Statistica Sinica **22** (2012), 1457-1477
doi:http://dx.doi.org/10.5705/ss.2011.088

JACKKNIFE EMPIRICAL LIKELIHOOD BASED CONFIDENCE INTERVALS FOR PARTIAL AREAS UNDER ROC CURVES

Gianfranco Adimari and Monica Chiogna

University of Padova

Abstract: The partial area under the ROC curve (partial AUC) summarizes the accuracy of a diagnostic or screening test over a relevant region of the ROC curve and represents a useful tool for the evaluation and comparison of tests. In this paper, we propose a jackknife empirical likelihood method for making inference on partial AUCs. Following the idea in Jing, Yuan, and Zhou (2009), we combine the empirical likelihood function with suitable jackknife pseudo-values obtained from a nonparametric estimator of the normalized partial AUC. This leads to a jackknife empirical likelihood function for normalized partial AUCs, for which a Wilks-type result is obtained. Such a pseudo-likelihood can be used, in a standard way, to construct confidence intervals or perform tests of hypotheses. We also give some simulation results that indicate that the jackknife empirical likelihood based confidence intervals compare favorably with alternatives in terms of coverage probability. The proposed method is extended to inference on the difference between two partial AUCs. Finally, an application to the Wisconsin Breast Cancer Data is discussed.

Key words and phrases: Diagnostic tests, jackknife pseudo-values, nonparametric statistical methods, pseudo-likelihoods.

1. Introduction

The evaluation of the ability of a diagnostic or a screening test to separate diseased from non-diseased subjects is a crucial issue in modern medicine. In fact, before applying a test in a clinical setting, rigorous statistical assessment of its performance in discriminating the diseased state from the non-diseased state is required.

The accuracy of a test, at a chosen threshold level c, can be evaluated by its sensitivity and specificity, defined as the probabilities that the test correctly identifies the diseased and non-diseased subjects, respectively. Let Y and X denote the results of a continuous-scale diagnostic (or screening) test for a diseased and a non-diseased subject, respectively. Let G and F be the cumulative distribution functions of Y and X, respectively, and assume that a value of the test greater than the threshold c indicate a positive test result, i.e., presence of disease. Then the sensitivity, or true positive rate, at the threshold c, $\mathrm{TPR}(c)$,

is defined as $\Pr\{Y > c\} = 1 - G(c)$. The corresponding specificity, or 1 minus the false positive rate, $1-\text{FPR}(c)$, is $1 - \Pr\{X > c\} = F(c)$.

Receiver Operating Characteristic (ROC) analysis is a commonly used methodology for representing the trade-offs between FPR and TPR in a two-group classification task. The ROC curve plots $\{\text{FPR}(c), \text{TPR}(c)\}$ for all possible thresholds c. We can also write the ROC curve as a function of $p = 1 - F(c)$ as follows: $\text{ROC}(p) = 1 - G(S^{-1}(p))$, where $S^{-1}(\cdot)$ denotes the inverse function of $1 - F(\cdot)$. An uninformative test is then represented by a straight line from the $(0, 0)$ vertex to $(1, 1)$, while a straight line from the $(0, 1)$ vertex to $(1, 1)$ indicates an ideally performing test.

Typically, the best threshold is not known when a test is under evaluation, and it may vary depending on the setting in which the test is implemented. A commonly used summary measure that aggregates performance information across a range of possible thresholds is the area under the ROC curve (AUC), $\text{AUC} = \int_0^1 \text{ROC}(t)dt$. However, the AUC also summarizes the test performance over values $\{\text{FPR}(c), \text{TPR}(c)\}$ of no practical interest. For instance, when screening for a certain disease for which the subsequent confirmatory test and/or treatments have large cost, the region of the curve corresponding to low false positive rates is of primary interest. On the other hand, when testing for a serious disease, it is critical to maintain a high TPR, because false negative test results may have serious consequences. Hence, in this case, interest is in the region of the ROC curve that corresponds to acceptable high TPRs.

A summary index for the ROC curve restricted to a relevant range of false positive rates is the partial AUC, defined as

$$\theta = \theta(p_1, p_2) = \int_{p_1}^{p_2} ROC(t)dt,$$

where the interval (p_1, p_2) denotes the false positive rates of interest. For an uninformative test, the partial AUC is $(p_2 + p_1)(p_2 - p_1)/2$, for a perfect test it is $\theta(p_1, p_2) = p_2 - p_1$. This suggests normalizing the partial area to provide an index with maximum of 1:

$$\tau = \tau(p_1, p_2) = \frac{\theta(p_1, p_2)}{p_2 - p_1}.$$

An analogous definition of the partial AUC corresponding to true positive rates is straightforward, see, e.g. Dodd and Pepe (2003). In what follows we focus on the partial AUC restricted to relevant FPR values. It is worth noting that the techniques developed here can be carried over to restrictions on the range of TPR values.

Methods for estimating and comparing partial AUCs are available, both in a parametric approach (see, e.g., McClish (1989); Thompson and Zucchini (1989);

Jiang, Metz, and Nishikawa (1996)) and in a nonparametric one (Dodd and Pepe (2003); Wieand et al. (1989); Zhang et al. (2002); Liu, Schisterman, and Wu (2005); He and Escobar (2008)). In particular, given the observations x_i, $i = 1, \ldots, m$, and y_j, $j = 1, \ldots, n$, the test values for m normal and n diseased subjects, a simple nonparametric estimator of $\theta(p_1, p_2)$ is

$$\hat{\theta} = \hat{\theta}(p_1, p_2) = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} I(y_j > x_i) I(\hat{q}_2 \leq x_i \leq \hat{q}_1),$$

where $I(\cdot)$ denotes the indicator function and $\hat{q}_2$, $\hat{q}_1$ are the empirical counterparts of $q_2 = S^{-1}(p_2)$ and $q_1 = S^{-1}(p_1)$. This estimator, generated by the consideration that

$$\theta(p_1, p_2) = \int_{S^{-1}(p_2)}^{S^{-1}(p_1)} (1 - G(y)) dF(y) = \Pr\{Y > X,\ X \in (S^{-1}(p_2), S^{-1}(p_1))\},$$

is discussed in Zhang et al. (2002); Dodd and Pepe (2003); Liu, Schisterman, and Wu (2005); He and Escobar (2008). It is proved to be asymptotically normal under suitable conditions. Moreover, various estimators of its variance are proposed. In particular, He and Escobar (2008) derive a simple estimator of the variance of $\hat{\theta}$, using the fact that $\hat{\theta} = (m_*/m)\hat{\tau}$. Here, $m_* = \sum_{i=1}^{m} I(\hat{q}_2 \leq x_i \leq \hat{q}_1)$ and

$$\hat{\tau} = \frac{1}{m_* n} \sum_{x_i \in [\hat{q}_2, \hat{q}_1]} \sum_{j=1}^{n} I(y_j > x_i)$$

is the estimator of the normalized partial area τ. It follows that normal approximation can be used to obtain nonparametric confidence intervals for (or to perform nonparametric hypotheses tests on) partial areas or normalized partial areas, in a simple classical way.

However, it is well known that normal approximation based confidence intervals have often low accuracy in samples with small to moderate sizes and that, to overcome this drawback, methods based on such suitable transformations, as the logit, or methods based on pseudo-likelihoods, are generally used. Methods based on the empirical likelihood belong to the latter class.

The empirical likelihood function, first introduced by Owen (1988, 1990) has found many applications. It is a nonparametric tool that allows one to obtain pseudo-likelihoods in several contexts and, in particular, for parameters that are determined by estimating equations. By an emulation of its parametric counterpart, the empirical likelihood function is obtained by maximization of a nonparametric likelihood supported on the data, subject to some constraints. In most cases, those constraints are linear; then, the maximization problem is easily

solved by using Lagrange multipliers. This leads to an explicit expression for the empirical likelihood function and, hence, for (minus twice) the empirical log likelihood ratio, for which a Wilks-type theorem is generally proved. It follows that the empirical likelihood can be used, in a standard way, to obtain nonparametric confidence regions or perform hypotheses tests (see Owen (2001)). However, when the constraints in the underlying maximization problem are not linear, significant computational difficulties arise that may diminish its attractiveness. This happens, for example, when one tries to calculate the empirical likelihood for parameters estimated by U-statistics (see Wood, Do, and Broom (1996)).

To overcome computational difficulties, Jing, Yuan, and Zhou (2009) recently introduced the so-called jackknife empirical likelihood method, which works by combining jackknife pseudo-values and empirical likelihood. The method is quite simple to use in practice, allows one to obtain a family of pseudo-likelihoods for which a Wilks-type theorem is established, and seems to be effective in handling inference on parameters estimated by one and two-sample U-statistics. As an application, Jing, Yuan, and Zhou (2009) consider inference on the AUC, and simulation results seem to suggest that the jackknife empirical likelihood based confidence intervals for the AUC compare favorably with intervals obtained by alternative approaches. As well, Gong and Peng (2010) propose a smoothed jackknife empirical likelihood technique to construct confidence intervals for the ROC curve.

In this paper we extend the use of the jackknife empirical likelihood to inference on partial AUCs. Following the idea in Jing, Yuan, and Zhou (2009), in Section 2 we combine the empirical likelihood function with suitable jackknife pseudo-values obtained from the estimator $\hat{\tau}$. This leads to a jackknife empirical likelihood function for normalized partial AUCs, for which a Wilks-type result is obtained. In Section 3, this approach is extended to the comparison of two normalized partial AUCs. The finite-sample accuracy of the confidence intervals produced by the method is investigated by a simulation study in Section 4, where a comparison with two alternative methods is also performed. An application to data is presented in Section 5 and some final remarks are in Section 6.

2. Inference for a Single Partial AUC

Take F and G to be distribution functions with continuous densities. Let f be the density of F and assume that $f(F^{-1}(1-p_2)) > 0$, $f(F^{-1}(p_1)) > 0$, and that f is bounded in some neighborhood of $F^{-1}(1-p_2)$ and $F^{-1}(p_1)$. Moreover, assume that $m/(m+n)$ converges to some positive and finite constant κ as $N = m + n$ increases to $+\infty$. Let $0 < \tau_0 < 1$ be the true value of τ and let $N_* = m_* + n$. Then, as $N \to +\infty$, $\sqrt{N_*}(\hat{\tau} - \tau_0)$ is asymptotically normal, with zero mean and some asymptotic variance σ^2 (see He and Escobar (2008), Appendix C).

Starting from the estimator $\hat{\tau}$, consider the jackknife pseudo-values

$$V_h^* = N_*\hat{\tau} - (N_* - 1)\hat{\tau}_{-h}, \qquad h = 1, \ldots, N_*,$$

where $\hat{\tau}_{-h}$ denotes the estimate derived by deleting the hth observation in the sample obtained by pooling all the diseased subjects and all those normal subjects for which the test value x lies between $\hat{q}_2$ and $\hat{q}_1$. In what follows, when convenient, we denote by $\hat{\tau}_{-x_k}$ the estimate derived after deleting a normal subject,

$$\hat{\tau}_{-x_k} = \frac{1}{(m_* - 1)n} \sum_{\substack{x_i \in [\hat{q}_2, \hat{q}_1] \\ x_i \neq x_k}} \sum_{j=1}^{n} I(y_j > x_i),$$

and by $\hat{\tau}_{-y_l}$ the estimate derived after deleting a diseased subject,

$$\hat{\tau}_{-y_l} = \frac{1}{m_*(n - 1)} \sum_{x_i \in [\hat{q}_2, \hat{q}_1]} \sum_{\substack{j=1 \\ j \neq l}}^{n} I(y_j > x_i).$$

Let $A = \sum_{x_i \in [\hat{q}_2, \hat{q}_1]} \sum_{j=1}^{n} I(y_j > x_i)$. It is easy to see that

$$\hat{\tau}_{-x_k} = \frac{A - \sum_{j=1}^{n} I(y_j > x_k)}{(m_* - 1)n}, \qquad \hat{\tau}_{-y_l} = \frac{A - \sum_{x_i \in [\hat{q}_2, \hat{q}_1]} I(y_l > x_i)}{m_*(n - 1)},$$

so that $\sum_{l=1}^{n} \hat{\tau}_{-y_l} = A/m_*$, $\sum_{k=1}^{m_*} \hat{\tau}_{-x_k} = A/n$ and $(1/N_*) \sum_{h=1}^{N_*} \hat{\tau}_{-h} = \hat{\tau}$. Hence,

$$\frac{1}{N_*} \sum_{h=1}^{N_*} V_h^* = \hat{\tau}. \tag{2.1}$$

Moreover, the jackknife variance

$$s_*^2 = \frac{1}{N_* - 1} \sum_{h=1}^{N_*} (V_h^* - \hat{\tau})^2 = (N_* - 1) \sum_{h=1}^{N_*} (\hat{\tau} - \hat{\tau}_{-h})^2$$

is a consistent estimator of the asymptotic variance of $\sqrt{N_*}\hat{\tau}$, since it asymptotically equals the estimator derived in He and Escobar (2008) (see, in particular, Appendix B of the paper).

If the jackknife is performed on the sample obtained by pooling all the $N = n + m$ diseased and normal subjects, only some of the above given quantities change. In particular, for the pseudo-values we have

$$V_h = N\hat{\tau} - (N - 1)\hat{\tau}_{-h}, \qquad h = 1, \ldots, N,$$

and $\hat{\tau}_{-x_k} = \hat{\tau}$ when deleting a normal subject for which $x_k \notin [\hat{q}_2, \hat{q}_1]$. Therefore, we still have $(1/N)\sum_{h=1}^{N} \hat{\tau}_{-h} = \hat{\tau}$ and

$$\frac{1}{N}\sum_{h=1}^{N} V_h = \hat{\tau}. \tag{2.2}$$

Moreover, since $\sum_{h=1}^{N}(\hat{\tau} - \hat{\tau}_{-h})^2 = \sum_{h=1}^{N_*}(\hat{\tau} - \hat{\tau}_{-h})^2$, the jackknife variance

$$s^2 = \frac{1}{N-1}\sum_{h=1}^{N}(V_h - \hat{\tau})^2 = (N-1)\sum_{h=1}^{N}(\hat{\tau} - \hat{\tau}_{-h})^2 = \frac{N-1}{N_*-1}s_*^2$$

is a consistent estimator for the asymptotic variance of $\sqrt{N}\hat{\tau}$. Note that in both jackknife schemes, the sample quantiles $\hat{q}_1$ and $\hat{q}_2$ are kept as fixed quantities, i.e., they do not change with the jackknife samples. Therefore, strictly speaking, these are not "genuine" jackknife schemes.

Equations (2.1) and (2.2) provide the estimating functions

$$\sum_{h=1}^{N^*}(V_h^* - \tau), \tag{2.3}$$

$$\sum_{h=1}^{N}(V_h - \tau), \tag{2.4}$$

for inference about τ_0, based on jackknife pseudo-values. Then, following the idea in Jing, Yuan, and Zhou (2009), by combining (2.3) or (2.4) with Owen's empirical likelihood, we obtain a jackknife empirical likelihood function for τ that can be used to construct confidence intervals or perform hypotheses tests.

To explain the method, in the following we refer to the estimating function (2.4) for the jackknife scheme performed on N units. This scheme is more general since it is useful also when the method is extended to the comparison of two partial areas. However, the jackknife empirical likelihood for τ works also when obtained from (2.3). See Remark 1 below.

Let

$$L(\tau) = \max_{w_1,\ldots,w_N} \prod_{h=1}^{N} w_h, \quad \text{subject to} \quad \sum_{h=1}^{N} w_h = 1 \quad \text{and} \quad \sum_{h=1}^{N}(V_h - \tau)w_h = 0,$$

be the empirical likelihood function for τ based on (2.4). Here $(w_1, \ldots, w_N)$ denotes a generic multinomial distribution on the pseudo-sample $V_1, \ldots, V_N$. Let $V_{(1)} = \min_h V_h$ and $V_{(N)} = \max_h V_h$. It is well known that $L(\tau)$ attains its maximum N^{-N} at $\tau = \hat{\tau}$. Moreover, when $\tau \in (V_{(1)}, V_{(N)})$, an explicit expression

for $L(\tau)$ can be obtained via Lagrangian multipliers. More precisely, when $\tau \in (V_{(1)}, V_{(N)})$, we have

$$L(\tau) = N^{-N} \prod_{h=1}^{N} \frac{1}{1 + \lambda(V_h - \tau)},$$

where the multiplier λ satisfies

$$\frac{1}{N} \sum_{h=1}^{N} \frac{V_h - \tau}{1 + \lambda(V_h - \tau)} = 0.$$

Outside the interval bounded by $V_{(1)}$ and $V_{(N)}$ it is necessary to set $L(\tau) = 0$. The function $l(\tau) = -2 \log\{L(\tau)/N^{-N}\}$, that is (minus twice) the jackknife empirical log likelihood ratio function, represents a pseudo-log likelihood ratio function for inference about τ. It remains to check whether a Wilks-type theorem still holds. Since $\hat{\tau}$ is not an ordinary two-sample U-statistic, Theorem 2 in Jing, Yuan, and Zhou (2009) does not apply here. However, our Theorem 1 below gives a theoretical justification to the jackknife empirical likelihood method. More precisely, Theorem 1 shows that $l(\tau_0) \xrightarrow{d} \chi_1^2$. Therefore, for instance, the set $\mathcal{C} = \{\tau : l(\tau) \le c_\gamma\}$ is an approximate confidence interval for τ_0, with nominal coverage $1 - \gamma$, if c_γ is such that $\Pr\{\chi_1^2 > c_\gamma\} = \gamma$. Clearly, the corresponding approximate confidence interval for θ_0 can be readily obtained from $\mathcal{C}$.

Theorem 1. *Under the assumptions made at the beginning of the section, as* $N \to +\infty$,

(i) $\max_{h=1,\ldots,N} |V_h| = O_p(1)$;

(ii) $\Pr\{V_{(1)} < \tau_0\} \to 1$ *and* $\Pr\{V_{(N)} > \tau_0\} \to 1$;

(iii) $l(\tau_0) \xrightarrow{d} \chi_1^2$.

Proof. (i) Clearly, $V_h = \hat{\tau}_{-h} + N(\hat{\tau} - \hat{\tau}_{-h})$. Thus $|V_h| \le |\hat{\tau}_{-h}| + N|\hat{\tau} - \hat{\tau}_{-h}|$. Moreover,

$$\hat{\tau} - \hat{\tau}_{-h} = \frac{j/n - \hat{\tau}}{m_* - 1} \quad \text{for some suitable integer } j \text{ such that } 0 \le j \le n,$$

or $\hat{\tau} - \hat{\tau}_{-h} = 0$, if $\hat{\tau}_{-h}$ is obtained by deleting a normal subject. Alternatively,

$$\hat{\tau} - \hat{\tau}_{-h} = \frac{j/m_* - \hat{\tau}}{n - 1} \quad \text{for some suitable integer } j \text{ such that } 0 \le j \le m_*,$$

if $\hat{\tau}_{-h}$ is obtained by deleting a diseased subject. It follows that

$$|\hat{\tau} - \hat{\tau}_{-h}| \le \max\{\frac{1}{m_* - 1}, \frac{1}{n - 1}\}.$$

Hence, we have $\max_{h=1,\ldots,N} |V_h| = O_p(1)$.

(ii) Let $\hat{\tau}_-^{\min}$ and $\hat{\tau}_-^{\max}$ denote, respectively, the smallest and the largest among the $\hat{\tau}_{-h}$'s values, with $h = 1, \ldots, N$. Clearly, we have $V_{(1)} = N\hat{\tau} - (N-1)\hat{\tau}_-^{\max}$ and $V_{(N)} = N\hat{\tau} - (N-1)\hat{\tau}_-^{\min}$. Therefore,

$$V_{(1)} - \hat{\tau}_-^{\min} = N(\hat{\tau} - \hat{\tau}_-^{\max}) + (\hat{\tau}_-^{\max} - \hat{\tau}_-^{\min}) \leq N(\hat{\tau} - \hat{\tau}_-^{\max}),$$
$$V_{(N)} - \hat{\tau}_-^{\max} = N(\hat{\tau} - \hat{\tau}_-^{\min}) - (\hat{\tau}_-^{\max} - \hat{\tau}_-^{\min}) \geq N(\hat{\tau} - \hat{\tau}_-^{\min}),$$

where the equalities hold only if $\hat{\tau}_-^{\min} = \hat{\tau}_-^{\max} = \hat{\tau}$, i.e., only if $\hat{\tau} = 0$ or $\hat{\tau} = 1$. Since $0 < \tau_0 < 1$ by assumption, we have that $\Pr\{0 < \hat{\tau} < 1\} \to 1$ as $N \to +\infty$. It follows that $V_{(1)} < \hat{\tau}_-^{\min}$ and $V_{(N)} > \hat{\tau}_-^{\max}$, with probability tending to one as $N \to +\infty$. Let $\varepsilon_0 > 0$ and $\beta > 0$ be fixed and suppose that $\Pr\{\hat{\tau}_-^{\min} < \tau_0 + \varepsilon_0\} \to 1 - \beta$, as $N \to +\infty$. Then, for any $0 < \varepsilon \leq \varepsilon_0$, one would have $\lim_{N\to+\infty} \Pr\{\hat{\tau}_-^{\min} \geq \tau_0 + \varepsilon\} \geq \beta$, so that $\lim_{N\to+\infty} \Pr\{\hat{\tau} \geq \tau_0 + \varepsilon\} \geq \beta$. This contradicts the consistency of $\hat{\tau}$. Hence, $\lim_{N\to+\infty} \Pr\{\hat{\tau}_-^{\min} \leq \tau_0\} = 1$, as $N \to +\infty$. Analogously, it is possible to show that $\lim_{N\to+\infty} \Pr\{\hat{\tau}_-^{\max} \geq \tau_0\} = 1$.
(iii) It is easy to see that

$$\frac{1}{N}\sum_{h=1}^{N}(V_h - \tau_0)^2 = \frac{s^2(N-1)}{N} + O_p(\frac{1}{N}).$$

Since $s^2(N-1)/N$ is a consistent estimator of the asymptotic variance of $\sqrt{N}(\hat{\tau} - \tau_0)$, the result follows by (i), (ii), and an application of Theorem 2.1 in Adimari and Guolo (2010).

Remark 1. Starting from the relation $V_h^* = \hat{\tau}_{-h} + N^*(\hat{\tau} - \hat{\tau}_{-h})$, with $h = 1, \ldots, N^*$, the proof of Theorem 1 can be easily adapted to show that the χ_1^2 calibration is adequate even if the jackknife empirical likelihood for the partial AUC is derived from the estimating function (2.3) and, hence, is based on the "parsimonious" jackknife scheme performed only over N^* units. In this case, when $V_{(1)}^* < \tau < V_{(N^*)}^*$,

$$L(\tau) = N^{*-N^*} \prod_{h=1}^{N^*} \frac{1}{1 + \lambda(V_h^* - \tau)},$$

where λ satisfies

$$\frac{1}{N}\sum_{h=1}^{N^*} \frac{V_h^* - \tau}{1 + \lambda(V_h^* - \tau)} = 0,$$

and $l(\tau) = -2\log\{L(\tau)/N^{*-N^*}\}$.

3. Comparing Two Partial AUCs

The method described in Section 2 can be easily extended to the comparison of two diagnostic or screening tests, when the comparison is requested in terms of some relevant portion of the AUC and paired data are available.

Consider two continuous-scale tests, both performed on the same n diseased and m non-diseased subjects. We use the superscripts a and b to distinguish the two tests. Let Y^a and Y^b denote the results of the tests for a diseased subject and X^a and X^b the results for a normal subject. In such an experimental setting, for each of the $N = n + m$ subjects two observations are available: y_j^a and y_j^b for the diseased subject j, $j = 1, \ldots, n$, and x_i^a and x_i^b for the non-diseased subject i, $i = 1, \ldots, m$. For the two tests, let $\hat{\tau}^a$ and $\hat{\tau}^b$ be the estimates of the normalized partial AUCs $\tau^a(p_1, p_2)$ and $\tau^b(p_1, p_2)$, both referred to the same range of false positive rates. Such estimates are derived from the observations $x_1^a, \ldots, x_m^a, y_1^a, \ldots, y_n^a$ and $x_1^b, \ldots, x_m^b, y_1^b, \ldots, y_n^b$, respectively.

Let δ denote the difference between the normalized partial AUCs, that is $\delta = \tau^a - \tau^b$. Such a difference can be estimated by

$$\hat{\delta} = \hat{\tau}^a - \hat{\tau}^b = \frac{1}{N} \sum_{h=1}^{N} (V_h^a - V_h^b), \tag{3.1}$$

where the values V_h^a and V_h^b, $h = 1, \ldots, N$, are the jackknife pseudo-values derived from $\hat{\tau}^a$ and $\hat{\tau}^b$, respectively. If the distribution functions F^a, F^b, G^a, G^b, and the densities f^a and f^b satisfy regularity conditions such as those given at the beginning of Section 2 for the case of a single test, then $\sqrt{N}(\hat{\delta} - \delta_0)$ is asymptotically normal, as $N \to +\infty$. Here, $\delta_0 = \tau_0^a - \tau_0^b$ denotes the true parameter value, with $0 < \tau_0^a < 1$ and $0 < \tau_0^b < 1$. Moreover, the asymptotic variance of $\sqrt{N}(\hat{\delta} - \delta_0)$ is consistently estimated by

$$s_\delta^2 = \frac{1}{N-1} \sum_{h=1}^{N} (V_h^a - V_h^b - \hat{\delta})^2,$$

because

$$s_\delta^2 = \frac{1}{N-1} \sum_{h=1}^{N} (V_h^a - \hat{\tau}^a)^2 + \frac{1}{N-1} \sum_{h=1}^{N} (V_h^b - \hat{\tau}^b)^2 - \frac{2}{N-1} \sum_{h=1}^{N} (V_h^a - \hat{\tau}^a)(V_h^b - \hat{\tau}^b),$$

and $(1/(N-1)) \sum_{h=1}^{N} (V_h^a - \hat{\tau}^a)(V_h^b - \hat{\tau}^b)$ is a consistent estimator of the asymptotic covariance of $\hat{\tau}^a$ and $\hat{\tau}^b$ (see also He and Escobar (2008)).

Observe that (3.1) provides an estimating function for inference on δ based on the jackknife pseudo-values

$$\sum_{h=1}^{N} (V_h^a - V_h^b - \delta). \tag{3.2}$$

Then, by combining (3.2) with the empirical likelihood, we easily obtain a jackknife empirical likelihood function for the difference between two normalized partial AUCs. Let $W_h = V_h^a - V_h^b$. As in Section 2, we have

$$L(\delta) = N^{-N} \prod_{h=1}^{N} \frac{1}{1 + \lambda(W_h - \delta)},$$

where λ satisfies

$$\frac{1}{N} \sum_{h=1}^{N} \frac{W_h - \delta}{1 + \lambda(W_h - \delta)} = 0$$

when $W_{(1)} < \delta < W_{(N)}$. Outside $(W_{(1)}, W_{(N)})$, we have $L(\delta) = 0$. Moreover, a Wilks-type result can be proved for (minus twice) the jackknife empirical log likelihood ratio function $l(\delta) = -2\log\{L(\delta)/N^{-N}\}$.

Theorem 2. *Under the assumptions made in this section, as* $N \to +\infty$,

$$l(\delta_0) \xrightarrow{d} \chi_1^2.$$

Proof. Since $|W_h| \leq |V_h^a| + |V_h^b|$, by Theorem 1 (i) we have that

$$\max_{h=1,\ldots,N} |W_h| = O_p(1). \tag{3.3}$$

On the other hand, $W_h = N(\hat{\tau}^a - \hat{\tau}^b) - (N-1)(\hat{\tau}^a_{-h} - \hat{\tau}^b_{-h}) = N\hat{\delta} - (N-1)\hat{\delta}_{-h}$. Thus, denoting by $\hat{\delta}_-^{\min}$ and $\hat{\delta}_-^{\max}$ the smallest and the largest among the $\hat{\delta}_{-h}$'s values, respectively, we have $W_{(1)} - \hat{\delta}_-^{\min} \leq N(\hat{\delta} - \hat{\delta}_-^{\max})$ and $W_{(N)} - \hat{\delta}_-^{\max} \geq N(\hat{\delta} - \hat{\delta}_-^{\min})$, where the equalities hold only if $\hat{\tau}^a = 0$ or $\hat{\tau}^a = 1$ and $\hat{\tau}^b = 0$ or $\hat{\tau}^b = 1$. It follows that $W_{(1)} < \hat{\delta}_-^{\min}$ and $W_{(N)} > \hat{\delta}_-^{\max}$, with probability tending to one as $N \to +\infty$. Then, similarly to the proof of Theorem 1 (ii), we can show that

$$\lim_{N\to+\infty} \Pr\{W_{(1)} < \delta_0\} = 1 \quad \text{and} \quad \lim_{N\to+\infty} \Pr\{W_{(N)} > \delta_0\} = 1 \tag{3.4}$$

as $N \to +\infty$. Finally, we have $(1/N)\sum_{h=1}^{N}(W_h - \delta_0)^2 = s_\delta^2(N-1)/N + O_p(1/N)$. Since $s_\delta^2(N-1)/N$ is a consistent estimator of the asymptotic variance of $\sqrt{N}(\hat{\delta} - \delta_0)$, the result follows by (3.3), (3.4), and an application of Theorem 2.1 in Adimari and Guolo (2010).

As a consequence of Theorem 2, the set $\{\delta : l(\delta) \leq c_\gamma\}$ is an approximate confidence interval for δ_0, with nominal coverage $1-\gamma$, if c_γ is such that $\Pr\{\chi_1^2 > c_\gamma\} = \gamma$. The derivation of the corresponding confidence interval for the difference between the (non normalized) partial AUCs is immediate.

It is, of course, the case that other strategies are available to compare diagnostic tests, for example based on the comparison of the entire ROC curves or

of the total AUCs. For the case of paired data, we refer the interested reader to Venkatraman and Begg (1996) and references therein.

4. Simulation Results

In this section, we report the results of a simulation study carried out to assess the finite-sample accuracy of the confidence intervals for partial AUCs obtained by using the technique discussed in Section 2. In the study, we also compare the performance of the proposed jackknife empirical likelihood method with the normal approximation method and the normal approximation method after logit transformation.

To fix the simulation setting, we took the area under the ROC curve of relevance to extend up to some defined maximum value of FPR. Thus we fixed p_1 at 0 and considered various values of p_2.

We took two models for the data: a Gaussian model, $\mathrm{N}(\alpha, \omega)$, and a Gamma model, $\mathrm{Ga}(\alpha, \omega)$. Random samples for the Y and the X values were generated, respectively, from a $\mathrm{N}(\alpha, 4)$ and a $\mathrm{N}(0, 1)$ in the first case, and from a $\mathrm{Ga}(\alpha, 5)$ and a $\mathrm{Ga}(1, 1)$ in the second, for some values of the parameter α. The simulation study used the "parsimonious" jackknife scheme performed over N^* units, and each simulation experiment was based on 5,000 replications.

Simulation results are given in Tables 1−6. For three levels of the nominal coverage $1 - \gamma$, Tables 1−6 report the estimated coverage probabilities of the (two-sided) confidence intervals for the (normalized) partial AUC, obtained by three methods: (minus twice) the jackknife empirical log likelihood ratio $l(\tau)$ (JEL), the asymptotic normality of $\hat{\tau}$ (AN), and the asymptotic normality of $t(\hat{\tau})$, with $t(\cdot)$ being the logit transformation (ANL). In this last case, we set $\rho = t(\tau) = \log(\tau/(1-\tau))$, $\hat{\rho} = t(\hat{\tau})$ and $\rho_0 = t(\tau_0)$, so that $\sqrt{N_*}(\hat{\rho} - \rho_0)$ is asymptotically normally distributed with mean zero and variance $\sigma^2/[\tau_0^2(1-\tau_0)^2]$. This asymptotic distribution was used to construct confidence intervals on the ρ scale, then converted back to the τ scale by the inverse transformation t^{-1}.

Tables 1−3 refer to the Gaussian model, whereas Tables 4-6 refer to the Gamma model. Each table corresponds to a chosen value for the parameter α. For each considered pair (α, p_2), the tables give the corresponding "true" partial AUC $\theta_0 = \theta(0, p_2)$ and the "true" normalized partial AUC $\tau_0 = \tau(0, p_2)$. Then, for each target pair (θ_0, τ_0), the simulation results are shown for two different settings for the sample sizes. The setting with the smallest sample sizes has been chosen so as to always present actual coverage probabilities reasonably close to the nominal ones. Clearly, sample sizes matching this criterion strongly depend on the target pair (θ_0, τ_0). In particular:

Table 1. Estimated coverage probabilities of the confidence intervals for the partial AUC, obtained by the jackknife empirical likelihood method (JEL), the normal approximation method (AN), and the normal approximation method after logit transformation (ANL). Gaussian model, $\alpha = 1$. Figures in bold do not differ from the nominal level by more than simulation error.

			$1-\gamma$		
			0.99	0.95	0.90
		JEL	**0.988**	**0.944**	**0.902**
	$m = 110, n = 15$	AN	0.966	0.930	0.875
		ANL	0.999	0.983	0.930
$p_2 = 0.05$					
$\theta_0 = 0.020,\ \tau_0 = 0.396$		JEL	**0.990**	**0.946**	**0.894**
	$m = 135, n = 50$	AN	0.985	0.942	**0.893**
		ANL	0.995	0.956	**0.906**
		JEL	**0.988**	**0.949**	**0.898**
	$m = 60, n = 15$	AN	0.969	0.928	0.885
		ANL	0.993	0.970	0.931
$p_2 = 0.10$					
$\theta_0 = 0.043,\ \tau_0 = 0.426$		JEL	**0.991**	**0.953**	**0.905**
	$m = 85, n = 50$	AN	**0.987**	**0.947**	**0.899**
		ANL	0.993	0.959	0.911
		JEL	0.993	**0.955**	**0.907**
	$m = 30, n = 15$	AN	0.979	0.926	0.881
		ANL	0.999	0.973	0.921
$p_2 = 0.25$					
$\theta_0 = 0.118,\ \tau_0 = 0.473$		JEL	**0.992**	**0.955**	**0.906**
	$m = 55, n = 50$	AN	**0.988**	**0.949**	**0.899**
		ANL	0.994	0.959	0.913
		JEL	0.994	**0.954**	0.910
	$m = 20, n = 15$	AN	0.974	0.927	0.883
		ANL	0.998	0.972	0.920
$p_2 = 0.50$					
$\theta_0 = 0.260,\ \tau_0 = 0.520$		JEL	0.993	0.956	**0.901**
	$m = 45, n = 50$	AN	**0.988**	**0.948**	**0.896**
		ANL	0.993	0.960	**0.906**

- small values of p_2 require a high value of m, as each method needs to estimate an extreme quantile q_2 of the distribution of the test result for non-diseased subjects. This requirement holds regardless of the "true" value of the partial AUC. In fact, when $p_2 = 0.05$, the value of m in the tables is at least 110;
- both m and n tend to be larger when the "true" normalized partial AUC τ_0 approaches 1. For example, in the Gaussian case with $p_2 = 0.10$, the smallest sample sizes are $m = 60$ and $n = 15$ when $\tau_0 = 0.426$, but they are $m = 110$ and $n = 70$ for $\tau_0 = 0.94$.

Table 2. Estimated coverage probabilities of the confidence intervals for the partial AUC, obtained by the jackknife empirical likelihood method (JEL), the normal approximation method (AN), and the normal approximation method after logit transformation (ANL). Gaussian model, $\alpha = 5$. Figures in bold do not differ from the nominal level by more than simulation error.

			$1-\gamma$		
			0.99	0.95	0.90
		JEL	0.986	0.938	**0.900**
	$m = 115, n = 20$	AN	0.957	0.935	0.867
		ANL	0.995	0.972	0.919
$p_2 = 0.05$					
$\theta_0 = 0.038,\ \tau_0 = 0.768$		JEL	**0.989**	0.943	**0.896**
	$m = 135, n = 50$	AN	0.977	0.958	0.889
		ANL	0.993	0.959	0.909
		JEL	**0.991**	**0.952**	0.885
	$m = 60, n = 25$	AN	0.971	0.911	0.878
		ANL	0.993	0.967	0.920
$p_2 = 0.10$					
$\theta_0 = 0.079,\ \tau_0 = 0.790$		JEL	**0.989**	**0.947**	**0.899**
	$m = 85, n = 50$	AN	0.980	0.936	0.889
		ANL	0.993	0.958	0.910
		JEL	0.986	**0.950**	**0.899**
	$m = 35, n = 25$	AN	0.956	0.925	0.865
		ANL	0.994	0.963	0.931
$p_2 = 0.25$					
$\theta_0 = 0.206,\ \tau_0 = 0.823$		JEL	**0.989**	**0.952**	**0.901**
	$m = 55, n = 50$	AN	0.974	0.933	0.887
		ANL	**0.991**	0.959	0.912
		JEL	**0.988**	**0.945**	**0.901**
	$m = 25, n = 35$	AN	0.961	0.919	0.872
		ANL	**0.990**	0.960	0.910
$p_2 = 0.50$					
$\theta_0 = 0.425,\ \tau_0 = 0.851$		JEL	**0.988**	**0.947**	**0.900**
	$m = 45, n = 50$	AN	0.968	0.927	0.880
		ANL	**0.991**	0.956	0.915

In the tables, figures in bold indicate that the estimated coverage probabilities $1 - \hat{\gamma}$ do not differ from the corresponding nominal levels by more than the simulation error, in the sense that the nominal coverages lie in the intervals with limits $(1 - \hat{\gamma}) \pm 2\sqrt{\hat{\gamma}(1 - \hat{\gamma})/5000}$.

Overall, simulation results indicate that, in terms of coverage probabilities, the jackknife empirical likelihood method outperforms the normal approximation method and the normal approximation method after logit transformation. In fact, for the JEL approach, only 29% of the estimated coverage levels differ

Table 3. Estimated coverage probabilities of the confidence intervals for the partial AUC, obtained by the jackknife empirical likelihood method (JEL), the normal approximation method (AN), and the normal approximation method after logit transformation (ANL). Gaussian model, $\alpha = 8$. Figures in bold do not differ from the nominal level by more than simulation error.

			$1-\gamma$		
			0.99	0.95	0.90
		JEL	**0.987**	0.941	**0.900**
	$m = 155, n = 70$	AN	0.953	0.904	0.859
		ANL	**0.991**	0.964	0.924
$p_2 = 0.05$					
$\theta_0 = 0.047,\ \tau_0 = 0.930$		JEL	**0.989**	**0.953**	**0.903**
	$m = 200, n = 125$	AN	0.973	0.931	0.884
		ANL	**0.992**	0.961	0.917
		JEL	0.981	**0.953**	0.890
	$m = 110, n = 70$	AN	0.939	0.908	0.875
		ANL	**0.988**	**0.954**	0.922
$p_2 = 0.10$					
$\theta_0 = 0.094,\ \tau_0 = 0.940$		JEL	**0.989**	**0.951**	**0.900**
	$m = 150, n = 125$	AN	0.996	0.926	0.884
		ANL	**0.992**	0.962	0.916
		JEL	0.986	**0.944**	**0.899**
	$m = 70, n = 110$	AN	0.957	0.913	0.866
		ANL	0.986	0.956	**0.907**
$p_2 = 0.25$					
$\theta_0 = 0.238,\ \tau_0 = 0.952$		JEL	**0.988**	**0.945**	**0.897**
	$m = 120, n = 125$	AN	0.955	0.913	0.870
		ANL	0.998	0.958	**0.907**
		JEL	0.982	**0.949**	**0.893**
	$m = 55, n = 130$	AN	0.951	0.911	0.869
		ANL	0.985	**0.954**	**0.904**
$p_2 = 0.50$					
$\theta_0 = 0.481,\ \tau_0 = 0.962$		JEL	**0.988**	**0.950**	**0.898**
	$m = 110, n = 150$	AN	0.955	0.915	0.874
		ANL	**0.989**	0.963	**0.900**

from the corresponding nominal levels by more than the simulation error. This percentage increases to roughly 62% for the ANL method and to 87% for the AN approach. As expected, results get uniformly better as sample sizes increase.

Methods for the construction of confidence intervals may be also compared in terms of the average length of the produced intervals. Clearly, such a comparison requires that intervals have almost exactly the desired coverage. However, the methods here are first-order asymptotically equivalent, so that similar average lengths are to be expected when the estimated coverage probabilities are close to

Table 4. Estimated coverage probabilities of the confidence intervals for the partial AUC, obtained by the jackknife empirical likelihood method (JEL), the normal approximation method (AN), and the normal approximation method after logit transformation (ANL). Gamma model, $\alpha = 0.8$. Figures in bold do not differ from the nominal level by more than simulation error.

			$1-\gamma$		
			0.99	0.95	0.90
		JEL	**0.988**	0.956	**0.903**
	$m = 110, n = 15$	AN	0.974	0.935	0.886
		ANL	0.999	0.979	0.932
$p_2 = 0.05$					
$\theta_0 = 0.018,\ \tau_0 = 0.362$		JEL	**0.989**	**0.953**	0.911
	$m = 135, n = 50$	AN	0.985	**0.948**	**0.907**
		ANL	0.994	0.961	0.923
		JEL	**0.988**	**0.949**	**0.901**
	$m = 60, n = 15$	AN	0.973	0.934	0.887
		ANL	0.998	0.973	0.932
$p_2 = 0.10$					
$\theta_0 = 0.043,\ \tau_0 = 0.425$		JEL	**0.988**	**0.954**	0.911
	$m = 85, n = 50$	AN	0.985	**0.948**	**0.905**
		ANL	0.993	0.961	0.918
		JEL	0.994	0.960	0.911
	$m = 30, n = 15$	AN	0.978	0.939	**0.892**
		ANL	0.999	0.974	0.932
$p_2 = 0.25$					
$\theta_0 = 0.132,\ \tau_0 = 0.529$		JEL	**0.988**	**0.954**	**0.902**
	$m = 55, n = 50$	AN	0.986	**0.947**	**0.897**
		ANL	**0.991**	0.960	**0.908**
		JEL	**0.990**	0.959	0.912
	$m = 20, n = 15$	AN	0.974	0.934	0.890
		ANL	0.996	0.970	0.931
$p_2 = 0.50$					
$\theta_0 = 0.315,\ \tau_0 = 0.630$		JEL	**0.991**	**0.954**	**0.906**
	$m = 45, n = 50$	AN	**0.987**	**0.947**	**0.900**
		ANL	0.993	0.960	0.912

the nominal one. This expectation is matched in our simulation study. For example, in the normal case with $\alpha = 1$, $p_2 = 0.5$, $m = 45$, and $n = 50$, the average length of the 90% confidence intervals is 0.2199 (standard deviation: 0.0044) for JEL, 0.2225 (standard deviation: 0.0045) for AN, and 0.2190 (standard deviation: 0.0042) for ANL. Analogously, in the Gamma case with $\alpha = 0.8$, $p_2 = 0.25$, $m = 55$, and $n = 50$, the average length of the 90% confidence intervals is 0.2275 (standard deviation: 0.0129) for JEL, 0.2301 (standard deviation: 0.0115) for

Table 5. Estimated coverage probabilities of the confidence intervals for the partial AUC, obtained by the jackknife empirical likelihood method (JEL), the normal approximation method (AN), and the normal approximation method after logit transformation (ANL). Gamma model, $\alpha = 1.5$. Figures in bold do not differ from the nominal level by more than simulation error.

			$1-\gamma$		
			0.99	0.95	0.90
		JEL	0.983	0.943	**0.899**
	$m = 115, n = 20$	AN	0.969	0.921	0.880
		ANL	0.996	0.968	0.927
$p_2 = 0.05$					
$\theta_0 = 0.033,\ \tau_0 = 0.663$		JEL	0.986	0.942	0.887
	$m = 135, n = 50$	AN	0.979	0.933	0.881
		ANL	0.993	**0.955**	**0.902**
		JEL	0.985	**0.944**	**0.895**
	$m = 60, n = 25$	AN	0.968	0.930	0.888
		ANL	0.997	0.974	0.929
$p_2 = 0.10$					
$\theta_0 = 0.073,\ \tau_0 = 0.727$		JEL	**0.987**	**0.950**	**0.900**
	$m = 85, n = 50$	AN	0.978	0.937	0.890
		ANL	0.993	0.960	**0.907**
		JEL	0.982	**0.949**	**0.900**
	$m = 35, n = 25$	AN	0.959	0.922	0.881
		ANL	**0.992**	0.963	0.926
$p_2 = 0.25$					
$\theta_0 = 0.203,\ \tau_0 = 0.813$		JEL	**0.991**	**0.952**	**0.903**
	$m = 55, n = 50$	AN	0.973	0.934	0.886
		ANL	0.994	0.961	0.914
		JEL	0.981	**0.949**	**0.903**
	$m = 25, n = 35$	AN	0.957	0.921	0.878
		ANL	**0.992**	0.960	0.918
$p_2 = 0.50$					
$\theta_0 = 0.438,\ \tau_0 = 0.876$		JEL	**0.988**	**0.947**	**0.896**
	$m = 45, n = 50$	AN	0.965	0.921	0.876
		ANL	**0.992**	0.957	**0.903**

AN, and 0.2262 (standard deviation: 0.0108) for ANL.

5. An Illustration

To illustrate the application of the method developed in the previous sections, we utilize the Wisconsin Breast Cancer Data, publicy available at the UCI Machine Learning Repository (Asuncion and Newman (2007)). The construction of the dataset was motivated by the need to accurately diagnose breast masses on the basis, solely, of a Fine Needle Aspiration (FNA). The dataset collects

Table 6. Estimated coverage probabilities of the confidence intervals for the partial AUC, obtained by the jackknife empirical likelihood method (JEL), the normal approximation method (AN), and the normal approximation method after logit transformation (ANL). Gamma model, $\alpha = 2.3$. Figures in bold do not differ from the nominal level by more than simulation error.

			$1-\gamma$		
			0.99	0.95	0.90
		JEL	**0.988**	**0.951**	**0.904**
	$m = 155, n = 70$	AN	0.967	0.927	0.885
		ANL	0.996	0.964	0.917
$p_2 = 0.05$					
$\theta_0 = 0.043$, $\tau_0 = 0.868$		JEL	**0.991**	**0.947**	**0.898**
	$m = 200, n = 125$	AN	0.975	0.930	0.889
		ANL	0.993	0.956	**0.902**
		JEL	**0.990**	**0.952**	**0.906**
	$m = 110, n = 70$	AN	0.965	0.924	0.878
		ANL	**0.992**	0.962	0.916
$p_2 = 0.10$					
$\theta_0 = 0.090$, $\tau_0 = 0.905$		JEL	**0.988**	**0.948**	**0.901**
	$m = 150, n = 125$	AN	0.970	0.932	0.884
		ANL	0.993	**0.954**	**0.905**
		JEL	0.984	**0.946**	**0.897**
	$m = 70, n = 110$	AN	0.959	0.913	0.867
		ANL	**0.989**	**0.954**	**0.903**
$p_2 = 0.25$					
$\theta_0 = 0.236$, $\tau_0 = 0.945$		JEL	0.984	0.940	0.887
	$m = 120, n = 125$	AN	0.958	0.912	0.868
		ANL	**0.989**	**0.946**	**0.894**
		JEL	0.975	0.936	0.888
	$m = 55, n = 130$	AN	0.946	0.902	0.859
		ANL	0.986	**0.952**	**0.899**
$p_2 = 0.50$					
$\theta_0 = 0.484$, $\tau_0 = 0.968$		JEL	0.981	0.940	0.891
	$m = 110, n = 150$	AN	0.957	0.914	0.876
		ANL	**0.990**	**0.946**	**0.893**

various features (markers) which are computed from a digitized image of a FNA of a breast mass, describing characteristics of the cell nuclei present in the image. A total of 30 nuclear markers are computed on each of 569 samples, of which 357 are benign and 212 malignant. The dataset has been extensively used in the literature. The interested reader can refer to the UCI Machine Learning Repository documentation for retrieving information about the dataset creation, the description of its attributes, and a list of relevant papers using or citing this data set.

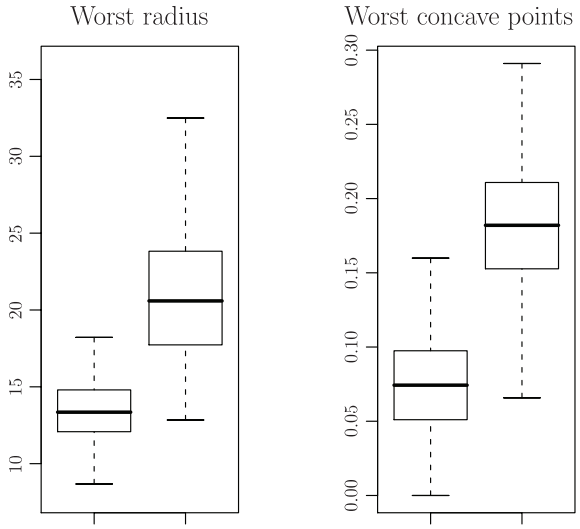

Figure 1. Boxplots of the distribution of the two markers in the groups (left: benign, right: malignant).

Here, we limit ourself to the comparison of two markers, i.e., the worst radius (mean of distances from center to points on the perimeter) and the worst concave points (mean number of concave portions of the contour) observed on each sample. Both markers behave, marginally, in a very similar fashion in the two classes (see Figure 1 for a boxplot of the distribution of the markers in the two groups). Moreover, they show a very similar discriminating capability: a simple linear discriminant analysis shows a predicted mean diagnostic accuracy (with leave-one-out cross-validation) of 89% for the first marker and of 93% for the second marker.

Figure 2 reports the ROC curves, which look quite comparable. In fact, the estimated AUCs are 0.970 for the worst radious, and 0.967 for the worst concave points, confirming the similarity of the two markers. Nevertheless, if discrimination between the markers has to be based only on their performances for specificities of at least 0.95, then one has to look at the normalized partial AUCs $\tau^a(0, 0.05)$ and $\tau^b(0, 0.05)$. In this case, nonparametric point estimates are 0.789 for the first marker and 0.767 for the second marker. Figure 3 shows the jackknife empirical log likelihood ratio function $l(\delta)$ for the difference δ between the two normalized partial AUCs $\tau^a(0, 0.05)$ and $\tau^b(0, 0.05)$. The plot highlights the set $\{\delta : l(\delta) \leq 3.84\}$, the approximate confidence interval for δ with nominal coverage 0.95. Note that the value 0 is not included in the interval, so that the hypothesis of no difference between the two parameters, i.e., $H_0 : \delta = 0$, is rejected at the 5% approximate level of significance. Therefore, the markers' performance over the region of the ROC space of interest is statistically different.

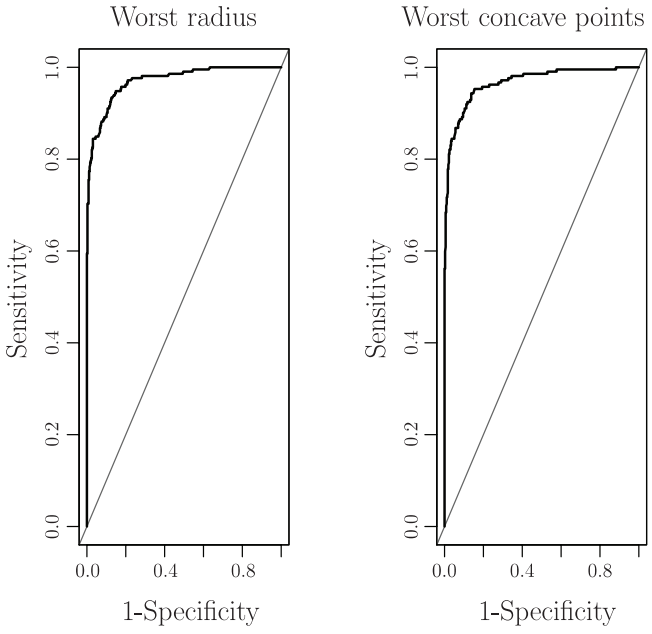

Figure 2. ROC curves for the markers.

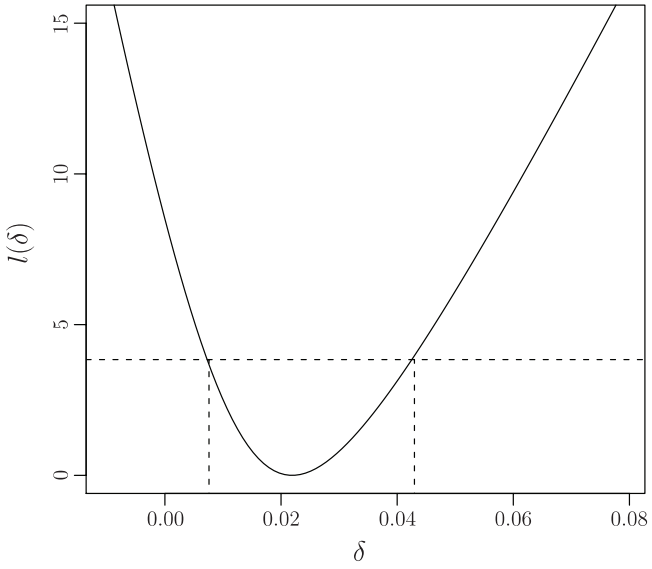

Figure 3. Minus twice jackknife empirical log likelihood ratio function $l(\delta)$, for the difference $\delta = \tau^a(0, 0.05) - \tau^b(0, 0.05)$. Vertical bars define the extremes of the approximate confidence interval for δ, with nominal coverage 0.95.

6. Conclusion

The partial AUC summarizes the accuracy of a diagnostic or screening test over a relevant region of the ROC curve and represents a useful tool for the evaluation and the comparison of tests.

In this paper, we propose a jackknife empirical likelihood method for making inference on partial AUCs. We combine the empirical likelihood function with suitable jackknife pseudo-values obtained from a nonparametric estimator of the normalized partial AUC, and derive a pseudo-likelihood that can be used to construct confidence intervals or perform tests of hypotheses. A theoretical justification of the proposed method is given. Moreover, our simulation results indicate that the jackknife empirical likelihood based confidence intervals compare favorably with alternatives in terms of coverage probability. Finally, the approach discussed in the paper is extended to inference on the difference between two partial AUCs, so that the method can also be used for comparing tests.

Overall, results in the paper seem to confirm that the jackknife empirical likelihood is a potentially useful tool in ROC analysis and, more generally, that is worthy of serious consideration in statistical inference, due to its relative simplicity. However, we remark that the jackknife empirical likelihood is not a genuine empirical likelihood function, but it is a pseudo-likelihood function which can be used as a surrogate to markedly reduce the computational burden. As a consequence, the jackknife empirical likelihood does not retain all features of the empirical likelihood function. In general, for example, the jackknife empirical likelihood based confidence intervals are not range preserving. Therefore, further studies, in particular comparing theoretical properties of jackknife empirical likeliood and empirical likelihood, would be highly desirable.

References

Adimari, G. and Guolo. A. (2010). A note on the asymptotic behaviour of empirical likelihood statistics. *Stat. Methods Appl.* **19**, 463–476.

Asuncion, A. and Newman, D. J. (2007). UCI Machine Learning Repository `http://www.ics.uci.edu/~mlearn/MLRepository.html`. University of California, School of Information and Computer Science, Irvine, CA.

Dodd, L. E. and Pepe, M. S. (2003). Partial AUC estimation and regression. *Biometrics* **59**, 614-623.

Gong, Y., Peng, L. and Qi, Y. (2010). Smoothed jackknife empirical likelihood method for ROC curve. *J. Multivariate Anal.* **101**, 1520-1531.

He, Y. and Escobar, M. (2008). Nonparametric statistical inference method for partial areas under receiver operating characteristic curves, with application to genomic studies. *Statist. Medicine* **27**, 5291-5308.

Jiang, Y., Metz, C. E. and Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* **201**, 745-750.

Jing, B.-Y., Yuan, J. and Zhou, W. (2009). Jackknife empirical likelihood. *J. Amer. Statist. Assoc.* **104**, 1224-1232.

Liu, A., Schisterman, E. F. and Wu, C. (2005). Nonparametric estimation and hypothesis testing on the partial area under receiver operating characteristic curves. *Comm. Statist. Theory Methods* **34**, 2077-2088.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Med. Decis. Making* **9**, 190-195.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall Ltd.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of roc curves. *Statist. Medicine* **8**, 1277-1290.

Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* **83**, 835-848.

Wieand, S., Gail, M. H., James, B. R. and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**, 585-592.

Wood, A. T. A., Do, K.-A. and Broom, B. M. (1996). Sequential linearization of empirical likelihood constraints with application to U-statistics. *J. Comput. Graph. Statist.* **5**, 365-385.

Zhang, D. D., Zhou, X.-H., Daniel H. Freeman, J. and Freeman, J. L. (2002). A non-parametric method for the comparison of partial areas under roc curves and its application to large health care data sets. *Statist. Medicine* **21**, 701-715.

Department of Statistical Sciences, University of Padova Via C. Battisti, 241–243, 35121 Padova, Italy.

E-mail: gianfranco.adimari@unipd.it

Department of Statistical Sciences, University of Padova Via C. Battisti, 241–243, 35121 Padova, Italy.

E-mail: monica.chiogna@unipd.it

(Received March 2011; accepted October 2011)