Statistics and Probability Letters 84 (2014) 81–87

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Two-sample extended empirical likelihood

Fan Wu*, Min Tsao

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

ARTICLE INFO

Article history:
Received 23 July 2013
Received in revised form 5 September 2013
Accepted 16 September 2013
Available online 27 September 2013

MSC:
primary 62G20
secondary 62E20

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

ABSTRACT

Jing (1995) and Liu et al. (2008) studied the two-sample empirical likelihood and showed that it is Bartlett correctable for the univariate and multivariate cases, respectively. We expand its domain to the full parameter space, and obtain a two-sample extended empirical likelihood which is more accurate and can also achieve the second-order accuracy of the Bartlett correction.

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; Liu et al., 2008; Liu and Yu, 2010; 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 in which the two populations are known to be normally distributed. Following DiCiccio et al. (1991), who showed the surprising result that the (one-sample) empirical likelihood for a smooth function of the mean is Bartlett correctable, Jing (1995) and Liu et al. (2008) 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 based on the Bartlett corrected empirical likelihood is only $O(n^{-2})$.

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 region, but the parameter space is $\mathbb{R}^d$. In this paper, we derive an extended version of the original two-sample empirical likelihood (OEL) by expanding its domain into $\mathbb{R}^d$ 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 $\mathbb{R}^d$, and hence is 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) and Liu et al. (2008). The first-order version of this EEL is substantially more accurate than the OEL, especially for small sample sizes. It is also easy to compute and competitive in accuracy to the second-order methods. We recommend it for two-sample empirical likelihood inference.

* Corresponding author.
E-mail address: fwu@uvic.ca (F. Wu).

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.09.016

2. Two-sample empirical likelihood

Let $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ be independent copies of random vectors $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$, respectively. Denote by μ_x and Σ_x the mean and covariance matrix of X, and by μ_y and Σ_y the mean and covariance matrix of Y, respectively. The unknown parameter of interest is the difference in means $\theta_0 = \mu_y - \mu_x \in \mathbb{R}^d$, and the parameter space is the entire $\mathbb{R}^d$. We will need the following three conditions later in the paper.

(C1) Σ_x and Σ_y are finite covariance matrices with full rank d.

(C2) $\limsup_{\|t\|\to\infty} |E[\exp\{it^T X\}]| < 1$ and $\limsup_{\|t\|\to\infty} |E[\exp\{it^T Y\}]| < 1$.

(C3) $E\|X\|^{15} < +\infty$ and $E\|Y\|^{15} < +\infty$.

Condition (C1) is needed to establish the first-order result for the EEL, and conditions (C2) and (C3) are needed for the second-order result. Denote by $p = (p_1, \ldots, p_m)$ and $q = (q_1, \ldots, q_n)$ 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; that is,

$$\theta(p, q) = \mu_y(q) - \mu_x(p).$$

The original two-sample empirical likelihood for a $\theta \in \mathbb{R}^d$, $L(\theta)$, is defined as

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

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

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

In order to develop our extended empirical likelihood, it is important to first investigate the domains of the original empirical likelihood ratio $L(\theta)$ and log-likelihood ratio $l(\theta)$. The domain of $L(\theta)$ is given by

$$D_\theta = \left\{\theta \in \mathbb{R}^d : \text{there exist } p \text{ and } q \text{ such that } \mu_x(p) = \sum_{i=1}^m p_i X_i,\right.$$
$$\left.\mu_y(q) = \sum_{j=1}^n q_j Y_j \text{ and } \theta = \theta(p, q) = \mu_y(q) - \mu_x(p)\right\}.$$

Since the "range" of $\mu_x(p)$ and $\mu_y(q)$ is the convex hull of the X_i and Y_i, respectively, D_θ is a bounded, closed and connected region in $\mathbb{R}^d$ without voids. Detailed discussions about this and other geometric properties of D_θ may be found in the proof of Lemma 1. One of these properties is that θ is an interior point of D_θ if and only if it can be expressed as $\theta = \theta(p, q) = \mu_y(q) - \mu_x(p)$ for some p and q with straightly positive elements. Correspondingly, a boundary point of D_θ can only be expressed as $\theta(p, q) = \mu_y(q) - \mu_x(p)$ where one or more elements of p and q are zero. This implies that $L(\theta) = 0$ if θ is a boundary point of D_θ and $L(\theta) > 0$ if θ is an interior point of D_θ. We define the domain of the empirical log-likelihood ratio $l(\theta)$ as

$$\Theta_n = \{\theta : \theta \in D_\theta \text{ and } l(\theta) < +\infty\},$$

which excludes the boundary points of D_θ. 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 $l(\theta)$". The extended version will be referred to as the "EEL $l^*(\theta)$".

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

$$l(\theta_0) = 2\left[\sum_{i=1}^m \log\{1 - f_m\lambda^T(X_i - \mu_x)\} + \sum_{j=1}^n \log\{1 + f_n\lambda^T(Y_j - \mu_y)\}\right], \tag{3}$$

where the multiplier $\lambda = \lambda(\theta_0)$ satisfies

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

and

$$\sum_{j=1}^n \frac{Y_j}{1 + f_n\lambda^T(Y_j - \mu_y)} - \sum_{i=1}^m \frac{X_i}{1 - f_m\lambda^T(X_i - \mu_x)} = \theta_0. \tag{5}$$