Impact of OVL Variation on AUC Bias Estimated by Non-parametric Methods

Carina Silva[1,2(✉)], Maria Antónia Amaral Turkman[2], and Lisete Sousa[2,3]

[1] Lisbon School of Health Technology, Polytechnic Institute of Lisbon, Lisbon, Portugal
carina.silva@estesl.ipl.pt
[2] Centro de Estatítica e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal
{antonia.turkman,lmsousa}@fc.ul.pt
[3] Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal

Abstract. The area under the ROC curve (AUC) is the most commonly used index in the ROC methodology to evaluate the performance of a classifier that discriminates between two mutually exclusive conditions. The AUC can admit values between 0.5 and 1, where values close to 1 indicate that the model of classification has a high discriminative power. The overlap coefficient (OVL) between two density functions is defined as the common area between both functions. This coefficient is used as a measure of agreement between two distributions presenting values between 0 and 1, where values close to 1 reveal total overlapping densities. These two measures were used to construct the *arrow plot* to select differential expressed genes. A simulation study using the bootstrap method is presented in order to estimate AUC bias and standard error using empirical and kernel methods. In order to assess the impact of the OVL variation on the AUC bias, samples from various continuous distributions were simulated considering different values for its parameters and for fixed OVL values between 0 and 1. Samples of dimensions 15, 30, 50 and 100 and 1000 bootstrap replicates for each scenario were considered.

Keywords: AUC · OVL · Arrow-plot · Bias

1 Introduction

Receiver operating curve (ROC) is a widespread methodology to evaluate the accuracy of binary classification systems, particularly in diagnostic tests [1]. The area under the ROC curve (AUC) is the index most commonly used to summarize the accuracy and recently used in genomic studies [2,3]. Another index

This work is partially financed by national funds through FCT Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020.

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12251, pp. 173–184, 2020.
https://doi.org/10.1007/978-3-030-58808-3_14

that recently has gained attention is the overlapping coefficient (OVL) [3,4], which is a measure of the similarity between two probability distributions. Silva-Fortes et al. (2012) [3] proposed a plot that uses both indices, the *arrow plot* (Fig. 1). This plot displays the overlapping coefficient (OVL) against the area under the ROC curve (AUC) for thousands of genes simultaneously using data from microarray experiments. The *arrow plot* allows to select different types of differentially expressed genes, namely up-regulated, down-regulated and genes with a mixture of both, called *special genes*, who may reveal a presence of different subclasses. Graphic analysis is quite intuitive, as it allows to obtain a global picture of the behavior of the genes and, based on its analysis, the user can choose the cutoff points for AUC and OVL, although this choice is arbitrary. In this approach AUC values near 1 or near 0 will be related by low OVL values, meaning that both probability distributions will not be overlaped.

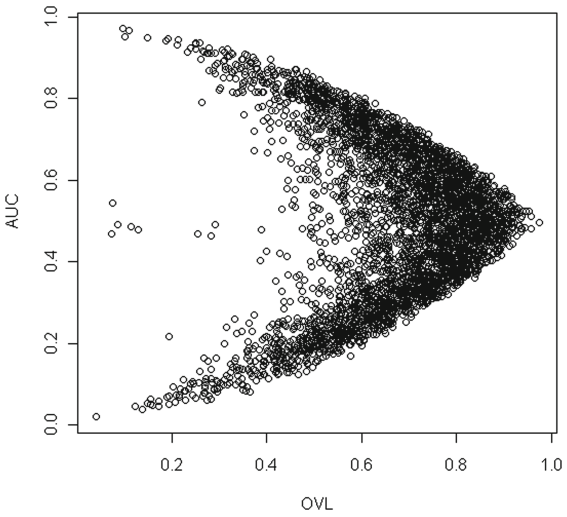

Fig. 1. Arrow plot [3]

In this study we present a simulation analysis to evaluate the impact of the OVL variation in AUC bias when non-parametric methods for their estimation are used. When microarray data analysis is conducted, thousands of AUCs and OVLs are produced, being computationally intensive and time consuming to perform a gene to gene analysis to evaluate if there is a need to make bias corrections. The goal is to understand where in the arrow plot is more likely the need to perform bias adjustments.

2 Methodology

A simulation study using a parametric bootstrap approach was performed using several distributional scenarios in order to reflect a range of distributional behaviours, particularly in genetic studies. AUC bias was analyzed considering non-parametric estimation for fixed OVL values. Consider $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ two independent random samples representing some characteristic in population 1 (e.g. control) and population 2 (e.g. cases) respectively. Let F_X be the distribution function of X_i, $i = 1, \ldots, n_1$ and G_Y is the distribution function of Y_j, $j = 1, \ldots, n_2$, and f_X and g_Y their respective density functions. Assume that, without loss of generality, that any cutoff point $c \in \mathbb{R}$, $F_X(c) > G_Y(c)$.

2.1 OVL

OVL is a coefficient used to measure the agreement between two distributions [5, 6], it ranges between 0 and 1, and closer to 1 means higher amount of overlapping area. OVL can detect any differences between two distributions, not only by mean value differences but also by variance differences. One good property of this index is that it is invariant through scale monotone transformations of the variables.

OVL can be expressed under several ways. Weitzman (1970) [5] proposed the expression:

$$\text{OVL} = \int_{\mathcal{C}} \min[f_X(c), g_Y(c)]dc. \tag{1}$$

Results can be extended to discrete distributions, replacing the integral by a summation.

To cover several levels of overlapping areas between the densities, we fixed the OVL values in 0.2, 0.4, 0.6 and 0.8. True OVL values were calculated using (1) for the distributions considered on the different simulated scenarios (see Table 1).

2.2 AUC

A ROC curve φ is defined by the sensitivity ($q(c) = 1 - F_Y(c)$) and specificity ($\text{p(c)} = F_X(c)$):

$$\varphi : [0, 1] \longrightarrow [0, 1]$$

$$\varphi(p) = 1 - F_Y(F_X^{-1}(1 - p)), \tag{2}$$

where F_X^{-1} is the inverse function of F_X defined by $F_X^{-1}(1 - p) = inf\{x \in W(F_X) : F_X(x) \geq 1 - p\}$ and $W(F_X) = \{x \in \mathbb{R} : 0 < F_X(x) < 1\}$ is the support of F_X when continuous distributions are considered.

The AUC is obtained integrating the ROC curve in its domain. It can be proved that it turns out to be equal to $P(X < Y)$, that is to say AUC ranges between 0.5 and 1, where values near 1 indicate a higher classification performance on the discrimination between the two populations. AUC is also invariant

under monotone scale transformations of the variables. However, in the *arrow plot* those values range between 0 and 1, because the same classification rule is applied whenever $F_X(c) > G_Y(c)$ or $F_X(c) < G_Y(c)$. This situation may produce ROC curves that are not proper, meaning that they may produce values for the AUC between 0 and 0.5 [3].

True values of the AUC were obtained accordingly with the distributional scenarios (Table 1):

Bi-Normal: When $X \frown N(\mu_1, \sigma_1)$ and $Y \frown N(\mu_2, \sigma_2)$, $\mu_1 < \mu_2$, the AUC is given by [7]:

$$AUC = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

where $\Phi(.)$ is the standard normal distribution.

Bi-Lognormal: When $X \frown LN(\mu_1, \sigma_1)$ and $Y \frown LN(\mu_2, \sigma_2)$, $\mu_1 < \mu_2$, the AUC is given by [8]:

$$AUC = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

where $\Phi(.)$ is the standard normal distribution.

Bi-Exponential: When $X \frown Exp(\lambda_1)$ and $Y \frown Exp(\lambda_2)$, $\lambda_1 < \lambda_2$, the AUC is given by [9]:

$$AUC = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

For non-parametric estimation of the AUC, it was used the empirical and kernel based methods:

Empirical AUC: The empirical estimator of the AUC corresponds to the Mann-Whitney statistic [10]:

$$\widehat{\text{AUC}} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(I[x_i < y_j] + \frac{1}{2} I[x_i = y_j] \right), \tag{3}$$

where I is the indicator function.

The empirical AUC estimates were obtained using `pROC` [11] package from `R`. From now on the notation *emp* is used whenever the empirical estimation of the AUC is mentioned.

Kernel AUC: Several authors have discussed the refinement of the non-parametric approach to produce smooth ROC curves [14]. Among the various non-parametric methodologies, an important method for estimating probability density functions is the kernel estimator. The kernel estimator of a density is given by [15]:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \forall x \in S, h > 0, \tag{4}$$

where K is the kernel function, h the bandwidth and S the support of X.

Several studies have demonstrated that the quality of the kernel estimator depends more on the choice of the bandwidth h, rather than on the choice of the functional form of the kernel [12]. In this work the Gaussian kernel will be used and three different bandwidth methods choice will be explored.

Lloyd (1997) showed that when a Gaussian kernel is considered, the AUC is estimated as [13]:

$$\widetilde{\text{AUC}} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Phi\left(\frac{y_j - x_i}{\sqrt{h_1^2 + h_2^2}}\right).$$

Silverman (1992) [12] considers the expression (5) as an optimal bandwidth when the kernel is Gaussian.

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \min\left(s, \frac{R}{1.34}\right) n^{-\frac{1}{5}}, \tag{5}$$

where s is the empirical standard deviation and R the interquartile range. Hereinafter, for our convenience, this method will be referred as $nrd0$.

Scott (1992) [16] considers the expression (6) when a Gaussian kernel is used. The notation nrd is used whenever this method is referred to.

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} s\, n^{-\frac{1}{5}}. \tag{6}$$

Hall et al. (1991) [17] proposed the plug-in method solve-the-equation for the optimal bandwidth. The notation SJ is used whenever this method is referred to.

2.3 Parametric Bootstrap Estimation

A Monte Carlo simulation study was used to compare bias, standard error (SE) and root mean squared error (RMSE) of the bootstrap AUC estimates obtained from four non-parametric methods ($emp, nrd0, nrd$ and SJ). In order to represent pairs of overlapping distributions with various degrees of separation and skewness (see Table 1), continuous datasets for control and experimental groups, with samples sizes of n: 15, 30, 50 and 100 on both groups, were considered. The bootstrap estimates were obtained from 1000 replicates in each scenario (see Fig. 2).

Table 1. Distributional scenarios for the simulation study.

Scenario	Control	Experimental	True OVL	True AUC
Bi-normal fixed μ	$N(0,1)$	$N(0,0.1)$	0.2	0.5
	$N(0,1)$	$N(0,0.4)$	0.4	0.5
	$N(0,1)$	$N(0,2.4)$	0.6	0.5
	$N(0,1)$	$N(0,1.5)$	0.8	0.5
Bi-normal fixed σ	$N(0,1)$	$N(2.55,1)$	0.2	0.96
	$N(0,1)$	$N(1.65,1)$	0.4	0.89
	$N(0,1)$	$N(1.04,1)$	0.6	0.77
	$N(0,1)$	$N(0.5,1)$	0.8	0.64
Bi-Lognormal	$LN(0,1)$	$LN(0,0.1)$	0.2	0.5
	$LN(0,1)$	$LN(1.65,1)$	0.4	0.87
	$LN(0,1)$	$LN(1.04,1)$	0.6	0.77
	$LN(0,1)$	$LN(0.5,1)$	0.8	0.64
Bi-Exponential	$Exp(1)$	$Exp(0.05)$	0.2	0.95
	$Exp(1)$	$Exp(0.15)$	0.4	0.87
	$Exp(1)$	$Exp(0.32)$	0.6	0.76
	$Exp(1)$	$Exp(0.58)$	0.8	0.63

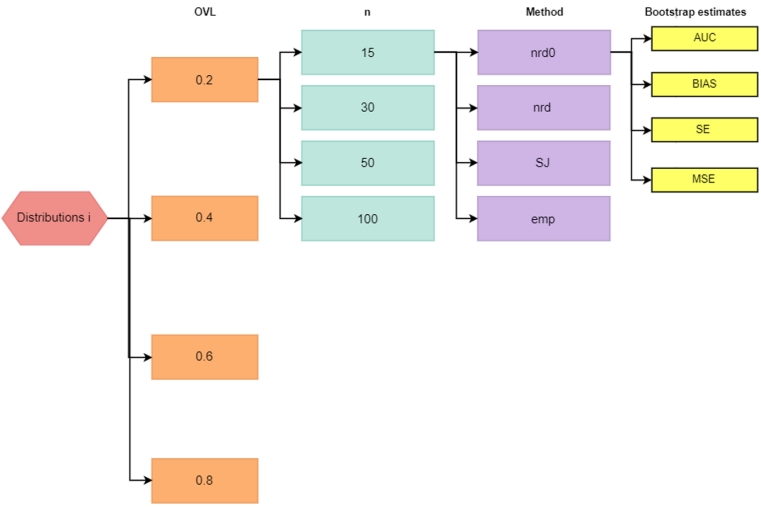

Fig. 2. Scheme of the simulation procedure. Total of 1024 estimates.

The bootstrap estimator of the AUC is:

$$\widehat{\mathrm{AUC}}_B = \frac{1}{1000}\sum_{i=1}^{1000}\widehat{\mathrm{AUC}}_i^*, \tag{7}$$

where $\widehat{\mathrm{AUC}}_i^*$ is the AUC estimate (empirical or kernel) in each bootstrap replicate.

The bootstrap estimator of the standard error of the AUC is given by:

$$\widehat{se}_B(\widehat{\mathrm{AUC}}) = \sqrt{\frac{1}{999}\sum_{i=1}^{1000}(\widehat{\mathrm{AUC}}_i^* - \widehat{\mathrm{AUC}}_B)^2}. \tag{8}$$

The bootstrap estimator of the bias of the bootstrap AUC is:

$$\widehat{\mathrm{bias}}_B(\widehat{\mathrm{AUC}}) = \widehat{\mathrm{AUC}}_B - \mathrm{AUC}, \tag{9}$$

where AUC corresponds to the true AUC value.
The bootstrap estimator of the root mean squared error (RMSE) of the AUC is given by:

$$\widehat{rmse}_B(\widehat{\mathrm{AUC}}) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\widehat{\mathrm{AUC}}_i^* - \mathrm{AUC})^2}. \tag{10}$$

3 Results and Discussion

Figure 3 and Fig. 4 depict the behaviour of the bootstrap AUC estimates along different OVL values, considering different sample sizes and non-parametric estimation methods for the distributional scenarios presented in Table 1. Analyzing Table 1 it is observed that AUC estimates are influenced by sample sizes and variability in results increases as OVL increase. However, when AUC values are obtained from distributions with equal mean values, which produce not proper ROC curves, the variability of the estimates is constant across different OVL values (Fig. 3a and Fig. 4a).

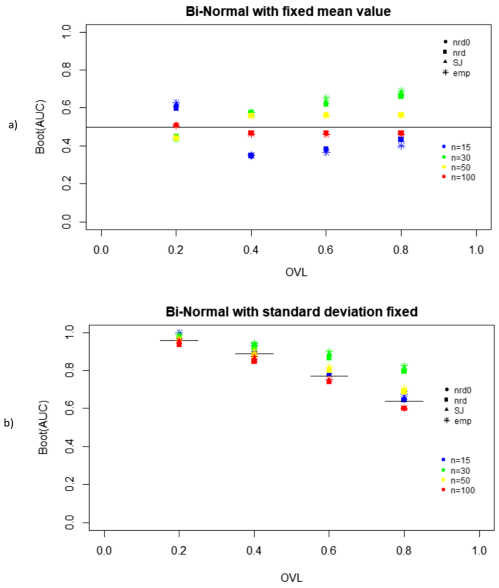

Fig. 3. Comparison of AUC bootstrap estimates against OVL values, considering different sample sizes and non-parametric estimation methods. a) Bi-normal distribution with fixed mean values. b) Bi-normal distribution with fixed standard deviation. True AUC values are represented by the horizontal lines.

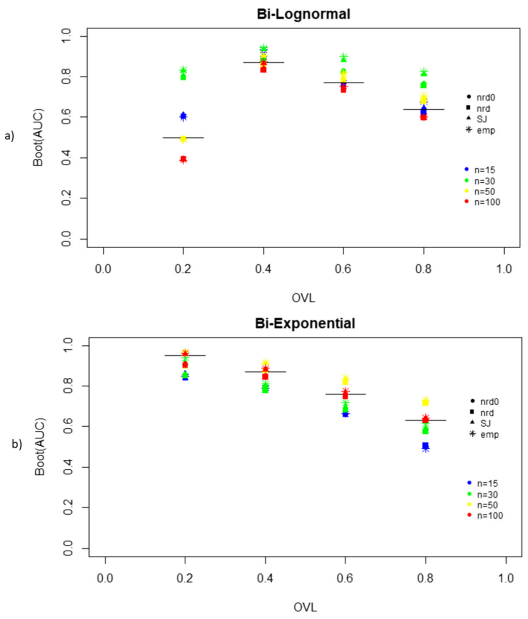

Fig. 4. Comparison of AUC bootstrap estimates against OVL values, considering different sample sizes and non-parametric estimation methods. a) Bi-lognormal distribution. b) Bi-exponential distribution. True AUC values are represented by the horizontal lines.

In general way, the estimated AUC bias is negligible considering all scenarios ($|\text{bias}| < 0.2$) (see Fig. 5).

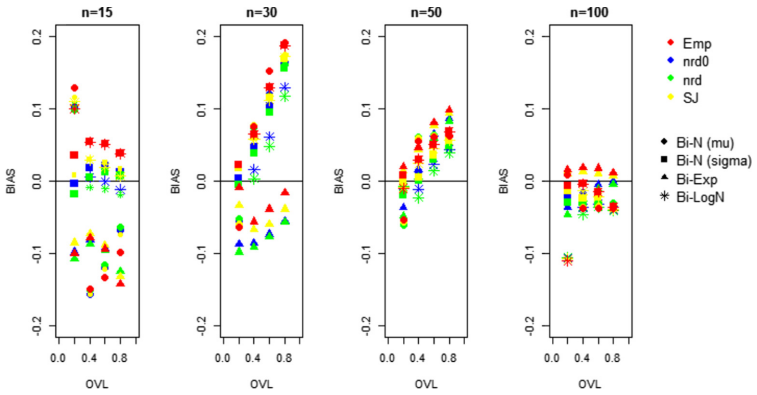

Fig. 5. Comparison of AUC bootstrap bias considering all distributions and non-parametric estimation methods for each sample dimension.

However, as expected, bias is lower for larger samples sizes. Considering small sample sizes, data simulated from bi-exponential distributions tend to underestimate the AUC, contrary to all the other distributions. For high OVL values bias tends to increase.

Precision tends to increase when OVL values decrease, unless if AUC values are obtained from not proper ROC curves (AUC ≈ 0.5), where precision tends to be lower when compared to all other scenarios for all OVL values (see Fig. 6).

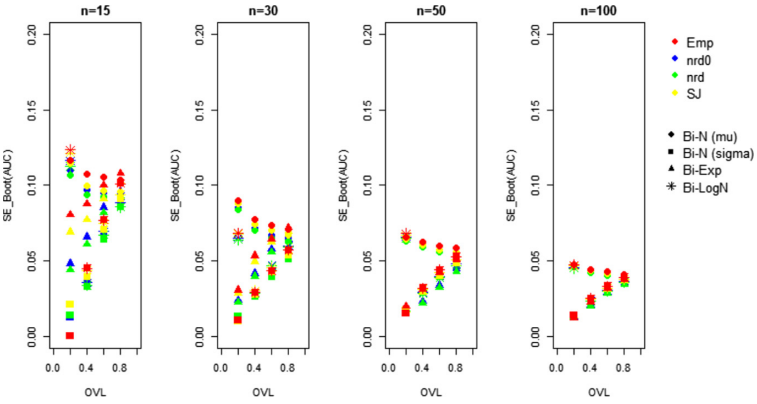

Fig. 6. Comparison of bootstrap estimates of the standard error of the AUC, considering all distributions and non-parametric methods for each sample size.

RMSE tends to be higher in lower samples sizes and increases as OVL increases (see Fig. 7).

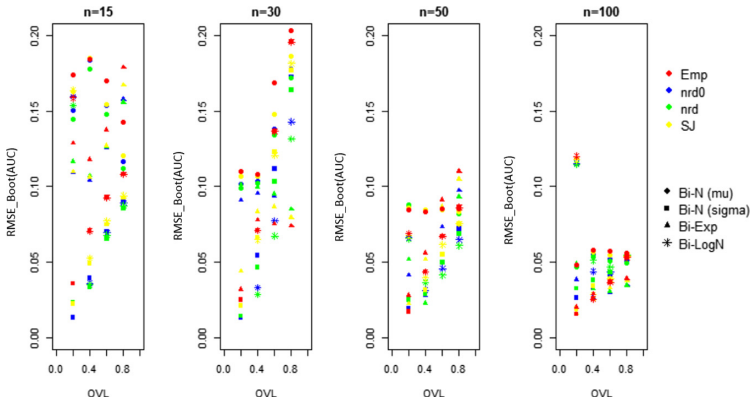

Fig. 7. Comparison of bootstrap estimates of the root of mean squared error of the AUC, considering all distributions and non-parametric methods for each sample size.

In Fig. 8 it is shown the area where bias correction should be considered.

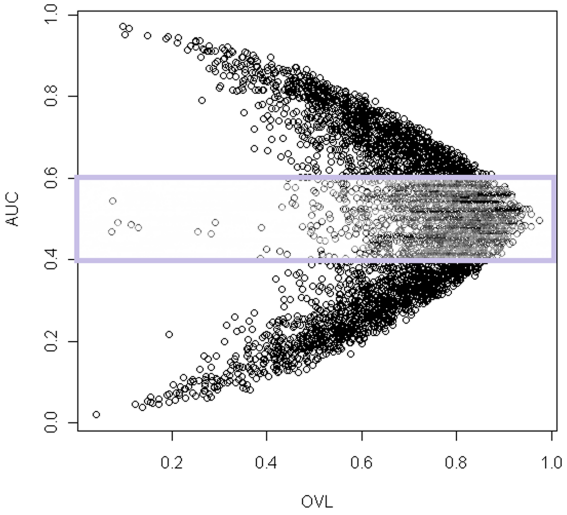

Fig. 8. Area of the Arrow plot where should be considered bias correction.

4 Conclusions

Overall, results of the simulation study suggest that for a broad range of pairs of distributions with several degrees of departure yield close estimates of AUC. Concern about bias or precision of the estimates should not be a major factor in choosing between non-parametric approaches, however there is an advantage in kernel methods since they produce smooth ROC curves. Precision is higher for small values of OVL, however this is not true when AUC values are around 0.5 and are obtained from distributions with the same mean value, leading to not proper ROC curves. This particular situation is related to "special genes" on the *arrow plot*, where greater attention should be paid to the possibility of bias correction.

The main issue with this study is that simulations were performed by using continuous distributions. Future research will include distributions related with count data, where not much research exists concerning the OVL index.

References

1. Lusted, L.: Introduction to Medical Decision Making. Charles C. Thomas, Springfield, Illinois (1968)
2. Parodi, S., Pistoia, V., Musseli, M.: Not proper ROC curves as new tool for the analysis of dierentially expressed genes in microarray experiments. BMC Bioinform. **9**, 410 (2008)
3. Silva-Fortes, C., Amaral Turkman, M.A., Sousa, L.: Arrow plot: a new graphical tool for selecting up and down regulated and genes differentially expressed on sample subgroups. BMC Bioinform. **13**, 147 (2012)
4. Wang, D., Tian, L.: Parametric methods for confidence interval estimation of overlap coefficients. Comput. Stat. Data Anal. **106**, 12–26 (2017)
5. Weitzman, M. S.: Measure of the overlap of income distribution of white and negro families in the United States. Technical report 22, U.S. Department of Commerce, Bureau of the Census, Washington, D.C. (1970)
6. Inman, H.F., Bradley, E.L.: The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Commun. Stat. Theor. Meth. **18**(10), 3851–3874 (1989)
7. Faraggi, D., Reiser, B.: Estimation of the area under the ROC curve. Stat. Med. **31**, 3093–3106 (2002)
8. Amala, R., Pundir, S.: Statistical Inference on AUC from a bi-lognormal ROC model for continuos data. Int. J. Eng. Sci. Innov. Technol. **1**(2), 283–295 (2012)
9. Vardhan, R.V., Pundir, S., Sameera, G.: Estimation of area under the ROC curve using Exponential and Weibull distributions. Bonfring Int. J. Data Min. **2**(2), 52–56 (2012)
10. Hanley, J.A., MacNeil, B.J.: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology **148**(3), 839–843 (1983)
11. Robin, X.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform. **12**, 77 (2011)
12. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1998)

13. Lloyd, C.J., Yong, Z.: Kernel estimators of the ROC curve are better than empirical. Stat. Probab. Lett. **44**(3), 221–228 (1999)
14. Zou, K.H., Hall, W.J., Shapiro, D.E.: Smooth non-parametric receiver-operating characteristic (ROC) curves for continuous diagnostic tests. Stat. Med. **16**, 2143–2156 (1997)
15. Rosenblatt, M.: Remarks on some nonparametric estimate of a density function. Ann. Math. Stat. **27**, 832–837 (1956)
16. Scott, D. W.: Kernel Density Estimators. Wiley Series in Probability and Statistics (1992)
17. Hall, P., Sheather, S.J., Jones, M.C., Marron, J.S.: On optimal data-based bandwith selection in kernel density estimation. Biometrika **78**(2), 263–269 (1991)