Journal of Multivariate Analysis 142 (2015) 1–15

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Two-sample extended empirical likelihood for estimating equations

Min Tsao *, Fan Wu

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

ARTICLE INFO

Article history:
Received 13 February 2015
Available online 5 August 2015

AMS 2000 subject classifications:
primary 62G20
secondary 62E20

Keywords:
Bartlett correction
Composite similarity transformation
Extended empirical likelihood
Estimating equation
Similarity transformation
Two-sample empirical likelihood

ABSTRACT

We propose a two-sample extended empirical likelihood for inference on the difference between two p-dimensional parameters defined by estimating equations. The standard two-sample empirical likelihood for the difference is Bartlett correctable but its domain is a bounded subset of the parameter space. We expand its domain through a composite similarity transformation to derive the two-sample extended empirical likelihood which is defined on the full parameter space. The extended empirical likelihood has the same asymptotic distribution as the standard one and can also achieve the second-order accuracy of the Bartlett correction. We include two applications to illustrate the use of two-sample empirical likelihood methods and to demonstrate the superior coverage accuracy of the extended empirical likelihood confidence regions.

1. Introduction

A two-sample problem is concerned with making inference for the difference between the corresponding parameters of two populations/models with two independent samples. The difference between two population means is a special case that has been extensively studied; when the sample sizes are not large and underlying distributions are normal, methods for the Behrens–Fisher problem or a two-sample t method can be used; when the sample sizes are large, non-parametric z based procedures can be used. Recently, the empirical likelihood method [15] has been successfully applied to this special case. See [9,12,11,23,24]. These empirical likelihood methods complement existing methods as they do not require strong conditions and are more accurate than normal approximation based methods when the underlying distributions are skewed. In particular, the extended two-sample empirical likelihood for the difference between two p-dimensional means [21] is defined on the whole of $\mathbb{R}^p$ and is more accurate than other empirical likelihood methods.

In this paper, we study empirical likelihood methods for the general two-sample problem concerning the difference between two p-dimensional parameters defined by general estimating equations. The main contribution of this paper is a new extended empirical likelihood for such a difference, which generalizes results of Tsao [20] and Tsao and Wu [21] to this two-sample problem. The empirical likelihood method was introduced by Owen [13,14]. It has since been applied to many problems in statistics; see [15] and references therein. In particular, Qin and Lawless [17] showed that the empirical likelihood is effective for inference on parameters defined by estimating equations. DiCiccio et al. [5] and Chen and Cui [2] proved that the empirical likelihood for estimating equations is Bartlett correctable; the Bartlett corrected empirical likelihood enjoys the second-order accuracy. Although there have been relatively few publications that apply empirical likelihood to the

* Corresponding author.
E-mail address: mtsao@uvic.ca (M. Tsao).

http://dx.doi.org/10.1016/j.jmva.2015.07.009
0047-259X/

general two-sample problem, it is well-suited for this problem as the formulation of the one-sample empirical likelihood for estimating equations can be readily extended to handle the two-sample case; see, e.g., [9,19,12,11,24,25]. In particular, Qin and Zhao [19] studied the standard two-sample empirical likelihood for the univariate version ($p = 1$) of the problem, and Zi et al. [25] considered the special case where the parameters are the coefficient-vectors of two linear models.

In Section 2, we study the standard two-sample empirical likelihood for estimating equations in the general multi-dimensional sitting where $p \geq 1$. Like its one-sample counterpart, this two-sample empirical likelihood also has an asymptotic chi-square distribution and is Bartlett correctable. Adopting the terminology in [21], we refer to this standard two-sample empirical likelihood as the two-sample original empirical likelihood (OEL) for estimating equations. The OEL suffers from a mismatch problem [21] in that it is only defined on a part of the parameter space. This problem affects the coverage accuracy of the OEL based confidence regions. To overcome this, in Section 3 we introduce a two-sample extended empirical likelihood (EEL) that is defined on the whole parameter space. The EEL is obtained by expanding the domain of the OEL to the full parameter space through a composite similarity mapping. We show that the EEL has the same asymptotic chi-square distribution as the OEL and that it can also achieve the second-order accuracy of the Bartlett correction. In Section 4, we discuss two applications of the two-sample OEL and EEL. The first application is concerned with the inference for the difference between two Gini indices, and the second application is concerned with that between coefficient vectors of two regression models. We also make use of these applications to compare the numerical accuracy of the OEL and EEL confidence regions and to illustrate the superior accuracy of the EEL.

Proofs of theoretical results on two-sample OEL and EEL are all relegated to the Appendix. Note that some of these results can be proved by slightly modifying the proofs of similar results for other empirical likelihoods in the literature. For brevity, we will not include detailed proofs for such results in the Appendix but will give relevant references containing similar proofs.

2. Two-sample original empirical likelihood (OEL) for estimating equations

We first describe the general two-sample problem for estimating equations as follows. Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$ be two random vectors with unknown parameters $\theta_{x_0} \in \mathbb{R}^p$ and $\theta_{y_0} \in \mathbb{R}^p$, respectively. Let $g(X, \theta_x)$ and $g(Y, \theta_y)$ be two q-dimensional estimating functions for θ_{x_0} and θ_{y_0} satisfying $E\{g(X, \theta_{x_0})\} = 0$ and $E\{g(Y, \theta_{y_0})\} = 0$, respectively. The unknown parameter of interest is the difference $\pi_0 = \theta_{y_0} - \theta_{x_0} \in \mathbb{R}^p$ and the parameter space is the entire $\mathbb{R}^p$. A more general version of this problem allows the estimating function for θ_{x_0} to be different from that for θ_{y_0}. For simplicity, we consider only the common case where the two estimating functions are the same. We assume that $\{X_1, \ldots, X_m\}$ are independent copies of X, $\{Y_1, \ldots, Y_n\}$ are independent copies of Y, and X_i and Y_j are independent.

We now generalize the one-sample OEL for estimating equations [17] to obtain a two-sample OEL for π_0 and study its asymptotic properties. We will need the following four conditions on $g(X, \theta_x)$ and $g(Y, \theta_y)$.

Condition 1. $E\{g(X, \theta_{x_0})\} = 0$ and $E\{g(Y, \theta_{y_0})\} = 0$, and $\text{var}\{g(X, \theta_{x_0})\} \in \mathbb{R}^{q\times q}$ and $\text{var}\{g(Y, \theta_{y_0})\} \in \mathbb{R}^{q\times q}$ are both positive definite.

Condition 2. $\partial g(X, \theta_x)/\partial\theta_x$ and $\partial g^2(X, \theta_x)/\partial\theta_x\partial\theta_x^T$ are continuous in θ_x, and for θ_x in a neighbourhood of θ_{x_0} they are each bounded in norm by an integrable function of X.

Condition 3. $\partial g(Y, \theta_y)/\partial\theta_y$ and $\partial g^2(Y, \theta_y)/\partial\theta_y\partial\theta_y^T$ are continuous in θ_y, and for θ_y in a neighbourhood of θ_{y_0} they are each bounded in norm by an integrable function of Y.

Condition 4. $\limsup_{\|t\|\to\infty} |E[\exp\{it^T g(X, \theta_x)\}]| < 1$ and $E\|g(X, \theta_x)\|^{15} < +\infty$; $\limsup_{\|t\|\to\infty} |E[\exp\{it^T g(Y, \theta_y)\}]| < 1$ and $E\|g(Y, \theta_y)\|^{15} < +\infty$.

Denote by $\bar{p} = (p_1, \ldots, p_m)$ and $\bar{q} = (q_1, \ldots, q_n)$ 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 θ_y and θ_x be points in $\mathbb{R}^p$ and denote by $\theta_y(\bar{q})$ and $\theta_x(\bar{p})$ values that satisfy

$$\sum_{i=1}^m p_i g(X_i, \theta_x(\bar{p})) = 0, \qquad \sum_{j=1}^n p_j g(Y_j, \theta_y(\bar{q})) = 0.$$

Let $\pi = \theta_y - \theta_x \in \mathbb{R}^p$ and let $\pi(\bar{p}, \bar{q}) = \theta_y(\bar{q}) - \theta_x(\bar{p})$. Then, the two-sample OEL for a possible value of the difference π, $L(\pi)$, is defined as

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

which is the maximum of the product of the one-sample OEL for θ_y and the one-sample OEL for θ_x taken over all pairs (θ_x, θ_y) that satisfies $\pi = \theta_y - \theta_x$. The corresponding two-sample empirical log-likelihood ratio for π is thus

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