Stat Comput
DOI 10.1007/s11222-010-9199-7

A two-sample empirical likelihood ratio test based on samples entropy

Gregory Gurevich · Albert Vexler

Received: 14 December 2009 / Accepted: 23 August 2010

Abstract Powerful entropy-based tests for normality, uniformity and exponentiality have been well addressed in the statistical literature. The density-based empirical likelihood approach improves the performance of these tests for goodness-of-fit, forming them into approximate likelihood ratios. This method is extended to develop two-sample empirical likelihood approximations to optimal parametric likelihood ratios, resulting in an efficient test based on samples entropy. The proposed and examined distribution-free two-sample test is shown to be very competitive with well-known nonparametric tests. For example, the new test has high and stable power detecting a nonconstant shift in the two-sample problem, when Wilcoxon's test may break down completely. This is partly due to the inherent structure developed within Neyman-Pearson type lemmas. The outputs of an extensive Monte Carlo analysis and real data example support our theoretical results. The Monte Carlo simulation study indicates that the proposed test compares favorably with the standard procedures, for a wide range of null and alternative distributions.

Keywords Empirical likelihood · Entropy · Likelihood ratio · Two-sample nonparametric tests

G. Gurevich
The Department of Industrial Engineering and Management, SCE—Shamoon College of Engineering, Beer Sheva 84100, Israel
e-mail: gregoryg@sce.ac.il

A. Vexler (✉)
Department of Biostatistics, New York State University at Buffalo, Buffalo, NY 14214, USA
e-mail: avexler@buffalo.edu

Published online: 07 September 2010

1 Introduction

Our objective in the present article is to propose and apply a simple approach that provides powerful nonparametric approximations to optimal likelihood ratio test statistics. We present a technique that utilizes the central idea of the empirical likelihood methodology where the empirical likelihood function consists of components that maximize this likelihood function and satisfy empirical constraints (e.g., Lazar and Mykland 1998; Owen 2001; Yu et al. 2010). The classical empirical likelihood method is based on empirical distribution functions. To approximate most powerful test-statistics stated by the Neyman-Pearson Lemma, we extend and adapt the density-based empirical likelihood method that was presented by Vexler and Gurevich (2010) in the context of goodness-of-fit testing. We apply the proposed approach to construct a powerful two-sample nonparametric likelihood ratio test based on samples entropy. Despite the fact that many statistical inference procedures have been developed to construct very efficient entropy-based tests for goodness-of-fit (e.g., Arizono and Ohta 1989; Dudewicz and Van Der Meulen 1981; Mudholkar and Tian 2002, 2004; Tusnady 1977; Vasicek 1976; Vexler and Gurevich 2010; Zhang 2002), to our knowledge, there does not exist an inference procedure for two-sample empirical likelihood ratio comparisons based on samples entropy.

In Sect. 2, we outline the empirical likelihood methodology for the sake of completeness and improving entropy-based test-statistics. In Sect. 3, we extend the density-based empirical likelihood approach to be applied to the standard two-sample problem. Section 3 also describes how to approximate optimal parametric likelihood ratios in a general nonparametric setting. Section 4 formulates and analyzes a new test. The proposed empirical likelihood approach is applied to a general case of two-sample testing (e.g., Lehmann

and Romano 2005). The classical problem of two-sample testing has various applications in different fields of statistics, engineering and management. Results of Monte Carlo analyses presented in Sect. 5 show that the proposed test is superior to the standard procedures in many various situations. A simulation study shows that the new test has high and stable power, detecting nonconstant shift alternatives in the two-sample problem, when the classical procedures break down completely. In Sect. 6, we illustrate advantages of the proposed test using data from a study on atherosclerotic process of coronary heart disease. Some concluding remarks are presented in Sect. 7.

2 Entropy-based empirical likelihood approximation to parametric likelihood functions

The likelihood principle is arguably the most important concept for inference in parametric models. Recently it has also been shown to be useful in nonparametric contexts. As an example, consider the goodness-of-fit testing problem where given a sample of k independent identically distributed observations $X_1, \ldots, X_k$, we want to test the hypothesis

$$H_0 : X_1, \ldots, X_k \sim F_0 \quad \text{versus} \quad H_1 : X_1, \ldots, X_k \sim F_1, \tag{2.1}$$

where F_0 and F_1 are some distributions with density functions $f_0(x)$ and $f_1(x)$, respectively. By virtue of the Neyman-Pearson Lemma, the most powerful test-statistic for (2.1) is the likelihood ratio

$$\frac{\prod_{i=1}^k f_1(X_i)}{\prod_{i=1}^k f_0(X_i)}, \tag{2.2}$$

where density functions $f_0(x)$ and $f_1(x)$ are assumed to be completely known (e.g., Lehmann and Romano 2005; Vexler et al. 2010). However, it is not always possible, or optimal, to use parametric likelihoods. For instance, when $f_0(x)$ and $f_1(x)$ depend on many unknown parameters (maximum likelihood estimators might be inconsistent in the case of multidimensional parameter estimation), or forms of $f_0(x)$ and $f_1(x)$ related to (2.1) cannot be assumed to be known. Thus, there has been much recent development of various empirical likelihood type approximations to parametric likelihood functions. The empirical likelihood (EL) method based on empirical distributions has been dealt with extensively (e.g., Owen 2001). The EL function has the form of $L_p = \prod_{i=1}^k p_i$, where the components $p_i, i = 1, \ldots, k$ maximize L_p and satisfy empirical constraints corresponding to hypotheses of interest. For example, if the null hypothesis is $H_0 : E(X_1) = 0$, then the values of p_i's in the H_0-empirical likelihood L_p should be chosen to maximize L_p given $\sum_{i=1}^k p_i = 1$ and $\sum_{i=1}^k p_i X_i = 0$, where the constraint $\sum_{i=1}^k p_i X_i = 0$ is an empirical version of $E(X_1) = 0$. Computation of $p_i, i = 1, \ldots, k$ is based on a simple exercise in Lagrange multipliers. This nonparametric approach is a result of consideration of the 'distribution functions'-based likelihood $\prod_{i=1}^k (F(X_i) - F(X_i-))$ over all distribution functions F (see, for details, Owen 2001). The density-based structure of the likelihood ratio has the main role in the Neyman-Pearson Lemma proof scheme. With this motivation, Vexler and Gurevich (2010) proposed to use the central idea of the EL technique to develop density-based empirical approximations to the likelihood $L_f = \prod_{i=1}^k f(X_i)$, where $f(x)$ is a density function. The authors introduced a method to construct nonparametric likelihood ratio test statistics for the problem (2.1), where the density function $f_1(x)$ is unknown, whereas $f_0(x)$ has a known parametric form. To outline this technique, we present the likelihood function L_f in the form of

$$L_f = \prod_{i=1}^k f(X_i) = \prod_{i=1}^k f(X_{(i)}) = \prod_{i=1}^k f_i$$

with $f_i = f(X_{(i)})$, and $X_{(1)} \le X_{(2)} \le \cdots \le X_{(k)}$ are the order statistics derived from $X_1, \ldots, X_k$. Following the maximum EL method, we can obtain estimated values of $f_i, i = 1, \ldots, k$ that maximize L_f and satisfy empirical constraints. Obviously, the equation $\int f(u)du = 1$ constrains values of $f_i, i = 1, \ldots, k$. To formalize this constraint, Vexler and Gurevich (2010) proposed the following result.

Proposition 2.1 *Assume $X_{(j)} = X_{(1)}$, if $j \le 1$, and $X_{(j)} = X_{(k)}$, if $j \ge k$. Then for all integer m, we have*

$$\begin{aligned}
&\sum_{j=1}^k \int_{X_{(j-m)}}^{X_{(j+m)}} f(u)du \\
&\quad = 2m \int_{X_{(1)}}^{X_{(k)}} f(u)du - \sum_{l=1}^{m-1} (m-l) \int_{X_{(k-l)}}^{X_{(k-l+1)}} f(u)du \\
&\qquad - \sum_{l=1}^{m-1} (m-l) \int_{X_{(l)}}^{X_{(l+1)}} f(u)du.
\end{aligned}$$

Denote

$$H_m = \frac{1}{2m} \sum_{j=1}^k \int_{X_{(j-m)}}^{X_{(j+m)}} f(x)dx.$$

Since $\int_{X_{(1)}}^{X_{(k)}} f(x)dx \le \int_{-\infty}^{+\infty} f(x)dx = 1$, Proposition 2.1 shows that $H_m \le 1$, as well as, one can expect that $H_m \approx 1$,

when $m/k \to 0$ as $m, k \to \infty$. While approximating $\int_{X_{(j-m)}}^{X_{(j+m)}} f(x)dx \cong (X_{(j+m)} - X_{(j-m)})f_j$, we represent the condition $H_m \leq 1$ in the empirical form of

$$\tilde{H}_m \leq 1, \quad \tilde{H}_m = \frac{1}{2m}\sum_{j=1}^{k}(X_{(j+m)} - X_{(j-m)})f_j. \tag{2.3}$$

Deriving $\partial/\partial f_i, i = 1, \ldots, k$, from the function $\log L_f + \lambda(1 - \tilde{H}_m)$ with the Lagrange multiplier λ, and then solving the resulting equation

$$\frac{\partial}{\partial f_i}(\log L_f + \lambda(1 - \tilde{H}_m)) = \frac{1}{f_i} - \lambda\frac{1}{2m}(X_{(i+m)} - X_{(i-m)}) = 0,$$

we obtain that the values

$$f_i = \frac{2m}{k(X_{(i+m)} - X_{(i-m)})}, \quad i = 1, \ldots, k,$$

maximize $\log L_f$, satisfying the constraint (2.3) (here $X_{(j)} = X_{(1)}$, if $j \leq 1$, and $X_{(j)} = X_{(k)}$, if $j \geq k$). Finally, the EL estimate of the likelihood has the form of $\prod_{i=1}^{k} 2m(k(X_{(i+m)} - X_{(i-m)}))^{-1}$. Therefore, the maximum EL method applied to approximate (2.2), with known $f_0(x)$ and unknown $f_1(x)$, forms the test-statistic

$$T_{mk} = \frac{\prod_{i=1}^{k} \frac{2m}{k(X_{(i+m)} - X_{(i-m)})}}{\prod_{i=1}^{k} f_0(X_i)}. \tag{2.4}$$

Note that the logarithm of the EL estimate of the parametric likelihood function is

$$\log\left(\prod_{i=1}^{k} \frac{2m}{k(X_{(i+m)} - X_{(i-m)})}\right) = -kH(m, k).$$

The statistic $H(m, k) = k^{-1}\sum_{i=1}^{k} \log(k(X_{(i+m)} - X_{(i-m)})/2m)$ was presented by Vasicek (1976), as an estimate of the entropy of the density $f(x)$, for some $m < k/2$, i.e. $H(m, k)$ estimates

$$\begin{aligned} H(f) &= E(-\log(f(X_1))) \\ &= -\int_{-\infty}^{+\infty} f(x)\log(f(x))dx \\ &= \int_0^1 \log\left(\frac{d}{dp}F^{-1}(p)\right)dp. \end{aligned}$$

Vasicek (1976), Arizono and Ohta (1989), Dudewicz and Van Der Meulen (1981) as well as Ebrahami et al. (1992) have demonstrated that the test statistic (2.4) with optimal values of m provides very powerful tests for normality, uniformity and exponentiality. Note that the authors obtained (2.4)-type test-statistics via estimation of the sample entropy (e.g., Vasicek 1976). The method of Vexler and Gurevich (2010) demonstrates the test statistic T_{mk} is an approximation to the optimal likelihood ratio. Thus, we expect directly that a test based on T_{mk} will provide highly efficient characteristics. The new EL approach leading to (2.4) can be applied to improve the entropy-based test-statistic as well as to extend the entropy-based methodology to general cases of testing problems.

The power of the tests based on the statistic T_{mk} strongly depends on values of m. The literature commonly defines optimal values of m (corresponding to a power perspective) by simulation studies assuming information on the alternative density function $f_1(x)$. This restricts applicability of (2.4)-type test-statistics to real-data problems. The empirical likelihood technique presented above completes the approximation to the parametric likelihood function in the form of

$$\min_{1 \leq m \leq k^{1-\delta}} \prod_{i=1}^{k} \frac{2m}{k(X_{(i+m)} - X_{(i-m)})}, \quad 0 < \delta < 1 \tag{2.5}$$

(Vexler and Gurevich 2010). The assertion of (2.5), where the operator min is applied, is a consequence of empirical likelihood considerations. The principles leading to the form (2.5) are outlined in the next section where the density based empirical likelihood approach is extended.

3 Entropy-based empirical likelihood approximation to parametric likelihood ratios

Finding an appropriate nonparametric likelihood ratio test statistic for the testing problem in the setting of two-sample distribution-free comparisons entails two issues: (1) in contrast to the density-based EL approach mentioned in Sect. 2, under the null hypothesis, parametric forms of relevant density functions cannot be assumed to be known; (2) it is reasonable to construct the target test statistic with an H_0-distribution function that is independent of distributions of observations, developing an exact test. (In this case, one can easily show that the issue 2 above declares against direct applications of the technique by Vexler and Gurevich (2010) to approximate separately the numerator and denominator of the relevant parametric likelihood ratio.) In this section, modifying and extending the EL technique mentioned in Sect. 2, we sketch the lines of arguments leading to the new test statistic. We start with a statement of the two-sample testing problem.

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_k$ be independent samples that consist of independent identically distributed observations from distributions F_X and F_Y with density functions $f_X(x)$ and $f_Y(y)$, respectively. We are interested to verify if

both the samples are from the same distribution. That is, we want to test for

$$H_0 : F_Y = F_X = F_Z \quad \text{versus} \quad H_1 : F_Y \neq F_X, \tag{3.1}$$

where distributions F_Z, F_X and F_Y are unknown. In this case, the likelihood ratio statistic based on all $n+k$ observations has the form of

$$\frac{\prod_{i=1}^{n} f_X(X_i) \prod_{j=1}^{k} f_Y(Y_j)}{\prod_{i=1}^{n} f_Z(X_i) \prod_{j=1}^{k} f_Z(Y_j)} = \frac{\prod_{i=1}^{n} f_{X,i} \prod_{j=1}^{k} f_{Y,j}}{\prod_{i=1}^{n} f_{ZX,i} \prod_{j=1}^{k} f_{ZY,j}}, \tag{3.2}$$

where a density function f_Z corresponds to the null hypothesis, $f_{X,i} = f_X(X_{(i)})$, $f_{Y,j} = f_Y(Y_{(j)})$, and $f_{ZX,i} = f_Z(X_{(i)})$, $f_{ZY,i} = f_Z(Y_{(j)})$, $i = 1, \ldots, n$, $j = 1, \ldots, k$; $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(k)}$ are the order statistics based on the observations $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_k$, respectively. We begin by applying the method of the density-based EL mentioned in Sect. 2 to estimate $f_{X,i}, i = 1, \ldots, n$. That is, we will derive values of $f_{X,i}, i = 1, \ldots, n$ that maximize the likelihood $\prod_{i=1}^{n} f_{X,i}$, satisfying an empirical constraint. Here the equation $\int f_X(u)du = 1$ constrains values of $f_{X,i}, i = 1, \ldots, n$. By virtue of Proposition 2.1, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \int_{X_{(i-m)}}^{X_{(i+m)}} \frac{f_X(u)}{f_Z(u)} f_Z(u)du \\
&= 2m \int_{X_{(1)}}^{X_{(n)}} \frac{f_X(u)}{f_Z(u)} f_Z(u)du \\
&\quad - \sum_{l=1}^{m-1} (m-l) \int_{X_{(n-l)}}^{X_{(n-l+1)}} \frac{f_X(u)}{f_Z(u)} f_Z(u)du \\
&\quad - \sum_{l=1}^{m-1} (m-l) \int_{X_{(l)}}^{X_{(l+1)}} \frac{f_X(u)}{f_Z(u)} f_Z(u)du.
\end{aligned} \tag{3.3}$$

Thus, since $\int_{X_{(1)}}^{X_{(n)}} f_X(u)du \leq \int_{-\infty}^{+\infty} f_X(u)du = 1$, we conclude

$$\Delta_m \leq 1, \quad \Delta_m = \frac{1}{2m} \sum_{i=1}^{n} \int_{X_{(i-m)}}^{X_{(i+m)}} \frac{f_X(u)}{f_Z(u)} f_Z(u)du, \tag{3.4}$$

and $\Delta_m \approx 1$ when $m/n \to 0$ as $m, n \to \infty$. In a similar manner to deriving the constraint (2.3), by applying the approximate analog to the mean-value integration theorem, we approximate Δ_m as

$$\begin{aligned}
\Delta_m &\cong \frac{1}{2m} \sum_{i=1}^{n} \frac{f_{X,i}}{f_{ZX,i}} \int_{X_{(i-m)}}^{X_{(i+m)}} f_Z(u)du \\
&= \frac{1}{2m} \sum_{i=1}^{n} (F_Z(X_{(i+m)}) - F_Z(X_{(i-m)})) \frac{f_{X,i}}{f_{ZX,i}} \\
&\cong \frac{1}{2m} \sum_{i=1}^{n} (F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})) \frac{f_{X,i}}{f_{ZX,i}},
\end{aligned}$$

where

$$F_{Z(n+k)}(u) = \frac{1}{n+k} \left(\sum_{i=1}^{n} I(X_i \leq u) + \sum_{j=1}^{k} I(Y_j \leq u) \right),$$

the empirical distribution function, estimates the distribution $F_Z(u)$ ($I(\cdot)$ is the indicator function). Thus, the condition (3.4) has the empirical form of

$$\begin{aligned}
&\tilde{\Delta}_m \leq 1, \\
&\tilde{\Delta}_m = \frac{1}{2m} \sum_{i=1}^{n} (F_{Z(n+k)}(X_{(i+m)}) \\
&\qquad - F_{Z(n+k)}(X_{(i-m)})) \frac{f_{X,i}}{f_{ZX,i}}.
\end{aligned} \tag{3.5}$$

Consequently, under the empirical constraint (3.5), the Lagrangian function of the relevant log EL is $\log(\prod_{i=1}^{n} f_{X,i}) + \lambda(1 - \tilde{\Delta}_m)$, where λ is a Lagrange multiplier. Then, values of $f_{X,i}$ satisfy the equation

$$f_{X,i} = \frac{2m}{n(F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)}))} f_{ZX,i},$$

maximizing $\sum_{i=1}^{n} \log(f_{X,i})$ given the constraint (3.5) (here $X_{(i)} = X_{(1)}$, if $i \leq 1$, and $X_{(i)} = X_{(n)}$, if $i \geq n$). This result implies that the entropy-based empirical likelihood estimator of the ratio $\prod_{i=1}^{n} f_{X,i}/f_{ZX,i}$ is

$$ELR_{X,m,n} = \prod_{i=1}^{n} \frac{2m}{n(F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)}))}. \tag{3.6}$$

The method mentioned above demonstrates the test statistic $ELR_{X,m,n}$ is an approximation to the optimal likelihood ratio. Thus, we expect directly that a test based on $ELR_{X,m,n}$ will provide highly efficient characteristics. The distribution of the statistic $ELR_{X,m,n}$ depends strongly on values of the integer parameter m. We improve the statistic $ELR_{X,m,n}$ in the context of eliminating dependence on the integer parameter m reconsidering the construction of the approximation to the likelihood ratio with respect to the EL concept. By virtue of the relevant arguments, outlined in Appendix A,

we modify the approximation (3.6) to be presented in the following form

$$ELR_{X,n} = \min_{a_n \le m \le b_n} \prod_{i=1}^{n} \frac{2m}{n(F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)}))},$$
$$a_n = n^{0.5+\delta}, b_n = \min\left(n^{1-\delta}, \frac{n}{2}\right),$$
$$\delta \in (0, 0.25). \tag{3.7}$$

Similarly, we obtain that the density-based EL estimator of the ratio $\prod_{j=1}^{k} f_{Y,j}/f_{ZY,j}$ from (3.2) has the form of

$$ELR_{Y,k} = \min_{a_k \le r \le b_k} \prod_{i=1}^{k} \frac{2r}{k(F_{Z(n+k)}(Y_{(i+r)}) - F_{Z(n+k)}(Y_{(i-r)}))},$$
$$a_k = k^{0.5+\delta}, b_k = \min\left(k^{1-\delta}, \frac{k}{2}\right), \delta \in (0, 0.25), \tag{3.8}$$

where the empirical distribution function

$$F_{Z(n+k)}(u) = \frac{1}{n+k}\left(\sum_{i=1}^{n} I(X_i \le u) + \sum_{j=1}^{k} I(Y_j \le u)\right)$$

and $Y_{(j)} = Y_{(1)}$, if $j \le 1$; $Y_{(j)} = Y_{(k)}$, if $j \ge k$. Therefore, (3.7) and (3.8) lead to the maximum EL ratio test-statistic

$$V_{nk} = ELR_{X,n} ELR_{Y,k} \tag{3.9}$$

that approximates the optimal likelihood ratio test statistic (3.2).

4 The proposed test

As stated in Sect. 3, our nonparametric test for (3.1) utilizes the EL approximation (3.9) to the parametric likelihood ratio (3.2). The proposed test is to reject the null hypothesis of (3.1) iff

$$\log(V_{nk}) > C, \tag{4.1}$$

where C is a test-threshold. (Similarly to Canner 1975, we will arbitrarily define $F_{Z(n+k)}(x) - F_{Z(n+k)}(y) = 1/(n+k)$, if $F_{Z(n+k)}(x) = F_{Z(n+k)}(y)$.)

The next proposition indicates that the test (4.1) is consistent as $n, k \to \infty, n/k \to \eta$, where a constant $\eta > 0$. To formulate the following result, we assume that F_X and F_Y, mentioned in the statement (3.1), are the continuous cumulative distribution functions with density functions f_X and f_Y, respectively.

Proposition 4.1 *If the expectations* $E(\log f_X(X_1))$, $E(\log f_X(Y_1))$, $E(\log f_Y(Y_1))$ *and* $E(\log f_Y(X_1))$ *are finite, then*

$$\frac{1}{n+k}\log(V_{nk}) \xrightarrow{P} \gamma, \quad as\ n, k \to \infty,$$

where

$$\gamma = -\frac{\eta}{1+\eta} E\left(\log\left(\frac{\eta}{1+\eta} + \frac{1}{1+\eta}\frac{f_Y(X_1)}{f_X(X_1)}\right)\right)$$
$$- \frac{1}{1+\eta} E\left(\log\left(\frac{1}{1+\eta} + \frac{\eta}{1+\eta}\frac{f_X(Y_1)}{f_Y(Y_1)}\right)\right)$$

and $n/k \to \eta, \eta > 0$ *is a constant.*

Proof See Appendix B. □

It is clear that, under the null hypothesis, the ratio $f_Y/f_X = 1$ that implies $\gamma = 0$. Under the alternative H_1, we have $E(f_Y(X_1)/f_X(X_1)) = E(f_X(Y_1)/f_Y(Y_1)) = 1$ that implies

$$\gamma \ge -\frac{\eta}{1+\eta}\left(\log\left(\frac{\eta}{1+\eta} + \frac{1}{1+\eta}E\frac{f_Y(X_1)}{f_X(X_1)}\right)\right)$$
$$- \frac{1}{1+\eta}\left(\log\left(\frac{1}{1+\eta} + \frac{\eta}{1+\eta}E\frac{f_X(Y_1)}{f_Y(Y_1)}\right)\right) = 0.$$

Thus, the consistency of the proposed density-based EL test (4.1) is given by Proposition 4.1.

Critical values of the proposed test Our test statistic is based on indicator functions involved in the definition of the empirical distribution function $F_{Z(n+k)}(u)$. Since $I(X > Y) = I(F_X(X) > F_X(Y))$, we have

$$P_{H_0}\{\log(V_{nk}) > C\}$$
$$= P_{X_1,\ldots,X_n,Y_1,\ldots,Y_k \sim UNIF(0,1)}\{\log(V_{nk}) > C\}.$$

Thus, the type I error of the test (4.1) can be calculated exactly, for all sample sizes n, k and $0 < \delta < 0.25$. Note that, a very substantial body of literature has now grown around the asymptotic distribution problems involving the Vasicek's entropy estimator and the analogous statistics (e.g., Dudewicz and Van Der Meulen 1981; van Es 1992). However, it is generally recognized that even the asymptotic distribution of the statistic T_{mk} by (2.4), which can depend on estimates of nuisance parameters of f_0, is analytically difficult. (We can also assume that various relevant tests based on large samples provide relatively equivalent and powerful outputs.) Certain lines of research, developed around nonparametric two-sample comparisons, display different advantages of exact decision rules with nonasymptotic critical values. Thus, following the recent literature related to nonparametric tests (e.g., Canner 1975; Hall and Welsh 1983; Mudholkar and Tian 2002, 2004), we will not attempt to provide here an analytical solution for the critical values for the test (4.1), whereas the Monte Carlo approach will be used to calculate these critical values. We conducted a broad Monte Carlo study to investigate the power of the proposed test under various alternatives (particularly considered in the next section)

Table 1 The critical values of the proposed test (4.1) with $\delta = 0.1$ at the different significance levels α

n	α	k												
		10	15	20	25	30	35	40	50	60	80	100	150	200
10	0.01	11.535	12.222	12.912	13.712	14.380	14.604	15.208	15.940	17.228	18.592	19.986	22.435	24.955
	0.03	10.482	11.242	11.958	12.695	13.413	13.589	14.213	14.933	16.187	17.524	18.830	21.317	23.824
	0.05	9.763	10.497	11.223	11.910	12.605	12.776	13.389	14.120	15.367	16.683	17.975	20.489	22.988
	0.1	9.042	9.748	10.464	11.123	11.811	11.942	12.535	13.250	14.490	15.806	17.083	19.594	22.092
	0.15	8.601	9.320	10.021	10.659	11.337	11.428	12.038	12.735	13.955	15.256	16.538	19.049	21.534
15	0.01		12.899	13.517	14.161	14.893	15.095	15.773	16.567	17.846	19.195	20.443	23.126	25.597
	0.03		11.880	12.531	13.199	13.882	14.069	14.726	15.509	16.790	18.110	19.406	21.968	24.501
	0.05		11.128	11.779	12.432	13.145	13.271	13.942	14.670	15.947	17.257	18.524	21.133	23.629
	0.1		10.406	11.066	11.721	12.402	12.497	13.125	13.853	15.069	16.372	17.680	20.241	22.728
	0.15		9.997	10.661	11.315	11.966	12.039	12.675	13.362	14.588	15.867	17.157	19.686	22.173
20	0.01			14.072	14.724	15.411	15.671	16.240	17.135	18.253	19.694	21.139	23.626	26.238
	0.03			13.146	13.785	14.443	14.660	15.241	16.064	17.248	18.619	19.997	22.571	25.151
	0.05			12.424	13.071	13.732	13.897	14.503	15.259	16.464	17.818	19.167	21.737	24.301
	0.1			11.724	12.381	13.042	13.144	13.778	14.490	15.703	16.996	18.317	20.839	23.388
	0.15			11.325	11.984	12.635	12.716	13.348	14.021	15.238	16.517	17.819	20.328	22.841
25	0.01				15.485	16.108	16.206	16.777	17.525	18.885	20.242	21.650	24.396	26.797
	0.03				14.473	15.083	15.217	15.817	16.540	17.848	19.204	20.577	23.202	25.728
	0.05				13.741	14.402	14.509	15.080	15.808	17.079	18.429	19.765	22.321	24.885
	0.1				13.050	13.725	13.800	14.388	15.107	16.345	17.658	18.961	21.482	23.996
	0.15				12.643	13.309	13.378	13.978	14.675	15.912	17.209	18.502	20.993	23.495
30	0.01					16.567	16.842	17.400	18.250	19.432	20.882	22.171	24.878	27.458
	0.03					15.686	15.871	16.446	17.173	18.448	19.884	21.140	23.790	26.351
	0.05					14.989	15.158	15.753	16.459	17.719	19.093	20.360	22.983	25.477
	0.1					14.321	14.465	15.052	15.761	16.998	18.339	19.604	22.157	24.622
	0.15					13.918	14.041	14.640	15.337	16.565	17.884	19.126	21.671	24.135
35	0.01						16.869	17.497	18.209	19.462	20.895	22.212	25.089	27.589
	0.03						15.970	16.549	17.296	18.529	19.903	21.232	23.974	26.503
	0.05						15.277	15.868	16.581	17.804	19.159	20.501	23.144	25.647
	0.1						14.567	15.175	15.866	17.113	18.431	19.723	22.301	24.805
	0.15						14.151	14.759	15.439	16.686	17.983	19.272	21.826	24.321
40	0.01							18.110	18.848	20.094	21.531	22.804	25.477	28.240
	0.03							17.191	17.935	19.152	20.507	21.837	24.478	27.140
	0.05							16.517	17.215	18.444	19.768	21.094	23.710	26.316
	0.1							15.812	16.484	17.742	19.043	20.356	22.913	25.458
	0.15							15.396	16.082	17.322	18.621	19.905	22.440	24.974
50	0.01								19.593	20.803	22.209	23.454	26.205	28.810
	0.03								18.621	19.870	21.257	22.522	25.220	27.793
	0.05								17.935	19.200	20.547	21.807	24.443	27.002
	0.1								17.220	18.463	19.822	21.061	23.678	26.221
	0.15								16.798	18.021	19.367	20.620	23.220	25.731
60	0.01									21.958	23.349	24.769	27.353	30.097
	0.03									21.026	22.442	23.780	26.403	29.031
	0.05									20.349	21.748	23.093	25.654	28.245
	0.1									19.650	21.034	22.341	24.911	27.474
	0.15									19.217	20.588	21.892	24.443	27.011

Table 1 (*Continued*)

n	α	k											
		10	15	20	25	30	35	40	50	60	80	100	150	200
80	0.01										24.689	26.033	28.703	31.416
	0.03										23.787	25.112	27.819	30.423
	0.05										23.084	24.413	27.091	29.683
	0.1										22.376	23.678	26.328	28.884
	0.15										21.926	23.226	25.863	28.416
100	0.01											27.389	30.147	32.602
	0.03											26.438	29.172	31.720
	0.05											25.729	28.379	30.970
	0.1											25.004	27.656	30.220
	0.15											24.533	27.194	29.743
150	0.01												32.770	35.426
	0.03												31.822	34.451
	0.05												31.068	33.682
	0.1												30.319	32.894
	0.15												29.864	32.411
200	0.01													37.930
	0.03													37.025
	0.05													36.295
	0.1													35.509
	0.15													35.019

and with different values of δ appeared in the definition (4.1) via (3.7)–(3.9). The Monte Carlo power of the proposed test was not found to be significantly depend on values of δ. In this article we focus on $\delta = 0.1$ applied to the definition (4.1). Table 1 presents Monte Carlo roots C_α of the equations $P_{X_1,\ldots,X_n,Y_1,\ldots,Y_k\sim UNIF(0,1)}\{\log(V_{nk}) > C_\alpha\} = \alpha$, based on 55,000 samples of size n and k, for different values of α.

5 Monte Carlo study

In this section, we investigate the power properties of the proposed test (4.1) comparing with the commonly used two-sample Kolmogorov-Smirnov (KS) test (Birnbaum and Hall 1960; Massey 1951), Wilcoxon rank sum test and t-test. Different statistical publications have introduced various nonparametric two-sample tests, commonly comparing them with the KS test. Thus, Monte Carlo results conducted in this section can be linked to those presented in the literature. To evaluate the properties of test (4.1), we conduct the following Monte Carlo simulations. For different values of n, k and δ included in the definition (4.1), 25,000 pairs of samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_k$ were generated from various alternative distributions corresponding to the testing problem (3.1). The powers of the tests are shown in Table 2, at the $\alpha = 0.05$ level of significance. Table 2 does not display results of all simulations that we executed. We present typical situations with respect to the designs A-H depicted in Table 2.

The power demonstrated by the proposed test is relatively the same for different values of $\delta \in (0, 0.25)$. Table 2 confirms that for relatively small and average sample sizes n and k the test (4.1) is a very efficient decision rule. The classical Wilcoxon test and t-test are known to be very powerful policies for the standard two-sample problem with a constant shift in location, especially when data follow normal distributions (the designs C and D presented in Table 2 show these cases). In these cases the proposed test presented the powers that are less than those of the classical procedure, especially when $F_X = N(0, 1)$, $F_y = N(0.5, 1)$. Under the design D, the density-based EL test provided characteristics that are approximately equivalent to those of the Wilcoxon's test and t-test. (Under the designs C and D, the t-test is expected to be most powerful.) The literature indicates that, for many applications, the standard two-sample problem with just a constant shift in location is far from realistic (e.g., Albers et al. 2001). The simulation study shows that the new test has high and stable power, detecting nonconstant shift alternatives, when the classical procedures break down completely (e.g., designs A, E, F and H). Note that entropy-

Table 2 The Monte Carlo powers of the tests: (4.1) with different values of δ, KS test, Wilcoxon rank sum test, and t-test; at $\alpha = 0.05$. For each sample sizes (n, k), the designs A-H display simulation studies based on samples from $A : F_X = N(0,1)$, $F_y = \text{Unif}[-1,1]$; $B : F_X = \text{Exp}(1)$, $F_y = \text{LogNorm}(0,1)$; $C : F_X = N(0,1)$, $F_y = N(0.5,1)$; $D : F_X = N(0,1)$, $F_y = N(1.5,1)$; $E : F_X = N(0,1)$, $F_y = N(0,1.5^2)$; $F : F_X = N(0,1)$, $F_y = N(0,0.5^2)$; $G : F_X = \text{Exp}(1)$, $F_y = \text{Exp}(1.5)$; $H : F_X = \text{Beta}(0.7,1)$, $F_y = \text{Exp}(2)$

Design	n	k	Proposed test (4.1) δ						KS test	Wilcoxon test	t-test
			0.025	0.05	0.1	0.12	0.15	0.2			
A											
	45	45	0.9533	0.9490	0.9565	0.9486	0.9525	0.9582	0.1461	0.0564	0.0529
	15	25	0.2852	0.3138	0.2782	0.2726	0.2613	0.2577	0.0901	0.0631	0.0501
	25	15	0.4069	0.3673	0.3717	0.3475	0.3531	0.3421	0.0622	0.0426	0.0490
	15	15	0.2388	0.2390	0.2152	0.2240	0.2022	0.1957	0.0349	0.0442	0.0466
B											
	45	45	0.5872	0.5865	0.5960	0.5828	0.5848	0.5794	0.3388	0.4936	0.4814
	10	10	0.1318	0.1316	0.1324	0.1263	0.1290	0.1313	0.1176	0.1256	0.0792
C											
	45	45	0.4999	0.4989	0.5291	0.5140	0.5209	0.5290	0.5131	0.6294	0.6503
	15	15	0.2019	0.2046	0.2178	0.2086	0.2076	0.2094	0.1405	0.2373	0.2592
D											
	25	15	0.9757	0.9778	0.9819	0.9776	0.9810	0.9804	0.9927	0.9913	0.9935
	15	15	0.9374	0.9392	0.9471	0.9511	0.9531	0.9524	0.8849	0.9698	0.9786
E											
	45	45	0.5026	0.4994	0.4881	0.4531	0.4427	0.4141	0.1285	0.0523	0.0499
	35	35	0.3825	0.3887	0.3553	0.3502	0.3377	0.3054	0.0734	0.0515	0.0485
F											
	45	45	0.9428	0.9398	0.9343	0.9251	0.9187	0.9021	0.3756	0.0593	0.0515
G											
	45	45	0.3168	0.3174	0.3507	0.3340	0.3394	0.3417	0.3041	0.3761	0.4564
	15	15	0.1347	0.1313	0.1475	0.1420	0.1410	0.1422	0.0859	0.1401	0.1483
H											
	45	45	0.5114	0.4704	0.4498	0.4447	0.4442	0.4460	0.0630	0.0575	0.1323

based tests for goodness-of-fit are very powerful for detecting a change towards small variance (Dudewicz and Van Der Meulen 1981). It seems that the test (4.1) has also a high level of the power when observed samples have different variances. In accordance with Table 2, almost for all considered situations, the test (4.1) can be recommended to be applied.

6 A coronary heart disease example

We illustrate the proposed approach using data from a study on atherosclerotic process of coronary heart disease (CHD). Free radicals have been implicated in the atherosclerotic process of CHD. Well-developed laboratory methods may make available a large number of biomarkers of individual oxidative stress and antioxidant status. Such markers are able to quantify different phases of the oxidative stress and antioxidant status of an individual. As pointed by many researchers (e.g., Schisterman et al. 2001; Vexler et al. 2006), among all the biomakers, TBARS were found to be a good discriminating between individuals with and without CHD. A population-based sample of randomly selected residents of Erie and Niagara counties of New York State, USA, 35–79 years of age, was the focus of this study. To illustrate the proposed approach, we obtain a subset of this dataset with $k = 25$ cases (say, Y's) and $n = 50$ controls (say, X's), where X's and Y's are values of TBARS corresponding to individuals with and without CHD, respectively. Figure 1 depicts

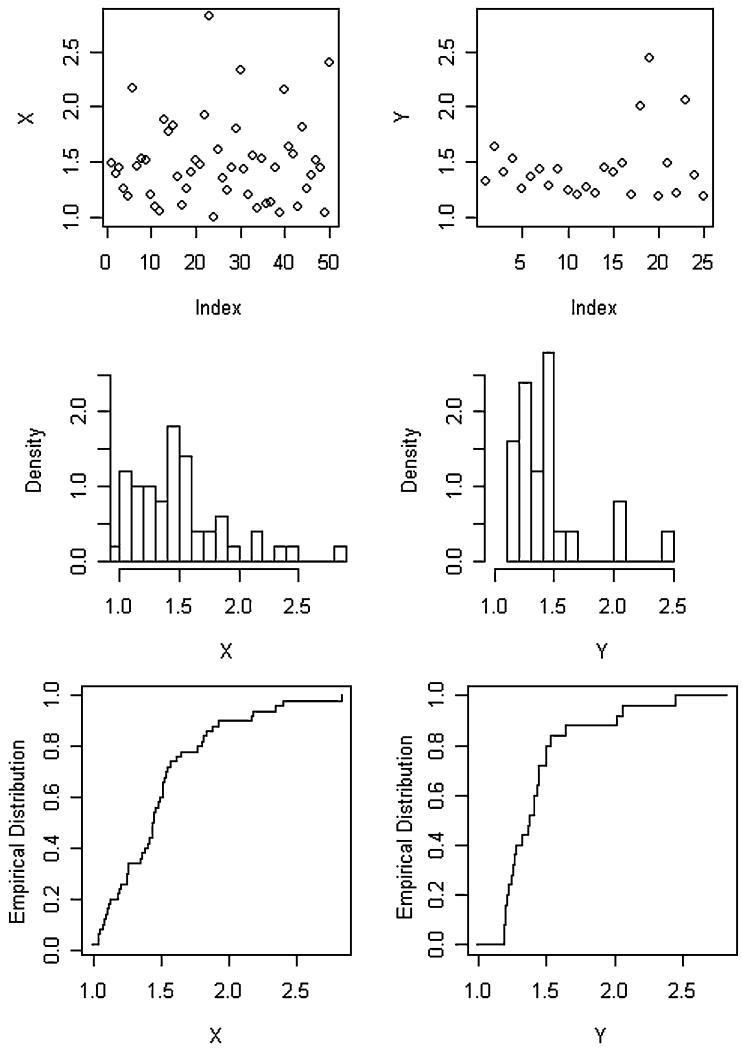

Fig. 1 The graphical presentation of the dataset from the CHD example

the observations with mean$(X) = 1.4935$, sd$(X) = 0.3876$ and mean$(Y) = 1.44384$, sd$(Y) = 0.30779$.

(Note that, the Shapiro-Wilk test for normality provides p-values 0.0002750, 0.06199, 3.891e-05 and 0.0005234 based on X, $\log(X)$, Y and $\log(Y)$, respectively.)

In the diagnostic medicine literature, the area under the so-called receiver operating characteristic (ROC) curve is an important measure of diagnostic accuracy of biomarkers (see, for example, a list of publications mentioned in Vexler et al. 2006). Values of the area under the ROC curve close to 1 indicate that the marker has high diagnostic accuracy, while a value of 0.5 indicates a noninformative marker that does no better than a random (fair) coin toss. Figure 2 shows that the nonparametric estimator of the area under the ROC curve has a value that is not significantly different from 0.5. Therefore, this method based on the dataset of this example does not indicate a difference between distributions of X and Y.

The two-sample KS test, Wilcoxon rank sum test (two sided), t-test (two sided) report p-values $= 0.3953$, 0.5589, 0.5491, respectively (i.e., the null hypothesis of (3.1) is not rejected by the classical procedures). The value of the test-statistic from (4.1) with $\delta = 0.1$ is 17.1645. Thus, in accordance with Table 1, the proposed decision rule rejects the null hypotheses of (3.1) with a p-value <0.01. This result confirms the epidemiological fact that the biomarker TBARS is distributed differently with respect to individuals with and without CHD.

7 Concluding remarks

We have outlined a general approach for constructing density-based empirical likelihood ratio tests that utilize samples entropy. The proposed development of the structures related to entropy-based tests differs from considerations mentioned in the literature. The new method improves entropy-based tests in the context of practical applications, since only tables of critical points are required for its implementation. (In the context of goodness-of-fit

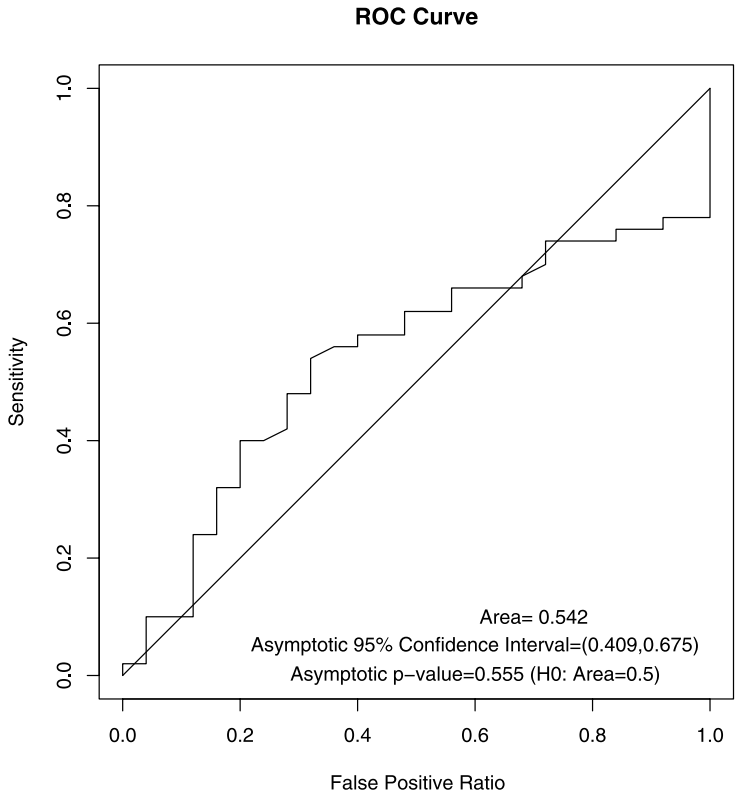

Fig. 2 The ROC curve analysis

tests, the statistical literature has well addressed entropy-based test-statistics that are defined up to an integer factor, advantageous values of which are unknown when the sample size is finite.) The proposed technique can provide powerful nonparametric approximations to parametric likelihood ratio tests related to various statements of hypothesis testing. Nonparametric procedures obtained using the approximations to relevant parametric likelihood ratios can be directly expected to be highly efficient. While considering the approach of this article, k-sample entropy-based tests can be easily constructed. We have only extensively examined the two-sample nonparametric test. We showed that the proposed approach is versatile and leads to the powerful nonparametric test that, in many cases, is superior to the standard procedures. The Monte Carlo study, performed in this article, demonstrates that the new test has high and stable power, detecting nonconstant shift alternatives, when the classical procedures may break down completely. In the context of the problem (3.1), the proposed test results in a relatively small power loss in comparison to Wilcoxon's test and the t-test when the constant shift is dominant and a large power gain otherwise.

The proposed approach can be applied to create a test for (3.1) with one-sided alternatives. (In this case, the two-sample Mann-Whitney-Wilcoxon test is a common procedure.) In a subsequent article, we plan to address the use of information regarding ordered alternative hypotheses in the context of the nonparametric approximations of parametric likelihood functions, which needs substantial mathematical details.

Further studies are needed to test the approach in other contexts. We hope that this article will stimulate future theoretical and applied research on this topic.

Acknowledgements The authors are indebted to Professors Govind S. Mudholkar and Alan D. Hutson for many helpful discussions and comments. We are grateful to the Editor and the Associate Editor and an anonymous referee for their helpful comments that clearly improved this article. This work was partially supported by the Internal Funding Program of the Shamoon College of Engineering (SCE).

Appendix A: The approximation (3.7) to the likelihood ratio

Assume M defines a set of appropriate values of the integer parameter m in (3.6). If there exists an integer $r \in M$ such that $(1/2r)\sum_{i=1}^{n}\int_{X_{(i-r)}}^{X_{(i+r)}} f_X(u)du \geq 1$, then $\int_{X_{(1)}}^{X_{(n)}} f_X(u)du \geq 1$ by virtue of (3.3). This is unacceptable. Therefore, we require that $f_{X,i}, i = 1, \ldots, n$ are subject to

(3.5), for all possible $m \in M$, i.e. for a fix integer $m_0 \in M$, the approximation to the H_1-likelihood $\prod_{i=1}^{n} f_{X,i}$ satisfies

$$\max_{f_{X,1},\ldots,f_{X,n} \text{ subject to } \tilde{\Delta}_m \le 1, \text{ for all } m \in M} \prod_{i=1}^{n} f_{X,i} \le \max_{f_{X,1},\ldots,f_{X,n} \text{ subject to } \tilde{\Delta}_m \le 1 \text{ with } m = m_0 \in M} \prod_{i=1}^{n} f_{X,i}.$$

Applying the operator $\min_{m_0}$ to both the sides of this inequality, we conclude that

$$\max_{f_{X,1},\ldots,f_{X,n} \text{ subject to } \tilde{\Delta}_m \le 1, \text{for all } m \in M} \prod_{i=1}^{n} f_{X,i} \le \min_{m_0} \max_{f_{X,1},\ldots,f_{X,n} \text{ subject to } \tilde{\Delta}_m \le 1 \text{ with } m = m_0 \in M} \prod_{i=1}^{n} f_{X,i}. \tag{A.1}$$

Now, since the constraint (3.5) is an approximation to (3.4), assuming $f_{X,i}, i = 1, \ldots, n$ satisfy (3.5) with $m = r$, we can write

$$\begin{aligned} 1 \ge \tilde{\Delta}_r &= \frac{1}{2r} \sum_{i=1}^{n} \left(F_{Z(n+k)}\left(X_{(i+r)}\right) - F_{Z(n+k)}\left(X_{(i-r)}\right) \right) \frac{f_{X,i}}{f_{ZX,i}} \\ &\approx \frac{1}{2r} \sum_{i=1}^{n} \int_{X_{(i-r)}}^{X_{(i+r)}} f_X(u)du \\ &\approx \int_{X_{(1)}}^{X_{(n)}} f_X(u)du \approx \frac{1}{2s} \sum_{i=1}^{n} \int_{X_{(i-s)}}^{X_{(i+s)}} f_X(u)du \approx \tilde{\Delta}_s, \end{aligned}$$

i.e., $f_{X,i}, j = 1, \ldots, n$ can be expected to satisfy (3.5), for $m = s$ too. Thus, we can expect that, for some m_0,

$$\begin{aligned} &\max_{f_{X,1},\ldots,f_{X,n} \text{ subject to } \tilde{\Delta}_m \le 1, \text{for all } m \in M} \prod_{i=1}^{n} f_{X,i} \\ &\approx \max_{f_{X,1},\ldots,f_{X,n} \text{ subject to } \tilde{\Delta}_m \le 1 \text{ with } m = m_0 \in M} \prod_{i=1}^{n} f_{X,i} \\ &\ge \min_{m_0} \max_{f_{X,1},\ldots,f_{X,n} \text{ subject to } \tilde{\Delta}_m \le 1 \text{ with } m = m_0 \in M} \prod_{i=1}^{n} f_{X,i}. \end{aligned} \tag{A.2}$$

The deductions (A.1) and (A.2) imply that the maximum EL should be defined as

$$\min_{m \in M} \max_{f_{X,1},\ldots,f_{X,n} \text{ subject to } \tilde{\Delta}_m \le 1} \prod_{i=1}^{n} f_{X,i}. \tag{A.3}$$

In (3.3), we have the remainder term

$$\begin{aligned} &- \sum_{l=1}^{m-1} (m-l) \int_{X_{(n-l)}}^{X_{(n-l+1)}} f_X(u)du \\ &\quad - \sum_{l=1}^{m-1} (m-l) \int_{X_{(l)}}^{X_{(l+1)}} f_X(u)du \\ &= - \sum_{l=1}^{m-1} (m-l) \left(\int_{X_{(n-l)}}^{X_{(n-l+1)}} f_X(u)du \right. \\ &\quad \left. + \int_{X_{(l)}}^{X_{(l+1)}} f_X(u)du \right) \\ &= - \sum_{l=1}^{m-1} (m-l) \left(F_X(X_{(n-l+1)}) - F_X(X_{(n-l)}) \right. \\ &\quad \left. + F_X(X_{(l+1)}) - F_X(X_{(l)}) \right) \\ &\approx - \sum_{l=1}^{m-1} (m-l) (F_{Xn}(X_{(n-l+1)}) - F_{Xn}(X_{(n-l)}) \\ &\quad + F_{Xn}(X_{(l+1)}) - F_{Xn}(X_{(l)})) \\ &= \frac{m(m-1)}{n}, \end{aligned}$$

$$F_{Xn}(u) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le u).$$

In the context of the constraint (3.4), to minimize the influence of the remainder term above, we need to require that $(m(m-1)/n)/(2m)$ should be vanished as $n \to \infty$, say, $m = n^{1-\delta}$ and $M = \{m : m \le n^{1-\delta}\}, \delta > 0$. This conditional bound on m can be directly associated with restrictions used to show the consistency of entropy based tests for goodness-of-fit (e.g., Vasicek 1976; Tusnady 1977; Vexler and Gurevich 2010). In this article, Proposition 4.1 presents an asymptotic efficiency of the proposed test statistic. The proof scheme of this proposition utilizes the lower bound $m \ge n^{0.5+\delta}$ to preserve the consistency of the density-based EL test. Thus, we specify the set $M = \{m : n^{0.5+\delta} \le m \le n^{1-\delta}\}, 0 < \delta < 0.25$ in the definition (A.3).

Appendix B: Proof of Proposition 4.1

By virtue of the definitions (3.6)–(3.9), (4.1), we can reformulate the test statistic as

$$\log(V_{nk}) = \log \min_{n^{0.5+\delta} \le m < n^{1-\delta}} V^1_{nkm} + \log \min_{k^{0.5+\delta} \le r < k^{1-\delta}} V^2_{nkr}, \tag{B.1}$$

where

$$\log\left(V_{nkm}^{1}\right) = \log\prod_{i=1}^{n}\frac{2m}{n(F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)}))} = -\sum_{i=1}^{n}\log\frac{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})}{2m/n},$$

$$\log\left(V_{nkr}^{2}\right) = \log\prod_{j=1}^{k}\frac{2r}{k(F_{Z(n+k)}(Y_{(j+r)}) - F_{Z(n+k)}(Y_{(j-r)}))} = -\sum_{j=1}^{k}\log\frac{F_{Z(n+k)}(Y_{(j+r)}) - F_{Z(n+k)}(Y_{(j-r)})}{2r/k}$$

and $0 < \delta < 0.25$.

Note that

$$F_{Z(n+k)}(u) = \frac{1}{n+k}\left(\sum_{i=1}^{n} I(X_i \le u) + \sum_{j=1}^{k} I(Y_j \le u)\right) = \frac{1}{n+k}(nF_{Xn}(u) + kF_{Yk}(u)),$$

where $F_{Xn}(x) = n^{-1}\sum_{i=1}^{n} I(X_i \le x)$ and $F_{Yk}(x) = k^{-1}\sum_{i=1}^{k} I(Y_i \le x)$ are the empirical distribution functions based on $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_k$, respectively.

For the term $\log(V_{nkm}^{1})$ in (B.1), we have

$$\begin{aligned}
&-\sum_{i=1}^{n}\log\frac{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})}{2m/n}\\
&= -\sum_{i=1}^{n}\log\frac{F_{nk}^{*}(X_{(i+m)}) - F_{nk}^{*}(X_{(i-m)})}{F_X(X_{(i+m)}) - F_X(X_{(i-m)})}\\
&\quad + \sum_{i=1}^{n}\log\frac{F_{nk}^{*}(X_{(i+m)}) - F_{nk}^{*}(X_{(i-m)})}{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})}\\
&\quad - \sum_{i=1}^{n}\log\frac{F_X(X_{(i+m)}) - F_X(X_{(i-m)})}{2m/n},
\end{aligned} \tag{B.2}$$

where $F_{nk}^{*}(x) = (1/(n+k))(nF_X(x) + kF_Y(x))$. Defining $F^{*}(x) = (\eta F_X(x) + F_Y(x))/(1+\eta)$, we consider the first term in the right side of (B.2)

$$\begin{aligned}
&\sum_{i=1}^{n}\log\frac{F_{nk}^{*}(X_{(i+m)}) - F_{nk}^{*}(X_{(i-m)})}{F_X(X_{(i+m)}) - F_X(X_{(i-m)})}\\
&= \sum_{i=1}^{n}\log\frac{F_{nk}^{*}(X_{(i+m)}) - F_{nk}^{*}(X_{(i-m)})}{X_{(i+m)} - X_{(i-m)}}\\
&\quad - \sum_{i=1}^{n}\log\frac{F_X(X_{(i+m)}) - F_X(X_{(i-m)})}{X_{(i+m)} - X_{(i-m)}}.
\end{aligned} \tag{B.3}$$

Following the proof scheme of Theorem 1 presented by Vasicek (1976), we apply some reorganization to represent

$$(n+k)^{-1}\sum_{i=1}^{n}\log\frac{F_{nk}^{*}(X_{(i+m)}) - F_{nk}^{*}(X_{(i-m)})}{X_{(i+m)} - X_{(i-m)}} = \frac{n}{n+k}(2m)^{-1}\sum_{j=1}^{2m} S_j, \tag{B.4}$$

where

$$S_j = \sum_{i=1}^{n}\log\frac{F_{nk}^{*}(X_{(i+m)}) - F_{nk}^{*}(X_{(i-m)})}{X_{(i+m)} - X_{(i-m)}} \times \{F_{Xn}(X_{(i+m)}) - F_{Xn}(X_{(i-m)})\}, \quad i \equiv j \,(\mathrm{mod}\, 2m).$$

Assume $X_{(i-m)}, X_{(i+m)}$ belong to an interval in which

$$f_{nk}^{*}(x) = \frac{dF_{nk}^{*}(x)}{dx} = \frac{n}{n+k}f_X(x) + \frac{k}{n+k}f_Y(x)$$

is a positive and continuous function. Then there exists $X_i^{*} \in (X_{(i-m)}, X_{(i+m)})$ such that

$$\frac{F_{nk}^{*}(X_{(i+m)}) - F_{nk}^{*}(X_{(i-m)})}{X_{(i+m)} - X_{(i-m)}} = f_{nk}^{*}(X_i^{*}).$$

This rewrites

$$S_j = \sum_{i=1}^{n}\log f_{nk}^{*}(X_i^{*})\{F_{Xn}(X_{(i+m)}) - F_{Xn}(X_{(i-m)})\}, \quad i \equiv j \,(\mathrm{mod}\, 2m).$$

To approximate the function $f_{nk}^{*}(x)$, define the density function

$$f^{*}(x) = \frac{dF^{*}(x)}{dx} = \frac{\eta}{1+\eta}f_X(x) + \frac{1}{1+\eta}f_Y(x).$$

Since $n/k \to \eta$, we have the inequality $(1-\varepsilon)f^{*}(X_i^{*}) \le f_{nk}^{*}(X_i^{*}) \le (1+\varepsilon)f^{*}(X_i^{*})$, for each $\varepsilon > 0$ and sufficiently large n and k. This implies

$$S_j^{-\varepsilon} \le S_j \le S_j^{\varepsilon},$$

$$S_j^{\varepsilon} = \sum_{i=1}^{n}\log((1+\varepsilon)f^{*}(X_i^{*})) \times \{F_{Xn}(X_{(i+m)}) - F_{Xn}(X_{(i-m)})\}, \quad i \equiv j \,(\mathrm{mod}\, 2m),$$

for sufficiently large n and k. That is, S_j^{ε} is a Stieltjes sum of the function $\log((1+\varepsilon)f^{*}(x))$ with respect to the measure

F_{Xn} over the sum of intervals of continuity of $f_X(x)$ and $f_Y(x)$ in which $f^*(x) > 0$.

Since $X_{(i+m)} - X_{(i-m)} \to 0$ a.s. uniformly over $m \in [n^{0.5+\delta}, n^{1-\delta}]$ and $F_{Xn}(x) \to F_X(x)$ uniformly over x as $n \to \infty$, following the proof of Theorem 1 presented in Vasicek (1976, p. 56), we obtain that S_j^{ε} converges in probability to

$$\int_{-\infty}^{\infty} \log\left((1+\varepsilon) f^*(x)\right) dF_X(x) = E\left(\log\left((1+\varepsilon) f^*(X_1)\right)\right)$$

as $n \to \infty$, uniformly over $m \in [n^{0.5+\delta}, n^{1-\delta}]$ and over j. Consequently,

$$E\left(\log\left((1-\varepsilon) f^*(X_1)\right)\right) \le (2m)^{-1} \sum_{j=1}^{2m} S_j \le E\left(\log\left((1+\varepsilon) f^*(X_1)\right)\right),$$

as $n \to \infty$, uniformly over $m \in [n^{0.5+\delta}, n^{1-\delta}]$. Thus, by (B.4), we have

$$(n+k)^{-1} \sum_{i=1}^{n} \log \frac{F_{nk}^*(X_{(i+m)}) - F_{nk}^*(X_{(i-m)})}{X_{(i+m)} - X_{(i-m)}} \xrightarrow{P} \frac{\eta}{1+\eta} E(\log f^*(X_1)) \tag{B.5}$$

as $n \to \infty, k \to \infty, n/k \to \eta, \eta > 0$, uniformly over $m \in [n^{0.5+\delta}, n^{1-\delta}]$.

Similarly, we obtain

$$(n+k)^{-1} \sum_{i=1}^{n} \log \frac{F_X(X_{(i+m)}) - F_X(X_{(i-m)})}{X_{(i+m)} - X_{(i-m)}} \xrightarrow{P} \frac{\eta}{1+\eta} E(\log f_X(X_1)) \tag{B.6}$$

as $n, k \to \infty, n/k \to \eta, \eta > 0$, uniformly over $m \in [n^{0.5+\delta}, n^{1-\delta}]$. Applying (B.5), (B.6) to (B.3), we conclude

$$\begin{aligned}(n+k)^{-1} \sum_{i=1}^{n} \log \frac{F_{nk}^*(X_{(i+m)}) - F_{nk}^*(X_{(i-m)})}{F_X(X_{(i+m)}) - F_X(X_{(i-m)})} &\xrightarrow{P} \frac{\eta}{1+\eta} E(\log f^*(X_1)) - \frac{\eta}{1+\eta} E(\log f_X(X_1)) \\ &= \frac{\eta}{1+\eta} E\left(\log \frac{f^*(X_1)}{f_X(X_1)}\right) \\ &= \frac{\eta}{1+\eta} E\left(\log\left(\frac{\eta}{1+\eta} + \frac{1}{1+\eta} \frac{f_Y(X_1)}{f_X(X_1)}\right)\right)\end{aligned} \tag{B.7}$$

as $n, k \to \infty, n/k \to \eta, \eta > 0$, uniformly over $n^{0.5+\delta} \le m \le n^{1-\delta}$.

Considering the term

$$\sum_{i=1}^{n} \log \frac{F_{nk}^*(X_{(i+m)}) - F_{nk}^*(X_{(i-m)})}{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})}$$

of (B.2), we note that, by virtue of the Theorem of Kolmogorov (e.g., Serfling 1980, p. 62), for each $0 < \varepsilon < \delta/4$, $P(\sup_{-\infty<x<\infty} |F_X(x) - F_{Xn}(x)| > n^{-0.5+\varepsilon}) \to 0$ as $n \to \infty$ and $P(\sup_{-\infty<x<\infty} |F_Y(x) - F_{Yk}(x)| > k^{-0.5+\varepsilon}) \to 0$ as $k \to \infty$. Therefore $P(\sup_{-\infty<x<\infty} |F_Y(x) - F_{Yk}(x)| > (2n/\eta)^{-0.5+\varepsilon}) \to 0$ as $n, k \to \infty, n/k \to \eta$, and hence $P(\sup_{-\infty<x<\infty} |F_{nk}^*(x) - F_{Z(n+k)}(x)| > n^{-0.5+2\varepsilon}) \to 0$ as $n, k \to \infty$. Thus, we can focus on situations, when $\sup_{-\infty<x<\infty} |F_{nk}^*(x) - F_{Z(n+k)}(x)| \le n^{-0.5+2\varepsilon}$ and we can use the inequality $F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)}) \ge n(n+k)^{-1} 2m/n = 2m/(n+k)$, by virtue of the definition of $F_{Z(n+k)}$. This leads to the inequalities

$$\begin{aligned}&\frac{1}{(n+k)} \sum_{i=1}^{n} \log \frac{F_{nk}^*(X_{(i+m)}) - F_{nk}^*(X_{(i-m)})}{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})} \\ &\quad \le \frac{1}{(n+k)} \times \sum_{i=1}^{n} \log \frac{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)}) + n^{-0.5+\delta/2}}{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})} \\ &\quad \le (n+k)^{-1} \sum_{i=1}^{n} \log\left(1 + \frac{n^{-0.5+\delta/2}}{2m/(n+k)}\right) \\ &\quad \le (n+k)^{-1} \sum_{i=1}^{n} \frac{n^{-0.5+\delta/2}}{2n^{0.5+\delta}/(n+k)} \\ &\quad = \frac{n}{2} n^{-1-\delta/2} \to 0 \quad \text{as } n \to \infty;\end{aligned}$$

$$\begin{aligned}&\frac{1}{(n+k)} \sum_{i=1}^{n} \log \frac{F_{nk}^*(X_{(i+m)}) - F_{nk}^*(X_{(i-m)})}{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})} \\ &\quad \ge \frac{1}{(n+k)} \times \sum_{i=1}^{n} \log \frac{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)}) - n^{-0.5+\delta/2}}{F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})} \\ &\quad \ge (n+k)^{-1} \sum_{i=1}^{n} \log\left(1 - \frac{n^{-0.5+\delta/2}}{2m/(n+k)}\right) \\ &\quad \ge -(n+k)^{-1} \sum_{i=1}^{n} \frac{2n^{-0.5+\delta/2}}{2n^{0.5+\delta}/(n+k)} \\ &\quad = -\frac{n}{n^{1+\delta/2}} \to 0 \quad \text{as } n \to \infty.\end{aligned}$$

(To obtain the inequalities above we utilized the fact $m \ge n^{0.5+\delta}$.)

Thus, we get

$$(n+k)^{-1}\sum_{i=1}^{n}\log\frac{F^*_{nk}(X_{(i+m)})-F^*_{nk}(X_{(i-m)})}{F_{Z(n+k)}(X_{(i+m)})-F_{Z(n+k)}(X_{(i-m)})}\xrightarrow{P}0 \tag{B.8}$$

uniformly over $n^{0.5+\delta}\le m\le n^{1-\delta}$ as $n,k\to\infty, n/k\to\eta, \eta>0$.

Consider the term $-\sum_{i=1}^{n}\log\frac{F_X(X_{(i+m)})-F_X(X_{(i-m)})}{2m/n}$ of (B.2). We apply the proof of Lemma 1 presented in Vasicek (1976, p. 55) to conclude

$$-(n+k)^{-1}\sum_{i=1}^{n}\log\frac{F_X(X_{(i+m)})-F_X(X_{(i-m)})}{2m/n}\ge 0 \quad\text{and}$$
$$-(n+k)^{-1}\sum_{i=1}^{n}\log\frac{F_X(X_{(i+m)})-F_X(X_{(i-m)})}{2m/n}\xrightarrow{P}0, \tag{B.9}$$

uniformly over $m\in[n^{0.5+\delta}, n^{1-\delta}]$ as $n,k\to\infty$. Therefore, by (B.1), (B.2), (B.7)–(B.9),

$$(n+k)^{-1}\log\left(V^1_{nkm}\right)\xrightarrow{P}-\frac{\eta}{1+\eta}E\left(\log\left(\frac{\eta}{1+\eta}+\frac{1}{1+\eta}\frac{f_Y(X_1)}{f_X(X_1)}\right)\right)$$

uniformly over $n^{0.5+\delta}\le m\le n^{1-\delta}$ as $n,k\to\infty, n/k\to\eta, \eta>0$. Because the derivation of the asymptotic result

$$(n+k)^{-1}\log(V^2_{nkr})=-\sum_{j=1}^{k}\log\frac{F_{Z(n+k)}(Y_{(j+r)})-F_{Z(n+k)}(Y_{(j-r)})}{2r/k}$$
$$\xrightarrow{P}-\frac{1}{1+\eta}E\left(\log\left(\frac{1}{1+\eta}+\frac{\eta}{1+\eta}\frac{f_X(Y_1)}{f_Y(Y_1)}\right)\right)$$

(uniformly over $n^{0.5+\delta}\le m\le n^{1-\delta}$ as $n,k\to\infty$) is quite similar, the proof of Proposition 4.1 is complete.

References

Albers, W., Kallenberg, W.C.M., Martini, F.: Data-driven rank tests for classes of tail alternatives. J. Am. Stat. Assoc. **96**, 685–696 (2001)

Arizono, I., Ohta, H.: A test for normality based on Kullback-Leibler information. Am. Stat. **43**, 20–22 (1989)

Birnbaum, Z.W., Hall, R.A.: Small sample distributions for multi-sample statistics of the Smirnov type. Ann. Math. Stat. **31**, 710–720 (1960)

Canner, P.L.: A simulation study of one-and two-sample Kolmogorov-Smirnov statistics with a particular weight function. J. Am. Stat. Assoc. **70**, 209–211 (1975)

Dudewicz, E.J., Van Der Meulen, E.C.: Entropy-based tests of uniformity. J. Am. Stat. Assoc. **76**, 967–974 (1981)

Ebrahami, N., Habibullah, M., Soofi, E.S.: Testing exponentiality based on Kullback-Leibler information. J. R. Stat. Soc. B **54**, 739–748 (1992)

Hall, P., Welsh, A.H.: A test for normality based on the empirical characteristic function. Biometrika **70**, 485–489 (1983)

Lazar, N.A., Mykland, P.A.: An evaluation of the power and conditionality properties of empirical likelihood. Biometrika **85**, 523–534 (1998)

Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer, New York (2005)

Massey, F.: The distribution of the maximum deviation between two sample cumulative step functions. Ann. Math. Stat. **22**, 125–128 (1951)

Mudholkar, G.S., Tian, L.: An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test. J. Stat. Plan. Inference **102**, 211–221 (2002)

Mudholkar, G.S., Tian, L.: A test for homogeneity of ordered means of inverse Gaussian populations. J. Stat. Plan. Inference **118**, 37–49 (2004)

Owen, A.B.: Empirical Likelihood. Chapman and Hall/CRC, London (2001)

Schisterman, E.F., Faraggi, D., Browne, R., Freudenheim, J., Dorn, J., Muti, P., Armstrong, D., Reiser, B., Trevisan, M.: TBARS and cardiovascular disease in a population-based sample. J. Cardiovasc. Risk **8**, 219–225 (2001)

Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)

Tusnady, G.: On asymptotically optimal tests. Ann. Stat. **5**, 385–393 (1977)

van Es, B.: Estimating functionals related to a density by a class of statistics based on spacings. Scand. J. Stat. **19**, 61–72 (1992)

Vasicek, O.: A test for normality based on sample entropy. J. R. Stat. Soc. B **38**, 54–59 (1976)

Vexler, A., Gurevich, G.: Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. Comput. Stat. Data Anal. **54**, 531–545 (2010)

Vexler, A., Liu, A., Schisterman, E.F., Wu, C.: Note on distribution-free estimation of maximum linear separation of two multivariate distributions. J. Nonparametr. Stat. **18**, 145–158 (2006)

Vexler, A., Wu, C., Yu, K.F.: Optimal hypothesis testing: from semi to fully Bayes factors. Metrika **71**, 125–138 (2010)

Yu, J., Vexler, A., Tian, L.: Analyzing incomplete data subject to a threshold using empirical likelihood methods: an application to a pneumonia risk study in an ICU setting. Biometrics **66**, 123–130 (2010)

Zhang, J.: Powerful goodness-of-fit tests based on the likelihood ratio. J. R. Stat. Soc. B **64**, 281–294 (2002)