Stat Papers (2019) 60:2161–2183
https://doi.org/10.1007/s00362-017-0915-7

REGULAR ARTICLE

Minimum distance estimation of the binormal ROC curve

Alicja Jokiel-Rokita[1] · **Rafał Topolnicki**[1]

Received: 6 July 2016 / Revised: 23 February 2017 / Published online: 25 May 2017

Abstract The receiver operating characteristic (ROC) curve describes the performance of a diagnostic test, which classifies individuals into one of two categories. Many parametric, semiparametric and nonparametric estimation methods have been proposed for estimating the ROC curve and its functionals. In this paper the minimum distance estimation of the binormal ROC curve is considered. A modification of the estimator considered in the paper of Davidov and Nov (J Stat Plan Inference 142(4):872–877, 2012) and some new estimators are proposed. We compare the accuracy of the new estimators with known minimum distance estimators of the binormal ROC curve and we conclude that our estimators generally perform better than their competitors.

Keywords Receiver operating characteristic (ROC) curve · Binormal model · Semiparametric estimation · Minimum distance estimation (MDE) · Bayesian bootstrap (BB)

1 Introduction

The receiver operating characteristic (ROC) curve is commonly used to describe the accuracy of a medical or other diagnostic test, which classifies individuals into "non-diseased" and "diseased" categories. It is defined as a plot of the true positive rate against the false positive rate, or sensitivity versus 1-specificity, for various threshold values. Over the years, it has been widely applied in many fields including

✉ Alicja Jokiel-Rokita
alicja.jokiel-rokita@pwr.edu.pl

1 Faculty of Pure and Applied Mathematics, Wroclaw University of Science and Technology, Wrocław, Poland

biosciences, data mining, experimental psychology, finance, geosciences, machine learning, medicine, radiology, sociology and others. For comprehensive review of the literature, see Zhou et al. (2002), Pepe (2003), Krzanowski and Hand (2009) and Gonçalves et al. (2014).

More precisely, let X and Y be the test results from a non-diseased population and a diseased population, respectively. Let F be a continuous cumulative distribution function (cdf) of the random variable X, and G—a continuous cdf of the random variable Y. The ROC curve is defined as a plot of $1 - G(c)$ versus $1 - F(c)$ for $-\infty \leq c \leq \infty$, or equivalently as a plot

$$ROC(t) = 1 - G(F^{-1}(1 - t)), \tag{1}$$

against t, for $t \in [0, 1]$.

A special feature of the ROC curve is that it is invariant to any increasing transformation of the data, i.e. if $X' = h(X)$, and $Y' = h(Y)$, for some increasing transformation h, then the ROC curve corresponding to the distribution functions F and G is the same as the ROC curve corresponding to the distribution function F' and G' of the random variables X' and Y', respectively.

In this paper we consider the problem of estimation of the ROC curve in the binormal model, i.e. we assume that after some increasing transformation h, the random variables X' and Y' are normally distributed. Without loss of generality we can assume that $X' \sim \mathcal{N}(0, 1)$ and $Y' \sim \mathcal{N}(\mu, \sigma^2)$. In this case the ROC curve has a simple parametric form

$$ROC(t) = \Phi\Big(\frac{\mu}{\sigma} + \frac{1}{\sigma}\Phi^{-1}(t)\Big), \tag{2}$$

where Φ is the cumulative distribution function of the standard normal distribution. Thus, in the binormal model, the estimation of the ROC curve reduces to the estimation of the parameters μ and σ. The most common arguments in favor of using the binormal estimator are presented in Hanley (1988). Swets (1986) and Hanley (1988, 1996) also argue that the binormal estimator is robust.

Many different techniques have been proposed to solve the problem of semiparametric estimation of the ROC curve. For estimating ROC curves from discrete or grouped response data, the most commonly used procedure is that proposed by Dorfman and Alf (1969). Metz et al. (1998) developed an algorithm called LABROC, which groups continuous data into a finite number of ordered categories and then uses the maximum likelihood algorithm from Dorfman and Alf (1969). Hsieh and Turnbull (1996) proposed a generalized least squares procedure for grouped data and a minimum distance estimator (MDE), which does not require grouping data. MDE of the binormal ROC curve was also considered in the papers of Davidov and Nov (2009, 2012). In the papers of Zou and Hall (2000), Cai and Moskowitz (2004), Zhou and Lin (2008) maximum likelihood and pseudo-likelihood approach to estimate the binormal ROC curve was considered. Techniques based on regression were also proposed (see for example Lloyd 2002; Cai and Pepe 2002; Qin and Zhang 2003; Wan and Zhang 2007). Bayesian approach to the semiparametric estimation of the ROC

curve was considered in the papers of Branscum et al. (2008), Erkanli et al. (2006), Gu et al. (2008), Gu and Ghosal 2009. The paper Gonçalves et al. (2014) overviews some developments on the estimation of the ROC curve with the particular emphasis on some frequentist and Bayesian methods which have been mostly employed in the medical setting.

This paper deals with minimum distance estimation of the binormal ROC curve. To the best of our knowledge, a minimum distance approach to estimating the binormal ROC curve parameters was considered only by Hsieh and Turnbull (1996) and Davidov and Nov (2009, 2012). In the paper of Davidov and Nov (2009) the central idea was to estimate the unknown function h (a transformation of X and Y to normal random variables) in two different ways; only one of the two estimates depended on the unknown parameters μ and σ of the binormal ROC curve. Then, they estimated μ and σ by the values that minimized a certain norm of the difference between the estimates of the function h. In this paper we do not develop this idea. A different approach is presented in the papers of Hsieh and Turnbull (1996) and Davidov and Nov (2012). They took into consideration two different measures of distance between the empirical and the theoretical ordinal dominance curve (ODC), the curve closely related to the ROC curve. Davidov and Nov (2012) showed that their MDE is consistent and asymptotically normally distributed and it outperforms Hsieh and Turnbull's original, grouped-data estimator, but it has not been compared with the Hsieh and Turnbull's MDE estimator.

In this paper we compare the accuracy of the known MDE's given by Hsieh and Turnbull (1996) and Davidov and Nov (2012). We obtain that the MDE given by Hsieh and Turnbull (1996) outperforms, in some sense, MDE given by Davidov and Nov (2012). Both of the estimators are obtained by minimization of distance measures between the unknown binormal and empirical ROC curve. Empirical ROC curve, as a step function, often gives unsatisfactory nonparametric estimators of the ROC curve in the case of small sample sizes. Therefore, the second purpose of this work is to introduce modifications of these known measures of distance by replacing the underlaying empirical ROC curve by its continuous nonparametric counterparts. Another modification of Davidov and Nov (2012) approach stems from widening the domain taken into account when the distance between empirical and binormal ROC curve is calculated. In this paper, a total of seven new estimators in binormal model are introduced and their performances are compared in the simulation study.

The paper is organized as follows. In Sect. 2 we recall the MDE's of the binormal ROC curve parameters considered in the papers of Hsieh and Turnbull (1996) and Davidov and Nov (2012). Then we propose a modification of the Davidov and Nov estimator, and some new MDE's by replacing the empirical ROC curve by the Bayesian bootstrap estimator of the ROC curve (see Gu and Ghosal 2008) in measures of distance considered by Hsieh and Turnbull (1996) and Davidov and Nov (2012). We prove the consistency of the estimators proposed. We also recall two smooth nonparametric estimators of the ROC curve, namely the kernel estimator considered by Lloyd (1998), and the estimator proposed by Jokiel-Rokita and Pulit (2013), which we also use to obtain MDE's of the binormal ROC curves. Results from simulation studies are provided in Sect. 3. In Sect. 4 real data analysis is discussed. The paper ends with some concluding remarks in Sect. 5.

2 Minimum distance estimation of the ROC curve

In this section, we recall some known methods and provide some new methods of estimation of the parameters μ and σ in the binormal model, basing on the minimum distance concept. Minimum distance estimation has been studied extensively beginning with the work of Wolfowitz (1957). The concept of minimum distance estimation of the binormal ROC curve parameters was introduced in framework of estimation of binormal ordinal dominance curve (ODC) given by $D(t) = F(G^{-1}(t))$, $t \in [0, 1]$. The ODC curve is closely related to the ROC curve and in the binormal model it has the following parametric form

$$D(t) = \Phi(\mu + \sigma\Phi^{-1}(t)).$$

However, in course of this paper, we find more convenient to construct all estimators of the unknown parameters μ and σ in the direct reference to the ROC curves. Therefore all results originally established for ODC curves will be rephrased in terms of ROC curves.

2.1 Minimum distance estimator of Hsieh and Turnbull

Assume that independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from distributions with cdf's F and G, respectively, are available. Denote by F_m and G_n the empirical distribution functions of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, respectively, and the empirical quantile function by $G_n^{-1}(t) = \inf\{y : G_n(y) \geq t\}$. The empirical ROC curve is defined as

$$ROC_{mn}(t) = 1 - G_n(F_m^{-1}(1-t)), \quad t \in (0, 1), \tag{3}$$

while the empirical ODC curve is given by

$$D_{mn}(t) = F_m(G_n^{-1}(t)), \quad t \in (0, 1).$$

In the paper of Hsieh and Turnbull (1996), MDE's of the ROC curve parameters are derived by finding the ODC curve that fits most closely to the empirical ODC curve using a L_2 norm criterion. We adopt the original idea introduced by Hsieh and Turnbull (1996). More precisely, for $\theta = (\mu, \sigma)^T$, let us denote by

$$\xi_{mn}(\theta) = ROC_{mn}(t) - \Phi\Big(\frac{\mu}{\sigma} + \frac{1}{\sigma}\Phi^{-1}(t)\Big), \tag{4}$$

and

$$\|\xi_{mn}(\theta)\| = \int_0^1 \xi_{mn}^2(\theta)dt \tag{5}$$

the L_2-distance measure between $ROC(t)$ and $ROC_{mn}(t)$.

The MDE $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma})^T$ of the parameter θ is defined by

$$\|\xi_{mn}(\widehat{\theta})\| = \inf_{\theta\in\Theta} \|\xi_{mn}(\theta)\|, \tag{6}$$

where $\Theta = \{\theta = (\mu, \sigma)' : \mu \in \mathbb{R}, \sigma > 1\}$, as in the paper of Hsieh and Turnbull (1996). The restriction that $\sigma > 1$ is not unreasonable if one thinks of the healthy response as "noise" and the diseased response as "noise plus signal". However, we can avoid this restriction if we modify the distance criterion (5) so that the integral is over a closed interval excluding 0 and 1. In the sequel, we will denote the MDE estimator $\widehat{\theta}$ by $\widehat{\theta}_{HT} = (\widehat{\mu}_{HT}, \widehat{\sigma}_{HT})$. Using the theory developed by Millar (1984), Hsieh and Turnbull (1996) proved the asymptotic normality of their MDE of the parameter θ, but did not provide any concrete procedure to compute them. In Sect. 3, we describe an algorithm, used in the simulation study, to obtain the estimates $\hat{\theta}_{HT}$.

2.2 Minimum distance estimator of Davidov and Nov

Hsieh and Turnbull (1996) also proposed (in Remark 1), as an object for future research, to modify their measure of distance by applying the Φ^{-1} transformation to both $D_{mn}(t)$ and $D(t)$ which, in terms of the ROC curve, leads to following counterpart

$$\nu_{mn}(\theta) = \Phi^{-1}(ROC_{mn}(t)) - \left(\frac{\mu}{\sigma} + \frac{1}{\sigma}\Phi^{-1}(t)\right) \tag{7}$$

of $\xi_{mn}(\theta)$. Davidov and Nov (2012) followed on this suggestion and considered estimation of the parameter θ based on minimization of the following objective function

$$\|\nu_{mn}(\theta)\| = \int_a^b \nu_{mn}^2(\theta)dt, \tag{8}$$

where the integration endpoints $0 < a < b < 1$ ensures that the last integral is finite. Namely, they considered the MDE

$$\widehat{\theta}_{DN} := (\widehat{\mu}_{DN}, \widehat{\sigma}_{DN}) = \arg\min_{\mu,\sigma} \int_a^b \left[\Phi^{-1}(ROC_{mn}(t)) - \left(\frac{\mu}{\sigma} + \frac{1}{\sigma}\Phi^{-1}(t)\right)\right]^2 dt, \tag{9}$$

where

$$a = \min\{i/m : ROC_{mn}(i/m) > 0, i = 1, \ldots, m\}, \tag{10}$$

$$b = \max\{i/m : ROC_{mn}(i/m) < 1, i = 1, \ldots, m\}. \tag{11}$$

The minimization problem given by (9) is convex and quadratic in μ and σ and, unlike (6), it enjoys a closed-form solution

$$\widehat{\mu}_{DN} = \widehat{\sigma}_{DN}\widehat{S}_1 - \widehat{S}_3, \qquad \widehat{\sigma}_{DN} = \frac{\widehat{S}_4 - \widehat{S}_3^2}{\widehat{S}_2 - \widehat{S}_1\widehat{S}_3},$$

where

$$\widehat{S}_1 = \frac{1}{b-a}\int_a^b \Phi^{-1}(ROC_{mn}(t))dt, \tag{12}$$

$$\widehat{S}_2 = \frac{1}{b-a}\int_a^b \Phi^{-1}(ROC_{mn}(t))\Phi^{-1}(t)dt, \tag{13}$$

$$\widehat{S}_3 = \frac{1}{b-a}\int_a^b \Phi^{-1}(t)dt, \tag{14}$$

$$\widehat{S}_4 = \frac{1}{b-a}\int_a^b \left(\Phi^{-1}(t)\right)^2 dt. \tag{15}$$

Please note that since we employed the ROC instead of the ODC curve, the formulas (12)–(15) differ from corresponding Davidov and Nov's (2012) formulas.

The integration endpoints a, b were introduced to ensure that $\Phi^{-1}(ROC_{mn}(t)) \neq \pm\infty$ and hence that optimization problem (9) is well-defined. However, the selection of the upper integral limit according to Eq. (11) causes that the difference between the empirical ROC curve and the true (binormal) ROC curve on the interval $[b, c]$, where $c := \min\{i/m : ROC_{mn}(i/m) = 1, i = 1, \ldots, m\}$ (on the last step of the ROC_{mn}) is not taken into account. We think that this loss of information influences the accuracy of estimates for small samples sizes m and n. Hence, we propose a modification of the minimum distance estimator considered by Davidov and Nov by choosing the upper limit of integration just before the last jump of the empirical ROC curve. Since $ROC_{mn}(t)$ is right-continuous, we take

$$b'_m = \sup\{t \in [0, 1] : ROC_{mn}(t) < 1\} - \varepsilon_m, \tag{16}$$

where $\varepsilon_m < 1/m$ is a positive constant, which guarantees that $\Phi^{-1}(ROC_{mn}(t)) < \infty$. Moreover, thanks to the right continuity of the empirical ROC curve, there is no need to introduce any modification for the lower integration endpoint (the lowest possible value is already provided by formula (10)).

The estimates of the parameters μ and σ computed with b'_m instead of b in (12)–(15) will be denoted by $\hat{\mu}_{DNM}$ and $\hat{\sigma}_{DNM}$, respectively. It is clear, that those modified estimators are consistent and asymptotically normal as the original estimators of Davidov and Nov (see Davidov and Nov 2012, Theorems 1 and 2), under the same assumptions.

2.3 Minimum distance estimators of the binormal ROC curve parameters based on BB estimator of the ROC curve

In the paper of Gu and Ghosal (2008) the Bayesian bootstrap (BB) for the nonparametric estimation of the ROC curve and its functionals has been proposed (see also Gu et al. 2008). In this approach stochastic empirical distribution functions, introduced by Rubin (1981), are employed. Let $U_1, \ldots, U_{m-1}$ be iid uniform $\mathcal{U}(0, 1)$ random variables, independent of data. Rubin's stochastic empirical distribution function, say $F_m^{(b)}$, based on the sample $X_1, \ldots, X_m$, is defied as follows

$$F_m^{(b)}(x) = \begin{cases} 0, & \text{when } x < X_{(1)}, \\ U_{(i)}, & \text{when } X_{(i)} \leq x < X_{(i+1)}, 1 \leq i \leq m-1, \\ 1, & \text{when } x \geq X_{(m)}, \end{cases} \tag{17}$$

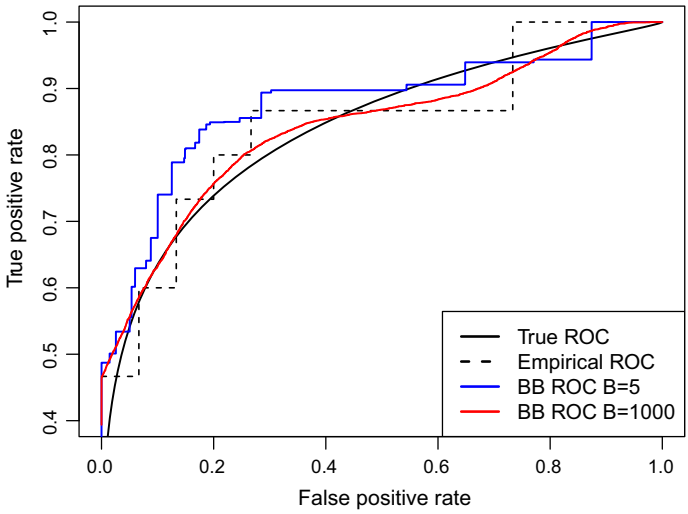

Fig. 1 Comparison of empirical and BB estimates of ROC curve with the true binormal ROC curve for $\mu = 1.8$ and $\sigma = 1.5$. The estimates are based on samples of sizes $m = n = 15$.

where $U_{(i)}$ denotes i-th order statistic of the vector $(U_1, \ldots, U_{m-1})$. The function $F_m^{(b)}$ is a step function which at each point $X_{(i)}$, $i = 1, \ldots, m$, jumps up by the random value $U_{(i)} - U_{(i-1)}$, where $U_{(0)} = 0, U_{(m)} = 1$. Let $G_n^{(b)}$ be Rubin's stochastic empirical distribution function based on the observations $Y_1, \ldots, Y_n$ from the second sample. In order to get a ROC curve estimator, say $ROC_{mn}^{(b)}$, we proceed in the same way as in the case of empirical ROC curve given by (3), and plug in Rubin's stochastic empirical distribution function $G_n^{(b)}$ and quantile function $F_m^{(b)-1}$ into (1). Next the BB estimate of the ROC curve is obtained by averaging over a large number of $ROC_{mn}^{(b)}$ realizations, i.e.

$$ROC_{mn}^{BB}(t) = \frac{1}{B} \sum_{b=1}^{B} ROC_{mn}^{(b)}(t).$$

The estimator ROC_{mn}^{BB} is a bandwidth-free nonparametric estimator and, because of averaging over two random variations, is "smoother" than ROC_{mn}. The BB estimates of the ROC curve for two different values of B, based on the samples of equal sizes $n = m = 15$, together with the empirical and the true ROC curve, are presented in Fig. 1. As can be seen, that even when we average over a small number of realizations, we obtain "smoother" estimate than the empirical ROC curve.

Remark 1 An efficient three-step procedure for computing BB estimates, which does not require inverting the stochastic empirical distribution function (17), was proposed by Gu et al. (2008). In the first step auxiliary variables Z_j are defined, based on BB resampling distribution,

$$Z_j = 1 - F^{\#}(Y_j) = 1 - \sum_{i=1}^{m} p_i I(X_i \leq Y_j),$$

where $(p_1, \ldots, p_m) \sim Dirichlet(m; 1, \ldots, 1)$ independent of others. In the second step a random realization of ROC curve, $ROC^{\#}_{mn}$, is generated as randomized distribution function of $Z_1, \ldots, Z_n$; we have

$$ROC^{\#}_{mn}(t) = \sum_{j=1}^{n} q_j I(Z_j \leq t),$$

where $(q_1, \ldots, q_n) \sim Dirichlet(n; 1, \ldots, 1)$ independent of others. In the last step the BB estimate of ROC curve is obtained by averaging over the ensemble of random ROC curves $ROC^{BB}_{mn}(t) = mean(ROC^{\#}_{mn}(t))$. A convenient method for generating $(p_1, \ldots, p_m) \sim Dirichlet(m; 1, \ldots, 1)$ was also proposed by Gu.

Let us assume that

$$\sup\{x : F(x) = 0\} = \sup\{x : G(x) = 0\} := \alpha$$

and

$$\inf\{x : F(x) = 1\} = \inf\{x : G(x) = 1\} := \beta.$$

Moreover, throughout this section we assume that the sample sizes m, n are such that $m = m(n)$ and $n/m \to \lambda \in (0, \infty)$ as $n \to \infty$, and that the following two conditions are satisfied

(C1) The continuous cdf F is twice differentiable on (α, β), the derivative $F' = f \neq 0$ on (α, β), and for some $\gamma > 0$,

$$\sup_{x \in (\alpha, \beta)} \left\{ F(x)(1 - F(x)) |f'(x)/f^2(x)| \right\} \leq \gamma.$$

(C2) Let cdf's F and G satisfy Condition 1, and additionally

$$\sup_{x \in (\alpha, \beta)} \left\{ F(x)(1 - F(x)) \left| \frac{g'(x)}{f^2(x)} \right| \right\} < \infty, \quad \sup_{x \in (\alpha, \beta)} \left\{ F(x)(1 - F(x)) \left| \frac{g(x)}{f(x)} \right| \right\} < \infty.$$

Using the theory of Kiefer processes, Gu and Ghosal (2008) proved some strong approximation results and asymptotic properties of the Bayesian bootstrap ROC curve estimator. In particular, its rate of convergence to the true ROC curve was shown to be $n^{-1/2}$.

We will consider minimum distance estimation of the binormal ROC curve parameters by replacing the empirical ROC curve with corresponding BB estimator $ROC^{BB}_{mn}(t)$ in measure (8). Since jumps of $ROC^{BB}_{mn}(t)$ are random we can choose the integration limits in (12)–(15) to be closer to 0 and 1 then in the original procedure. Namely we define

$$a'_m = \inf \left\{ t \in [0, 1] : ROC^{BB}_{mn}(t) > 0 \right\}, \tag{18}$$

$$b'_m = \sup \left\{ t \in [0, 1] : ROC^{BB}_{mn}(t) < 1 \right\} - \varepsilon_m, \tag{19}$$

where $\varepsilon_m < 1/m$ is a positive constant, which need to be introduced due to right continuity of ROC_{mn}^{BB} function (analogously to (16)). To be more specific, we consider the MDE

$$\begin{aligned}\widehat{\theta}_{DNB} &= (\widehat{\mu}_{DNB}, \widehat{\sigma}_{DNB}) \\ &:= \arg\min_{\mu,\sigma} \int_{a'_m}^{b'_m} \left[\Phi^{-1}(ROC_{mn}^{BB}(t)) - \left(\frac{\mu}{\sigma} + \frac{1}{\sigma}\Phi^{-1}(t)\right)\right]^2 dt.\end{aligned}$$

Using the same approach as in Sect. 2.2, one can show that the solution to the optimization problem above is given by

$$\hat{\mu}_{DNB} = \hat{\sigma}_{DNB}\tilde{S}_1 - \tilde{S}_3, \qquad \hat{\sigma}_{DNB} = \frac{\tilde{S}_4 - \tilde{S}_3^2}{\tilde{S}_2 - \tilde{S}_1\tilde{S}_3}, \tag{20}$$

where

$$\begin{aligned}\tilde{S}_1 &= \frac{1}{b'_m - a'_m}\int_{a'_m}^{b'_m} \Phi^{-1}(ROC_{mn}^{BB}(t))dt, \\ \tilde{S}_2 &= \frac{1}{b'_m - a'_m}\int_{a'_m}^{b'_m} \Phi^{-1}(ROC_{mn}^{BB}(t))\Phi^{-1}(t)dt,\end{aligned}$$

are the counterparts of Eqs. (12), (13), respectively. Similarly, $\tilde{S}_3$ and $\tilde{S}_4$ are computed by changing the integration domain from (a, b) to (a'_m, b'_m) in Eqs. (14)–(15).

The following lemma can be proved in an analogous manner to Lemma 1 in Davidov and Nov (2012).

Lemma 1 *Under the above assumptions, $a'_m \to 0$ and $b'_m \to 1$ a.s., as $m \to \infty$.*

Denote

$$ROC_{mn}^{DNB}(t) = \Phi\left(\frac{\hat{\mu}_{DNB}}{\hat{\sigma}_{DNB}} + \frac{1}{\hat{\sigma}_{DNB}}\Phi^{-1}(t)\right).$$

Theorem 1 *Under assumptions (C1)–(C2), $\hat{\mu}_{DNB} \to \mu$ and $\hat{\sigma}_{DNB} \to \sigma$ in probability, as $n \to \infty$, and hence the estimator ROC_{mn}^{DNB} of the binormal ROC curve converges pointwise to the true ROC curve on* $(0, 1)$.

A proof of Theorem 1 is given in in Appendix.

We will also consider an estimator of the parameter ϑ, which combines the minimum distance concept of Hsieh and Turnbull with the BB nonparametric estimator of the ROC curve. In this method, Eq. (4) is modified by replacing the empirical $ROC_{mn}(t)$ curve with the Bayesian bootstrap estimator $ROC_{mn}^{BB}(t)$ which gives

$$\xi_{mn}^{BB}(\theta) = ROC_{mn}^{BB}(t) - \Phi\left(\frac{\mu}{\sigma} + \frac{1}{\sigma}\Phi^{-1}(t)\right),$$

and the corresponding L_2-distance measure is

$$\|\xi_{mn}^{BB}(\theta)\| = \int_0^1 \xi_{mn}^{BB^2}(\theta)dt. \tag{21}$$

The minimum distance estimate $\hat{\theta}_{HTB} = (\hat{\mu}_{HTB}, \hat{\sigma}_{HTB})$ of the parameter θ is defined as the value which minimizes (21), i.e.

$$\|\xi_{mn}^{BB}(\hat{\theta}_{HTB})\| = \inf_{\theta} \|\xi_{mn}^{BB}(\theta)\|.$$

2.4 Minimum distance estimators of the binormal ROC curve parameters based on smooth nonparametric estimators of the ROC curve

The empirical ROC curve retains many properties of the empirical distribution function. It is uniformly convergent to the theoretical curve (Hsieh and Turnbull 1996), but it is also not continuous and not very accurate for small sample sizes. The idea behind semiparametric procedures of Hsieh and Turnbull, as well as Davidov and Nov, is to minimize a distance between binormal ROC curve given by (2), and the empirical one. In this section we propose MDE's of the binormal curve by replacing the empirical ROC curve, in measures (5) and (8), by its continuous nonparametric counterparts. Consequently, each considered nonparametric estimator of the ROC curve leads to two new semiparametric minimum distance estimators.

2.4.1 Kernel estimator of the ROC curve

Lloyd (1998) used the kernel smoothing technique to obtain a smooth ROC curve estimator given by

$$ROC_{mn}^{K}(t) = 1 - G_n^K(F_m^{K^{-1}}(1-t)), \quad t \in [0,1], \tag{22}$$

where

$$F_m^K(x) = \frac{1}{m}\sum_{j=1}^{m} \mathcal{K}\left(\frac{x - X_j}{h_m}\right), \quad G_n^K(x) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{K}\left(\frac{x - Y_j}{h_n}\right)$$

are standard kernel estimators with kernel function K, $\mathcal{K}(v) = \int_{-\infty}^{v} K(z)dz$ and bandwidth parameters h_n and h_m. Lloyd and Yong (1999) showed that estimator (22) has better mean squared error properties than the empirical ROC curve. In the problem of kernel density estimation, choosing between many available kernel functions is of secondary importance as all give comparable results, but more care needs to be taken over the selection of bandwidth. Therefore, in the kernel ROC curve estimation the main emphasis is put on the bandwidth selection (Zhou and Harezlak 2002, Hall and Hyndman 2003). In the Simulation study (Sect. 3), the Gaussian kernel is employed and the bandwidth parameter h_m is chosen according to

$$h_m = 0.9 \min(s_x, iqr_x/1.34) m^{-1.5},$$

where s_x and iqr_x are the standard deviation and the interquartile range for non-diseased population, respectively. The bandwidth parameter h_n for diseased population was determined in the same way. This method of bandwidth selection was recommended by Silverman (1986) as it works 'very well for a wide range of densities', which is reasonable in our case, since we have no information about samples distribution.

Kernel estimator (22) of the ROC curve allows us to introduce two new minimum distance estimators of the binormal ROC curve parameters which will be denoted by $\hat{\theta}_{HTK}$ and $\hat{\theta}_{DNK}$. The first one employs the $ROC^K_{mn}(t)$ instead of the empirical ROC curve in Eq. (4), while the latter—in Eq. (7), e.g.

$$\hat{\theta}_{HTK} = (\hat{\mu}_{HTK}, \hat{\sigma}_{HTK}) = \arg\min_{\mu,\sigma} \int_0^1 \left[ROC^K_{mn}(t) - \Phi\Big(\frac{\mu}{\sigma} + \frac{1}{\sigma}\Phi^{-1}(t)\Big)\right]^2 dt,$$
$$\hat{\theta}_{DNK} = (\hat{\mu}_{DNK}, \hat{\sigma}_{DNK}) = \arg\min_{\mu,\sigma} \int_{a'}^{b'} \left[\Phi^{-1}(ROC^K_{mn}(t)) - \Big(\frac{\mu}{\sigma} + \frac{1}{\sigma}\Phi^{-1}(t)\Big)\right]^2 dt,$$

where the integration limits a' and b' are the counterparts of Eqs. (18)–(19), where $ROC^{BB}_{mn}(t)$ is replaced with $ROC^K_{mn}(t)$.

2.4.2 Estimator of the ROC curve by smoothing the sample distribution functions

In the paper of Jokiel-Rokita and Pulit (2013), the authors proposed to estimate the ROC curve using the plug in method with smoothed sample distribution functions. Let $X_{1:m} \le X_{2:m} \le \cdots \le X_{m:m}$ and $Y_{1:n} \le Y_{2:n} \le \cdots \le Y_{n:n}$ denote order statistics from the samples $\boldsymbol{X}_m$ and $\boldsymbol{Y}_n$, respectively. We set

$$X_{0:m} = 2L - X_{1:m}, \quad X_{(m+1):m} = 2U - X_{m:m},$$
$$Y_{0:n} = 2L - Y_{1:n}, \quad Y_{(n+1):n} = 2U - Y_{n:n},$$

where L, U are random variables such that $L \le \min\{X_{1:m}, Y_{1:n}\}$ and $U \ge \max\{X_{m:m}, Y_{n:n}\}$ almost surely. Denote

$$Q_j(\boldsymbol{X}_m) = \frac{X_{(j-1):m} + X_{j:m}}{2}, \quad j = 1, 2, \ldots, m+1,$$
$$R_j(\boldsymbol{X}_m) = Q_{j+1}(\boldsymbol{X}_m) - Q_j(\boldsymbol{X}_m) = \frac{X_{(j+1):m} - X_{(j-1):m}}{2}, \quad j = 1, 2, \ldots, m,$$
$$Q_j(\boldsymbol{Y}_n) = \frac{Y_{(j-1):n} + Y_{j:n}}{2}, \quad j = 1, 2, \ldots, n+1,$$
$$R_j(\boldsymbol{Y}_n) = Q_{j+1}(\boldsymbol{Y}_n) - Q_j(\boldsymbol{Y}_n) = \frac{Y_{(j+1):n} - Y_{(j-1):n}}{2}, \quad j = 1, 2, \ldots, n.$$

With this notation we define the estimators of the distribution functions F, G by

$$F_m^S(x) = \frac{1}{m}\sum_{j=1}^{m} T\left(\frac{x - Q_j(\boldsymbol{X}_m)}{R_j(\boldsymbol{X}_m)}\right),$$
$$G_n^S(x) = \frac{1}{n}\sum_{j=1}^{n} T\left(\frac{x - Q_j(\boldsymbol{Y}_n)}{R_j(\boldsymbol{Y}_n)}\right),$$

respectively, where

$$T(x) = \begin{cases} 0, & \text{for } x < 0, \\ r(x), & \text{for } 0 \le x \le 1, \\ 1, & \text{for } x > 1, \end{cases} \tag{23}$$

where $r\colon [0,1] \to [0,1]$ is a continuous, strictly increasing function such that $r(0) = 0$, $r(1) = 1$, e.g. $r(x) = x$. The inverse function of $F_m^S(t)$ on $[L, U]$ can be written as

$$F_m^{S^{-1}}(t) = \begin{cases} L, & \text{for } t = 0, \\ r^{-1}(mt - (k-1))R_k(\boldsymbol{X}_m) + Q_k(\boldsymbol{X}_m), & \text{for} \frac{k-1}{m} < t \le \frac{k}{m}, k = 1, \ldots, m. \end{cases}$$

It is clear that $F_m^{S^{-1}}(t)$ is continuous and strictly increasing on $[0, 1]$. Since $G_n^S(t)$ is continuous and strictly increasing on $[L, U]$, it follows that the composition $G_n^S(F_m^{S^{-1}}(t))$ is continuous and strictly increasing on $[0, 1]$. Hence we can define the continuous and strictly increasing nonparametric ROC curve estimator by

$$ROC_{mn}^S(t) = 1 - G_n^S(F_m^{S^{-1}}(1-t)), \quad t \in [0, 1]. \tag{24}$$

An appropriate choice of the function r, appearing in formula (23), can guarantee differentiability of the estimator (e.g. if function r is differentiable and $r'_+(0) = r'_-(1) = 0$). Simultaneously, determination of the estimator (24) remains as easy as in the case of the empirical ROC curve.

Minimum distance estimators of the parameter θ, based on the nonparametric ROC curve estimator ROC_{mn}^S applied in (4) and (7) instead of the estimator ROC_{mn}, will be denoted by $\hat{\theta}_{HTS}$ and $\hat{\theta}_{DNS}$, respectively.

3 Simulation study

A simulation experiment was conducted in order to

- Investigate the accuracy of the original minimum distance estimators considered by Davidov and Nov (2012) in comparison with their modification proposed in Sect. 2.2,
- Compare the accuracy of the minimum distance estimators of the binormal ROC curve parameters proposed by Hsieh and Turnbull (1996) with those considered by Davidov and Nov (2012) (answer the question: which measure of distance provides more accurate estimators),

- Compare the accuracy of the minimum distance estimators considered by Hsieh and Turnbull (1996) and Davidov and Nov (2012) with their counterparts obtained by replacing the empirical ROC curve with BB estimator or with the smooth nonparametric estimators of the ROC curve (the kernel estimator and the estimator proposed by Jokiel-Rokita and Pulit 2013).

An important index connected with the ROC curve is the area under the curve, commonly denoted by

$$AUC = \int_0^1 ROC(t)dt. \tag{25}$$

It can be easily shown that in the model considered $AUC = P(X < Y)$. We considered binormal ROC curves which values of AUC were 0.75 and 0.85 and assumed that $X \sim \mathcal{N}(0, 1)$ and Y is normally distributed with standard deviation $\sigma \in \{1, 4/3, 2\}$ and mean value μ follows according to $\mu = \sqrt{1+\sigma^2}\Phi^{-1}(\text{AUC})$. For each ROC curve, 5000 data sets with $m = n \in \{15, 20, 100\}$ were generated. Next, for each data set, four nonparametric ROC curve estimators were computed: the empirical ROC curve $\widehat{ROC}_{mn}$, the smoothed estimator ROC^S_{mn} according to Eq. (24) with linking function $r(x) = x$, the kernel estimator ROC^K_{mn} given by formula (22), and the Bayesian bootstrap estimator ROC^{BB}_{mn} averaged over $B = 1000$ realizations.

All nonparametric estimators were calculated on regular grid with intervals length of 0.0001. For kernel estimator we additionally used four times denser support grid, in order to compute the inverse of the cdf estimator $F_m^{K\,-1}$ with sufficient accuracy. As it was tested, further increase of the grid density virtually did not alter the simulation results. Then semiparametric minimum distance estimators were calculated based on nonparametric ones. In study, nine distinct semiparametric estimators were considered: five based on minimum distance approach considered by Davidov and Nov (2012) (shortly D–N estimators) and four based on the measure of distance considered by Hsieh and Turnbull (1996) (shortly H–T estimators). For all D–N estimators, except the original DN, the integration endpoints were calculated according to equation (19) with proper nonparametric ROC estimator plugged in. In practice, due to the finite distance between grid points, there is no need to introduce the ε_n constant.

In Hsieh and Turnbull approach one need to numerically minimize the L_2-distance between the binormal ROC curve and considered nonparametric estimator. For the binormal model this problem corresponds to minimization of a function of two variables μ and σ. In simulations the Nelder–Mead method was employed to minimize the objective function and initial values of unknown parameters were calculated using corresponding DNM estimator.

The performance of estimators introduced in previous section is studied in two ways: by comparing the estimates of binormal parameters and by looking at the deviation of estimated ROC curve from it's true shape. In Table 1 estimated bias and MSE of parameters μ and σ are listed for four binormal models (with $\sigma = 1$ and $\sigma = 2$ and for two values of AUC: 0.75 and 0.85). In practice one is more interested in estimation of the ROC curve than the parameters of binormal model. Hence, in order to examine overall goodness of fit of the ROC curve estimator the mean integrated square error (MISE)

Table 1 Estimated bias and MSE (in parentheses) of the estimators of the binormal ROC curve parameters μ and σ

	σ		$n = m = 15$				$n = m = 20$				$n = m = 100$			
			$\hat{\mu}$		$\hat{\sigma}$		$\hat{\mu}$		$\hat{\sigma}$		$\hat{\mu}$		$\hat{\sigma}$	
AUC = 0.75	1	DN	0.248	(1.009)	0.365	(0.984)	0.160	(0.427)	0.224	(0.346)	0.022	(0.030)	0.026	(0.017)
		DNM	0.218	(0.557)	0.350	(0.484)	0.156	(0.292)	0.229	(0.221)	0.027	(0.030)	0.031	(0.016)
		DNS	0.028	(0.211)	0.055	(**0.119**)	0.051	(0.149)	0.061	(**0.074**)	0.012	(0.028)	0.013	(**0.013**)
		DNK	0.151	(0.423)	0.126	(0.362)	0.119	(0.243)	0.103	(0.199)	0.025	(0.030)	0.026	(0.016)
		DNB	0.145	(0.330)	0.147	(0.259)	0.112	(0.185)	0.089	(0.133)	0.020	(**0.027**)	**−0.001**	(0.020)
		HT	0.114	(0.392)	0.126	(0.287)	0.083	(0.228)	0.081	(0.162)	0.010	(0.029)	0.008	(0.020)
		HTS	**0.018**	(**0.183**)	**0.037**	(0.127)	**0.025**	(**0.146**)	**0.026**	(0.098)	**0.007**	(0.028)	0.005	(0.019)
		HTK	0.114	(0.390)	0.126	(0.286)	0.083	(0.227)	0.081	(0.162)	0.009	(0.029)	0.008	(0.020)
		HTB	0.197	(0.486)	0.290	(0.363)	0.142	(0.273)	0.203	(0.194)	0.018	(0.030)	0.029	(0.021)
	2	DN	0.662	(5.162)	0.793	(5.185)	0.348	(1.579)	0.409	(1.486)	0.025	(0.100)	0.032	(0.089)
		DNM	0.484	(2.316)	0.616	(2.432)	0.301	(1.045)	0.369	(0.967)	0.030	(0.096)	0.039	(0.085)
		DNS	**−0.078**	(0.452)	**−0.058**	(0.491)	**−0.058**	(0.374)	**−0.048**	(0.368)	−0.031	(0.084)	**−0.031**	(**0.071**)
		DNK	0.317	(2.060)	0.332	(2.544)	0.210	(0.899)	0.221	(1.002)	0.029	(0.096)	0.037	(0.085)
		DNB	0.376	(2.011)	0.482	(2.612)	0.253	(0.884)	0.324	(1.081)	0.049	(0.100)	0.070	(0.095)
		HT	0.326	(1.838)	0.412	(2.229)	0.201	(0.822)	0.258	(0.906)	0.022	(0.090)	0.041	(0.094)
		HTS	−0.135	(**0.358**)	−0.143	(**0.453**)	−0.111	(**0.288**)	−0.127	(**0.356**)	−0.029	(**0.074**)	−0.032	(0.077)
		HTK	0.329	(1.968)	0.416	(2.404)	0.201	(0.818)	0.259	(0.902)	**0.021**	(0.090)	0.041	(0.094)
		HTB	0.465	(2.215)	0.628	(2.730)	0.301	(0.967)	0.416	(1.082)	0.040	(0.094)	0.073	(0.100)
AUC = 0.85	1	DN	0.555	(2.651)	0.641	(2.127)	0.389	(1.553)	0.407	(1.024)	0.054	(0.051)	0.047	(0.028)
		DNM	0.442	(1.422)	0.561	(1.158)	0.321	(0.679)	0.361	(0.439)	0.061	(0.049)	0.053	(0.026)
		DNS	**0.011**	(0.215)	0.097	(**0.119**)	0.053	(0.191)	0.092	(**0.087**)	0.045	(0.045)	0.049	(**0.020**)
		DNK	0.228	(0.934)	0.195	(0.755)	0.190	(0.465)	0.134	(0.312)	0.054	(0.047)	0.044	(0.026)
		DNB	0.109	(0.472)	0.084	(0.356)	0.085	(0.234)	0.024	(0.160)	**−0.005**	(0.040)	−0.037	(0.031)

Table 1 continued

σ		$n = m = 15$				$n = m = 20$				$n = m = 100$			
		$\hat{\mu}$		$\hat{\sigma}$		$\hat{\mu}$		$\hat{\sigma}$		$\hat{\mu}$		$\hat{\sigma}$	
	HT	0.241	(0.850)	0.256	(0.565)	0.181	(0.432)	0.166	(0.247)	0.025	(0.041)	0.024	(0.027)
	HTS	−0.036	(**0.185**)	**0.032**	(0.129)	**−0.014**	(**0.143**)	**0.003**	(0.092)	0.009	(**0.035**)	**0.005**	(0.023)
	HTK	0.242	(0.848)	0.257	(0.563)	0.181	(0.432)	0.166	(0.247)	0.025	(0.041)	0.024	(0.027)
	HTB	0.384	(1.041)	0.440	(0.698)	0.294	(0.539)	0.314	(0.307)	0.045	(0.045)	0.055	(0.030)
2	DN	1.066	(10.041)	1.109	(7.524)	0.970	(9.465)	0.839	(5.630)	0.073	(0.189)	0.042	(0.125)
	DNM	0.825	(4.513)	0.928	(3.729)	0.760	(4.453)	0.685	(2.824)	0.078	(0.174)	0.046	(0.117)
	DNS	−0.340	(0.569)	**−0.173**	(0.564)	**−0.193**	(0.543)	**−0.136**	(0.485)	**−0.043**	(0.146)	−0.048	(0.106)
	DNK	0.332	(3.520)	0.335	(3.361)	0.400	(3.812)	0.283	(2.675)	0.069	(0.174)	0.038	(0.120)
	DNB	**0.310**	(3.368)	0.370	(3.384)	0.329	(3.548)	0.254	(2.651)	0.052	(0.186)	**0.023**	(0.147)
	HT	0.469	(3.071)	0.550	(2.881)	0.478	(3.265)	0.429	(2.278)	0.067	(0.144)	0.055	(0.121)
	HTS	−0.447	(**0.490**)	−0.308	(**0.510**)	−0.354	(**0.374**)	−0.324	(**0.440**)	−0.106	(**0.094**)	−0.123	(**0.101**)
	HTK	0.466	(2.995)	0.548	(2.831)	0.476	(3.258)	0.428	(2.266)	0.066	(0.144)	0.056	(0.121)
	HTB	0.678	(3.793)	0.774	(3.566)	0.647	(3.948)	0.603	(2.745)	0.103	(0.156)	0.096	(0.129)

Results are based on 5000 simulation replications per model and method. For each binormal model and for all considered sample sizes the best result is highlighted with bold font

Table 2 Simulated mean integrate square error, multiplied by 100, for AUC = 0.75

Estimator	$n = m = 15$			$n = m = 20$			$n = m = 100$		
	$\sigma = 1$	$\sigma = \frac{4}{3}$	$\sigma = 2$	$\sigma = 1$	$\sigma = \frac{4}{3}$	$\sigma = 2$	$\sigma = 1$	$\sigma = \frac{4}{3}$	$\sigma = 2$
ROC_{mn}	1.807	1.652	1.455	1.424	1.250	1.120	0.284	0.256	0.231
ROC_{mn}^{S}	1.395	1.274	1.195	1.161	1.010	0.966	0.271	0.242	0.228
ROC_{mn}^{K}	1.788	1.634	1.437	1.414	1.241	1.111	0.284	0.256	0.230
ROC_{mn}^{BB}	1.445	1.355	1.241	1.163	1.036	0.968	0.249	0.229	0.211
DN	1.512	1.366	1.218	1.141	0.985	0.912	0.198	0.181	0.176
DNM	1.344	1.209	1.093	1.039	0.895	0.843	0.193	0.177	0.172
DNS	1.329	**1.072**	**1.002**	**0.862**	**0.780**	**0.789**	**0.185**	**0.173**	**0.171**
DNK	1.479	1.370	1.293	1.119	0.981	0.941	0.194	0.178	0.172
DNB	**1.215**	1.183	1.143	0.975	0.897	0.881	0.199	0.183	0.173
HT	1.403	1.275	1.156	1.104	0.947	0.884	0.209	0.187	0.176
HTS	1.222	1.108	1.041	0.999	0.854	0.820	0.207	0.184	0.174
HTK	1.404	1.275	1.154	1.104	0.948	0.883	0.209	0.187	0.176
HTB	1.335	1.231	1.130	1.062	0.922	0.868	0.207	0.187	0.175

Results are based on 5000 simulation replications per model and method. In each column result with lowest MISE is given with bold font. MISE's for nonparametric estimators are given for completeness

$$\text{MISE} = E\Big(\int_0^1 \big(ROC(t) - \widehat{ROC}(t)\big)^2 dt\Big),$$

was estimated, where $\widehat{ROC}(t)$ stands for the considered ROC curve estimator. In Table 2 the estimated values of MISE (multiplied by 100, for brevity) are collected for three values of σ, AUC=0.75, and different sample sizes. Results corresponding to AUC=0.85 are given in Table 3. MISE's are presented for both semiparametric and nonparametric ROC curves estimates for comparison.

As can be seen from Table 1, there are quite big differences in accuracy between the original (DN) and the modified (DNM) minimum distance estimators of Davidov and Nov, even though the latter requires only a marginal modification in the computational procedure. For $m = n = 10$ and $m = n = 15$ estimated mean square errors of the DNM estimators of parameters μ and σ are significantly smaller (sometimes even by half) than the corresponding estimated errors of the original DN estimators. The bias for $\hat{\vartheta}_{DNM}$ is also smaller than the one for $\hat{\vartheta}_{DN}$, but the difference between them is less prominent. For large samples size, $m = n = 100$, when formulas (11) and (16) yields virtually the same integration endpoints, the DN and DNM procedures give almost the same biases and mean square errors, as expected. The DNM estimator outperforms the original Davidov and Nov (2012) estimator (DN) also in terms of mean integrated square error. The results given in Tables 2 and 3 indicate a reduction of MISE by approximately 10% in the case of small sample sizes and 3% for $m = n = 100$.

Table 3 Same as in Table 2, but for AUC = 0.85

Methods	$n = m = 15$			$n = m = 20$			$n = m = 100$		
	$\sigma = 1$	$\sigma = \frac{4}{3}$	$\sigma = 2$	$\sigma = 1$	$\sigma = \frac{4}{3}$	$\sigma = 2$	$\sigma = 1$	$\sigma = \frac{4}{3}$	$\sigma = 2$
ROC_{mn}	1.379	1.183	0.953	1.078	0.916	0.756	0.223	0.192	0.162
ROC^S_{mn}	0.971	0.892	0.892	0.790	0.700	0.709	0.201	0.176	0.178
ROC^K_{mn}	1.364	1.169	0.939	1.070	0.909	0.749	0.223	0.192	0.162
ROC^{BB}_{mn}	1.091	0.942	0.786	0.872	0.748	0.644	0.192	0.169	0.147
DN	1.242	1.025	0.811	0.909	0.753	0.632	0.159	0.140	0.128
DNM	1.037	0.860	**0.690**	0.781	0.653	**0.556**	0.153	0.135	0.124
DNS	0.811	**0.719**	0.738	**0.593**	**0.539**	0.566	**0.145**	0.131	0.127
DNK	1.046	0.978	0.949	0.802	0.732	0.712	0.153	0.137	0.125
DNB	**0.771**	0.767	0.750	0.645	0.620	0.617	0.161	0.150	0.131
HT	1.048	0.884	0.721	0.811	0.679	0.576	0.156	0.134	0.120
HTS	0.857	0.776	0.767	0.681	0.594	0.592	0.151	**0.129**	0.122
HTK	1.048	0.883	0.719	0.811	0.679	0.576	0.157	0.134	0.120
HTB	0.988	0.834	0.693	0.779	0.653	0.564	0.156	0.134	**0.120**

We find interesting to examine the accuracy of the estimates obtained by minimization of two distinct measures (5) and (8). In the case of small sample sizes $m = n = 15$ and $m = n = 20$, the HT procedure performs much better in terms of bias and mean square error than DNM, and hence also outperforms the DN, regardless of AUC and true value of parameter σ (cf. Table 1). For $m = n = 100$, the bias of $\hat{\mu}_{HT}$ remains much lower than the corresponding bias of $\hat{\mu}_{DN}$ and $\hat{\mu}_{DMN}$, while the differences in MSE between these estimators are reduced. Simultaneously, the HT method gives also smaller bias of the estimator of σ in comparison to DN and DNM procedures but in some cases it yields greater MSE. These conclusions also holds to a great extend when DNS estimator, based on smoothed nonparametric ROC curve, is compared with corresponding HTS estimator. At the same time, inspection of the results collected in Tables 2 and 3 reveals that estimators based on D–N approach, aside from the original DN, yielded better fit to the true ROC curve in terms of MISE than these originating from H–T procedure—in all models, expect one, estimates that gave the lowest MISE were obtained utilizing the distance measure considered by Davidov and Nov (2012).

Based on simulations, we may also address the influence of replacing the empirical ROC curve with other nonparametric estimators on the accuracy of estimated binormal ROC curve. In all considered models, semiparametric estimators based on smoothed empirical ROC curve, $ROC^S_{mn}(t)$, performed better than their counterparts based on empirical curve $ROC_{mn}(t)$ for both employed distance measures. The bias and MSE of $\hat{\mu}_{DNS}$ and $\hat{\sigma}_{DNS}$ are considerably smaller than of $\hat{\mu}_{DNM}$ and $\hat{\sigma}_{DNM}$, respectively. Similar conclusions can be drawn when compare HTS with original HT procedure. For small sample sizes, the mean square error for estimates of both parameters decreases, by factor of 4.5 on average, when underlaying empirical ROC curve is replaced with it's smoothed counterpart (24). Naturally, the advantage of estimates based on $ROC^S_{mn}(t)$

Table 4 Estimated parameters for Tupikowski's kidney cancer data for hemoglobin level (HB) and fibrinogen concentration (FC)

	HL			FC		
	$\hat{\mu}$	$\hat{\sigma}$	AUC	$\hat{\mu}$	$\hat{\sigma}$	AUC
DN	0.899	1.391	0.7002	0.709	1.134	0.6805
DNM	0.837	1.067	0.7165	0.688	0.998	0.6868
DNS	0.884	1.192	0.7149	0.782	1.117	0.6990
DNK	0.881	1.082	0.7250	0.786	1.085	0.7030
DNB	0.998	1.187	0.7399	0.859	1.111	0.7173
HT	0.857	1.301	0.6992	0.629	1.025	0.6699
HTS	0.853	1.333	0.6957	0.675	1.085	0.6763
HTK	0.915	1.337	0.7081	0.689	1.058	0.6820
HTB	0.877	1.135	0.7190	0.699	0.931	0.6955

over those based on $ROC_{mn}(t)$ decreases when sample size increases. However, no significant improvement of parameters estimates is observed when kernel or BB methods are employed. In the case of methods based on Davidov and Nov approach, when one minimizes the objective function given by (9), the estimated biases and MSE's of the estimators $\hat{\theta}_{DNK}$ and $\hat{\theta}_{DNB}$ are only slightly reduced with comparison to DNM method. Furthermore, for HTK and HTB methods even some increase of bias and MSE is observed in comparison to original minimum distance procedure of Hsieh and Turnbull. Replacing the underlaying empirical ROC curve with it's smoothed counterpart leads also to decrease of mean integrated square error of both semiparametric and nonparametric estimators. For eighteen binormal models considered in Tables 2 and 3 the DNS method always outperform the DN and in fifteen cases it yields smaller MISE than DNM estimator. In fact, for AUC = 0.75, the DNS estimator achieves the lowest MISE among all considered in 8 out of 9 comparisons. The HTS estimator exceeds the HT also in 15 out of 18 comparisons. Some improvement of estimates is observed when bootstrap estimator is employed (DNB and HTB methods). Consequently, simulation study shows that replacing empirical ROC curve (3) with its smoothed counterpart (24) significantly improves the minimum distance estimates of the binormal ROC curve.

4 Real data analysis

To illustrate all considered semiparametric estimators, we apply them to data analysed in the paper of Tupikowski et al. (2012). In the dataset the effectiveness of combined treatment of interferon alpha and metronomic cyclophosphamide in patients with metastatic kidney cancer was studied in terms of hemoglobin level (HL) and serum fibrinogen concentration (FC). The dataset contains 31 observations in total; 14 with and 17 without clinical response. Low value of HL or FC level has been recognized as a negative predictor of treatment response and associated with short survival. The estimates of the binormal ROC curves parameters for HL and FC as predictive factors are given in Table 4 for all considered methods. The estimated values of AUC are also tabulated. Interestingly, while the estimates of the parameters μ and σ vary

between methods, the estimates of AUC are close to each other, and differ only by 7% for both HL and FC.

5 Conclusions and some prospects

In this article seven new estimators of binormal ROC curve in semiparametric setting have been proposed. New estimators originate from the minimum distance concept applied to the ROC curve estimation by Hsieh and Turnbull (1996) and recently revisited by Davidov and Nov (2012). In the original MDE procedures one minimizes some distance measures between the binormal ROC curve, characterized by two parameters μ and σ, and the empirical ROC curve. In our methods we propose to replace the ROC_{mn} estimator, which is not continuous and not very accurate for small sample sizes, with other nonparametric estimators of the ROC curve. Procedures involving kernel, Bayesian bootstrap and smoothed ROC curve estimators were considered. Moreover, for estimators based on the Davidov and Nov (2012) approach, the role of appropriate integration limits was emphasized.

The small-sample performance of the proposed estimators was investigated numerically and compared with original procedures of Davidov and Nov (2012) and Hsieh and Turnbull (1996). The biggest improvement, both in terms of the parameters accuracy and MISE, was observed for estimators based on the smoothed ROC_{mn}^S nonparametric ROC curve estimator (see Sect. 2.4.2). For samples of small sizes, we observed that replacing the ROC_{mn} with ROC_{mn}^S in minimum distance procedures can reduce the MSE of the estimators of μ and σ parameters by an order of magnitude, and by factor of 4.5 on average. The goodness of fit of the estimator of the ROC curve to the true ROC curve is also improved as indicated by lower mean integrated square error. Employing the BB estimator does not improve the performance of MDE's so much, while using the kernel estimators sometimes leads to even less accurate semiparametric ROC curves estimates.

In the future research we are going to examine the asymptotic equivalence of the estimators considered. Especially, the asymptotic properties of DNS and HTS estimators needs further investigation since as these methods clearly outperforms the others. In fact, the smoothed nonparametric estimator of the ROC curve, introduced by Jokiel-Rokita and Pulit (2013), seems to be very promising method and theoretical investigation of its asymptotic properties is of our interest. We are also going to study robustness of the considered estimators on model misspecification.

Appendix

Proof of Theorem 1 The idea behind the proof is the same as in the proof of Theorem 1 in Davidov and Nov (2012). Let S_i, $i = 1, \ldots, 4$, be the deterministic counterparts

of $\tilde{S}_i$, obtained by substituting $ROC(t)$ for $ROC_{mn}^{BB}(t)$, and the values 0 and 1 for the lower and the upper integration limit, respectively, i.e., for example

$$S_1 = \int_0^1 \Phi^{-1}(ROC(t))dt.$$

Convergence $\tilde{S}_3 \to S_3$ and $\tilde{S}_4 \to S_4$ in probability, as $n \to \infty$, can be easily derived from Lemma 1. We will show that $\tilde{S}_1 \to S_1$ in probability. In very similar fashion one can show that $\tilde{S}_2 \to S_2$ in probability, hence, by definition (20), and Continuous Mapping Theorem, the theorem will be proved.

From Lemma 1, the coefficient $1/(b'_m - a'_m)$ converges to 1 a.s., therefore it can be omitted. We have,

$$\begin{aligned}
&\left| \int_{a'_m}^{b'_m} \Phi^{-1}(ROC_{mn}^{BB}(t))dt - \int_0^1 \Phi^{-1}(ROC(t))dt \right| \\
&\quad \le \left| \int_{a'_m}^{b'_m} \Phi^{-1}(ROC_{mn}^{BB}(t))dt - \int_{a'_m}^{b'_m} \Phi^{-1}(ROC(t))dt \right| \\
&\qquad + \left| \int_{a'_m}^{b'_m} \Phi^{-1}(ROC(t))dt - \int_0^1 \Phi^{-1}(ROC(t))dt \right|. \qquad (26)
\end{aligned}$$

The second term of the right-hand side of the above inequality converges to 0 a.s., as it was indicated in Lemma 1, hence it also converges to 0 in probability. Therefore it remains to show that the first term of the above inequality converges to 0 in probability. Using the same arguments as in the original paper of Davidov and Nov (2012), one can show that

$$\begin{aligned}
&\left| \int_{a'_m}^{b'_m} \Phi^{-1}(ROC_{mn}^{BB}(t))dt - \int_{a'_m}^{b'_m} \Phi^{-1}(ROC(t))dt \right| \\
&\quad \le \max\{\dot{\Phi}^{-1}(ROC(a'_m)), \dot{\Phi}^{-1}(ROC(b'_m))\} \sup_{0 \le t \le 1} |ROC_{mn}^{BB}(t) - ROC(t)|, \quad (27)
\end{aligned}$$

where $\dot{\Phi}^{-1}(x) = (d/dx)\Phi^{-1}(x)$. Note that the first factor of the right side of inequality (27) depends on the integration limits, while the second—depends on nonparametric ROC curve estimator. In fact, the rate of convergence of

$$\sup_{0 \le t \le 1} |ROC_{mn}^{BB}(t) - ROC(t)|$$

is $O_P(1/\sqrt{m})$, what can be deduced from Theorem 4.1 of Gu and Ghosal (2008). We will show that although $\dot{\Phi}^{-1}(ROC(a'_m))$ converges to ∞ as m increases, it converges to 0 after being multiplied by $1/\sqrt{m}$; the corresponding proof for $\dot{\Phi}^{-1}(ROC(b'_m))$ is very similar and hence it is omitted. Let

$$a_m^{(b)} = \inf\{t \in [0, 1] : ROC_{mn}^{(b)} > 0\},$$

then the lower integration limit, defined by (19), can be expressed in terms of $a_m^{(b)}$ as

$$a'_m = \inf\{t \in [0,1] : \frac{1}{B}\sum_{b=1}^{B} ROC_{mn}^{(b)}(t) > 0\} = \min_{b=1,\ldots,B}\{a_m^{(b)}\}. \tag{28}$$

The definition of $a_m^{(b)}$ may be equivalently written as

$$\begin{aligned} a_m^{(b)} = 1 - F^{(b)}(Y_{n:n}) = [1 - F(Y_{n:n})] + [F(Y_{n:n}) - F_m(Y_{n:n})] \\ + [F_m(Y_{n:n}) - F^{(b)}(Y_{n:n})], \end{aligned} \tag{29}$$

where F_m is the empirical distribution function based on $X_1, \ldots, X_m$. As in the proof of Theorem 1 in Davidov and Nov (2012), we can show that the rate of convergence of the first term of (29) is $\Omega_P(1/\sqrt{m})$. The notation Ω_P is the equivalent of O_P for an asymptotic lower bound, i.e., $Q_n = \Omega_P(R_n)$ if R_n/Q_n is bounded in probability. By the Dvoretzky–Kiefer–Wolfowitz inequality, the term in second bracket in (29) converges in probability to 0 exponentially. We will show that the expression in third bracket in (29) converges in probability to 0 faster than $1/m$, hence $a_m^{(b)} = O_P(1/\sqrt{m})$. For given samples, let K denote the number of observations in $X_1, \ldots, X_m$ which are not greater than $Y_{n:n}$: $K = \sum_{i=1}^{m} I_{\{X_j \le Y_{n:n}\}}$. By definition (17) and properties of the empirical distribution function, the following inequality holds

$$\begin{aligned} P(|F_m(Y_{n:n}) - F^{(b)}(Y_{n:n})| > \varepsilon) &= P\left(\left|\frac{K}{m} - U_{(K)}\right| > \varepsilon\right) \\ &= \sum_{k=0}^{m} P\left(|U_{(k)} - \frac{k}{m}| > \varepsilon | K = k\right) P(K = k) \\ &\le \sum_{k=0}^{m} P\left(|U_{(k)} - \frac{k}{m}| > \varepsilon\right). \end{aligned}$$

Since $U_{(k)}$ is k-th order statistic from the uniform distribution $\mathcal{U}(0,1)$, it has beta distribution $B(k, m-k)$ with expected value equal to k/m. A suitably tight upper bound for the last probability can be obtained using the following inequality (see Mitzenmacher and Upfal 2005, p. 59)

$$P(|U_{(k)} - E[U_{(k)}]| > tE[(U_{(k)} - E[U_{(k)}])^4]^{1/4}) \le \frac{1}{t^4}.$$

We have

$$\lambda_k := E[(U_{(k)} - E[U_{(k)}])^4] = \frac{3k(m-k)[(k+2)m^2 - (k+6)km + 6k^2]}{m^4(m+1)(m+2)(m+3)},$$

and

$$\sum_{k=0}^{m} P\left(|U_{(k)} - \frac{k}{m}| > \varepsilon\right) \le \frac{1}{\varepsilon^4}\sum_{k=0}^{m}\lambda_k = \frac{1}{\varepsilon^4}\frac{(m-1)^2}{10m^3} = O\left(\frac{1}{m}\right).$$

Therefore, due to decomposition (29), we have $a_m^{(b)} = O_P(1/\sqrt{m})$, and combining this with relation (28), we conclude that $a'_m = O_P(1/\sqrt{m})$. Using the same approach as Davidov and Nov (2012) in their proof of Theorem 1, we can show that $\dot{\Phi}^{-1}(ROC(a'_m)) = o_P(\sqrt{m})$ which completes the proof that $\tilde{S}_1 \to S_1$ in probability, as $n \to \infty$, and thus theorem is proved.

References

Branscum AJ, Johnson WO, Hanson TE, Gardner IA (2008) Bayesian semiparametric ROC curve estimation and disease diagnosis. Stat Med 27:2474–2496

Cai T, Moskowitz CS (2004) Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. Biostatistics 5(4):573–586

Cai T, Pepe MS (2002) Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. J Am Stat Assoc 97(460):1099–1107

Davidov O, Nov Y (2009) Minimum-norm estimation for binormal receiver operating characteristic (ROC) curves. Biometrical J 51(6):1030–1046

Davidov O, Nov Y (2012) Improving an estimator of Hsieh and Turnbull for the binormal ROC curve. J Stat Plan Inference 142(4):872–877

Dorfman DD, Alf E (1969) Maximum likelihood estimation of parameters of signal detection theory and determination of confidence interval - rating method data. J Math Psychol 6:487–496

Erkanli A, Sung M, Costello EJ, Angold A (2006) Bayesian semi-parametric ROC analysis. Stat Med 25:3905–3928

Gonçalves L, Subtil A, Oliveira MR, De Zea Bermudez P (2014) ROC curve estimation: an overview. REVSTAT Stat J 12(1):1–20

Gu J, Ghosal S (2008) Strong approximations for resample quantile process and applications to ROC methodology. J Nonparametr Stat 20(3):229–240

Gu J, Ghosal S (2009) Bayesian ROC curve estimation under binormality using a rank likelihood. J Stat Plan Inference 139:2076–2083

Gu J, Ghosal S, Roy A (2008) Bayesian bootstrap estimation of ROC curve. Stat Med 27:5407–5420

Hall PG, Hyndman RJ (2003) Improved methods for bandwidth selection when estimating ROC curves. Stat Prob Lett 64(2):181–189

Hanley JA (1988) The robustness of the "binormal" assumptions used in fitting ROC curves. Med Decis Mak 8:197–203

Hanley JA (1996) The use of binormal model for parametric ROC analysis of quantitative diagnostic tests. Stat Med 15:1575–1585

Hsieh F, Turnbull B (1996) Nonparametric and semiparametric estimation of the receiver operating characteristic curve. Ann Stat 24(1):25–40

Jokiel-Rokita A, Pulit M (2013) Nonparametric estimation of the ROC curve based on smoothed empirical distribution function. Stat Comput 23:703–712

Krzanowski W, Hand D (2009) ROC curves for continuous data, volume 111 of $C\&H/CRC$ monographs on statistics & applied probability. Chapman and Hall/CRC, Boca Raton

Lloyd CJ (1998) Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. J Am Stat Assoc 93(444):1356–1364

Lloyd CJ (2002) Estimation of a convex ROC curve. Stat Prob Lett 59(1):99–111

Lloyd C, Yong Z (1999) Kernel estimators of the ROC curve are better than empirical. Stat Prob Lett 44(3):221–228

Metz CE, Herman BA, Shen J-H (1998) Maximum likelihood estimation of receiver characteristic (ROC) curves from continuosly-distributed data. Stat Med 17:1033–1053

Millar PW (1984) A general approach to the optymality of minimum distance estimators. Trans Am Math Soc 286:377–418

Mitzenmacher M, Upfal E (2005) Probability and computing: randomized algorithms and probabilistic analysis. Cambridge University Press, New York

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, Oxford

Qin J, Zhang B (2003) Using logistic regression procedures for estimating receiver operating characteristic curves. Biometrika 90(3):585–596

Rubin DB (1981) The Bayesian bootstrap. Ann Stat 9(1):130–134

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London

Swets JA (1986) Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psychol Bull 99:181–198

Tupikowski K, Dembowski J, Kołodziej A, Niezgoda T, Debiński P, Małkiewicz B, Szydełko T, Kowal P, Zdrojowy R (2012) C133 interferon alpha and metronomic cuclophsphamide for metastatic kidney cancer. Eur Urol Suppl 11(4):113–113

Wan S, Zhang B (2007) Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. Stat Med 26:2565–2586

Wolfowitz J (1957) The minimum distance method. Ann Math Stat 28(1):75–88

Zhou XH, Harezlak J (2002) Comparison of bandwidth selection methods for kernel smoothing of ROC curves. Stat Med 21:2045–2055

Zhou X-H, Lin H (2008) Semi-parametric maximum likelihood estimates for ROC curves of continuous-scale tests. Stat Med 27:5271–5290

Zhou XH, Obuchowski NA, McClish DK (2002) Statistical methods in diagnostic medicine. Wiley, New York

Zou KH, Hall WJ (2000) Two transformation models for estimating an ROC curve derived from continuous data. J Appl Stat 27(5):621–631