81</td>
</tr>
<tr>
<td></td>
<td>( E_{5,30} )</td>
<td>5.25</td>
<td>6.33</td>
<td>7.71</td>
<td>8.60</td>
<td>9.63</td>
<td>11.08</td>
<td>13.02</td>
<td>17.12</td>
<td>26.78</td>
</tr>
<tr>
<td>+\infty</td>
<td>( \chi^2_5 )</td>
<td>4.35</td>
<td>5.13</td>
<td>6.06</td>
<td>6.63</td>
<td>7.29</td>
<td>8.12</td>
<td>9.24</td>
<td>11.07</td>
<td>15.09</td>
</tr>
</tbody>
</table>

of size 10,000 from the distribution. A +\( \infty \) at, say \((n, \alpha) = (15, 0.30)\), indicates that at \( n = 15 \) the atom for the distribution of \( l(\theta_0) \) is more than 30%. At \( n = 30 \), although the atom for the distribution of \( l(\theta_0) \) is less than 1%, the upper percentiles are still much bigger than that of the \( \chi^2 \). If we compute a 95% confidence region for \( \theta_0 \) using the \( \chi^2 \) critical value, we would obtain a confidence region with a coverage level of about 75%.

For a reasonably large sample size \((n \geq 20)\), the bound on coverage level is very close to 1 when the dimension \( k \) is not high. This suggests that in practice all commonly used confidence levels can be achieved by properly calibrated empirical likelihood ratio confidence regions. For the above example, at \( n = 30 \), the atom of \( l(\theta_0) \) is small enough to admit an empirical likelihood ratio confidence region with a coverage level of more than 95%. Thus the serious undercoverage problem observed above is caused by the inaccurate asymptotic Chi-square calibration rather than an impossible confidence level, one that exceeds the bound on the coverage level. Various asymptotic methods have been proposed to improve the coverage level for the empirical likelihood ratio confidence regions. These include the \( F \) calibration and the bootstrap calibration (Owen, 2001) and the Bartlett correction. See Owen (2001) for detailed discussions.

In practice, when the sample size is not large, the \( t \) critical values have often been used with asymptotically normal estimators to construct confidence intervals instead of the \( z \) values. When the quality of the normal approximation is a concern, the use of the \( t \) values provides some protection against undercoverage. For empirical likelihood inference, the \( E \) distribution may be similarly used in place of the limiting Chi-square distribution.
Suppose \( l(\theta_0) \) has an asymptotic \( \chi^2_k \) distribution. For an \( \alpha > a(k, n) \), the 100(1 − \( \alpha \))% confidence region

\[
\{ \theta : l(\theta) \leq e_{\alpha, k, n} \},
\]

(4.19)
is the \( E \) calibrated confidence region for \( \theta_0 \). The \( E_{k,n} \) distribution converges to the \( \chi^2_k \) distribution. Thus an \( E \) calibrated confidence region is consistent in that its coverage level converges to the nominal level as the sample size \( n \) goes to infinity. For finite sample sizes, the \( E \) calibration may be substantially more accurate than the Chi-square calibration; if we use the \( E \) calibration to compute a 95% confidence region for \( \theta_0 \) in the above example at \( n = 30 \), the resulting coverage level would be about 87%.

Comparing to asymptotic methods of calibration, in particular, the Chi-square calibration, the \( E \) calibration has the following advantages:

1. **Respecting bounds:** Because of its recognition of the least upper bound through the atoms of the \( E \) distributions, the \( E \) calibration allows us to avoid going over the least upper bound when deciding on a nominal level.

2. **Better coverage level:** From the size of the atom standpoint, the \( E \) distributions are “between” the asymptotic Chi-square distributions and other finite sample distributions of the empirical loglikelihood ratios. They provide better approximations to the finite sample distributions than the asymptotic Chi-square distributions and are exact in the case of the normal mean.

   An \( E \) calibrated region gives consistently higher coverage probability than the corresponding Chi-square calibrated region due to the larger critical values of the \( E \) distributions, thereby consistently correcting the undercoverage problem of the Chi-square calibration. In small sample and high dimension situations, corrections provided by the \( E \) calibration can be more than 10%.

   In spite of the substantial corrections by the \( E \) calibration, we have found that the coverage level of the \( E \) calibrated region still tends to be smaller than the nominal level. In this sense, the \( E \) calibration represents the “minimum” amount of correction to the Chi-square calibration one should make without overcorrecting for the normal mean.

3. **Easy to implement:** When compared to the Bartlett correction and bootstrap calibration, the \( E \) calibration is very easy to implement. It does not involve more
computation or derivation than Chi-square calibration. One just looks up the $E$ distribution table for the desired critical value and computes the empirical likelihood ratio confidence region like using the Chi-square calibration.

Although always more accurate than the Chi-square calibration, the relative accuracy of the $E$ calibration to that of the estimated Bartlett correction, Bootstrap calibration and $F$ calibration depends on the parameter of interest and more importantly the sample size $n$ and dimension $k$. At small sample sizes ($n \leq 20$), the $E$ calibration is in general more accurate than the Bartlett correction and $F$ calibration, especially when the nominal level and/or $k$ are high. At moderately large sample sizes of 20 to 40, the $E$ calibration is comparable to the Bartlett correction. Its difference with the $F$ calibration also becomes small due to that the atom of the finite sample distribution of $l(\theta_0)$ becomes small, and the continuous $F$ approximation to the distribution becomes more accurate. The $E$ calibration and the bootstrap calibration are comparable in many cases, and they both are more accurate than other methods for small sample sizes. The good accuracy of the bootstrap calibration is in part due to the fact that it too takes into consideration the atom of $l(\theta_0)$ through that of the bootstrap distribution of $l(\hat{\theta}_0)$.

To illustrate the relative performance of the $E$ calibration and the Bartlett correction, we revisit examples considered by DiCiccio, Hall and Romano (1991). We only consider the $\chi^2$ case and the $t_5$ case since for the normal case $E$ calibration is exact. Table 3 contains the simulated coverage probabilities of $E$ calibrated confidence intervals at sample sizes used in their paper. Comparing these to that of the estimated and theoretical Bartlett-

<p>| Table 3: Simulated coverage probabilities of the $E$ distribution based intervals |
|---------------------------------|-----|-----|-----|-----|</p>
<table>
<thead>
<tr>
<th>Nominal level</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$ data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>0.7618</td>
<td>0.8599</td>
<td>0.9157</td>
<td>0.9701</td>
</tr>
<tr>
<td>$n = 40$</td>
<td>0.7815</td>
<td>0.8770</td>
<td>0.9355</td>
<td>0.9789</td>
</tr>
<tr>
<td>$t_5$ data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 15$</td>
<td>0.7827</td>
<td>0.8858</td>
<td>0.9395</td>
<td>0.9825</td>
</tr>
<tr>
<td>$n = 30$</td>
<td>0.7918</td>
<td>0.8922</td>
<td>0.9423</td>
<td>0.9876</td>
</tr>
</tbody>
</table>

corrected intervals in DiCiccio, Hall and Romano (1991, p1058), the $E$ calibrated interval is as accurate as the estimated Bartlett-corrected interval but is less accurate than the
theoretical Bartlett-corrected interval. Considering that the sample sizes used here are more to the advantage of the estimated Bartlett correction and $k = 1$, the accuracy of the $E$ calibration is quite competitive.

Note that in the above discussion on $E$ calibration, we have implicitly assumed that the dimension of the limiting Chi-square distribution of $l(\theta_0)$, $k$, is the same as that of the estimating function $m(Y, \theta)$. In practice, this may not be the case. In order to be consistent, the $E$ calibration as formulated in (4.19) calibrates $l(\theta_0)$ by using the $E_{k,n}$ distribution. It has all the advantages stated above when the dimension of $m(Y, \theta)$, which we now denote by $d$, equals $k$. In cases where $k < d$, the $E$ calibration (4.19) respects the bound on coverage level $b(k, n)$, which by Theorem 4 may be higher than the least upper bound $b(d, n)$. In such cases, one needs to check the least upper bound $b(d, n)$ to ensure that the nominal level is not higher. In two sample problems that we had considered at the end of our discussion on bounds, suppose $k = k_1$. The $E$ calibration uses $E_{k_1,n_1}$ to calibrate the loglikelihood ratio. It respects the upper bound $b(k_1, n_1)$ which may be higher than $\min\{b(k_1, n_1), b(k_2, n_2)\}$ or $b(k_1, n_1) \times b(k_2, n_2)$. The latter bounds should also be checked against the nominal level to ensure it is possible.

For cases where $l(\theta_0)$ is infinity when the origin is not in the convex hull of some non-iid observations, a non-trivial upper bound on the coverage level may be difficult to obtain. Nevertheless, numerical evidence suggest that the undercoverage problem of the Chi-square calibration affects such cases as well. The $E$ calibration (4.19) remains consistent. It may be used for such cases since its use, like the use of the $t$ values instead of the asymptotic $z$ values, provides some protection against the undercoverage of the asymptotic Chi-square calibration at no extra cost. Finally, the Chi-square calibration is based entirely on asymptotic considerations whereas the $E$ calibration is based on asymptotic as well as finite sample considerations. From a philosophical standpoint, the $E$ and the Chi-square calibrations do not have any fundamental difference in that they all amount to approximating the unknown finite sample distributions of $l(\theta_0)$ with some known distributions. In this perspective, the preffered method is the one that utilizes the more accurate approximation to the finite sample distribution of $l(\theta_0)$. In small sample situations, the $E$ distributions provide more accurate approximation for many different formulations of $l(\theta_0)$, partly due to its ability to accommodate an atom at the infinity. Thus $E$ calibration may be used for a general $l(\theta_0)$ in small sample situations.
Acknowledgments

References


