Evaluation of Biomarker using Two Parameter Bi-exponential ROC Curve

Sudesh Pundir
Department of Statistics, Pondicherry University, Puducherry, India
sudeshpundir19@gmail.com

R Amala
Department of Statistics, Pondicherry University, Puducherry, India
amalar.statistics@gmail.com

Abstract

Receiver Operating Characteristic (ROC) Curve is used for assessing the ability of a biomarker/screening test to discriminate between non-diseased and diseased subject. In this paper, the parametric ROC curve is studied by assuming two-parameter exponential distribution to the biomarker values. The ROC model developed under this assumption is called bi-exponential ROC (EROC) model. Here, the research interest is to know how far the biomarker will make a distinction between diseased and non-diseased subjects when the gold standard is available using parametric EROC curve and its Area Under the EROC Curve (AUC). Here, the standard error is used as an estimate of the precision of the accuracy measure AUC. The properties of EROC curve that explains the behavior of the EROC curve are also discussed. The AUC along with its asymptotic variance and confidence interval are derived.

Keywords: Two parameter bi-exponential *ROC* model, *AUC* and variance of *AUC*, Monte Carlo simulation.

1. Introduction

1.1 Diagnostic Accuracy

In a medical diagnosis, a subject is categorized into either *non-diseased* or *diseased* group (a binary classification) by using some clinical measurement based on the selected cut-off *t*. If the clinical measurement is 'greater than or equal to' *t*, then the subject is labeled as diseased and if the measurement is 'less than' *t*, then the subject is labeled as non-diseased. The clinical measurements are often called as test results or test scores or Biomarker. The *accuracy* of a biomarker is defined as "its ability to distinguish the diseased group from non-diseased group". The purpose of evaluating the potentiality of a biomarker in diagnosing a disease is to filter out the patients as those belonging to 'high risk' and 'no risk' for disease during the initial stage of medical diagnosis/screening process, because it is not necessary for all the in-patient to undergo a gold standard test (e.g. endoscopy) as it proves to be a costly, time consuming process and invasive.

1.2 *ROC* curve and Diagnostic Accuracy

The accuracy of a binary classification can be visualized as well as quantified by a renowned statistical technique called *Receiver Operating Characteristic* (*ROC*) curve. It is a plot of two probabilities namely *Sensitivity* versus *1- Specificity* for various threshold *t* where the *sensitivity* can be defined as likelihood of classifying a diseased subject correctly and the *specificity* can be defined as likelihood of classifying a non-diseased subject correctly. The measure of accuracy explained by the plotted *ROC* curve is

quantified by *Area Under the ROC curve* (*AUC*). Hence, the *ROC* plot reports the accuracy visually and the *AUC* reports the accuracy numerically.

1.3 Nomenclature

Let the biomarker values of diseased subject by the random variable *Y* with Probability Density Function (PDF), $g_Y(y)$ and Cumulative Distribution Function (CDF), $G_Y(y)$. Similarly, let the biomarker values of non-diseased subject by the random variable *X* with PDF, $f_X(\mathrm{x})$ and CDF, $F_X(\mathrm{x})$. Assume that *X* and *Y* are independent, continuous and $Y > X$ this is due to the fact that higher values of the biomarker indicates a condition of disease in an individual which in turn implies mean of *Y* is greater than mean of *X* for a better discrimination of subjects.

Sensitivity of the biomarker can be evaluated using, $\bar{G}_Y(t) = P(Y > t)$, which is the probability of correctly categorizing a diseased subject when a cut-off *t* for the classification is given. It is also known as "*True Positive Rate*" (*TPR*). Similarly, the *Specificity* of the biomarker can be evaluated using, $\bar{F}_X(t) = P(X \le t)$, which is the probability of correctly categorizing a non-diseased subject for a given *t*. and it is also known as False Negative Rate (FNR). *1- Specificity* is called as "*False Positive Rate*" (*FPR*).

Then *ROC* curve is defined as a plot of *TPR*, $\bar{G}_Y(t)$ on the vertical axis versus the *FPR*, $\bar{F}_X(t)$ on the horizontal axis for different values of *t*, where $-\infty < t < \infty$. In other words, the mathematical model representing the *ROC* curve takes the form

$$ROC(p) = \bar{G}_Y \circ \bar{F}_X^{-1}(p); 0 \le p \le 1 \tag{1.1}$$

For an appropriate diagnostic test, the *ROC* curve should lie very close to upper left corner of the unit square.

1.4 Properties

Once the ROC curve is plotted, it is important to study the properties of it, in order to highlight some key understanding from the plot. It is known that a typical parametric *ROC* curve must satisfy the basic three properties viz. monotonicity, invariance to monotone increasing transformation and the slope defined at a particular threshold *t* (Krzanowski and Hand, 2002). Recently, Hughes and Bhattacharya (2013) have given a quite interesting property known as asymmetry property of the ROC curve for few ROC models viz. Bi-Exponential, Bi-Normal and Bi-Gamma. In this section, we have more generally discussed the properties satisfied by a parametric ROC curve.

1. $ROC(t)$ is monotonically increasing function i.e. $\frac{dROC(t)}{d\bar{F}_X(t)} > 0.$

2. $ROC(t)$ is said to be concave, if $\frac{d^2ROC(t)}{d\bar{F}_X^2(t)} < 0$ and convex, if $\frac{d^2ROC(t)}{d\bar{F}_X^2(t)} > 0.$

3. The slope of the *ROC* curve at any operating point is equal to the ratio of PDF of diseased to PDF of non-diseased at cut-off point '*t*' (Krzanowski and Hand, 2002) is given by

$$slope = \frac{g(t)}{f(t)} \tag{1.2}$$

4. If f(x) and g(y) denote the continuous PDF for non-diseased and diseased groups respectively. Let $KL(f,g)$ denote the *Kullback – Leibler* (*KL*) divergence between the distributions of non-diseased and diseased group with f(*x*) as the comparison distribution and g(*y*) as the reference distribution (Hughes and Bhattacharya, 2013). Then

$$KL(f,g) = \int_D f(x) \ln\left[\frac{f(x)}{g(y)}\right] dz \tag{1.3}$$

where z is the common range of x and y i.e. {LL=max[LL(x), LL(y)], UL=[min(UL(x), UL(y)] where LL-Lower Limit, UL – Upper limit. *D* is based on *z*, let us represent *x* and *y* by *z*.

$$KL(f,g) = \int_D f(z) \ln\left[\frac{f(z)}{g(z)}\right] dz. \tag{1.4}$$

Similarly, $KL(g,f)$ denote the *KL* divergence between the distribution of diseased and non-diseased population with g(y) as the comparison distribution and f(x) as the reference distribution, then

$$KL(g,f) = \int_D g(y) \ln\left[\frac{g(y)}{f(x)}\right] dz \tag{1.5}$$

where z is the common range of x and y i.e. {LL=max[LL(x), LL(y)], UL=[min(UL(x), UL(y)] where LL-Lower Limit, UL – Upper limit. *D* is based on *z*, let us represent *x* and *y* by *z*. Hence we have,

$$KL(g,f) = \int_D g(z) \ln\left[\frac{g(z)}{f(z)}\right] dz. \tag{1.6}$$

It is to be noted that $KL(f,g)$ and $KL(g,f)$ are positive and $KL(f,g) = KL(g,f) = 0$, if and only if $f(x) = g(y)$. These two measures tell us about the asymmetry of *ROC* curve about the negative diagonal of the *ROC* plot. If $KL(f,g) < KL(g,f)$, then the *ROC* curve is said to be *TPR* asymmetric and if $KL(f,g) > KL(g,f)$ then the ROC curve is said to be FPR asymmetric.

5. ROC(t) is invariance with respect to any monotonically increasing transformation.

Result: 1 (**Proper *ROC* curve)**: The concavity property of *ROC* curve implies Proper. A ROC curve is said to be a proper *ROC* curve if it never crosses the chance line (the line connecting the co-ordinates [0, 0] and [1, 1]). Otherwise, *TPR* is a strictly increasing function over the range of all possible *FPR*.

Proof: Consider any two points 'x' and 'y' (say) where $0 < x,\ y < 1$ on the *FPR*.

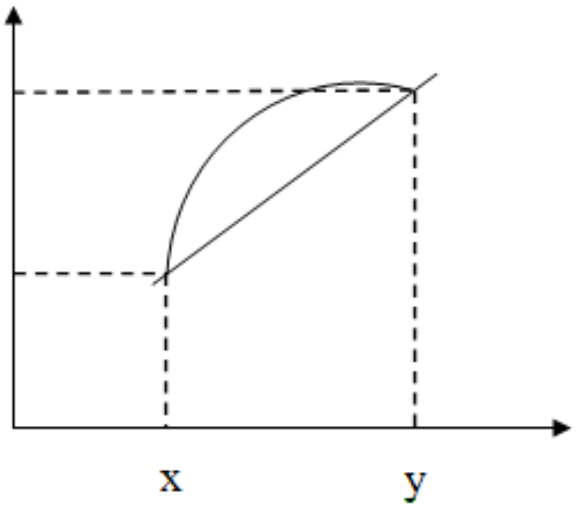

Fig. 1 A proper *ROC* curve in the interval [x, y]

By the definition of concavity, the line segment connecting the point on the *ROC* curve parallel to x and y never lies above the curve. If we take the extreme point i.e., x=0 and y=1, it becomes the chance line which never lies above the curve. Hence we have proved that the concavity property of *ROC* curve implies it is also proper.

Area under the *ROC* curve is the frequently used measure for quantifying the biomarker or performance of a diagnostic test. It is defined as the probability that in a randomly selected pair of non-diseased and diseased subjects, the biomarker value of diseased subject is higher than the non-diseased subject. The analytical expression for *AUC* is given by

$$AUC = P(\mathrm{Y} > \mathrm{X}) = \int_0^1 \bar{G}_Y(t)\, d\bar{F}_X(t). \qquad (1.7)$$

The evaluation of *AUC* and its inference are the crucial part of *ROC* curve analysis.

The classic 'bi-normal' *ROC* model consists in assuming normal distribution to the biomarker values from diseased and non-diseased groups while modeling the *ROC* curve. Many authors have encountered intensive study on parametric bi-normal *ROC* curve in a diversified directions such as by using Bayesian approach (O. Malley *et al.* 2002), regression modeling (Zhang, and Pepe 2012), pooling the biomarkers when drawing sample is expensive (Mumford *et al.* 2006), limit of detection when the biomarkers are unobserved (Perkins, Schisterman, and Vexler 2006), etc., especially for bi-normal *ROC* model. A very similar work of analyzing the *ROC* curves based on other distributional assumptions are bi-lomax *ROC* curve using Lomax distribution (Campbell and Ratnaparkhi 1993), bi-logistic *ROC* model using logistic distribution (Oglive, and Creelman 1968) and proper bi-gamma model using gamma distribution (Dorfman *et al.* 1996) for rating data. The *ROC* curve modeling for two discrete distributions viz. Uniform, Triangular and two continuous distributions such as Normal and Beta (Marzban

2004), bi-exponential *ROC* model using one parameter exponential distribution (Betinec 2008 and Pundir and Amala 2014). Betinec has provided an analysis of ROC curves based on Exponential distribution to compare two classification methods namely Linear discriminant analysis and Support Vector machine. Bi-generalized exponential *ROC* model (Hussian 2011), bi- lognormal *ROC* model makes use of lognormal distribution (Amala and Pundir 2012), bi-rayleigh (Pundir and Amala 2012a) and left truncated bi-rayleigh *ROC* model using Rayleigh distribution (Pundir and Amala 2015). Pundir and Amala 2014d have studied the constant shape bi-weibull *ROC* curve using Weibull distribution and a review of all parametric *ROC* curves have been presented by Pundir and Amala 2014a) and bi-variate bi-lognormal ROC model (Pundir and Amala 2015).

In this paper, the bi-exponential *ROC* curve is studied by assuming two parameter exponential distribution along with its properties, asymptotic variance and confidence interval for *AUC*. This paper is organized as follows: In Section 2, Bi-exponential *ROC* model, its properties and Maximum Likelihood Estimation of parameters are discussed. Section 3, provides estimation of *AUC*, asymptotic distribution and confidence interval for *AUC*. In Section 4, the proposed theory is validated by using Monte Carlo Simulation.

2. Two Parameter Bi-Exponential *Roc* Model

Let Z be a random variable that follows two parameter exponential distribution with scale parameter λ and location parameter γ which is denoted by Z~exp (λ, γ). It possess the PDF as

$$f_Z(z,\lambda,\gamma) = \lambda e^{-\lambda(z-\gamma)}; z > \gamma, \lambda > 0. \tag{2.1}$$

The CDF of Z is given by

$$F_Z(z) = P(Z \le z) = 1 - e^{-\lambda(z-\gamma)}, \lambda > 0, z > \gamma. \tag{2.2}$$

Let us assume that X and Y are independent and exponentially distributed with different parametric values (λ_x, γ_x) and (λ_y, γ_y) respectively. Notationally, $X \sim \exp(\lambda_x, \gamma_x)$ and $Y \sim \exp(\lambda_y, \gamma_y)$ with the constraint $\gamma_y > \gamma_x$, $\lambda_x > \lambda_y$. This restrictive condition is because to satisfy the assumption of higher values of biomarker values of the diseased subjects. The *ROC* model developed under this assumption is represented by "*EROC*".

The *FPR* of *EROC* curve at the threshold 't' is found to be

$$\bar{F}_X(t) = P(X > t) = \int_t^\infty \lambda_x exp[-\lambda_x(x - \gamma_x)]\, dx$$
$$= exp[-\lambda_x(t - \gamma_x)]. \tag{2.3}$$

The *TPR* of *EROC* curve at the threshold 't' is found to be

$$\bar{G}_Y(t) = P(Y > t) = \int_t^\infty \lambda_y exp[-\lambda_y(y - \gamma_y)]\, dy$$
$$= exp[-\lambda_y(t - \gamma_y)] \tag{2.4}$$

where the parameters can be estimated from the sample data by any of the standard estimation procedure. If *X* and *Y* are two independent random variables of size '*m*' and '*n*' respectively, with PDF given in (2.1), the MLE of parameters (Johnson, Kotz and Balakrishnan, 2004) are determined as follows:

$$\hat{\lambda}_x = \frac{m}{\sum_{i=1}^{m}\left(x_i - x_{(1)}\right)}, \hat{\lambda}_y = \frac{n}{\sum_{j=1}^{n}\left(y_j - y_{(1)}\right)}, \hat{\gamma}_x = x_{(1)} \quad \text{and} \quad \hat{\gamma}_y = y_{(1)} \tag{2.5}$$

where $x_{(1)} = min(x_1, x_2, x_3 ... x_m)$ and $y_{(1)} = min(y_1, y_2, y_3 ... y_n)$. By substituting the above estimates in (2.3) and (2.4), one would get the estimates of $\overline{F}_X(t)$ and $\overline{G}_Y(t)$. By plotting $\overline{F}_X(t)$ on the horizontal axis and $\overline{G}_Y(t)$ on the vertical axis, one will get the *EROC* curve.

Aliter:

One can also obtain an analytical form of the EROC curve as follows: From (2.3), we get the expression for threshold '**t**' as

$$t = \gamma_x - \frac{\ln \overline{F}_X(t)}{\lambda_x}. \tag{2.6}$$

Since, *ROC* model is *TPR* as a function of *FPR*. By substituting (2.6) in (2.4), we can get the two parameter Bi-exponential *ROC* model as

$$EROC(t) = exp\left[-\lambda_y\left((\gamma_x - \gamma_y) - \frac{\ln \overline{F}_X(t)}{\lambda_x}\right)\right]$$

$$= exp[\lambda_y(\gamma_y - \gamma_x)]\,\overline{F}_X(t)^{\frac{\lambda_y}{\lambda_x}};\ 0 \le \overline{F}_X(t) \le 1,\ \lambda_x > \lambda_y,\ \gamma_y > \gamma_x. \tag{2.7}$$

Plotting *EROC(t)* along y-axis and $\overline{F}_X(t)$ along x-axis for different values of *t*, we get an estimate of *EROC* curve. Now, we will discuss some of the properties of two parameter exponential *ROC* curve.

2.1 Properties and Characteristics

1. *EROC* curve is monotonically increasing in nature for $\lambda_y > \lambda_x$.

Proof: Since, the first derivative of *EROC* curve with respect to $\overline{F}_X(t)$ is positive i.e.

$$\frac{d}{d\overline{F}_X(t)} EROC(\mathrm{t}) = exp[\lambda_y(\gamma_y - \gamma_x)]\frac{\lambda_y}{\lambda_x}\left[\overline{\mathrm{F}}_X(t)\right]^{\left(\frac{\lambda_y}{\lambda_x}-1\right)} > 0. \tag{2.8}$$

EROC curve is monotonically increasing in nature.

Aliter:

Consider two *FPR* values $\bar{F}_X(t_1)$ and $\bar{F}_X(t_2)$ such that $\bar{F}_X(t_1) < \bar{F}_X(t_2)$. Now raising the power $\frac{\lambda_y}{\lambda_x}$ and multiplying the constant $\exp[\lambda_y(\gamma_y - \gamma_x)]$, the inequality remains the same and hence

$$exp[\lambda_y(\gamma_y - \gamma_x)]\left(\bar{F}_X(t_1)\right)^{\frac{\lambda_y}{\lambda_x}} < exp[\lambda_y(\gamma_y - \gamma_x)]\left(\bar{F}_X(t_2)\right)^{\frac{\lambda_y}{\lambda_x}}$$

$$EROC(t_1) < EROC(t_2). \tag{2.9}$$

Hence, the *EROC* curve is monotonically increasing.

2. *EROC* curve is concave when $\lambda_y > \lambda_x$ and proper as long as it is concave.

   Proof: The second derivative of *EROC(t)* is given by

   $$\frac{d^2}{d\bar{F}_X^2(t)} EROC(\mathrm{t}) = \frac{\lambda_y}{\lambda_x}\left(\frac{\lambda_y}{\lambda_x} - 1\right) exp[\lambda_y(\gamma_y - \gamma_x)]\left[\bar{\mathrm{F}}_X(t)\right]^{\left(\frac{\lambda_y}{\lambda_x} - 2\right)} < 0. \tag{2.10}$$

   The ratio $\frac{\lambda_y}{\lambda_x}$ will always be less than one since we assumed that $\lambda_y > \lambda_x$ and hence the term $\left(\frac{\lambda_y}{\lambda_x} - 1\right) < 0$, $exp[\lambda_y(\gamma_y - \gamma_x)] > 0$ and $\left[\bar{\mathrm{F}}_X(t)\right]^{\left(\frac{\lambda_y}{\lambda_x} - 2\right)} > 0$ since $0 \leq \bar{F}_X(t) \leq 1$. On the whole, we will get $\frac{d^2}{d\bar{F}_X^2(t)} EROC(\mathrm{t}) < 0$ for $\lambda_y > \lambda_x$. Hence, *EROC* curve is concave in nature and by using result 1, it is also proper.

   Though normal distribution is thought to fit many real world datasets (Hanley, 1988), it is not concave in nature in [0,1] i.e. the Bi-Normal ROC curve may lies below the chance line which in turn reduces the AUC. Therefore, we prefer a model that estimates ROC curve for biomarker which is concave in nature.

3. The slope of the EROC curve at the threshold 't' is found as

   $$slope(t) = \frac{\lambda_y}{\lambda_x} exp\{\mathrm{t}(\lambda_x - \lambda_y) + (\gamma_y\lambda_y - \gamma_x\lambda_x)\}. \tag{2.11}$$

4. *EROC* curve is *TNR* asymmetric.

   Proof: The *KL* divergence between the distribution of diseased and non-diseased group with $f(x)$ as the comparison distribution and $g(x)$ as the reference distribution has been derived as

   $$KL(f, g) = exp[\lambda_x(\gamma_x - \gamma_y)]\left(\frac{\lambda_y}{\lambda_x} - 1 + \lambda_x(\gamma_x - \gamma_y) + \ln\left(\frac{\lambda_x}{\lambda_y}\right)\right). \tag{2.12}$$

Similarly, the *KL* divergence between the distribution of non-diseased and diseased group with $g(x)$ as the comparison distribution and $f(x)$ as the reference distribution has been given as

$$KL(\mathrm{g},\mathrm{f}) = \frac{\lambda_x}{\lambda_y} - 1 + \lambda_x(\gamma_y - \gamma_x) + \ln\left(\frac{\lambda_y}{\lambda_x}\right). \tag{2.13}$$

It is found that $KL(f,g) > KL(g,f)$. A numerical check for this condition is also presented in simulation studies. These two divergence measures would be zero, if the non-diseased and diseased groups are identical. Hence, we have proved that, the *EROC* curve is *TNR* asymmetric.

5. EROC curve is invariance with respect to monotonically increasing transformation.

3. *AUC* of *EROC* Curve and Its Asymptotic Confidence Interval

Let X and Y be two independent and continuous random variables representing non-diseased and diseased group respectively, following two parameter exponential distribution individually with respective parameters (λ_x, γ_x) and (λ_y, γ_y). According to the relationship between γ_x and γ_y, *AUC* of *EROC* takes two different forms which are given as follows.

(i) If $\gamma_x > \gamma_y$, then the *AUC* is obtained as

$$\begin{aligned} P(Y > X) &= \int_{\gamma_y}^{\infty}\int_{\gamma_x}^{y} \lambda_x exp[-\lambda_x(x-\gamma_x)]\lambda_y exp[-\lambda_y(y-\gamma_y)]dxdy \\ &= 1 - \frac{\lambda_y}{\lambda_x+\lambda_y} exp\left[\lambda_x(\gamma_x-\gamma_y)\right]. \end{aligned} \tag{3.1}$$

(ii) If $\gamma_x < \gamma_y$, then the *AUC* is obtained as

$$\begin{aligned} P(Y > X) &= \int_{\gamma_x}^{\infty}\int_{x}^{\infty} \lambda_x exp\left[-\lambda_x(x-\gamma_x)\right]\lambda_y exp\left[-\lambda_y(y-\gamma_y)\right]dxdy \\ &= \frac{\lambda_x}{\lambda_x+\lambda_y} exp\left[\lambda_y(\gamma_y-\gamma_x)\right]. \end{aligned} \tag{3.2}$$

Hence, the *AUC* takes the following form

$$AUC = \begin{cases} \dfrac{\lambda_x}{\lambda_x+\lambda_y}\exp\left[\lambda_y(\gamma_y-\gamma_x)\right] & for \quad \gamma_x > \gamma_y \\ 1-\dfrac{\lambda_y}{\lambda_x+\lambda_y}\exp\left[\lambda_x(\gamma_x-\gamma_y)\right] & for \quad \gamma_y > \gamma_x. \end{cases} \tag{3.3}$$

The MLE of *AUC* can be numerically be obtained by substituting the estimates from the sample data by using (2.5) with the help of invariant property of MLE (Casella and Berger, 2002).

Theorem 1: If $m \to \infty, n \to \infty$ then $\sqrt{m+n}(\hat{AUC} - AUC)$ tends to be normally distributed with mean zero and variance,

$$\tau = \frac{exp\{2\lambda_x(\gamma_x - \gamma_y)\}}{(\lambda_y + \lambda_x)^2}\left(\frac{\lambda_y^2}{m^2} + \frac{\lambda_x^2}{n^2} + \frac{\lambda_x^2\lambda_y^2}{m(\lambda_y+\lambda_x)^2}[-1+(\gamma_x-\gamma_y)(\lambda_y+\lambda_x)]^2 + \frac{\lambda_x^2\lambda_y^2}{n(\lambda_y+\lambda_x)^2}\right.$$
$$\left. - \frac{2\lambda_y^2\lambda_x[-1+(\lambda_y+\lambda_x)(\gamma_x-\gamma_y)]}{m(\lambda_y+\lambda_x)} + \frac{2\lambda_x^2\lambda_y}{n(\lambda_y+\lambda_x)}\right).$$

Proof: Let $L(\theta / x, y)$; $\theta = (\gamma_x, \gamma_y, \lambda_x, \lambda_y)'$ be the likelihood function of the sample observations from *X* and *Y* which is given by

$$\ln L = m\ln\lambda_x - \lambda_x \sum_{i=1}^{m}(x_i - \gamma_x) + n\ln\lambda_y - \lambda_y \sum_{j=1}^{n}(y_j - \gamma_y). \tag{3.4}$$

Asymptotic normality of MLE, states that a consistent solution of the likelihood equation is asymptotically normally distributed about the true value θ i.e. $\hat{\theta} \sim N\left(\theta, I^{-1}(\theta)\right)$.

$$\Rightarrow \sqrt{N}(\hat{\theta} - \theta) \to N(0, I^{-1}(\theta)). \tag{3.5}$$

The *I(θ)* is the Fisher Information matrix is given by

$$I(\theta) = -\begin{bmatrix} E\left(\frac{\partial^2 \ln L}{\partial\gamma_x^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial\gamma_x\partial\gamma_y}\right) & E\left(\frac{\partial^2 \ln L}{\partial\gamma_x\partial\lambda_x}\right) & E\left(\frac{\partial^2 \ln L}{\partial\gamma_x\partial\lambda_y}\right) \\ E\left(\frac{\partial^2 \ln L}{\partial\gamma_y\partial\gamma_x}\right) & E\left(\frac{\partial^2 \ln L}{\partial\gamma_y^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial\gamma_y\partial\lambda_x}\right) & E\left(\frac{\partial^2 \ln L}{\partial\gamma_y\partial\lambda_y}\right) \\ E\left(\frac{\partial^2 \ln L}{\partial\lambda_x\partial\gamma_x}\right) & E\left(\frac{\partial^2 \ln L}{\partial\lambda_x\partial\gamma_y}\right) & E\left(\frac{\partial^2 \ln L}{\partial\lambda_x^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial\lambda_x\partial\lambda_y}\right) \\ E\left(\frac{\partial^2 \ln L}{\partial\lambda_y\partial\gamma_x}\right) & E\left(\frac{\partial^2 \ln L}{\partial\lambda_y\partial\gamma_y}\right) & E\left(\frac{\partial^2 \ln L}{\partial\lambda_y\partial\lambda_x}\right) & E\left(\frac{\partial^2 \ln L}{\partial\lambda_y^2}\right) \end{bmatrix}. \tag{3.6}$$

then $I^{-1}(\theta)$ is calculated as

$$I^{-1}(\theta) = \begin{bmatrix} V(\hat{\gamma}_x) & Cov(\hat{\gamma}_x,\hat{\gamma}_y) & Cov(\hat{\gamma}_x,\hat{\lambda}_x) & Cov(\hat{\gamma}_x,\hat{\lambda}_y) \\ Cov(\hat{\gamma}_y,\hat{\gamma}_x) & V(\hat{\gamma}_y) & Cov(\hat{\gamma}_y,\hat{\lambda}_x) & Cov(\hat{\gamma}_y,\hat{\lambda}_y) \\ Cov(\hat{\lambda}_x,\hat{\gamma}_x) & Cov(\hat{\lambda}_x,\hat{\gamma}_y) & V(\hat{\lambda}_x) & Cov(\hat{\lambda}_x,\hat{\lambda}_y) \\ Cov(\hat{\lambda}_y,\hat{\gamma}_x) & Cov(\hat{\lambda}_y,\hat{\gamma}_y) & Cov(\hat{\lambda}_y,\hat{\lambda}_x) & V(\hat{\lambda}_y) \end{bmatrix}$$
$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \tag{3.7}$$

where

$$a_{11} = \frac{1}{m^2\lambda_x^2};\ a_{12} = a_{21} = a_{41} = a_{14} = 0,\ a_{13} = a_{31} = \frac{-1}{m},\ a_{23} = a_{32} = 0,\ a_{22} = \frac{1}{n^2\lambda_x^2}$$

$$a_{24} = a_{42} = \frac{-1}{n},\ a_{44} = \frac{\lambda_y^2}{n},\ a_{34} = a_{43} = 0 \text{ and } a_{33} = \frac{\lambda_x^2}{m}.$$

Since area under the EROC curve is a function of parameters $\theta = (\gamma_X,\ \gamma_Y,\ \lambda_X,\ \lambda_Y)'$, we will adopt the delta method for finding the approximate variance. $V(A\hat{U}C)$ is obtained as follows:

$$V(A\hat{U}C) = \begin{pmatrix} \frac{\partial AUC}{\partial \gamma_x} & \frac{\partial AUC}{\partial \gamma_y} & \frac{\partial AUC}{\partial \lambda_x} & \frac{\partial AUC}{\partial \lambda_y} \end{pmatrix} \begin{pmatrix} \frac{1}{m^2\lambda_x^2} & 0 & \frac{-1}{m} & 0 \\ 0 & \frac{1}{n^2\lambda_x^2} & 0 & \frac{-1}{n} \\ \frac{-1}{m} & 0 & \frac{\lambda_x^2}{m} & 0 \\ 0 & \frac{-1}{n} & 0 & \frac{\lambda_y^2}{n} \end{pmatrix} \begin{pmatrix} \frac{\partial AUC}{\partial \gamma_x} \\ \frac{\partial AUC}{\partial \gamma_y} \\ \frac{\partial AUC}{\partial \lambda_x} \\ \frac{\partial AUC}{\partial \lambda_y} \end{pmatrix}$$

$$= \frac{exp\left\{2\lambda_x(\gamma_x - \gamma_y)\right\}}{(\lambda_y + \lambda_x)^2}\left(\frac{\lambda_y^2}{m^2} + \frac{\lambda_x^2}{n^2} + \frac{\lambda_x^2\lambda_y^2}{\mathrm{m}(\lambda_y + \lambda_x)^2}[-1 + (\gamma_x - \gamma_y)(\lambda_y + \lambda_x)]^2 \right.$$
$$\left. + \frac{\lambda_x^2\lambda_y^2}{\mathrm{n}(\lambda_y + \lambda_x)^2} - \frac{2\lambda_y^2\lambda_x[-1 + (\lambda_y + \lambda_x)(\gamma_x - \gamma_y)]}{m(\lambda_y + \lambda_x)} + \frac{2\lambda_x^2\lambda_y}{n(\lambda_y + \lambda_x)} \right). \quad (3.8)$$

The estimate of variance is obtained by substituting the estimates of the parameters λ_x, γ_x and λ_y, γ_y. Hence,

$$\frac{\sqrt{N}\,(A\hat{U}C - AUC)}{\sqrt{V(A\hat{U}C)}} \to N(0,1). \quad (3.9)$$

where AUC is the true area under the ROC curve. Thus, it is proved that $A\hat{U}C \sim N(\mathrm{AUC}, \tau)$. Numerically, this result can be visualized from a numerical example presented in the simulation studies.

The standard error of $A\hat{U}C$ can be obtained by taking square root of $V(A\hat{U}C)$ in (3.8). The 100(1-α) % confidence interval is obtained by

$$[A\hat{U}C \pm Se(A\hat{U}C)Z_{\frac{\alpha}{2}}] \quad (3.10)$$

where α is the level of significance and $Z_{\alpha/2}$ is the critical value.

4. Numerical Example

(i) Simulation studies

In this Section, we observed the behavior of asymptotic variance of *AUC* by using Monte Carlo simulation. We have considered four different samples of size (m, n) = (30, 30) with different parametric values for λ_x and λ_y. The MLE of λ_x, γ_x, λ_y and γ_y can be obtained by using (2.5). The assumed parametric values, $\hat{AUC}$, $Se(\hat{AUC})$ and 95% confidence interval for $\hat{AUC}$ of *EROC* curve using asymptotic MLE and Monte Carlo methods are presented in Table 1.

Table 1: Estimated parameters, $\hat{AUC}$, $Se(\hat{AUC})$, and 95% confidence interval for $\hat{AUC}$ of *EROC* curve using asymptotic MLE and Monte Carlo methods

Description	Asymptotic MLE method	Monte Carlo method	$\bar{Y} - \bar{X}$
γ_x=0.15 γ_y=0.99 λ_x=4.50 λ_y=0.82 $\hat{AUC}(Se(\hat{AUC}))$ CI (95%)	0.1577 (0.0061) 0.9975 (0.0354) 5.4691 (0.9985) 0.8086 (0.1719) 0.9985 (0.0020) [0.9947, 1.0000]	0.1573 (0.0074) 1.0301 (0.0394) 4.8354 (0.9375) 0.8799 (0.1692) 0.9966 (0.0029) [0.9909, 1.0000]	1.7192
γ_x=0.15 γ_y=0.44 λ_x=2.00 λ_y=1.00 $\hat{AUC}(Se(\hat{AUC}))$ CI (95%)	0.1846 (0.0119) 0.4674 (0.0186) 2.8006 (0.5113) 1.7937 (0.3275) 0.8232 (0.0864) [0.6539, 0.9935]	0.1734 (0.0133) 0.4458 (0.0255) 2.6760 (0.5088) 1.3954 (0.2683) 0.8289 (0.0854) [0.6625, 0.9973]	0.9211
γ_x=0.16 γ_y=0.42 λ_x=1.80 λ_y=1.30 $\hat{AUC}(Se(\hat{AUC}))$ CI (95%)	0.1788 (0.0612) 0.4509 (0.0459) 1.8325 (0.3346) 1.3761 (0.2512) 0.7395 (0.1081) [0.5265, 0.9526]	0.1784 (0.0184) 0.4458 (0.0255) 1.9299 (0.3728) 1.3890 (0.2642) 0.7450 (0.1083) [0.5327, 0.9573]	0.3963
γ_x=0.18 γ_y=0.32 λ_x=2.5 λ_y=2.0 $\hat{AUC}(Se(\hat{AUC}))$ CI (95%)	0.1826 (0.0126) 0.3288 (0.0155) 2.6462 (0.4831) 2.1457 (0.3918) 0.6958 (0.1179) [0.4647, 0.9270]	0.1933 (0.0132) 0.3367 (0.0165) 2.6804 (0.5108) 2.1459 (0.4101) 0.6932 (0.1045) [0.4884, 0.8980]	0.23428

*CI : Confidence Interval

In Table 1, first column represents the assumed parametric values, second column represents the MLE of parameters with their standard errors given within the parenthesis along with the 95% asymptotic confidence interval for $\hat{AUC}$, the third column provides the Monte Carlo estimates of parameters with their standard errors given within parenthesis along with the 95% confidence interval for $\hat{AUC}$ and the final column represents the difference between the mean of diseased $(\bar{Y})$ and the mean of non-diseased $(\bar{X})$ samples i.e. $(\bar{Y} - \bar{X})$ As far as the discrimination is concerned, the measure $(\bar{Y} - \bar{X})$ indirectly tell us the degree of separation between the two groups.

We also observe that, as the measure $(\bar{Y} - \bar{X})$ or the deviation between the estimated parameters of non-diseased and diseased group increases, the *AUC* tends to increase which in turn decreases the standard error of *AUC*. We also notice that there is no major difference in the estimates of parameters and $\hat{AUC}$, obtained by asymptotic and Monte Carlo methods from Table 1. The *EROC* curves for different parametric values are plotted in Figure 1.

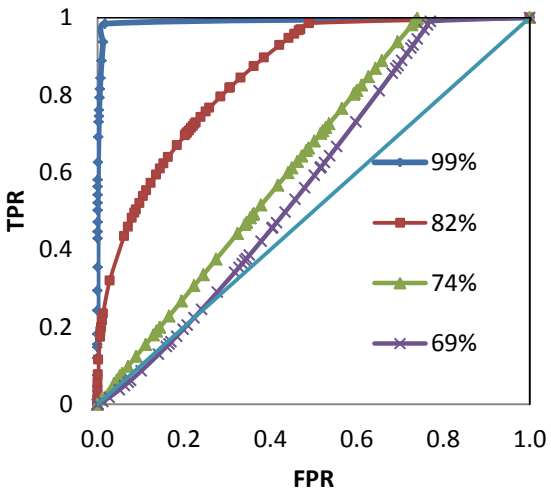

Fig. 1 *EROC* curve for different values of *AUC*

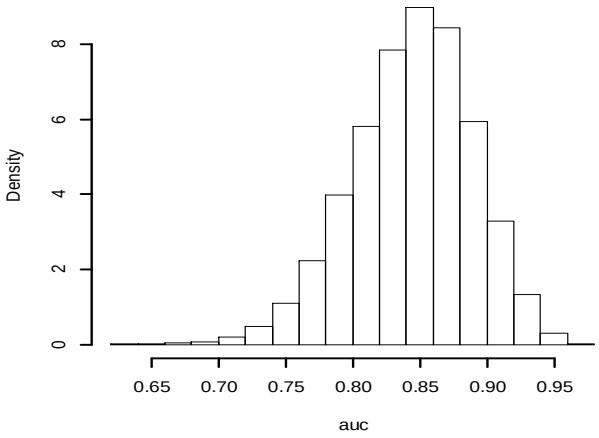

Fig. 2 Histogram of the AUC statistic

In support of the theorem1, we have plotted the values of the statistic, AUC in the following Figure.

Table 2: $\hat{AUC}$, $Se(\hat{AUC})$, **95% asymptotic confidence interval for** $\hat{AUC}$ **of EROC curve for different values of parameters**

Parametric values	Sample size					
	(5, 5)	(10, 10)	(30, 30)	(50, 50)	(80, 80)	(100, 100)
γ_x=0.15;	0.9985	0.9985	0.9985	0.9985	0.9985	0.9985
γ_y=0.99;	0.0046	0.0031	0.0018	0.0014	0.0011	0.0010
λ_x=4.5;	[0.9896,	[0.9925,	[0.9952,	[0.9960,	[0.9966,	[0.9968,
λ_y=0.82	1.0000]	1.0000]	1.0000]	1.0000]	1.0000]	1.0000]
γ_x=0.16;	0.8232	0.8232	0.8232	0.8232	0.8232	0.8232
γ_y=0.42;	0.2199	0.1520	0.0864	0.0667	0.0526	0.0470
λ_x=2.5;	[0.3922,	[0.5253,	[0.6539,	[0.6925,	[0.7200,	[0.7309,
λ_y=1.3	1.0000]	1.0000]	0.9925]	0.9539]	0.9263]	0.9154]
γ_x=0.16;	0.7395	0.7395	0.7395	0.7395	0.7395	0.7395
γ_y=0.42;	0.2778	0.1916	0.1087	0.0839	0.0662	0.0592
λ_x=1.80;	[0.1951,	[0.3640,	[0.5265,	[0.5751,	[0.6098,	[0.6236,
λ_y=1.30	1.0000]	1.0000]	0.9526]	0.9039]	0.8693]	0.8555]
γ_x=0.18;	0.6958	0.6958	0.6958	0.6958	0.6958	0.6958
γ_y=0.32;	0.3020	0.2080	0.1179	0.0910	0.0718	0.0640
λ_x=2.50;	[0.1039,	[0.2881,	[0.4647,	[0.5175,	[0.5552,	[0.5701,
λ_y=2.00	1.0000]	1.0000]	0.9270]	0.8743]	0.8366]	0.8217]

In Table 2, the data has been generated by assuming different parametric values for various sample sizes namely {(5, 5), (10, 10), (30, 30), (50, 50), (80, 80), (100, 100)}. The first element in each row represents the accuracy; second element is the standard error of $\hat{AUC}$, the third and fourth value being the lower and upper confidence limits. It is obvious that the asymptotic estimate of standard error holds good for large sample size. As we go through the columns, the standard error tends to decrease and the confidence intervals get narrow as the sample size increase.

(ii) Real life example

The proposed method is applied to Prostate cancer markers(PSA). PSA is a biomarker which is significant in detecting the prostate cancer. The data consisted of 50 randomly chosen individuals who were affected by prostate cancer and 50 non-diseased individuals who were participated in a lung cancer prevention trial (Etzioni , 1999). The two correlated prostate cancer biomarkers were considered namely total serum PSA (tPSA) and the ratio of (percent) free to total PSA (fPSA). Among these biomarkers tPSA has higher *AUC* than fPSA and hence it is preferred to assess the accuracy of diagnosis for prostate cancer.

The tPSA has been evaluated for the Goodness of Fit for two parameter exponential distribution using Kolmogrov-Smirnov, Anderson-Darling and Chi-Square test. The results are reported for significance level (α) 20, 10, 5, 2 and 1% in Table 3 from the software 'Easy Fit'. The EROC curve is plotted for tPSA and it is presented in Figure 2.

Table 3: Goodness of Fit test for tPSA biomarker

Group	Test	Statistic	p-value	Rank	α %
H	K-S	0.121	0.426	18	20, 10, 5, 2, 1
	χ^2	0.6441	0.958	3	20, 10, 5, 2, 1
	A-D	1.219	-	23	20, 10, 5, 2, 1
D	K-S	0.128	0.357	29	20, 10, 5, 2, 1
	χ^2	4.004	0.5488	24	20, 10, 5, 2, 1
	A-D	1.747	-	31	10, 5, 2, 1

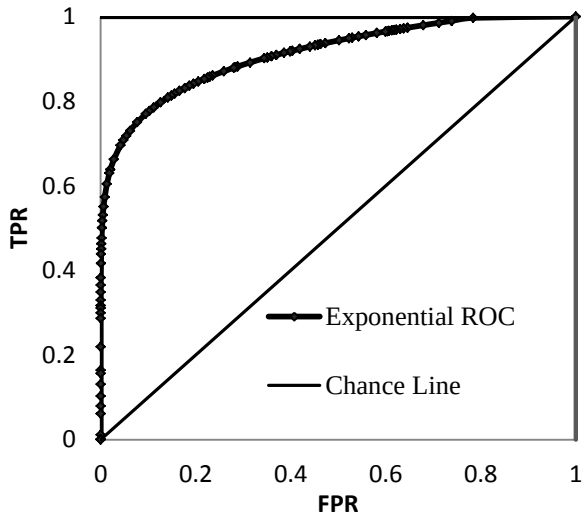

Fig. 2 EROC curve for tPSA biomarker

The *EROC* model showed that the marker 'tPSA' is able to identify the prostate cancer individual with an accuracy of 93% with the estimated Standard error, 0.055 with a confidence interval [0.8079, 1.000]. The *sensitivity* and *specificity* of tPSA by using *EROC* are 83% and 76% respectively at the threshold 2.865 ng/ml.

From *sensitivity* and *specificity*, we could infer that an individual who is having the "tPSA" marker value greater than 2.865 ng/ml is 83% likely to be detected with the prostate cancer. Similarly, an individual having the marker value less than 2.865 ng/ml is 76% likely to be not detected with the prostate cancer.

5. Conclusion

In this paper, we have also extended the one parameter Bi-Exponential *ROC* curve analysis to two parameter exponential *ROC* curve analysis. The properties of the two parameter bi-exponential *ROC* curve have been studied. It is found that, *EROC* is monotonically increasing, concavity an important property for a *ROC* to be proper, *TNR* asymmetric which is justified theoretically as well as graphically. The 95% asymptotic confidence interval for $\hat{AUC}$ have been derived. For the prostate cancer data, the EROC

model showed that the biomarker 'tPSA' is able to identify the prostate cancer individual with an accuracy of 91% with the estimated standard error, 0.055 with a confidence interval [0.8079, 1.000]. The sensitivity and specificity are 83% and 76% respectively at the threshold 2.865 ng/ml. The proposed EROC curve analysis can be adopted for assessing the accuracy of classification made by a particular biomarker provided the goodness of fit test is evaluated.

Acknowledgements

The authors wish to express their gratitude to UGC MRP and The Vice Chancellor & Registrar, Pondicherry University (through the University Research Fellowship) for the financial assistance to carry out this research work.

References

1. Amala, R. and Pundir, S. 2015, Detecting Diagnostic Accuracy of two Biomarker using Bi-variate Lognormal ROC curve, Journal of Applied Statistics, yet to be published.
2. Amala, R. and Pundir, S., 2015, ROC curve and AUC for a Left truncated sample from Rayleigh Distribution, *American Journal of Mathematical and Management Sciences* 34, 1-28.
3. Amala, R. Pundir, S., 2012, Statistical Inference on AUC from a Bi-Lognormal ROC model for Continuous data, International *Journal of Engineering Science and Innovative Technology*, 1(2), 283-295.
4. Betinec, M., 2008, Testing the difference of the ROC Curves in Biexponential model, *Tatra Mountains Mathematical Publications* 39, 215-223.
5. Campbell, G., Ratnaparkhi, M.V., 1993, An application of Lomax distributions in receiver operating characteristic (roc) curve analysis, *Communication in Statistics* 22(6), 1681–1687.
6. Casella, G., Berger, R. L., 2002, *Statistical Inference*, Cengage Learning: 2nd Edition, India, ISBN: 13: 978-053424-3126.
7. Dorfman, D.D., Berbaum, K.S., Metz, C.E., Lenth, R.V., Hanley, J.A., Dagga, H.A., 1996, Proper Receiver Operating Characteristics Analysis: The bigamma model, *Academic Radiology* 4, 138-149.
8. Etzioni, R., Pepe, M., Longton, G., Hu C. and Goodman, G., 1999, Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer, *Medical Decision Making*, 19(3), 242-251.
9. Hughes, G., Bhattacharya, B., 2013, Symmetry Properties of Bi-Normal and Bi-Gamma Receiver Operating Curves are Described by Kullback-Leibler Divergences, *Entropy* 15, 1342-1356.
10. Hussain, E., 2011, The ROC Curve Model from Generalized-Exponential Distribution, *Pakistan Journal of Statistics and Operations Research*, 7(2), 323-330.

11. Johnson, N. L., Kotz, S., Balakrishnan, N., 1994, *Continuous Univariate Distributions*, Volume 2, John Wiley and Sons, New York.

12. Krzanowski, W.J., Hand, D.J., 2002, *ROC curves for continuous data, Monographs on Statistics and Applied Probability*, CRC Press, Taylor and Francis Group, New York.

13. Marzban, C., 2004, The ROC curve and the Area under it as Performance Measures, *Whether and Forecasting* 19, 1106-1114.

14. Mumford,S. L., Schisterman, E. F. A., Vexler, A. and Liu, A. (2006), Pooling biospecimens and limits of detection: effects on ROC curve analysis, *Biostatistics*, 7(4), 585-598.

15. O' Malley, J. and Zou, K.H. 2006, Bayesian multivariate hierarchical transformation models for ROC analysis, *Statistics in Medicine*, 25(3), 459–479.

16. Oglive, J. C., Creelman, C.D., 1968, Maximum-likelihood Estimation of Receiver Operating Characteristic Curve Parameters, *Journal of Mathematical Psychology*, **5**, 377-391.

17. Perkins, N.J., Schisterman, E.F., and Vexler, A. 2006, Receiver operating characteristic curve inference from a sample with a limit of detection, *American Journal of Epidemiology* 165(3), 325-333.

18. Pundir, S. and Amala, R., 2014c, Standard error of Area under the Bi-Exponential ROC curve, *International Journal of Engineering Sciences and Research Technology* 3(8), 712-721.

19. Pundir, S., Amala, R. 2014b, Evaluation of Area Under the Constant Shape Bi-Weibull ROC Curve, *International Journal of Modern and Applied Statistical Methods*, 13(1), 305-328.

20. Pundir, S., Amala, R., 2012:b, A study on the comparison of Bi-Rayleigh ROC model with Bi-Gamma ROC model, Edited volume, *Application of Reliability Theory and Survival Analysis*, Bonfring Publication, Coimbatore, Tamil Nadu, 196-209.

21. Pundir, S., Amala, R., 2012a, A study on the Bi-Rayleigh ROC model, *Bonfring International Journal of Data Mining* **2**(2), 42-47.

22. Pundir, S., Amala, R., 2014a, Parametric Receiver Operating Characteristic Modeling for continuous data: A Glance, *Model Assisted Statistics and Application* 9(2), 121-135.

23. Zhang, Z. and Pepe, M.S. 2012, A linear regression framework for Receiver Operating Characteristic (ROC) curve analysis, *Journal of Biometrics and Biostatistics* 3(2), doi:10.4172/2155-6180.1000137.