Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 113–121

Estimation of Area under the Multi-Class ROC for Non-Normal Data

Arunima S. Kannan and R. Vishnu Vardhan
Department of Statistics, Ramanujan School of Mathematical Sciences,
Pondicherry University, Puducherry, India

Received: 24 March 2022; Revised: 20 May 2022; Accepted: 05 July 2022

Abstract

In modelling ROC curves, there are several bi-distributional ROC models available in the literature. These are developed in the context of normal and non-normal data patterns and in the framework of binary classification. However, in most of the practical data at hand may exhibit multi-model patterns or it may be of multi-class, then the existing bi-distributional ROC forms are not viable to apply and fit the curve. So, in this paper, we made an attempt to address the above mentioned situations using finite mixtures. We proposed a mixture Exponential ROC model and its measures like AUC, FPR, TPR and optimal cut-offs are derived. The methodology is supported with simulated and real data sets.

Key words: Area under the curve; Exponential distributions; Finite mixture; ROC curve.

AMS Subject Classifications: 62P10

1. Introduction

The Receiver Operating Characteristic (ROC) curve is a classification tool that is widely used in the field of diagnostic medicine. Classification of individuals into one of the predefined groups/populations will be based on a cut-off. For a given value of cut-off, one can define the pair of true-positive rates (TPRs) and false-positive rates (FPRs), using these the ROC curve is constructed. The summary measure of ROC, which assesses the performance of a particular diagnostic test, is the area under the curve (AUC) whose value lies between 0 and 1. Higher the AUC value, the better the diagnostic test's performance.

The initial work on the distributional approach to model the ROC curve was by Green *et al.* (1966) where data is assumed to follow the gaussian distribution. In later years, Dorfman and Alf (1968) gave the maximum likelihood estimates for the binormal ROC curve. Metz (1978) gave a detailed explanation about the basic principles of ROC curve and its measures. Estimation of the parameters of the binormal model was of prime focus by many researchers. Goddard and Hindberg (1990) proposed a ROC model that meets the criterion of non-normal data, namely the Bi-logistic ROC model. Farraggi and Reiser (2002) provided the parametric and non-parametric approach of estimating the AUC of the ROC curve. Over the years, many researchers have attempted to develop Bi-distributional

Corresponding Author: R. Vishnu Vardhan
Email: vrstatsguru@gmail.com

ROC curves; a few to mention are the Bi-Generalized Exponential ROC model by Hussain (2011), Vardhan *et al.* (2012) on Bi-Exponential and Bi-Weibull ROC model, Bi-Gamma ROC model by Hussain (2012). A detailed review of several bi-distributional ROC models was made by Balaswamy and Vishnu (2016).

In classification, one of the main issues is that in most of the data sets, we do not have the information about the group membership; there, we need to use appropriate statistical methods to figure out the homogeneous subsets. We can make the graphical depiction of unsupervised data, and it may exhibit unimodal or multi-model patterns that exist in the data. One of the most widely used methodologies that helps to sum up the multi-model patterns accompanied by their respective weights in the form of convex combination is the Finite Mixture Models (FMM). The general expression of the finite mixture distribution is given in equation (1).

$$g(x) = \sum_{i=1}^{k} \pi_i f_i(x) \tag{1}$$

where, π_i's are the mixture proportions or mixture weights such that $\pi_i > 0$; $\sum_{i=1}^{k} \pi_i = 1$ and f_i's are component distributions ; $i = 1, 2, ..k$.

The seminal work on mixture models using crabs data was by Pearson (1894) and a detailed study on mixture models was given by Lindsay (1995). Over the years, the practical applicability of FMM branched out to various fields like remote sensing, environmental studies, diagnostic medicine, survival analysis, social and psychological science (Peel and MacLahlan, 2000). But, most of the works reported in the literature were based on the normal distribution. However, there are several practical instances where data may not follow the normal distribution. In such situations, the existing normal mixture models do not support, hence there is a need to have mixture models for non-normal data. Here a brief review on Mixture Exponential is presented. Mendenhall and Hader (1958) estimated the parameters of mixed exponentially distributed failure time distribution. Jewell (1982) gives a detailed explanation of the mixture of exponential distributions and gives a practical algorithm for the maximum likelihood estimate. Wang and Wang (2014) proposed an EM Algorithm for the finite mixture of exponential distribution models. Literature has many applications with the use of mixture exponential distributions, recently, Polymenis (2020) used mixture of exponential distributions for assessing hazard rates from COVID-19.

This paper provides an approach to classify the non-normal data with hidden populations. Here it is assumed that the population follows a mixture of exponential distribution and derived the Mixture Exponential ROC and its measures. The rest of the paper is organized as follows. Section 2 discusses the proposed Mixture Exponential ROC. Section 3 provides numerical illustrations of the proposed methodology with simulated as well as real-life data sets. Section 4 concludes the paper with the summary.

2. Mixture exponential ROC

Let us assume that healthy population, $H \sim exp(\theta_0)$ and diseased population has two sub populations/mixture of populations of D_1 and D_2 such that, $D_1 \sim exp(\theta_1)$ and $D_2 \sim exp(\theta_2)$. Then the expressions for intrinsic measures of Mixture Exponential ROC (mixExp ROC) are defined as FPR of mixExp ROC (mixFPR) is given as

$$mixFPR = \pi_1 FPR_1 + \pi_2 FPR_2 \tag{2}$$

where

$$\begin{aligned} FPR_1 = P(S > t_1 \mid H) \;\; &; \;\; FPR_2 = P(S > t_2 \mid D_1) \\ FPR_1 = x(t_1) = e^{-\theta_0 t_1} \;\; &; \;\; FPR_2 = x(t_2) = e^{-\theta_1 t_2} \end{aligned} \tag{3}$$

where π_i's are mixing proportions/weights, t_1 and t_2 are the respective cut-off values for the classification of (H, D_1) and (D_1, D_2) respectively. Here FPR_1, FPR_2 are the false positive rate values of H and D_1 populations & D_1 and D_2 populations respectively. From equation (3) we can write t_1 and t_2 as

$$t_1 = -\frac{log(x(t_1))}{\theta_0} \;\; ; \;\; t_2 = -\frac{log(x(t_2))}{\theta_1} \tag{4}$$

TPR of mixExp ROC (mixTPR) is given as

$$mixTPR = \pi_1 TPR_1 + \pi_2 TPR_2 \tag{5}$$

where

$$\begin{aligned} TPR_1 = P(S > t_1 \mid D_1) \;\; &; \;\; TPR_2 = P(S > t_2 \mid D_2) \\ TPR_1 = y(t_1) = e^{-\theta_1 t_1} \;\; &; \;\; TPR_2 = y(t_2) = e^{-\theta_2 t_2} \end{aligned} \tag{6}$$

Here, TPR_1, TPR_2 are the true positive rate values of H and D_1 populations & D_1 and D_2 populations respectively. Substituting equation (4) in (6) we will get the mixture exponential ROC curve which be written as

$$mixROC = \pi_1 ROC_1 + \pi_2 ROC_2 \tag{7}$$

$$ROC_1 = x(t_1)^{\frac{\theta_1}{\theta_0}} \;\; ; \;\; ROC_2 = x(t_2)^{\frac{\theta_2}{\theta_1}} \tag{8}$$

By equating the pdf's of the distributions, the optimal cut-off can be obtained as

$$t_1 = \frac{log\theta_1 - log\theta_0}{\theta_1 - \theta_0} \;\; ; \; t_2 = \frac{log\theta_2 - log\theta_1}{\theta_2 - \theta_1} \tag{9}$$

accuracy can be expressed notationally as

$$mixAUC = \int_0^1 mixROC(t)dt = \pi_1 \frac{\theta_0}{\theta_0 + \theta_1} + \pi_2 \frac{\theta_1}{\theta_1 + \theta_2} \tag{10}$$

Youden's J index (Youden, 1950) is another way of summarising the performance of a diagnostic test.

Youden's J index is defined as

$$J = (TPR - FPR) \tag{11}$$

then maximum Youden's index is reported as

$$J = max_{(t)}\left(TPR(t) - FPR(t)\right) \tag{12}$$

where t denotes the classification threshold for which J is maximal. From the above equation, the optimal threshold can be estimated at the maximum Youden's Index value, since, the maximum distance between the curve and the chance line can be used to identify the optimal threshold and will be unique in nature. This optimal threshold classifies the individuals with a better accuracy and further it can be used to assign the status of the unspecified subjects/individuals. A value of J=1 sures that there are no false positives or false negatives, *i.e.* the test is perfect.

3. Numerical illustrations

3.1. Simulated data

Simulation studies are carried out at various parameter combinations by considering equal mixture weights. Using the parameters values given in Table 1, random samples are generated for n = (25, 50, 100, 200).

Table 1: Initial parameters

Case	π_1	π_2	θ_0	θ_1	θ_2
I	0.5	0.5	0.4	0.1	0.01
II	0.5	0.5	0.4	0.2	0.05
III	0.5	0.5	0.4	0.25	0.1
IV	0.5	0.5	0.4	0.4	0.4

The results pertaining to each case at every sample size in Table 2 and respective ROC curves are depicted in Figure 1. The parameter values are chosen in such a way that they exhibit worst and moderate classification scenarios. The estimation of parameters of the mixture distribution is carried out using EM algorithm in R software. It is a known fact that as higher the AUC, minimum will be the overlapping region, in turn giving out better percentage of correct classification. The estimated values of t_1 and t_2 are the optimal one, which are derived using the Youden's J. The interpretation of t_1 and t_2 goes like this:

Let S be the values/samples generated using each parameter combination

$$\text{The individual will be classified as} = \begin{cases} H, \text{ if } S \leq t_1 \\ D_1, \text{ if } t_1 < S \leq t_2 \\ D_2, \text{ if } S > t_2 \end{cases}$$

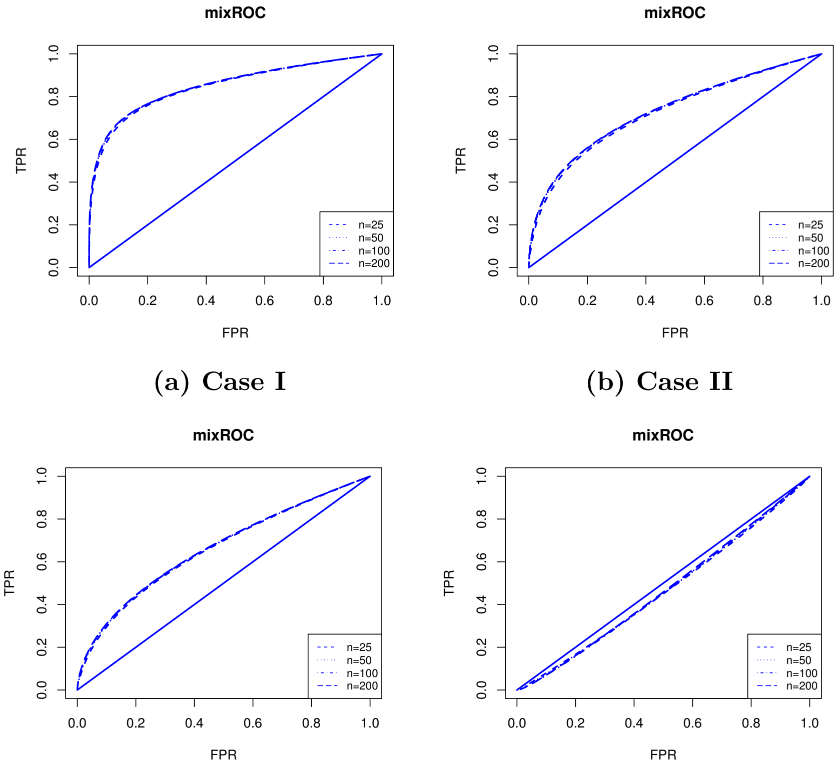
To have a better understanding of t_1 and t_2, FPR and TPR, let us consider an instance under case I from Table 2. For n=100, the $\hat{t_1} = 4.60711$; $\hat{t_2} = 25.43069$; $\widehat{mixFPR} = 0.12541$; $\widehat{mixTPR} = 0.70705$ and $\hat{J} = 0.58164$ results $\widehat{mixAUC} = 0.85340$. This means to that an individual will be classified in the following manner

$$\text{The individual will be classified as} = \begin{cases} H, \text{ if } S \leq 4.60711 \\ D_1, \text{ if } 4.60711 < S \leq 25.43069 \\ D_2, \text{ if } S > 25.43069 \end{cases}$$

Table 2: ROC curve estimates for simulated data

Case	n	$\widehat{\pi_1}$	$\widehat{\pi_2}$	$\widehat{t_1}$	$\widehat{t_2}$	$\widehat{J}$	$\widehat{mixFPR}$		$\widehat{mixTPR}$		$\widehat{mixAUC}$	
							$\widehat{FPR_1}$	$\widehat{FPR_2}$	$\widehat{TPR_1}$	$\widehat{TPR_2}$	$\widehat{AUC_1}$	$\widehat{AUC_2}$
I	25	0.50213	0.49787	4.56913	25.23253	0.57538	0.13001		0.70539		0.85036	
							0.16142	0.08043	0.62848	0.77137	0.79547	0.90595
	50	0.49731	0.50269	4.59730	25.44494	0.58133	0.12534		0.70667		0.85318	
							0.15910	0.07838	0.62913	0.77283	0.79784	0.90766
	100	0.49938	0.50062	4.60711	25.43069	0.58164	0.12541		0.70705		0.85340	
							0.15871	0.07795	0.62899	0.77323	0.79837	0.90821
	200	0.49976	0.50024	4.61104	25.45291	0.58233	0.12478		0.70710		0.85378	
							0.15800	0.07801	0.62933	0.77331	0.79917	0.90832
II	25	0.53123	0.46877	3.42247	9.14696	0.34870	0.21594		0.56464		0.72485	
							0.25757	0.15880	0.50218	0.63186	0.66176	0.79887
	50	0.50817	0.49183	3.41111	9.14908	0.35861	0.20937		0.56798		0.73139	
							0.24995	0.15892	0.50140	0.63015	0.66688	0.79854
	100	0.50680	0.49320	3.46106	9.23796	0.35864	0.20821		0.56685		0.73169	
							0.25005	0.15837	0.50063	0.62916	0.66671	0.79866
	200	0.49881	0.50119	3.46938	9.24755	0.36089	0.20754		0.56843		0.73312	
							0.24999	0.15823	0.50041	0.62900	0.66676	0.79885
III	25	0.53521	0.46479	3.08832	6.03442	0.23997	0.26023		0.50021		0.65824	
							0.31259	0.21777	0.48196	0.54426	0.61200	0.71325
	50	0.52410	0.47590	3.11203	6.04253	0.24393	0.25670		0.50063		0.66124	
							0.29402	0.21777	0.46093	0.54316	0.61165	0.71320
	100	0.51487	0.48513	3.13166	6.11228	0.24513	0.25448		0.49960		0.66246	
							0.28625	0.21780	0.45799	0.54262	0.61539	0.71332
	200	0.51262	0.48738	3.12554	6.10010	0.24725	0.25349		0.50073		0.66390	
							0.28586	0.21708	0.45705	0.54311	0.61515	0.71427
IV	25	0.56482	0.43518	2.47823	2.48022	0.02424	0.56973		0.59397		0.49789	
							0.58665	0.59789	0.62503	0.63822	0.49900	0.50069
	50	0.54976	0.45024	2.47846	2.48475	0.01853	0.55706		0.57559		0.49997	
							0.61440	0.58862	0.64271	0.61674	0.49925	0.49956
	100	0.53953	0.46047	2.47643	2.48859	0.01319	0.57909		0.59228		0.49943	
							0.62699	0.61702	0.64770	0.63862	0.49972	0.50064
	200	0.55491	0.44509	2.48864	2.49903	0.01036	0.57716		0.58752		0.50036	
							0.63807	0.62109	0.65242	0.63512	0.49983	0.50025

The cut-offs $\widehat{t_1}$ and $\widehat{t_2}$, are able to provide an $\widehat{FPR}$ of 12.54% and $\widehat{TPR}$ of 70.70%. So, if there are 100 samples in the data, these two cut-offs will be able to detect the true positives upto 70% and with a wrong classification of around 12%. In total, the accuracy of $\hat{t_1}$ and $\hat{t_2}$ is around 85%. In similar lines, the other combinations can be interpreted. From Figure 1, it is clear that the area under the curve is decreasing from case I to case IV, which is indicating that the accuracy of the classification is decreasing from case I to case IV. The curve of case IV is close to the diagonal line, results the worst classification.



(a) Case I

(b) Case II

(c) Case III

(d) Case IV

Figure 1: mixExp ROC curves of simulated data sets

3.2. Real data

The real data set represent the survival times of 121 patients with breast cancer obtained from a large hospital in a period from 1929 to 1938 (Lee and Wang, 2003). This data set has recently been studied by Yang *et al.* (2021). The p-value for K-S test for exponential distribution for this data is 0.06024 (Test statistic, $D = 0.12031$), which indicates that the data follows exponential distribution. We have $\theta_0 = 0.4$, $\theta_1 = 0.0280$ and $\theta_2 = 0.0202$. The estimated measures of mixExp ROC curve is given in Table 3 and respective ROC curve is depicted in Figure 2. As the curve is observed between the chance line and the left top corner and also connecting to the AUC= 0.7355, this indicates a moderate amount of classification with cut-offs $\hat{t_1}$ and $\hat{t_2}$.

From Table 3, $\hat{t_1} = 20.24715$; $\hat{t_2} = 46.32893$; $\widehat{mixFPR} = 0.24871$; $\widehat{mixTPR} = 0.6628$

Table 3: mixExp ROC curve estimates of breast cancer data

$\widehat{\pi_1}$	$\widehat{\pi_2}$	$\widehat{t_1}$	$\widehat{t_2}$	$\widehat{J}$	$\widehat{mixFPR}$		$\widehat{mixTPR}$		$\widehat{mixAUC}$	
					$\widehat{FPR_1}$	$\widehat{FPR_2}$	$\widehat{TPR_1}$	$\widehat{TPR_2}$	$\widehat{AUC_1}$	$\widehat{AUC_2}$
0.49918	0.50018	20.24715	46.32893	0.24908	0.24871		0.6628		0.7355	
					0.04951	0.3493	0.8418	0.3865	0.9458	0.5253

and results $\widehat{mixAUC} = 0.7355$. This means that an individual will be classified as follows.

$$\text{The individual will be classified as} = \begin{cases} \text{low survival rate, if } S \leq 20.24715 \\ \text{moderate survival rate, if } 20.24715 < S \leq 46.32893 \\ \text{high survival rate, if } S > 46.32893 \end{cases}$$

The cut-offs $\widehat{t_1}$ and $\widehat{t_2}$, are able to provide false positive rate of 24.87% and true positive rate of 66.28%. In total, the accuracy is around 73.55%, which indicates of a moderate classification. Further, 100 bootstrap samples are generated from the breast cancer data. The bootstrap estimates of important measures and their confidence intervals are reported in Table 4. The mixROC curves are also drawn for all the bootstrap samples and is shown in Figure 3. The curves clearly depict a moderate classification. From Table 4, it is observed

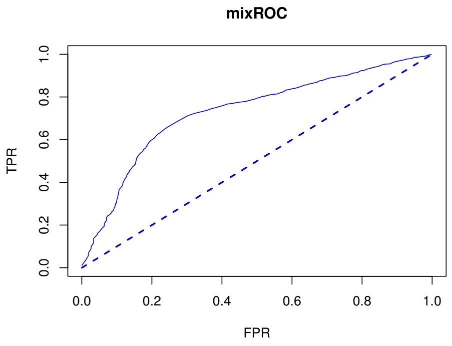


Figure 2: mixExp ROC Curve

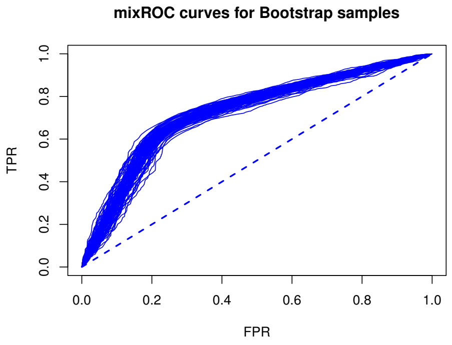


Figure 3: mixExp ROC Curves for 100 bootstrap samples

Table 4: Bootstrap estimates of breast cancer data

Bootstrap	$\widehat{mixFPR}_{boot}$	$\widehat{mixTPR}_{boot}$	$\hat{J}_{boot}$	$\widehat{mixAUC}_{boot}$
Estimates	0.248375	0.647383	0.399009	0.7230622
Variance	0.000271	0.000207	0.000309	0.000160269
95% Lower limit	0.2375	0.635805	0.388395	0.7164525
95% Upper limit	0.257676	0.658599	0.41017	0.730174

that the cut-offs provide reasonably low FPR = 0.248375 (0.2375, 0.257676) and a good level of TPR = 0.647383 (0.635805, 0.658599). This means that if there are 100 subjects then the cut-offs will be able to detect the class/status of around 65 subjects correctly, providing an accuracy of 0.7230 (0.7164525, 0.730174). Upon conducting 100 bootstraps and constructing the 95% confidence interval the outcomes revealed an observation that the width of the confidence interval is shorter indicating consistent estimates. Further the results of the bootstrap matches closely to the results in Table 3.

4. Summary

In this paper, we proposed an ROC model that follows exponential distribution with multi-class classification. Here we considered situation like (i) if we come across multi-model patterns in the diseased population and (ii) if there are more than two categories in the data. The proposed model addresses the above two situations and is dealt using the concept of finite mixtures. The model so constructed is named as *Mixture Exponential ROC Curve*. The measures such as mixAUC, mixFPR, mixTPR and optimal cut-offs are derived and supported with numerical illustrations. With respect to simulations, we tried to present the behaviour of the proposed ROC model by constructing the worst and moderate cases. Further the numerical illustrations is extended with breast cancer dataset. It is noticed that there were two sub populations in the diseased population. The overall AUC is observed to be 73.5 and optimal thresholds are 20.24 and 46.32. To summarize the work, and mixture exponential ROC model is proposed, and for the non-normal and multi-class data this model can be applied.

Acknowledgements

The authors are thankful to the referees for their valuable comments and suggestions which helped in improvising the paper.

References

Balaswamy, S. and Vishnu, R. V. (2016). An anthology on parametric ROC models. *Research Reviews:Journal of Statistics*, **5**, 32-46.

Dorfman, D. D. and Alf, E. (1968). Maximum likelihood estimation of parameters of signal detection theory - A direct solution. *Psychometrika*, **33**, 117-124.

Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, **21**, 3093-3106.

Goddard, M. and Hindberg, I. (1990). Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine*, **9**, 325-337.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* John Wiley & Sons.

Hussain, E. (2011). The ROC curve model from generalized-exponential distribution. *Pakistan Journal of Statistics and Operation Research*, **7**, 283-303.

Hussain, E. (2012). The bi-gamma ROC curve in a straightforward manner. *Journal of Basic & Applied Sciences*, **8**, 309-314.

Jewell, N. P. (1982). Mixtures of exponential distributions. *The Annals of Statistics*, **10**, 479–484.

Lee, E. T. and Wang, J. (2003). *Statistical Methods for Survival Data Analysis.* 3^{rd} Edition, John Wiley & Sons.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, **5**, 1–163.

Mendenhall, W. and Hader, R. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, **45**, 504-520.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**, 283–298.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, **185**, 71–110.

Peel, D. and MacLahlan, G. (2000). *Finite Mixture Models.* John Wiley & Sons.

Polymenis, A. (2020). An application of a mixture of exponential distributions for assessing hazard rates from covid-19. *Journal of Population Therapeutics and Clinical Pharmacology*, **27**, 58–63.

Vardhan, R. V., Pundir, S., and Sameera, G. (2012). Estimation of area under the ROC curve using exponential and weibull distributions. *Bonfring International Journal of Data Mining*, **2**, 52–56.

Wang, Y. and Wang, J. (2014). The EM Algorithm for The Finite Mixture of Exponential Distribution Models. *International Journal of Contemporary Mathematical Sciences*, **9**, 57-64.

Yang, Y., Tian, W., and Tong, T. (2021). Generalized mixtures of exponential distribution and associated inference. *Mathematics*, **9**, 1-22.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, **3**, 32-35.