Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Evaluating the relative merits of competing models based on empirical likelihood ratio test

Yan Fan & Yukun Liu

To cite this article: Yan Fan & Yukun Liu (2016) Evaluating the relative merits of competing models based on empirical likelihood ratio test, Journal of Applied Statistics, 43:14, 2595-2607, DOI: 10.1080/02664763.2016.1142944

To link to this article: http://dx.doi.org/10.1080/02664763.2016.1142944

JOURNAL OF APPLIED STATISTICS, 2016
VOL. 43, NO. 14, 2595–2607
http://dx.doi.org/10.1080/02664763.2016.1142944

Taylor & Francis
Taylor & Francis Group

Evaluating the relative merits of competing models based on empirical likelihood ratio test

Yan Fan[a] and Yukun Liu[b]

[a]School of Business Information, Shanghai University of International Business and Economics, Shanghai, People's Republic of China; [b]School of Statistics, East China Normal University, Shanghai, People's Republic of China

ABSTRACT

Competing models arise naturally in many research fields, such as survival analysis and economics, when the same phenomenon of interest is explained by different researcher using different theories or according to different experiences. The model selection problem is therefore remarkably important because of its great importance to the subsequent inference; Inference under a misspecified or inappropriate model will be risky. Existing model selection tests such as Vuong's tests [26] and Shi's non-degenerate tests [21] suffer from the variance estimation and the departure of the normality of the likelihood ratios. To circumvent these dilemmas, we propose in this paper an empirical likelihood ratio (ELR) tests for model selection. Following Shi [21], a bias correction method is proposed for the ELR tests to enhance its performance. A simulation study and a real-data analysis are provided to illustrate the performance of the proposed ELR tests.

ARTICLE HISTORY
Received 1 March 2015
Accepted 13 January 2016

KEYWORDS
Empirical likelihood; Kullback–Leibler information criterion; model selection; non-degenerate test; Vuong test

1. Introduction

Competing models arise naturally in many research fields, such as economics and survival analysis, when the same phenomenon of interest is explained by different researcher using different theories or according to different experiences. For example, lognormal modeling and exponential modeling for a lifetime data in survival analysis [6], and Keynesian and new classical explanations of unemployment in economics [18]. The comparison of competing models or model selection is therefore remarkably important because of its great importance to the subsequent inference; Inference under a misspecified or inappropriate model will be risky. Since Cox [6,7] first formulated this problem in terms of hypothesis testing rather than discrimination, it has attracted considerable attention in the literature. See [5,9,16,21,23,26] and references therein.

A natural way to achieve model selection is to first introduce a statistical measure of the closeness between two models, and then recommend the one closer to the underlying true model. The most popular closeness measure in model selection is Kullback–Leibler information criterion (KLIC; [1,2]). Cox 's [6] centered log-likelihood ratio

CONTACT Yukun Liu ✉ ykliu@sfs.ecnu.edu.cn

Supplemental data for this article can be accessed here. http://dx.doi.org/10.1080/02664763.2016.1142944

test, proposed under the assumption that one of the competing models is true, is in fact a KLIC of the alternative model from the null model. This test has been applied to the testing of linear and nonlinear regression models [3,17] and more-than-one alternatives [22]. However it will lose power if neither model is true, which is often the case.

Without any model assumption on data, Vuong [26] proposed Studentized tests based on log-likelihood ratios, which in essence compares the KLIC of the two models from the underlying true model. When constructing his tests, Vuong [26] differentiated non-overlapping and overlapping models for the competing models under test. For non-overlapping models, Vuong test [26] is a Student's t-test based on log-likelihood ratios, calibrated by the standard normal distribution. In the case of overlapping models, Vuong [26] proposed a two-step test: (1) test whether the log-likelihood ratio has variance zero; (2) if the decision of (1) is rejected, then apply the test proposed for non-overlapping models. The null hypothesis that the competing models are the same close to the truth is rejected only if both (1) and (2) are rejected.

In general, Vuong test has good power if the variance of the log-likelihood ratio is away from zero or the two competing models have clearly different KLICs from the true model. Otherwise, it may have severe size distortion in finite samples in both the overlapping case and the non-overlapping case [21]. By studying the asymptotic performance of Vuong test at a series of local alternatives, Shi [21] found that the size distortion is mainly due to the asymptotic bias in both the denominator and nominator of Vuong test. A modified Vuong test is consequently proposed by correcting both the biases, and is further enhanced with a simulation-intensive calibration method.

Vuong test and Shi's modified Vuong test are both Student-type tests, which necessitates a variance estimation of the likelihood ratio and has good performance if the likelihood ratio follows a normal distribution. When the variance is rather small, it is difficult to estimate it accurately, which will increase the variation of these tests, and therefore may increases both types I and II errors. Also they will lose power if the distribution of the likelihood ratio is far away from a normal distribution.

We propose in this paper a model selection test based on empirical likelihood (EL; [13,14]), which is a popular non-parametric tool for statistical inference [11,12,19]. Comparing the KLIC of the two competing models is equivalent to test whether the mean of the likelihood ratio is zero. This motivates our proposed EL ratio test for model selection. Further, similar to Shi's modification strategy, a bias-correction method is also proposed for the EL ratio test to enhance its finite-sample performances. A significance advantage of the proposed test over Student-type tests is that it neither involves a variance estimation, nor depends on the normality of the likelihood ratio, and is therefore expected to have better performance in more general situations. Our simulation results confirm this point. We found that Vuong test often inflates its type I error substantially, therefore its power is questionable. The proposed bias-corrected EL ratio test not only has the most accurate type I errors, but also is uniformly more powerful than Shi's test; The latter restrictively controls its type I error, but is somewhat conservative, and therefore may lose certain power. Another significant advantage of the proposed test is that since calibrated by its limiting χ_1^2 distribution, its critical values or p values are available anytime and anywhere; while in comparison, Shi's test is calibrated by a computation-intensive searching method, which is rather time-consuming.

We remark that the problem considered in this paper is to test the null hypothesis that two competing models under consideration have the same appropriateness for modeling given data. Even under the null hypothesis, we still have no idea whether one of the two models is true. This is an essential difference from the well-known goodness-of-fit testing problem, which assumes that the true model is contained in either the null or the alternative hypothesis. The goodness-of-fit testing problem is a fundamental research problem in statistics and has been extensively investigated by the means of divergence measure (see, e.g. [15]). If a non-parametric assumption is imposed on the alternative, non-parametric goodness-of-fit methods, such as density-based EL techniques, Kolmogorov-Smirnov type procedures, and kernel-based approaches, have been proposed in the literature [8,24]. However, these approaches generally do not apply to the problem considered in this paper where in generally none of the models under consideration is true.

The rest of the paper is organized as follows. We define notation and review Vuong and Shi's tests in Section 2. The proposed EL ratio test is presented in Section 3, together with its asymptotical properties. The size and power of the proposed test is investigated in Section 4 by comparing with existing tests. In Section 5, we analyze a real data-set to illustrate the usefulness of the EL ratio test. All proofs are postponed to the supplemental data for clarity.

2. EL ratio test

2.1. Problem formulation

Suppose we have n independent and identically distributed (iid) copies $\{(Y_i, X_i) : i = 1, 2 \ldots, n\}$ of a random vector (Y, X) and two competing parametric probability models $\mathscr{F} = \{f(y|x; \alpha) : \alpha \in \mathcal{A}\}$ and $\mathscr{G} = \{g(y|x; \beta) : \beta \in \mathcal{B}\}$. Given the data, we wish to know which model fits the conditional density function of Y given X better.

Following [1,2,26], we take the KLIC as a measure of distance between a candidate model and the true model or a goodness measure of a candidate model. Suppose the true conditional density function of Y given X is $q(y|x)$ with distribution $Q(y|x)$. We define the distance between a given distribution family and the true distribution to be the minimum KLIC,

$$\begin{aligned} d(q, \mathscr{F}) &= \inf_{\alpha \in \mathcal{A}} \mathbb{E}_0[\ln\{q(Y|X)\} - \ln\{f(Y|X; \alpha)\}] \\ &= \mathbb{E}_0[\ln\{q(Y|X)\} - \ln\{f(Y|X; \alpha^*)\}], \end{aligned}$$

where $\alpha^* = \arg\max_{\alpha \in A} \mathbb{E}_0[\ln\{f(Y|X; \alpha)\}]$, and $\mathbb{E}_0$ denotes expectation with respect to the true joint distribution of (Y, X). The value α^* is called the pseudo-true value of α. Similarly we have

$$d(q, \mathscr{G}) = \mathbb{E}_0[\ln\{q(Y|X)\} - \ln\{g(Y|X; \beta^*)\}],$$

where $\beta^* = \arg\max_{\beta \in \mathcal{B}} \mathbb{E}_0[\ln\{g(Y|X; \beta)\}]$ is the pseudo-true value of β.

In terms of hypothesis testing, the model selection problem is to test whether the two models have the same distance from the underlying true model, that is, $H_0 : d(q, \mathscr{F}) = d(q, \mathscr{G})$, or to testing

$$H_0 : \mathbb{E}_0\{\Lambda_i(\phi^*)\} = 0, \tag{1}$$

where $\Lambda_i(\phi^*) = d(q, \mathscr{G}) - d(q, \mathscr{F}) = \ln\{f(Y_i|X_i; \alpha^*)\} - \ln\{g(Y_i|X_i; \beta^*)\}$ and $\phi^* = (\alpha^{*\tau}, \beta^{*\tau})^\tau$. When H_0 is rejected, a positive $\Lambda_i(\phi^*)$ supports $\mathscr{F}$ since positive $\Lambda_i(\phi^*)$ implies that $d(q, \mathscr{F}) < d(q, \mathscr{G})$, that is, $\mathscr{F}$ is closer to the true than $\mathscr{G}$; Otherwise, we conclude that $\mathscr{G}$ provides better fit for the data.

2.2. EL ratio test

Under the hypothesis testing formulation in Equation (1), any hypothesis test for mean is applicable to the model selection problem by taking $\{\Lambda_i(\phi)\}$ as observations if ϕ is known. This strategy still works in the case of unknown ϕ if an appropriate estimate $\hat{\phi}$ is plugged in.

Let $\hat{\alpha} = \arg\max_{\alpha\in\mathcal{A}} \sum_{i=1}^n [\ln\{f(Y_i|X_i; \alpha)\}]$ and $\hat{\beta} = \arg\max_{\beta\in\mathcal{B}} \sum_{i=1}^n [\ln\{g(Y_i|X_i; \beta)\}]$ be the pseudo maximum likelihood estimators of α and β. We propose to test (1) by the empirical likelihood ratio test (ELR; [13,14])

$$\mathrm{ELR} = -2\sup\left\{\sum_{i=1}^n \log(np_i) : \; p_i \geq 0, \; \sum_{i=1}^n p_i = 1, \; \sum_{i=1}^n p_i Z_i = 0\right\}, \tag{2}$$

where $Z_i = \Lambda_i(\hat{\phi})$ and $\hat{\phi}^\tau = (\hat{\alpha}^\tau, \hat{\beta}^\tau)$. In determining the accompanying critical values, we find that the ELR test has two totally different limiting behaviors for two exclusive cases of the null hypothesis H_0: (a) H_0 is true and $\omega^2 = \mathbb{V}\mathrm{ar}(\Lambda(\phi^*)) > 0$ or $f(y|x, \alpha^*) \neq g(y|x, \beta^*)$ in a set with positive $Q(y, x)$ probability, and (b) H_0 is true and $\omega^2 = 0$ or $f(y|x, \alpha^*) = g(y|x, \beta^*)$ $Q(y, x)$-almost surely.

For ease of presentation, we define

$$A(\phi) = \mathbb{E}\{\nabla^2\Lambda_i(\phi^*)\}, \quad B(\phi) = \mathbb{E}[\nabla\Lambda_i(\phi^*)\{\nabla\Lambda_i(\phi^*)\}^\tau],$$

where ∇ is the differentiation operator $\partial/\partial\phi$ with $\phi^\tau = (\alpha^\tau, \beta^\tau)$. Parallelling to Assumptions 1–5 of [26], we make the following assumptions on the competing models under consideration.

(C1) The parameter spaces $\mathcal{A} \subset R^{d_1}$ and $\mathcal{B} \subset R^{d_2}$ are both compact.

(C2) (Differentiability and integrability)

  (i) For all (y, x) on their supports, $f(y|x, \alpha)$ and $g(y|x, \beta)$ are three times differentiable with respect to α and β, respectively.

  (ii) There exists a non-negative function $H(y, x)$ satisfying $\mathbb{E}\{H(Y, X)\} < \infty$ such that $|\log f(y|x, \alpha)|$, $|\log g(y|x, \beta)|$ and all their first three orders of derivatives with respect to α and β are controlled by $H(y, x)$.

(C3) As a function of α, $\mathbb{E}[\log\{f(Y|X, \alpha)\}]$ has a unique maximum on $\mathcal{A}$ at an interior point α^*; And as a function of β, $\mathbb{E}[\log\{g(Y|X, \beta)\}]$ has a unique maximum on $\mathcal{B}$ at an interior point β^*.

(C4) $A(\phi^*)$ is non-singular.

Let $\Sigma = A^{-1}BA^{-1}$ and $\Omega = -\Sigma^{1/2}A\Sigma^{1/2}$ with $A = A(\phi^*)$ and $B = B(\phi^*)$. The next theorem presents the limiting distributions of the ELR test.

Theorem 2.1: *Assume the conditions in* (C1–C4). *If case* (*a*) *is true,* [ELR$\xrightarrow{\mathrm{d}}\chi_1^2$ *where* $\xrightarrow{\mathrm{d}}$ *denotes convergence in distribution. If case* (*b*) *is true,* ELR$\xrightarrow{\mathrm{d}}(\xi^\tau\Omega\xi)^2/(\xi^\tau\Omega^2\xi)$, *where* ξ, *of the same length as* $(\alpha^\tau, \beta^\tau)$, *is a standard multivariate normal random vector.*

The relationship of the two competing models generally divides into three cases: (1) non-nested, that is, $\mathscr{F} \cap \mathscr{G} = \emptyset$, (2) nested, that is, $\mathscr{F} \subset \mathscr{G}$ or $\mathscr{G} \subset \mathscr{F}$, (3) overlapping, that is, $\mathscr{F} \cap \mathscr{G} \neq \emptyset$. When the two models are nested or overlapping and $q(y|x) \in \mathscr{F} \cap \mathscr{G}$, H_0 is equivalent to case (b). We need a complicated rejection principle to test (1) according to the second part of Theorem 2.1. When the two models are non-nested or overlapping but $q(y|x) \notin \mathscr{F} \cap \mathscr{G}$, H_0 is equivalent to case (a). The first part of Theorem 2.1 recommends rejecting (1) if $\mathtt{ELR} > \chi_1^2(1-\alpha)$, the $1-\alpha$ quantile of χ_1^2. We may adopt this method for all cases although it may increase type I error in case (b). When H_0 is rejected, we recommend model $\mathscr{F}$ if

$$R = \operatorname{sgn}(\bar{Z}) \cdot \sqrt{\mathtt{ELR}} < 0$$

and model $\mathscr{G}$ otherwise.

Instead of ELR, Vuong [26] proposed to test (1) using Student's t-test statistic

$$\mathtt{VT} = \sqrt{n}\bar{Z}/\hat{\omega}_n, \tag{3}$$

where $\bar{Z} = (1/n)\sum_{i=1}^n Z_i$ and $\hat{\omega}_n^2 = (1/(n-1))\sum_{i=1}^n (Z_i - \bar{Z})^2$. Under H_0, $\mathtt{VT} \xrightarrow{d} N(0,1)$ when case (a) is true; when case (b) is true, $n\hat{\omega}_n^2 \xrightarrow{d} \xi^\tau \Omega^2 \xi$ and $(n\bar{Z})^2 \xrightarrow{d} \xi^\tau \Omega \xi$. The totally different two limiting behaviors of Vuong test leads to two testing strategies: (1) one-step test, reject H_0 if $|\mathtt{VT}| > z_{1-\alpha/2}$ where z_α is the α quantile of $N(0,1)$ and $\alpha \in (0, 0.5)$; (2) two-step test, reject H_0 if both $n\hat{\omega}_n^2 > c(\alpha; \hat{\Omega})$ and $|\mathtt{VT}| > z_{1-\alpha/2}$, where $c(\alpha; \Omega)$ is the α quantile of $\xi^\tau \Omega^2 \xi$ and $\hat{\Omega}$ is a root-n consistent estimator of Ω given later.

The variance estimation may render the Vuong test to vibrate dramatically if the two models are quite competing and the corresponding ω^2 and $\hat{\omega}_n^2$ is very close to zero. This will result in size distortion, that is, the resulting type I errors are at a distance from the significance level. Vuong's two-step test also has such a problem. This calls for a bias-correction technique to improve the efficiency of Vuong test.

2.3. Bias-corrected ELR

By local asymptotical theory, Shi [21] found that the size distortion of Vuong's tests is mainly caused by the biases in both the numerator and the denominator of Vuong test statistic $\mathtt{VT}$. A bias-corrected numerator is given by

$$\breve{Z} = \bar{Z} + \operatorname{tr}(\hat{\Omega})/(2n).$$

where $\operatorname{tr}(\hat{\Omega}) = \operatorname{tr}(\hat{A}^{-1}\hat{B})$ and $\hat{A}$ and $\hat{B}$ are the estimates of A and B respectively.. The bias in the denominator can not be eliminated but diminished or adjusted. Shi [21] proposed to modify the denominator to be

$$\breve{\omega}_n^2 = \hat{\omega}_n^2 + c \cdot \operatorname{tr}(\hat{\Omega}^2)/n.$$

where c is a tuning parameter. With these preparations, Shi [21] propose to test (1) by her non-degenerate test (NDT) statistic

$$\mathtt{NDT} = \sqrt{n}\breve{Z}/\sqrt{\breve{\omega}_n^2}.$$

Both the constant c and accompanying critical values are determined by a computation-intensive critical value determining procedure. We refer the reader to Shi [21] for details.

In our theoretical analysis, we find that the ELR is equivalent to squared Vuong test up to an ignorable term. This implies that the ELR may also suffer from the size distortion problem, which is mainly due to the biases in the numerator and the denominator. We can tell clearly where the biases come from, because Vuong test statistic `VT` has a specific fraction form and both the denominator and the nominator have closed forms. However neither the ELR nor its signed root R has a closed form, therefore Shi [21]'s bias correction method does not apply to the ELR or its signed root R, because the source of bias is unclear. To be simple, we ignore the bias in the numerator and define a bias-corrected ELR test

$$R_c = R + 0.5 \cdot \mathrm{tr}(\hat{A}^{-1}\hat{B})/\sqrt{n\hat{\omega}_n^2}.$$

The proposed testing rule is to reject H_0 if $|R_c| > z_{1-\alpha/2}$. An immediate advantage of this test over Shi's test is its convenience of practical use because it needs neither a tuning parameter nor a computation-intensive critical-value determining procedure. What is more, our simulations (see section 4) indicate that the bias-corrected ELR test usually have comparable or even better testing performance than Vuong test and Shi's `NDT` test.

As pointed out by an anonymous referee, we may correct the bias of the ELR test in Equation (2) by the strategy of Chen [4] and Vexler *et al.* [25], who proposed bias corrections for the t-test and the ELR test for mean by carefully studying their Edgeworth expansions; The goal of the corrections is to improve the approximation accuracy of their type I errors from $O_p(n^{-1/2})$ to $o_p(n^{-1/2})$. However when this strategy is applied to the ELR test in Equation (2), we find that it is formidable to derive an Edgeworth expansion of R and a subsequent bias-corrected ELR test because the 'observations' Z_i's are not iid. Hence to be simple and easy to use, we choose to use R_c as the proposed test for model comparison.

3. Extension to moment-based models

The proposed ELR tests apply also to moment-based models. Suppose the two competing moment-based models are

$$\mathscr{F} = \{F : \mathbb{E}_F(m_f(X,\alpha)) = 0\}, \tag{4}$$
$$\mathscr{G} = \{F : \mathbb{E}_F(m_g(X,\beta)) = 0\}. \tag{5}$$

We define the profile empirical log-likelihood (up to a non-random constant) of α and β to be

$$L_f(\alpha) = \max_{\lambda_1} L_f(\alpha,\lambda_1) \quad \text{and} \quad L_g(\beta) = \max_{\lambda_2} L_g(\alpha,\lambda_2),$$

where $L_f(\alpha,\lambda_1) = -\sum_{i=1}^n \log\{1+\lambda_1^\tau m_f(x,\alpha)\}$ and $L_g(\alpha,\lambda_2) = -\sum_{i=1}^n \log\{1+\lambda_1^\tau m_g(x,\alpha)\}$. Denote

$$\hat{\alpha} = \arg\max L_f(\alpha,\lambda_1(\alpha)) \quad \text{with } \lambda_1(\alpha) = \arg\min_{\lambda_1} L_f(\alpha,\lambda_1)$$

and let $\hat{\lambda}_1 = \lambda_1(\hat{\alpha})$. We define $\hat{\beta}, \lambda_2(\beta), \hat{\beta}$ and $\hat{\lambda}_2$ similarly.

The empirical KL distance between the two moment models is

$$\begin{aligned}\hat{d}(\mathscr{F},\mathscr{G}) &= \frac{1}{n}\{L_f(\hat{\alpha}) - L_g(\hat{\beta})\}\\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\{1+\lambda_1^{\tau}(\alpha)m_f(X_i,\alpha)\} + \frac{1}{n}\min_{\beta}\sum_{i=1}^{n}\log\{1+\lambda_2^{\tau}(\beta)m_g(X_i,\beta)\}\\ &= \frac{1}{n}\sum_{i=1}^{n}[\log\{1+\hat{\lambda}_2^{\tau}m_g(X_i,\hat{\beta})\} - \log\{1+\hat{\lambda}_1^{\tau}m_f(X_i,\hat{\alpha})\}].\end{aligned}$$

If models $\mathscr{F}$ and $\mathscr{G}$ are the same appropriate for fitting the data, then $d(\mathscr{F},\mathscr{G})$ tends to be small; otherwise it should be at a distance from zero.

We define the pseudo true value of α to be

$$\begin{aligned}\alpha_* &= \arg\max \mathbb{E}\log\{1+\lambda_1^{\tau}(\alpha)m_f(X_i,\alpha)\}\\ \lambda_{1*}(\alpha) &= \arg\min_{\lambda_1} \mathbb{E}\log\{1+\lambda_1^{\tau}m_f(X_i,\alpha)\}\end{aligned}$$

and the pseudo value of λ_1 to be $\lambda_{1*} = \lambda_{1*}(\alpha_*)$. We define β_*, $\lambda_{2*}(\beta)$ and λ_{2*} in the same way. Then the KL distance between the two moment models is

$$d(\mathscr{F},\mathscr{G}) = \mathbb{E}_0[\log\{1+\lambda_{2*}^{\tau}m_g(X_i,\beta_*)\} - \log\{1+\lambda_{1*}^{\tau}m_f(X_i,\alpha_*)\}]$$

and the formal testing problem is

$$H_0 : d(\mathscr{F},\mathscr{G}) = 0 \longleftrightarrow H_1 : d(\mathscr{F},\mathscr{G}) \neq 0.$$

Let $\phi = (\alpha,\lambda_1,\beta,\lambda_2)^{\tau}$, $\phi_* = (\alpha_*,\lambda_{1*},\beta_*,\lambda_{2*})^{\tau}$ and $\hat{\phi} = (\hat{\alpha},\hat{\lambda}_1,\hat{\beta},\hat{\lambda}_2)^{\tau}$. Define

$$\Lambda_i(\phi) = \log\{1+\lambda_2^{\tau}m_g(X_i,\beta)\} - \log\{1+\lambda_1^{\tau}m_f(X_i,\alpha)\}.$$

The testing problem is equivalent to

$$H_0 : \mathbb{E}\{\Lambda_i(\phi_*)\} = 0 \longleftrightarrow H_1 : \mathbb{E}\{\Lambda_i(\phi_*)\} \neq 0.$$

The proposed EL and EL$_c$ tests apply to this problem directly by setting $Z_i = \Lambda_i(\hat{\phi})$.

4. Simulation study

In this section, we report Monte Carlo simulation results to evaluate the performance of the proposed two ELR tests, the sign root R of ELR (EL) and the bias-corrected sign root R_c of ELR (EL$_c$). This purpose is achieved by comparing them with three existing tests: One-step Vuong test (1-step VT), Two-step Vuong test (2-step VT) and [21]'s NDT.

Example 4.1: [(Normal regression; [21])] Suppose the true underlying data generating process is

$$Y = 1 + (a_1/\sqrt{d_1})\mathbf{1}_{d_1}^{\tau}X_{(1)} + (a_2/\sqrt{d_2})\mathbf{1}_{d_2}^{\tau}X_{(2)} + \varepsilon, \tag{6}$$

where X_1 and X_2 are d_1 and d_2-variate covariates, and $(X_1^{\tau}, X_2^{\tau}, \varepsilon)$ follows the (d_1+d_2+1)-variate standard normal distribution and $a_1, a_2 \in [0,1]$. With data generated from

Table 1. Simulated rejection probabilities (%) of the five tests under comparison.

Cases	EL	EL$_c$	1-StepVT	2-StepVT	NDT	Var. T
$\frac{n\mathbb{E}\{\Lambda_i(\phi^*)\}}{250} = 0$						
Base	(0.24, 7.52)	(1.70, 1.34)	(0.32, 8.24)	(0.32, 7.82)	(1.18, 1.08)	95.14
$d_2 = 19$	(0.00, 26.58)	(1.34, 0.88)	(0.00, 28.34)	(0.00, 18.02)	(0.96, 0.96)	64.36
$d_2 = 4$	(0.86, 3.32)	(1.78, 1.26)	(0.94, 3.82)	(0.94, 3.82)	(1.00, 0.78)	99.08
$n = 100$	(0.04, 10.78)	(1.20, 0.46)	(0.06, 12.50)	(0.04, 1.84)	(0.34, 0.18)	19.84
$n = 500$	(0.44, 6.62)	(1.80, 1.90)	(0.48, 7.12)	(0.48, 7.12)	(1.36, 1.66)	100.00
$\frac{n\mathbb{E}\{\Lambda_i(\phi^*)\}}{250} = \log(1.09)$						
Base	(16.58, 0.02)	(50.90, 0.00)	(18.10, 0.02)	(17.44, 0.00)	(42.78, 0.00)	80.76
$d_2 = 19$	(2.06, 0.86)	(36.52, 0.00)	(2.28, 1.06)	(1.60, 0.24)	(32.30, 0.00)	41.60
$d_2 = 4$	(41.34, 0.00)	(59.70, 0.00)	(43.74, 0.00)	(43.64, 0.00)	(47.66, 0.00)	95.08
$n = 100$	(19.46, 0.00)	(50.90, 0.00)	(22.02, 0.04)	(14.04, 0.00)	(37.16, 0.00)	53.82
$n = 500$	(15.10, 0.02)	(50.26, 0.00)	(15.74, 0.02)	(15.68, 0.02)	(43.68, 0.00)	87.22
$\frac{n\mathbb{E}\{\Lambda_i(\phi^*)\}}{250} = -\log(1.09)$						
Base	(0.00, 88.60)	(0.00, 39.72)	(0.00, 90.06)	(0.00, 74.56)	(0.00, 36.12)	77.76
$d_2 = 19$	(0.00, 98.68)	(0.00, 26.22)	(0.00, 98.96)	(0.00, 33.92)	(0.00, 29.90)	34.00
$d_2 = 4$	(0.00, 76.24)	(0.00, 53.92)	(0.00, 78.90)	(0.00, 78.24)	(0.00, 43.94)	94.58
$n = 100$	(0.00, 90.20)	(0.00, 38.08)	(0.00, 92.76)	(0.00, 29.70)	(0.00, 27.86)	30.78
$n = 500$	(0.00, 88.52)	(0.00, 43.26)	(0.00, 89.70)	(0.00, 82.94)	(0.00, 41.06)	86.86

Note: In each pair (p_1, p_2), p_1 denotes the probability of rejecting H_0 and supporting $\mathscr{F}$, and p_2 the probability of rejecting H_0 and supporting $\mathscr{G}$.

Equation (5), we wish to select between the following two models

$$\mathscr{F}: \ Y = \alpha_0 + \alpha_1^\tau X_{(1)} + e_1, \quad e_1|(X_{(1)}, X_{(2)}) \sim N(0, \sigma_1^2),$$
$$\mathscr{G}: \ Y = \beta_0 + \beta_1^\tau X_{(2)} + e_2, \quad e_2|(X_{(1)}, X_{(2)}) \sim N(0, \sigma_2^2),$$

where $\alpha_0, \alpha_1, \sigma_1^2, \beta_0, \beta_1$ and σ_2^2 are all unknown.

Our simulation settings in this example are the same as Example 1 of [21]. Given (a_1, a_2), we consider five cases. The base case is $n = 250, d_1 = 1$ and $d_2 = 9$. The rest four are variants of the base case, which are only different from the base case in $d_2 = 19, d_2 = 4, n = 100$ and $n = 500$, respectively. Three pairs of (a_1, a_2) are considered: (H1) $a_1 = a_2 = 0.25$; (H2) $a_1 = \sqrt{1.09^{250/n} - 1}, a_2 = 0$; (H3) $a_1 = 0, a_2 = \sqrt{1.09^{250/n} - 1}$. In this example, $\mathbb{E}\{\Lambda_i(\phi^*)\} = \frac{1}{2}\{\log(1 + a_1^2) - \log(1 + a_2^2)\}$. Therefore in case (H1), $\mathbb{E}\{\Lambda_i(\phi^*)\} = 0$ and the null hypothesis is true. In case (H2), $\mathbb{E}\{\Lambda_i(\phi^*)\} = \log(1.09) \times (250/n)$; the null hypothesis is violated and model $\mathscr{F}$ is true. Case (H3) is the opposite of case (H2).

We generated 5000 data-sets from Example 1 under each of the 15 settings and computed the simulated rejection probabilities (in percentage) of the tests under consideration. In addition, we also record the proportion (denoted Var.T) of rejecting the hypothesis that the variance of likelihood ratio is zero. The simulation results are reported in Table 1.

The first panel of Table 1 includes simulated type I errors of the five tests under comparison. Since the two competing models have equal goodness-of-fit for the data, it is ideally expected that the probability that the tests reject the null hypothesis and recommend either model should be at most 2.5 % at the 0.05 significance level. However only the bias-corrected ELR and [21]'s NDT make it, and the former has closer-to-nominal one-sided type I errors than the latter, which is somewhat conservative. The original ELR and the two Vuong tests often have severely-inflated one-sided type I errors particularly when the parameter dimension is large (case $d_2 = 19$) or the sample size is small (case $n = 100$). This also implies that the EL does have an unignorable bias and the EL$_c$ succeeds in correcting it, leading to rather accurate type I errors.

Simulated power comparisons are presented in the second and third panels of Table 1, corresponding respectively to (a_1, a_2) values given in cases (H2) and (H3). The power comparisons are only meaningful for the bias-corrected ELR and [21]'s NDT control their type I errors, because only they two control their type I errors. Case (H2) is designed such that model $\mathscr{F}$ is better than model $\mathscr{G}$. The EL$_c$ test has uniformly larger power than NDT in detecting the fact. We have similar observations in case (H3), which is designed such that model $\mathscr{G}$ is better than model $\mathscr{F}$.

Example 4.2: Suppose the two competing models are a Poisson model

$$\mathscr{F} = \{f(y|\mathbf{x}, \alpha) = e^{-\lambda}\lambda^y/y!, \quad \lambda = \exp(\alpha^\tau \mathbf{x})\}, \tag{7}$$

and a Geometric model

$$\mathscr{G} = \{g(y|\mathbf{x}, \beta) = (1-p)^{y_i}p, \quad p = e^{\beta^\tau \mathbf{x}}/(1 + e^{\beta^\tau \mathbf{x}})\}, \tag{8}$$

where y takes non-negative integer values. The covariate $\mathbf{x}$ is a 5-variate vector with its first component being 1 and the rest four are iid as the uniform distribution on $(0, 0.25)$. Assume that the data generating process is a two-component mixture distribution, $\pi f(y|\mathbf{x}, \alpha^*) + (1-\pi)g(y|\mathbf{x}, \alpha^*)$ where $\alpha^* = (1, 1, 1, 1, 1)$ and $\pi \in (0, 1)$ to be determined.

This example is designed to mimic the model selection setting for number of doctor visits in Section 5.1. We hope that the simulation results in this example can shed light on the performances of the tests under comparison and provide evidence for their relative efficiency.

Intuitively, when π is small, $\mathbb{E}\{\Lambda_i(\phi^*)\}$ tends to be negative, and the Geometric model fits better; while when π is large, $\mathbb{E}\{\Lambda_i(\phi^*)\}$ tends to be positive, and the Poisson model fits better. Based on extra-large samples (sample size 100,000), we found that $\mathbb{E}\{\Lambda_i(\phi^*)\} = 0$ when $\pi = 0.875$. In our simulation, the sample size is $n = 100$ and the simulation size is 2000. We generate data from the mixture models with π varying from 0.675 to 0.975 with increment 0.02. When $\pi < 0.875$, model $\mathscr{G}$ fits the data better, while when $\pi > 0.875$, model $\mathscr{F}$ fits the data better. Our simulation results are tabulated in Table 2. The powers of the two-step VT are not reported because they are almost the same as those of the one-step VT.

When $\pi = 0.875$ or the null hypothesis $H_0 : \mathbb{E}\{\Lambda_i(\phi^*)\} = 0$ holds, the powers (bold numbers) given in Table 2 are actually type I errors. We find that only the EL$_c$ controls its type I error below the significance level 5%; it type I error is also the closest to 5%. The Vuong tests have the largest and excessive type I error as was pointed out by Shi [21]. With the bias-correction strategy, the EL$_c$ and Shi's NDT tests reduce the type I errors of EL and Vuong's VT tests, respectively.

In view of power comparison, among all the tests the EL$_c$ test has the largest power when $\pi < 0.875$ or the Geometric model fits better, and has the smallest power $\pi > 0.875$ or the Poisson model fits better. Because of the comparison result in type I error, only the

Table 2. Simulated powers (unbolded numbers) and sizes (bolded numbers) in percentage in Example 4.2.

π	EL	EL_c	1-step VT	NDT
0.675	97.40	98.70	96.20	97.00
0.695	94.75	96.25	92.85	94.00
0.715	88.55	91.90	85.05	87.05
0.735	80.60	86.10	75.80	78.40
0.755	66.45	73.40	60.30	63.25
0.775	53.40	61.20	46.20	49.90
0.795	37.30	46.75	31.75	35.00
0.815	21.90	29.25	17.35	19.60
0.835	11.60	17.05	9.50	10.05
0.855	6.30	8.60	6.25	5.50
0.875	**5.90**	**4.80**	**7.80**	**5.45**
0.895	10.95	7.65	16.25	10.85
0.915	23.60	16.95	33.20	22.95
0.935	47.70	37.90	58.30	46.95
0.955	71.75	62.40	80.35	71.45
0.975	91.30	86.75	95.35	91.15

simulated power of the EL_c is trustable; those for the EL, VT and NDT tests may not be representative for their true testing power.

5. Real-data analysis

We illustrate the proposed bias-corrected ELR test by analyzing a data set taken from the first 12 annual waves (1984 through 1995) of the German Socioeconomic Panel data. This data set, studied by Greene [10] and Riphahn *et al.* [20] , consists of 27,326 observations on 25 variables including number of doctor visits in the last three months (Docvis), number of hospital visits in last calendar year (Hospvis), and numerous other socio-demographic variables such as age (Age), education (Edu), house income (Income) and having kids or not (Kids). We choose $y =$ Docvis or Hospvis, and $\mathbf{x} =$ (Age, Edu, Income, Kids). Following Example 14.10 of [10], we consider the model selection problem between the competing models (7) and (8) for the conditional probability of y given $\mathbf{x}$.

The histograms of Docvis and Hospvis are too skewed. Among all the 27,326 values, there 10,135 zeros in Docvis and 24,931 zeros in Hospvis. For a better presentation, in Figure 1 we only display the non-zero values in Docvis and Hospvis. The log likelihood ratios $\Lambda_i(\hat{\phi})$ for $y =$ Docvis or Hospvis, are also calculated and displayed in Figure 1.

We find that the variance estimate $\hat{\omega}^2 = 1276948.35$ (in the case of Docvis) and 86986.38 (in the case of Hospvis) are both larger than the accompanying critical value 583.82 and 1137.52 at the 5% significance level. Thus the one- and two-step Vuong tests will lead to the same decision. Table 3 presents the test statistics of the proposed ELR and bias-corrected ELR tests, one-step Vuong test and Shi's NDT test.

The critical values proposed by Shi [21] for the NDT test are 2.0229 and 2.1949 corresponding to the cases $y =$ Docvis and Hospvis, respectively. Clearly all the four tests conclude that (1) the Poisson model and the Geometric model do not have the same appropriateness for the data at the 5% significance level, and (2) the Geometric model fits better because the mean of $\Lambda_i(\hat{\phi})$ is negative and a negative $\mathbb{E}\{\Lambda_i(\phi^*)\}$ supports $\mathscr{G}$. Meanwhile the absolute value of the EL_c is much larger those of the VT and NDT tests. Given that all

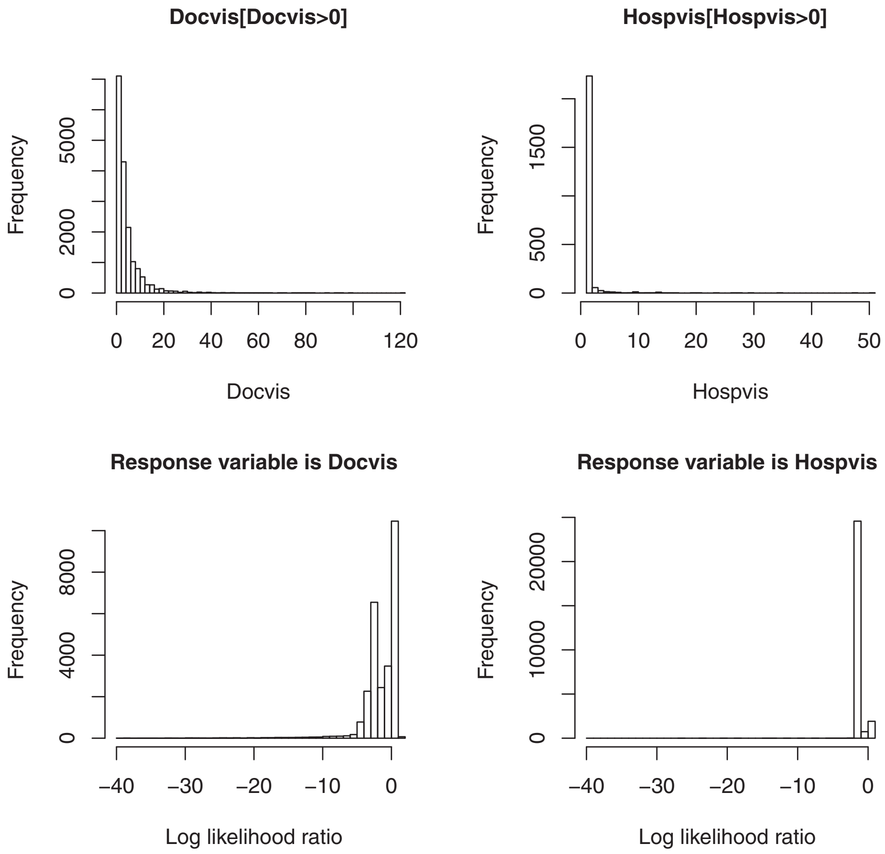

Figure 1. Histogram display of Docvis, Hospvis and the corresponding log likelihood ratios.

Table 3. Test statistics of the EL tests, the VT and NDT tests for the real data example.

Responsevariable	EL	EL_c	VT	NDT
Docvis	−147.9161	−147.9408	−46.7393	−46.7531
Hospvis	−270.5393	−270.6181	−91.0500	−91.08769

the critical values are all around 2, this implies that the proposed EL_c test provides much stronger evidence for the superiority of the Geometric model.

Two observations make the conclusion not surprising. On one hand, according to our simulation experience in Example 4.2, the EL_c test has the most accurate type I error and largest power in supporting the Geometric model among the four tests. On the other hand, the student-type test relies to some extent to the normality of the likelihood ratios, while the EL-type tests do not. However we observe from Figure 1 the likelihood ratios in the cases of $y =$ Docvis and Hospvis are both severely skewed and far away from being normally distributed. Therefore it is natural that the EL and EL_c tests have better powers than the VT and NDT in this real data analysis.

Acknowledgments

The authors thank the Editor and two anonymous referees for their valuable and constructive comments which significantly improved the quality of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Natural Science Foundations of China [grant numbers 11501354, 11371142, 11501208], Program of Shanghai Subject Chief Scientist [14XD1401600], and the 111 Project [B14019].

References

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, in *Proceedings of the 2nd International Symposium on Information Theory*, N. Petrov and F. Csadki, eds., Akademiai Kiado, Budapest, 1973, pp. 267–281.

[2] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Auto Cont. AC19 (1974), pp. 716–723.

[3] G. Aneuryn-Evans and N. Deaton, *Testing linear versus logarithmic regression models*, Rev. Econom. Stud. 47 (1980), pp. 275–291.

[4] L. Chen, *Testing the mean of skewed distributions*, J. Amer. Statist. Assoc. 90 (1995), pp. 767–772.

[5] X. Chen, H. Hong, and M. Shum, *Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models*, J. Econometrics 141 (2007), pp. 109–140.

[6] D. Cox, *Tests of separate families of hypothesis*, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1961.

[7] D. Cox, *Further results on tests of separate families of hypothesis*, J. R. Stat. Soc. Ser. B 24 (1962), pp. 406–424.

[8] J.H.J. Einmahl and I.W. McKeague, *Empirical likelihood based hypothesis testing*, Bernoulli 9 (2003), pp. 267–290.

[9] G. Gourieroux and A. Monfort, *Testing non-nested hypotheses*, in *Handbook of Econometrics*, R. Engle and D. McFadden, eds., Vol. 4. North-Holland, Amsterdam, 1994.

[10] W.H. Greene, *Econometric Analysis*, Prentice Hall, New York, 2002.

[11] K. Kitamura, *Comparing misspecified dynamic econometric models using nonparametric likelihood*, Department of Economics, University of Wisconsin, 2001. Available at http://tswww.ism.ac.jp/kitagawa/Japan-US/preprints/kitamura.pdf.

[12] K. Kitamura, *Empirical likelihood methods in economics: Theory and practice*, Foundation Discussion Paper No. 1569, Cowles Foundation, Yale University, 2006.

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

[14] A.B. Owen, *Empirical Likelihood*, Chapman and Hall/CRC, New York, 2001.

[15] L. Pardo, *Statistical Inference Based on Divergence Measures*, Chapman & Hall/CRC, Taylor & Francis Group, Florida, 2006.

[16] M.H. Pesaran, *On the general problem of model selection*, Review of Economic Studies 41 (1974), pp. 153–171.

[17] M.H. Pesaran and A.S. Deaton, *Testing non-nested nonlinear regression models*, Econometrica 46 (1978), pp. 677–694.

[18] M.H. Pesaran and M. Weeks, *Nonnested hypothesis testing: An overview*, A Companion to Theoretical Econometrics, BALTAGI, BADI H. Blackwell Blackwell Reference Online, 2003. Available at http://www.blackwellreference.com/subscriber/tocnode.html?id = g9781405106764_chunk_g978140510676418.

[19] J. Qin and J. Lawless, *Empirical likelihood and general equations*, Ann. Statist. 22 (1994), pp. 300–325.

[20] R.T. Riphahn, A. Wambach, and A. Million, *Incentive effects in the demand for health care: A bivariate panel count data estimation*, J. Appl.Econometrics 18 (2003), pp. 387–405.

[21] X. Shi, *A non-degenerate Vuong test*, Quant. Econ. 6 (2015), pp. 85–121.

[22] K.R. Sawyer, *Mutiple hypothesis testing*, J. R. Stat. Soc. Ser. B 46 (1984), pp. 419–424.

[23] C.-Y. Sin and H. White, *Information criteria for selection possible misspecified parametric models*, J. Econometrics 71 (1996), pp. 207–225.

[24] A. Vexler and G. Gurevich, *Empirical likelihood ratios applied to goodness-of-Fit tests based on sample entropy*, Comput. Statist. Data Anal. 54 (2010), pp. 531–545.

[25] A. Vexler, S. Liu, L. Kang, and G. Gurevich, *Modifications of the empirical likelihood interval estimation with improved coverage probabilities*, Comm. Statist. Simulation Comput. 38 (2009), pp. 2171–2183.

[26] Q.H. Vuong, *Likelihood ratio test for model selection and non-nested hypothesis*, Econometrica 57 (1989), pp. 307–333.