

Modelling Receiver Operating Characteristic Curves Using Gaussian Mixtures

Amay S. M. Cheam* and Paul D. McNicholas

Abstract

The receiver operating characteristic curve is widely applied in measuring the performance of diagnostic tests. Many direct and indirect approaches have been proposed for modelling the ROC curve, and because of its tractability, the Gaussian distribution has typically been used to model both populations. We propose using a Gaussian mixture model, leading to a more flexible approach that better accounts for atypical data. Monte Carlo simulation is used to circumvent the issue of absence of a closed-form. We show that our method performs favourably when compared to the crude binormal curve and to the semi-parametric frequentist binormal ROC using the famous LABROC procedure.

Keywords: Binormal curve; EM algorithm; Gaussian mixture; LABROC; mixture models; Monte Carlo; ROC curve.

1 Introduction

The receiver operating characteristic (ROC) curve has gained tremendous popularity since its use in the signal detection theory during World War II. This phenomenon can be justified by the necessity to evaluate the performance of a diagnostic test, as noted by Lusted (1971). Despite being a useful tool to evaluate the efficiency of a diagnostic test, the ROC curve also presents a practical way to select an optimal threshold and to compare different tests. However, the empirical ROC curve is not desirable for the simple reason that it violates certain theoretical properties. Many authors have proposed different ways to model the ROC curve to circumvent this issue. Approaches to modelling the ROC curve within the literature can be divided into two categories: direct and indirect.

The direct approach, which is less appealing, does not depend on any distributional hypotheses. The idea is to construct the ROC curve directly from the population scores, often in medical setting, are divided into two groups: diseased and non-diseased, without any

*Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada, N1G 2W1.
E-mail: acheam@uoguelph.ca

assumptions (Lloyd, 1998; Zhou and Harezlak, 2002). As mentioned previously, the empirical ROC curve violates certain theoretical properties; e.g., it is not necessarily monotonic increasing. To overcome this obstacle, some authors proposed non-parametric estimation of the density function of each population using kernel smoothing methods (Hall and Hyndman, 2003; Lloyd, 1998; Lopez-de Ullibarri et al., 2008; Qiu and Le, 2001; Zou et al., 1997). Hence, the problem is reduced to selection of an optimal bandwidth (Lloyd, 1998; Peng and Zhou, 2004; Zhou and Harezlak, 2002). Lloyd (1998) suggested using the bootstrap to minimize any distortion when smoothing the ROC curve.

The indirect approach assumes that each population follows a certain distribution and implicitly derives a functional form for the ROC curve. To construct a curve, parametric and semi-parametric methods have been proposed. One of the parametric methods assumes that diseased and non-diseased populations follow a family of distributions such as: the Gaussian, which is the obvious and simple choice; the gamma (Dorfman et al., 1997); and others (Zweig and Campbell, 1993). For the Gaussian assumption, Goddard and Hinberg (1990) pointed out it is not always an adequate choice in some scenarios like prostate cancer. The authors emphasized that an inconsiderate and careless application of the method is not recommended, because it depends strongly on distributional assumptions. Furthermore, Zhou et al. (2002) stressed the need to carefully verify the consistency of data with the assumptions. An alternative to the previous method is to specify a functional form of the ROC curve instead of assuming a distribution. For instance, both populations can be assumed to follow a logistic distribution with the same variance (Swets, 1986). England (1988) suggested an exponential model with two parameters. Both parametric methods are very similar because the distribution of the test scores entirely determines the shape of the ROC curve. The main advantages of a parametric method are simplicity, the smoothness of the curve, and an ability to work with a small number of parameters.

The semi-parametric method is more attractive in terms of flexibility due to the presence of non-parametric and parametric components. The binormal model (Green and Swets, 1966) is a good example; it assumes that both populations follow a Gaussian distribution after some monotone increasing transformation (Hanley, 1996). Hence, the problem is reduced to estimating the parameters, i.e., the slope and intercept. A range of solutions has been proposed using different techniques such as generalized least squares (Hsieh and Turnbull, 1996), maximum likelihood, pseudo-likelihood (Cai and Moskowitz, 2004; Zhou and Lin, 2008; Zou and Hall, 2000), and others. For example, to obtain a smooth binormal ROC curve, 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 (1968) for ordinal data. A variation of this method was suggested by Li et al. (1999), where they model the scores of a diagnostic test for non-diseased and diseased patients non-parametrically and parametrically, respectively. On the other hand, no functional relationship is assumed between these two distributions. Instead of directly modelling the distributions of the diagnostic scores of the two populations when the true status of the disease is known, another approach is to model the probability of knowing

the disease status of the diagnostic scores using logistic regression (Qin and Zhang, 2003). Evidently, like any estimation problem, the lack-of-fit can be an issue for the semi-parametric method. In addition to this estimation problem, the construction of confidence bands, for a given choice of both population distributions, is complicated.

Our motivation is to develop a method that can give an estimate of the ROC curve with more flexibility and smoothness, produce reliable confidence bands, and ensure the natural monotonicity property of the ROC curve. We propose a Gaussian mixture (GM) distribution to model both non-diseased and diseased populations. This enables us to capture more complex behaviour and distribution shapes than the traditional normality assumption. By combining Monte Carlo simulation and the GM distribution, our method generates an ensemble of replica ROC curves and computes summary measures, such as the area under curve (AUC) and the partial AUC (pAUC), based on the ensemble.

The remainder of the paper is organized as follows. In Section 2, we provide some background on ROC curves, followed by details of our proposed approach (Section 3). Results from simulation studies are provided in Section 4 and real data analyses are discussed in Section 5. In Section 6, some concluding remarks and possible extensions are discussed.

2 Background

The ROC curve is defined to be a plot of the true positive rate (TPR) against the false positive rate (FPR), or sensitivity versus $(1 - \text{specificity})$, for various threshold values. This is generally a curve in the unit square anchored at $(0, 0)$ and $(1, 1)$, and above the line joining those points. Let $X \sim F$ and $Y \sim G$ be two independent continuous variables or two diagnostic variables coming from two populations, non-diseased and diseased, respectively. By convention, a patient is considered diseased if the value of the score is greater than a specified threshold. Note that we borrow the notation of Gu et al. (2008) in some of what follows. For a given threshold value $c_t \in \mathbb{R}$,

$$\text{FP}(c_t) = \int_{-\infty}^{+\infty} f_X(x)I(x - c_t) dx = P(X > c_t), \quad (1)$$

$$\text{TP}(c_t) = \int_{-\infty}^{+\infty} g_Y(y)I(y - c_t) dy = P(Y > c_t), \quad (2)$$

where

$$I(u) = \begin{cases} 1, & \text{if } u > 0, \\ 0, & \text{if } u \leq 0. \end{cases}$$

Therefore, the ROC curve is obtained by

$$\{(t, R(t))\} = \{(\text{FP}(c_t), \text{TP}(c_t))\}, \quad (3)$$

where $t \in D \subset [0, 1]$.

When t is given, $c_t = \bar{F}^{-1}(t) = F^{-1}(1 - t)$, where $F^{-1}(\zeta) = \inf\{x : F(x) \geq \zeta\}$. If $\bar{F}^{-1}(t)$ exists, then the functional form of the ROC curve is given by

$$R(t) = TP(c_t) = \bar{G}(\bar{F}^{-1}(t)) = \bar{G}(c_t) = P(Y > c_t) = P(Y > \bar{F}^{-1}(t)), \quad (4)$$

where $\bar{F}(u) = P(X > u)$ and $\bar{G}(u) = P(Y > u)$ are known as survival functions of X and Y , respectively.

The AUC is an extensively used summary index to quantify the information given by an ROC curve. The AUC, A , and its estimate \hat{A} are defined as

$$A = \int_0^1 R(t)dt \quad \text{and} \quad \hat{A} = \int_0^1 \hat{R}(t)dt, \quad (5)$$

respectively, where $\hat{R}(t)$ is an estimate of $R(t)$.

3 Methodology

We refer to our approach, where a Gaussian mixture is used in conjunction with Monte Carlo simulation, as the MG method. The purpose of using the MG method is to produce a valid curve estimate and reliable confidence bands for any ROC curve. Using Gaussian mixtures leads to a more flexible model that accounts for data that might be considered atypical. Monte Carlo simulation is applied to circumvent the issue of the absence of a closed-form. Its properties enable the computation of confidence bands with ease. Because each pair (X, Y) constructs one ROC curve, the idea of our MG method is to generate an ensemble of replica ROC curves by simulating many pairs (X, Y) , where F and G are assumed to be Gaussian mixture densities. Accordingly, we have

$$f(\mathbf{x} | \Theta) = \sum_{k=1}^K \pi_k \phi(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (6)$$

where

$$\phi(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_k|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \quad (7)$$

is the k th Gaussian component density with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, $\pi_k > 0$ with $\sum_{k=1}^K \pi_k = 1$ are the mixing proportions, and $\Theta = (\pi_1, \dots, \pi_K, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K)$ is the collection of all model parameters. The density $g(\mathbf{y} | \Psi)$ is defined similarly, where Ψ is the collection of all model parameters.

The major difference between our parameter estimation approach and that of Gu et al. (2008) consists in the fact they use a Bayesian approach whereas we do not. They propose the Dirichlet process prior and then perform a bootstrap to resample. Like Gu et al. (2008), we can lay out each step of our MG method, which combines Gaussian mixture modelling and Monte Carlo simulation.

Step 1 (Parameter estimation for F and G): Let X and Y be the vectors of scores of non-diseased and diseased populations, respectively. Suppose both X and Y follow Gaussian mixture densities as in (6). Parameter estimation is carried out via an expectation-maximization (EM) algorithm Dempster et al. (1977) and we thereby obtain Θ_X and Θ_Y .

Step 2 (Generating the ensemble of random ROC curves): After obtaining the parameter estimates, we generate (\tilde{X}, \tilde{Y}) where $\tilde{X} \sim F(\Theta_X)$ and $\tilde{Y} \sim G(\Theta_Y)$. With the simulated ensemble, we compute $\{(t, \tilde{R}(t))\}$ and \tilde{A} . This gives only one ROC curve, and after repeating this step M times, via Monte Carlo simulation, we obtain

$$\{(t, \tilde{R}_1(t))\}, \dots, \{(t, \tilde{R}_M(t))\} \quad \text{and} \quad \tilde{A}_1, \dots, \tilde{A}_M.$$

Step 3 (Averaging the ensemble of random ROC curves): The MG estimate, denoted as $\hat{R}^{MG}(t)$, is obtained by averaging the random realizations of the ROC curves such that

$$\hat{R}^{MG}(t) = \text{mean}(\tilde{R}(t)),$$

where $t \in D \subset [0, 1]$. Similarly, we compute

$$\hat{A}^{MG} = \int_0^1 \hat{R}^{MG}(t) dt.$$

Remark 1 : The estimate $\hat{R}^{MG}(t)$ is much smoother than the empirical estimate because it is obtained by averaging over the ensemble of random realizations $\tilde{R}(t)$.

Remark 2 : When plotting the curves, it is useful to add bands indicating the region where 95% of the curves $\tilde{R}(t)$ lie. Therefore, to compute the confidence bands of the MG estimators of the ROC curves, we use the M $\{(t, \tilde{R}(t))\}$ in Step 2. By the fundamental theory of Monte Carlo, which is based on the strong law of large numbers and the central limit theorem, we compute the MG standard error of $\hat{R}(t)$. For a given t ,

$$s_t = \sqrt{\frac{1}{M-1} \sum_{l=1}^M (\tilde{R}_l(t) - \hat{R}^{MG}(t))^2}. \quad (8)$$

Hence, the upper and lower bounds of the $100(1 - \alpha)\%$ confidence interval can be written

$$\left[\hat{R}_{LB}^{MG}(t), \hat{R}_{UB}^{MG}(t) \right] = \left[\hat{R}^{MG}(t) - z_{1-\alpha} \frac{s_t}{\sqrt{M}}, \hat{R}^{MG}(t) + z_{1-\alpha} \frac{s_t}{\sqrt{M}} \right]. \quad (9)$$

4 Simulation Studies

In this section, we conduct several simulation studies to investigate the flexibility and the fit of the proposed MG approach by comparing it with existing procedures, i.e., the binormal

model and the semi-parametric frequentist binormal ROC using the LABROC4 software (Metz, 1990). We assume that $X \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$. The parameters are chosen randomly. To evaluate the accuracy of MG method, we compute the AUC using the trapezoidal rule and the Mann-Whitney U test.

To visualize the flexibility of the MG method, three cases of discrimination are examined: strong, moderate, and poor. From Figure ??, we observe that the diseased population is a bimodal distribution. This scenario is practically relevant because the diseased population may contain a subpopulation of patients at different stages of a disease. Our goal is to replicate the empirical curve but with more smoothness. When strong discrimination is present (Figure 2), we can observe that the MG curve performs as well as the commonly used LABROC method and significantly better than the crude binormal curve. Furthermore, for the two other cases, we notice that the MG curve follows the empirical curve closely when compared to the binormal and the LABROC approaches (Figures 3–6). The bands (dashed lines) in these graphs indicate the region covering 95% of our simulated curves; recall that the MG curve represents an averaging of these simulated curves. The crude binormal curve is obtained directly from the following equation without any monotonic transformation:

$$R(t) = \Phi(a + b\Phi^{-1}(t)), \quad (10)$$

where

$$a = \frac{\mu_D - \mu_N}{\sigma_D} \quad \text{and} \quad b = \frac{\sigma_N}{\sigma_D}.$$

It follows that the AUC of a binormal curve is given by

$$A = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right). \quad (11)$$

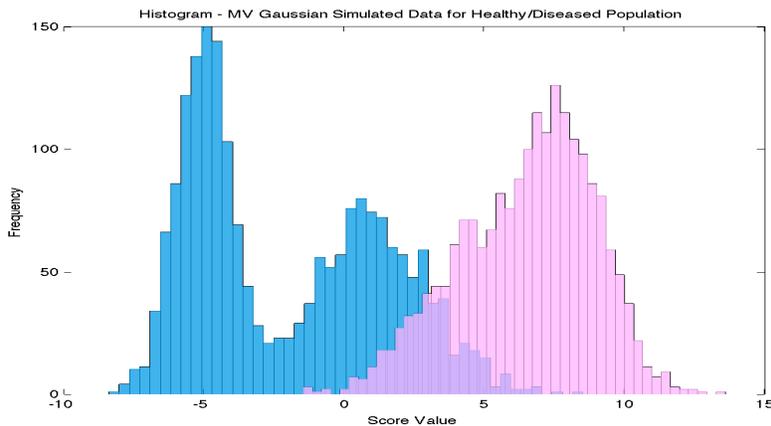


Figure 1: Histogram of the simulated multivariate Gaussian data with strong discrimination.

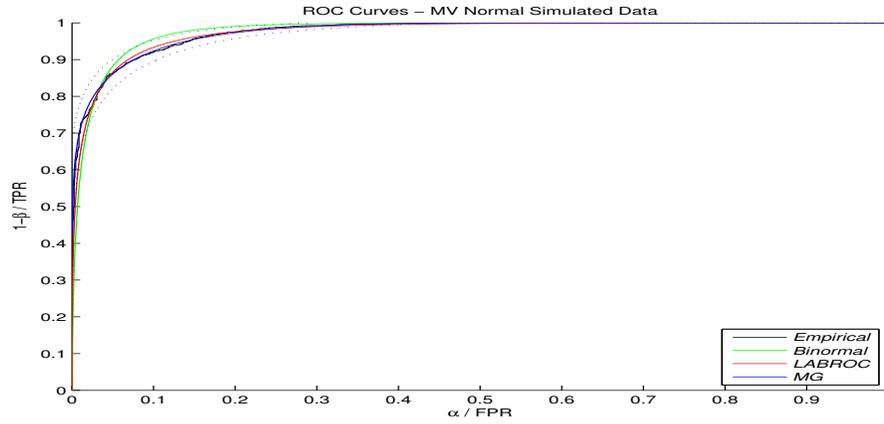


Figure 2: ROC curve for the simulated multivariate Gaussian data with strong discrimination (Figure ??).

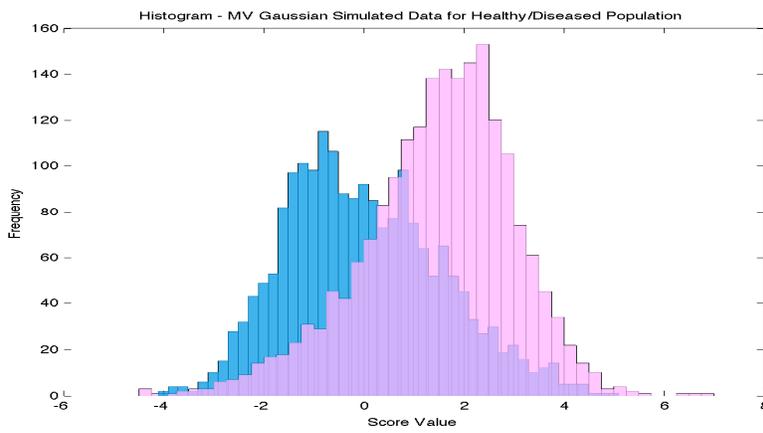


Figure 3: Histogram of the simulated multivariate Gaussian data with moderate discrimination.

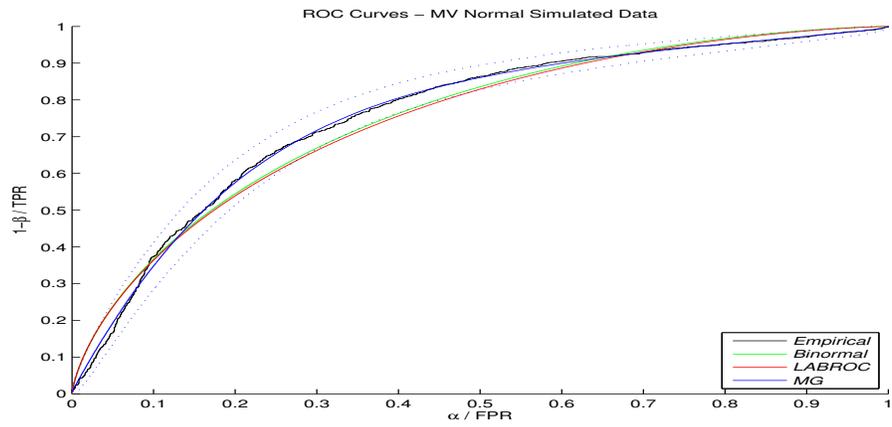


Figure 4: ROC curve for the simulated multivariate Gaussian data with moderate discrimination (Figure 3).

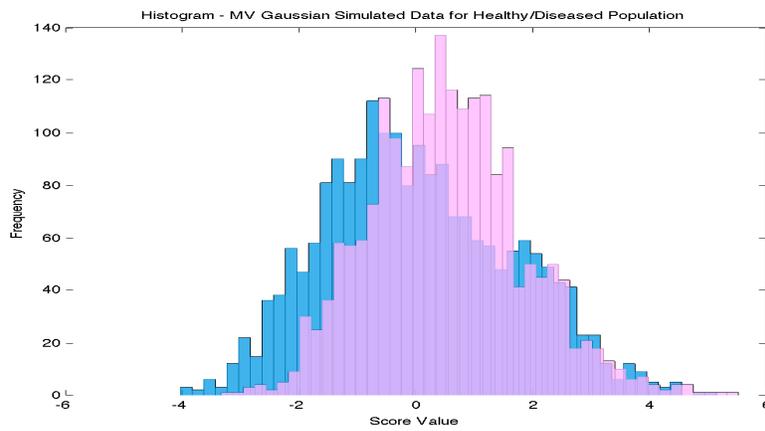


Figure 5: Histogram of the simulated multivariate Gaussian data with poor discrimination.

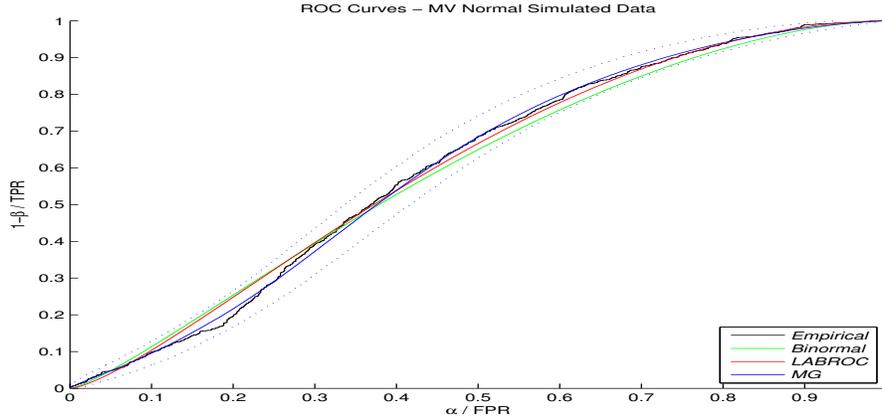


Figure 6: ROC curve for the simulated multivariate Gaussian data with poor discrimination (Figure 5).

Another way to compare the MG method against the two binormal methods is to calculate a summary measure such as the AUC (Table 1). We observe that our MG method obtained an AUC close to the empirical AUC and performed relatively well compared to the binormal and LABROC procedures. These simulation studies show that our approach is at least as accurate as the two classical binormal methods and sometimes superior.

Table 1: AUC values for the simulated multivariate Gaussian data, with the model closest to the empirical curve in bold font.

Level of Discrimination	ROC Model	AUC	
		Trapezoidal	Mann-Whitney
Strong	Empirical	0.9760	0.9759
	Binormal	0.9787	0.9787*
	LABROC	0.9750	0.8337
	MG	0.9764	0.9757
Moderate	Empirical	0.7599	0.7597
	Binormal	0.7521	0.7521*
	LABROC	0.7479	0.7163
	MG	0.7593	0.7588
Poor	Empirical	0.6008	0.6006
	Binormal	0.5550	0.5950*
	LABROC	0.6024	0.6053
	MG	0.6013	0.6011

*Using the binormal crude AUC equation

5 Real Data Analyses

Having observed favourable results on the simulation studies, we illustrate our newly proposed method on publicly available case-control cancer data published by Wieand et al. (1989). This data set has been used extensively in the literature to illustrate newly developed methods for ROC curves; accordingly, we selected it to compare our method to the current state-of-the-art. This study examined two biomarkers: a cancer antigen (CA 125) and a carbohydrate antigen (CA 19-9). The data consist of 90 selected cases representing patients with pancreatic cancer as well as 51 controls without cancer but with pancreatitis. We used the two biomarkers to illustrate the application of our methodology. From Figures 8 and 10, we observe that our MG method undoubtedly outperforms the crude binormal approach. The poor performance of the binormal curve can be explained by the unsuitability of the normality assumption for these data (see Figures 7 and 9). Compared to the LABROC procedure, our MG method performs relatively well in terms of replication and closeness to the empirical curve. Without any monotonic transformation, the MG method outperforms the crude binormal ROC and performs as well as LABROC. As suspected, the AUC of our approach is closer to the empirical than the binormal for both biomarkers (Table 2). For CA 125, our MG method obtains a summary index closer to the empirical curve than the LABROC. For biomarker CA 19-9, LABROC gives a better AUC using the trapezoidal rule; however, when using the Wilcoxon-Mann-Whitney statistic, our MG method performs better.

Table 2: AUC values for pancreatic cancer data using two biomarkers, with the model closest to the empirical curve in bold font.

Biomarker	ROC Model	AUC	
		Trapezoidal	Mann-Whitney
Pancreas CA 125	Empirical	0.7143	0.7056
	Binormal	0.5924	0.5924*
	LABROC	0.6946	0.6808
	MG	0.7147	0.7143
Pancreas CA 19-9	Empirical	0.8651	0.8614
	Binormal	0.6774	0.6776*
	LABROC	0.8625	0.7793
	MG	0.8569	0.8565

*Using the binormal crude AUC equation

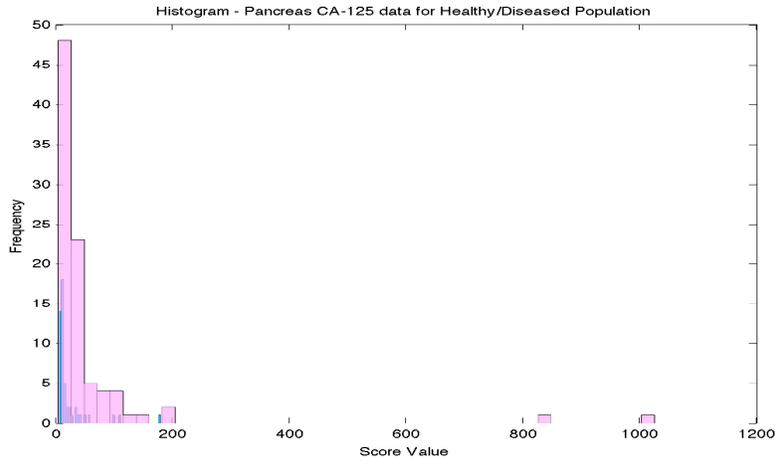


Figure 7: Histogram of the pancreatic cancer data using biomarker CA 125.

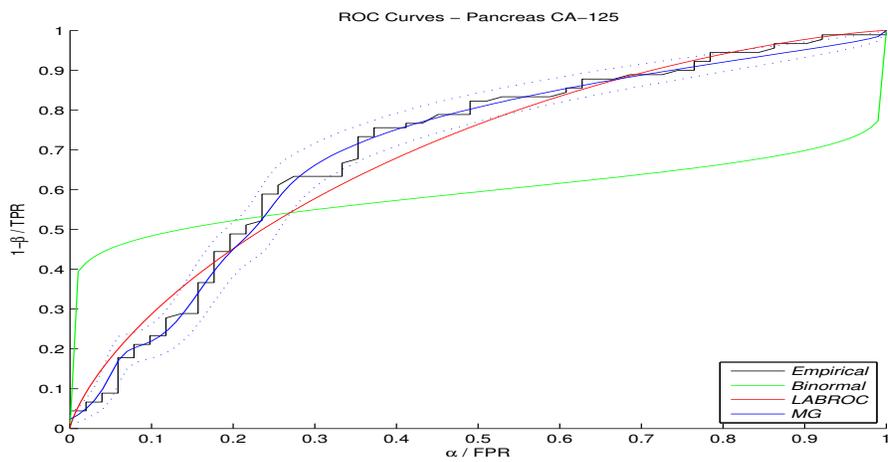


Figure 8: ROC curve for the pancreatic cancer data using biomarker CA 125.

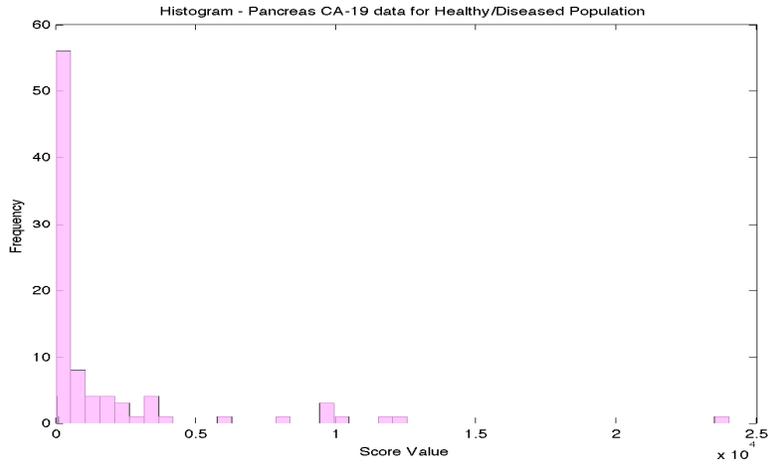


Figure 9: Histogram of the pancreatic cancer data using biomarker CA 19-9.

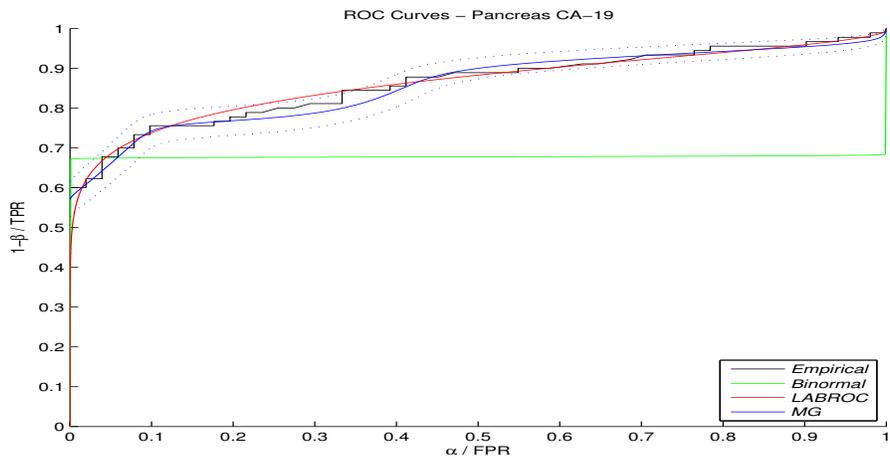


Figure 10: ROC curve for the pancreatic cancer data using biomarker CA 19-9.

6 Concluding Remarks

In this paper, we have outlined a methodology to estimate the ROC curve with more flexibility and smoothness than provided by existing approaches. The proposed method utilizes the Gaussian mixture in conjunction with Monte Carlo simulation, and we refer to our approach as the MG method. The performance of the MG method was illustrated via several simulation studies and real data on pancreatic cancer. We found that our MG curve performed favourably when compared to the crude binormal curve in term of flexibility and fitting. Even without any monotonic transformation, the MG produced similar, if not better, results than the LABROC procedure. Furthermore, the MG method does not require any assumptions other than that the populations follow a Gaussian mixture. An interesting avenue for future work is to extend this approach using a non-Gaussian mixture model instead of a Gaussian mixture model.

References

- Cai, T. and C. S. Moskowitz (2004). Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics* 5, 573–586.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39, 1–38.
- Dorfman, D. D. and E. J. Alf (1968). Maximum likelihood estimation of parameters of signal detection theory - a direct solution. *Psychometrika* 33, 117–124.
- Dorfman, D. D., K. S. Berbaum, C. E. Metz, R. V. Lenth, J. A. Hanley, and H. Dogga (1997). Proper receiver operating characteristic analysis: the bigamma model. *Academic Radiology* 4, 138–149.
- England, W. L. (1988). An exponential model used for optimal threshold selection on ROC curves. *Medical Decision Making* 8, 120–131.
- Goddard, M. J. and I. Hinberg (1990). Receiver operating characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine* 9, 325–337.
- Green, D. M. and J. Swets (1966). *Signal Detection Theory and Psychophysics*. New York: John Wiley and Sons.
- Gu, J., S. Ghosal, and A. Roy (2008). Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine* 27, 5407–5420.
- Hall, P. G. and R. J. Hyndman (2003). Improved methods for bandwidth selection when estimating ROC curves. *Statistics & Probability Letters* 64, 181–189.

- Hanley, J. A. (1996). The use of binormal model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine* 15, 1575–1585.
- Hsieh, F. and B. W. Turnbull (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics* 24, 25–40.
- Li, G., R. C. Tiwari, and M. T. Wells (1999). Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika* 86, 487–502.
- Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* 93, 1356–1364.
- Lopez-de Ullibarri, I., R. Cao, C. Cadarso-Suarez, and M. J. Lado (2008). Non-parametric estimation of conditional ROC curves: application to discrimination tasks in computerized detection of early breast cancer. *Computational Statistics & Data Analysis* 52, 2623–2631.
- Lusted, L. B. (1971). Signal detectability and medical decision-making. *Science* 171, 1217–1219.
- Metz, C. E. (November 1990). Metz roc software. <http://metz-roc.uchicago.edu/MetzROC/software>. Accessed July 2013.
- Metz, C. E., B. A. Herman, and J.-H. Shen (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine* 17, 1033–1053.
- Peng, L. and X.-H. Zhou (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference* 118, 129–143.
- Qin, J. and B. Zhang (2003). Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika* 90, 585–596.
- Qiu, P. and C. Le (2001). ROC curve estimation based on local smoothing. *Journal of Statistical Computation and Simulation* 70, 55–69.
- Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin* 99, 100–117.
- Wieand, S., M. H. Gail, B. R. James, and K. L. James (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 76, 585–592.
- Zhou, X.-H. and J. Harezlak (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine* 21, 2045–2055.

- Zhou, X.-H. and H. Lin (2008). Semi-parametric maximum likelihood estimates for ROC curves of continuous-scale tests. *Statistics in Medicine* 27, 5271–5290.
- Zhou, X.-H., N. A. Obuchowski, and D. K. McClish (2002). *Statistical Methods in Diagnostic Medicine*. New York: Wiley.
- Zou, K. H. and W. J. Hall (2000). Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* 27, 621–631.
- Zou, K. H., W. J. Hall, and D. E. Shapiro (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostics tests. *Statistics in Medicine* 16, 2143–2156.
- Zweig, M. H. and G. Campbell (1993). Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 39, 561–577.