Two-sample extended empirical likelihood

Fan Wu[1] and Min Tsao

Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada V8W 3R4

Abstract

The two-sample empirical likelihood is Bartlett correctable (Jing, 1995). We extend it to the full parameter space and obtain a two-sample extended empirical likelihood which is more accurate than the original and can also achieve the second-order accuracy of the Bartlett correction.

AMS 2000 subject classifications: Primary 62G20; secondary 62E20.

Keywords: Two-sample empirical likelihood; Extended empirical likelihood; Bartlett correction; Composite similarity mapping.

1. INTRODUCTION

The empirical likelihood introduced by Owen (1988, 1990) is a versatile non-parametric method of inference with many applications (Owen, 2001). One problem which the empirical likelihood method has been successfully applied to is the two-sample problem (Jing, 1995; Wu and Yan, 2012) where the parameter of interest θ is the difference between the means of two populations. The well-known Behrens-Fisher problem is a special two-sample problem where the two populations are known to be normally distributed. Following DiCiccio, Hall and Romano (1991) who showed the surprising result that the (one-sample) empirical likelihood for a smooth function of the mean is Bartlett correctable, Jing (1995) proved that the two-sample empirical likelihood for θ is also Bartlett correctable. The coverage error of a confidence region based on the original empirical likelihood is $O(n^{-1})$, but that of the Bartlett corrected empirical likelihood is only $O(n^{-2})$.

[1] Corresponding author; email address: fwu@uvic.ca

For a one-sample empirical likelihood, there is a mismatch between its domain and the parameter space in that it is defined on only a part of the parameter space. This mismatch is a main cause of the undercoverage problem associated with empirical likelihood confidence regions (Tsao, 2013). The two-sample empirical likelihood for θ also has the mismatch problem as it is defined on a bounded interval but the parameter space is $\mathbb{R}$. In this paper, we derive an extended version of the original two-sample empirical likelihood (OEL) by mapping its domain onto $\mathbb{R}$ through the composite similarity mapping of Tsao and Wu (2013). The resulting two-sample extended empirical likelihood (EEL) for θ is defined on the entire $\mathbb{R}$ and hence free from the mismatch problem. Under mild conditions, this EEL has the same asymptotic properties as the OEL. It can also attain the second order accuracy of the two-sample Bartlett corrected empirical likelihood (BEL) of Jing (1995). The first order version of the EEL is more accurate than the OEL. It is also easy to compute and surprisingly more accurate than the second order BEL. We recommend it for two-sample empirical likelihood inference.

2. Two-sample empirical likelihood

Let $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_m\}$ be two independent random samples from two unknown distributions F and G, respectively. Let $\mu_x = E(X_i)$ and $\mu_y = E(Y_j)$. The unknown parameter of interest is the difference in mean $\theta_0 = \mu_y - \mu_x$ and the parameter space is $\mathbb{R}$. Denote by $p = (p_1, ..., p_m)$ and $q = (q_1, ..., q_n)$ two probability vectors satisfying $p_i \geq 0$, $q_j \geq 0$, $\sum_{i=1}^m p_i = 1$ and $\sum_{i=1}^n q_j = 1$. Let $\mu_x(p) = \sum_{i=1}^m p_i X_i$ and $\mu_y(q) = \sum_{j=1}^n q_j Y_j$, and denote by $\theta(p,q)$ their difference, more precisely, $\theta(p,q) = \mu_y(q) - \mu_x(p)$. For a $\theta \in \mathbb{R}$, Jing (1995) defined the two-sample empirical likelihood $L(\theta)$ as

$$L(\theta) = \max_{(p,q):\theta(p,q)=\theta} \left(\prod_{i=1}^m p_i \right) \left(\prod_{j=1}^n q_i \right). \tag{1}$$

The corresponding two-sample empirical log-likelihood ratio for θ is thus

$$l(\theta) = -2 \max_{(p,q):\theta(p,q)=\theta} \left(\sum_{i=1}^m \log(mp_i) + \sum_{j=1}^n \log(nq_i) \right). \tag{2}$$

To differentiate between the $l(\theta)$ in (2) and the extended version of $l(\theta)$ in the next section, we will refer to the $l(\theta)$ in (2) as the original two-sample empirical log-likelihood ratio or simply "OEL $l(\theta)$".

Let $N = m + n$, $f_m = N/m$ and $f_n = N/n$. Without loss of generality, assume that $m \geq n$. By the method of Lagrangian multipliers, we have

$$l(\theta) = 2 \left[\sum_{i=1}^{m} \log\{1 - f_m \lambda(X_i - \mu_x)\} + \sum_{j=1}^{n} \log\{1 + f_n \lambda(Y_j - \mu_y)\} \right] \tag{3}$$

where the multiplier $\lambda = \lambda(\theta)$ satisfies

$$\sum_{i=1}^{m} \frac{X_i - \mu_x}{1 - f_m \lambda(X_i - \mu_x)} = 0 \quad \text{and} \quad \sum_{j=1}^{n} \frac{Y_j - \mu_y}{1 + f_n \lambda(Y_j - \mu_y)} = 0, \tag{4}$$

and

$$\sum_{j=1}^{n} \frac{Y_j}{1 + f_n \lambda(Y_j - \mu_y)} - \sum_{i=1}^{m} \frac{X_i}{1 - f_m \lambda(X_i - \mu_x)} = \theta. \tag{5}$$

Under the assumption that F and G have finite variances, Jing (1995) showed that

$$l(\theta_0) \xrightarrow{D} \chi_1^2 \quad \text{as} \ n \to +\infty. \tag{6}$$

Hence, the $100(1-\alpha)\%$ OEL confidence interval for θ_0 is

$$\mathcal{C}_{1-\alpha} = \{\theta : \theta \in \mathbb{R} \text{ and } l(\theta) \leq c_\alpha\} \tag{7}$$

where c_α is $(1-\alpha)$th quantile of the χ_1^2 distribution. The coverage error of $\mathcal{C}_{1-\alpha}$ is $O(n^{-1})$, that is

$$P(\theta_0 \in \mathcal{C}_{1-\alpha}) = P(l(\theta_0) \leq c_\alpha) = 1 - \alpha + O(n^{-1}). \tag{8}$$

Under a stronger assumption that F and G have finite fourth moments, Jing (1995) also showed the OEL $l(\theta_0)$ is Bartlett correctable, that is

$$P(\theta_0 \in \mathcal{C}'_{1-\alpha}) = P(l(\theta_0) \leq c_\alpha(1 + \eta N^{-1})) = \alpha + O(n^{-2}) \tag{9}$$

where $\mathcal{C}'_{1-\alpha} = \{\theta : l(\theta) \leq c_\alpha(1 + \eta N^{-1})\}$ is the Bartlett corrected empirical likelihood (BEL) confidence interval and η is the Bartlett correction constant in Theorem 2 of Jing (1995). Detailed proofs for (6) and (9) are in a supplement to Jing (1995) available from Professor Bing-Yi Jing. See also Wu and Yan (2012) and Qin (1994) about two-sample empirical likelihood methods.

3. Two-sample extended empirical likelihood

The two-sample OEL $l(\theta)$ also suffers from the mismatch problem between its domain and the parameter space. To see this, since $\min\{X_i\} \leq \mu_x(p) \leq \max\{X_i\}$ and $\min\{Y_j\} \leq \mu_y(q) \leq \max\{Y_j\}$, we have

$$\min\{Y_j\} - \max\{X_i\} \leq \theta(p, q) \leq \max\{Y_j\} - \min\{X_i\}.$$

It follows that the domain of the OEL $l(\theta)$, $\Theta_n = \{\theta : l(\theta) < +\infty\}$, is

$$\Theta_n = (\min\{Y_j\} - \max\{X_i\}, \max\{Y_j\} - \min\{X_i\}).$$

Since the parameter space is $\mathbb{R}$, the mismatch can be expressed as $\Theta_n \subset \mathbb{R}$ which is a main cause of the undercoverage problem of empirical likelihood confidence regions (Tsao, 2013; Tsao and Wu, 2013). To overcome the mismatch, we extend the OEL $l(\theta)$ by expanding its domain to the entire $\mathbb{R}$.

For simplicity, we also assume that $m \geq n$ and further that $m/n = O(1)$ so that $O(n^{-1})$, $O(m^{-1})$ and $O(N^{-1})$, for example, are all interchangeable. A point estimator for θ_0 is $\hat{\theta} = \bar{Y} - \bar{X}$ where $\bar{X} = m^{-1} \sum X_i$ and $\bar{Y} = n^{-1} \sum Y_j$ are the sample means. It is easy to verify that $\hat{\theta}$ is the maximum empirical likelihood estimator for θ_0. Following Tsao and Wu (2013), we define the composite similarity mapping $h_N^C : \Theta_n \to \mathbb{R}$ centred on $\hat{\theta}$ as

$$h_N^C(\theta) = \hat{\theta} + \gamma(N, l(\theta))(\theta - \hat{\theta}) \tag{10}$$

where function $\gamma(n, l(\theta))$ is the expansion factor given by

$$\gamma(N, l(\theta)) = 1 + \frac{l(\theta)}{2N}. \tag{11}$$

Theorem 1 below gives two key properties of mapping h_N^C.

Theorem 1. *Suppose the variances of F and G are finite and positive. Then, $h_N^C : \Theta_n \to \mathbb{R}$ defined by (10) and (11) satisfies (i) it has a unique fixed point at $\hat{\theta}$ and (ii) it is a bijective mapping from Θ_n to $\mathbb{R}$.*

Since $h_N^C : \Theta_n \to \mathbb{R}$ is bijective, it has an inverse function which we denote by $h_N^{-C}(\theta) : \mathbb{R} \to \Theta_n$. For any $\theta \in \mathbb{R}$, let $\theta' = h_N^{-C}(\theta) \in \Theta_n$. The extended empirical log-likelihood ratio EEL $l^*(\theta)$ is given by

$$l^*(\theta) = l(h_N^{-C}(\theta)) = l(\theta_0'), \tag{12}$$

which is defined for θ values throughout $\mathbb{R}$. Hence the EEL $l^*(\theta)$ is free from the mismatch problem of the OEL $l(\theta)$. Theorem 2 shows that EEL $l^*(\theta_0)$ has the same asymptotic chi-square distribution as the OEL $l(\theta_0)$.

Theorem 2. *Suppose the variances of F and G are finite and positive. Then, the EEL $l^*(\theta_0)$ defined by (12) satisfies*

$$l^*(\theta_0) \xrightarrow{D} \chi_1^2 \quad \text{as} \;\; n \to +\infty. \tag{13}$$

By Theorem 2, the $100(1-\alpha)\%$ EEL confidence interval for θ_0 is

$$\mathcal{C}^*_{1-\alpha} = \{\theta : \theta \in \mathbb{R} \text{ and } l^*(\theta) \leq c_\alpha\}, \tag{14}$$

which has a coverage error of $O(n^{-1})$. The expansion factor in (11) is a convenient choice which also gives good numerical results. There are, however, other choices available under which Theorems 1 and 2 also hold. This provides an opportunity to optimize the choice of expansion factor to obtain the second order accuracy. Theorem 3 below gives such an optimal choice.

Theorem 3. *Suppose F and G have finite and positive fourth moments. Let $l_2^*(\theta)$ be the EEL defined by the composite similarity mapping (10) with the following expansion factor*

$$\gamma_2(N, l(\theta)) = 1 + \frac{\eta}{4N}\left[2l(\theta)\right]^{\delta(n)} \tag{15}$$

where $\delta(n) = O(n^{-1/2})$ and η is the Bartlett correction factor for two-sample problem in (9). Then, we have

$$l_2^*(\theta_0) = l(\theta_0)\left[1 - \eta/N + O_p(n^{-3/2})\right] \tag{16}$$

and

$$P(l_2^*(\theta_0) \leq c) = P(\chi_1^2 \leq c) + O(n^{-2}) \tag{17}$$

Replacing EEL $l^*(\theta)$ in (14) with $l_2^*(\theta)$ gives an EEL confidence interval which, by (17), has a coverage error of $O(n^{-2})$. Because of this, we call $l_2^*(\theta)$ the *second order* EEL or EEL$_2$. Correspondingly, we call the EEL $l^*(\theta)$ defined by expansion factor (11) the *first order* EEL or EEL$_1$.

There are two lemmas that are needed for the proofs of Theorems 1, 2 and 3. In order to limit the length of the paper, we have not included them here. The lemmas, their proofs and the proofs of the theorems are given in the appendix of a technical report available on request from the authors.

4. Numerical examples

We now compare the coverage accuracy of 95% confidence intervals based on the OEL, BEL and EEL with two numerical examples. Comparisons based on 90% and 99% confidence intervals give similar conclusions and they can be found in the appendix in the aforementioned technical report.

In the first example, F and G are both standard normal distributions $N(0, 1)$. In the second example, they are χ_1^2 and $N(0, 1)$, respectively. Simulated coverage probabilities for these examples are given in Tables 1 and 2, respectively. Each simulated probability in the tables is based on 10,000 pairs of random samples whose sizes are indicated by the row and column headings, respectively. The BEL and EEL$_2$ were computed by using the estimated Bartlett correction factor from page 317 in Jing (1995). We summarize the tables with the following observations: (1) EEL$_1$ is consistently more accurate than OEL. Surprisingly, it is also more accurate than the second order BEL and EEL$_2$ for small and moderate sample sizes ($n, m \leq 20$) and competitive in accuracy when sample sizes are larger. (2) EEL$_2$ is more accurate than OEL and BEL for small and moderate sample size. It is comparable to BEL when one or both sample sizes are large.

To conclude, the EEL$_1$ is easy-to-compute and is the most accurate overall. Hence, we recommend EEL$_1$ for two-sample problems.

Table 1: Coverage probabilities of OEL, EEL$_1$, BEL & EEL$_2$ confidence intervals: both F and G are $N(0,1)$

		n=10	n=20	n=30	n=40
m=10	OEL	92.2	92.2	92.1	92.0
	EEL$_1$	94.5	93.6	93.3	93.0
	BEL	92.9	92.7	92.8	92.8
	EEL$_2$	93.4	93.2	93.3	93.4
m=20	OEL	92.2	93.8	94.3	94.5
	EEL$_1$	93.8	95.0	95.1	95.1
	BEL	92.8	94.3	94.6	94.8
	EEL$_2$	93.3	94.4	94.7	94.9
m=30	OEL	92.0	93.8	94.6	94.4
	EEL$_1$	93.4	94.7	95.3	95.2
	BEL	92.9	94.2	94.8	94.7
	EEL$_2$	93.5	94.2	94.8	94.7
m=40	OEL	92.4	94.2	94.7	94.6
	EEL$_1$	93.2	95.0	95.4	95.2
	BEL	93.1	94.6	95.0	94.9
	EEL$_2$	93.6	94.7	95.0	94.8

Table 2: Coverage probabilities of OEL, EEL$_1$, BEL & EEL$_2$ confidence intervals: F is χ_1^2 and G is $N(0,1)$

		n=10	n=20	n=30	n=40
m=10	OEL	89.4	91.5	91.1	91.4
	EEL$_1$	92.2	93.3	92.4	92.7
	BEL	90.7	92.5	92.0	92.6
	EEL$_2$	91.7	93.3	92.9	93.4
m=20	OEL	89.7	92.2	92.7	93.5
	EEL$_1$	91.7	93.5	93.7	94.1
	BEL	90.8	93.0	93.4	93.9
	EEL$_2$	91.7	93.3	93.6	94.1
m=30	OEL	89.2	92.6	93.3	93.4
	EEL$_1$	90.5	93.5	94.0	94.0
	BEL	90.1	93.3	93.8	93.8
	EEL$_2$	90.8	93.4	93.8	93.8
m=40	OEL	88.6	92.4	93.2	93.7
	EEL$_1$	89.7	93.1	93.9	94.2
	BEL	89.4	93.0	93.8	94.0
	EEL$_2$	90.1	93.2	93.8	94.0

References

References

[1] DiCiccio, T. J., Hall, P. and Romano, J. P. (1991). Empirical likelihood is Bartlett Correctable. *Ann. Statist.* **19**, 1053–1061.

[2] Jing, Bing-Yi (1995). Two-sample empirical likelihood method. *Statistics & Probability Letters.* **24**, 315–319.

[3] Owen, A. B. (1988). Empirical likelihood ratio confidence regions for single functional. *Biometrika.* **75**, 237-249.

[4] Owen, A. B. (1990). Empirical likelihood confidence regions. *Ann. Statist.* **18**, 90–120.

[5] Qin, J. (1994). Semi-parametric likelihood ratio confidence intervals for the difference of two sample means. *Ann. Insti. Statist. Math.* **46**, 117–126.

[6] Owen, A. B. (2001). *Empirical Likelihood.* Chapman & Hall/CRC, London.

[7] Tsao, M. (2013). Extending the empirical likelihood by domain expansion. *Canadian J. Statist.* **41**, 257–274.

[8] Tsao, M. & Wu, F. (2013). Empirical likelihood on the full parameter space. Accepted for publication by the *Annals of Statistics.*

[9] Wu, C. & Yan, Y. (2012). Empirical likelihood inference for two-sample problems. *Stat. Interface.* **5**, 345–354.