Memorial Sloan-Kettering Cancer Center
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2013 *Paper* 27

Mixtures of Receiver Operating Characteristic Curves

Mithat Gonen*

*Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org

http://biostats.bepress.com/mskccbiostat/paper27

Mixtures of Receiver Operating Characteristic Curves

Mithat Gonen

Abstract

Rationale and Objectives: ROC curves are ubiquitous in the analysis of imaging metrics as markers of both diagnosis and prognosis. While empirical estimation of ROC curves remains the most popular method, there are several reasons to consider smooth estimates based on a parametric model.

Materials and Methods: A mixture model is considered for modeling the distribution of the marker in the diseased population motivated by the biological observation that there is more heterogeneity in the diseased population than there is in the normal one. It is shown that this model results in an analytically tractable ROC curve which is itself a mixture of ROC curves.

Results: The use of CK-BB isoenzyme in diagnosis of severe head trauma is used as an example. ROC curves are fit using the direct binormal method, ROCKIT and the Box-Cox transformation as well as the proposed mixture model. The mixture model generates an ROC curve that is much closer to the empirical one than the other methods considered.

Conclusions: Mixtures of ROC curves can be helpful in fitting smooth ROC curves in datasets where the diseased population has higher variability than can be explained by a single distribution.

Mixtures of Receiver Operating Characteristic Curves

ABSTRACT

Rationale and Objectives: ROC curves are ubiquitous in the analysis of imaging metrics as markers of both diagnosis and prognosis. While empirical estimation of ROC curves remains the most popular method, there are several reasons to consider smooth estimates based on a parametric model.

Materials and Methods: A mixture model is considered for modeling the distribution of the marker in the diseased population motivated by the biological observation that there is more heterogeneity in the diseased population than there is in the normal one. It is shown that this model results in an analytically tractable ROC curve which is itself a mixture of ROC curves.

Results: The use of CK-BB isoenzyme in diagnosis of severe head trauma is used as an example. ROC curves are fit using the direct binormal method, ROCKIT and the Box-Cox transformation as well as the proposed mixture model. The mixture model generates an ROC curve that is much closer to the empirical one than the other methods considered.

Conclusions: Mixtures of ROC curves can be helpful in fitting smooth ROC curves in datasets where the diseased population has higher variability than can be explained by a single distribution.

Keywords: Finite mixture, E-M algorithm, ROC curves

INTRODUCTION

Receiver operating characteristic (ROC) curves have long become the standard way to describe the diagnostic accuracy of imaging methodologies. Initial applications of ROC curves in radiology focused on ordinal imaging metrics mostly based on reader evaluations (1-4). As quantitative imaging markers became more widely available, ROC curves were adapted in their evaluation (5-7). The use of ROC curves in evaluating predictive and prognostic models is quickly becoming standard as well (8-10). Several recent articles illustrate the recent methodological developments and the widening reach of ROC curves (11-13).

Given the depth and breadth of the applications of ROC curves in several fields it is no surprise that methodology for estimating ROC curves has proliferated. Many standard techniques are well-covered in several books that are largely or entirely devoted to the use of ROC curves (14-17). In addition software is available for major statistical packages like SAS (14), R (18) and Stata (16).

An ROC curve can be thought of as a measure of how far apart the distributions in the diseased and non-diseased populations are. In many applications the distribution of test results in the diseased population is more heterogenous than it is in the normal population. This can be explained by the many different ways that the disease manifests itself in the human body and the attempts by the test to capture this diversity. While normal biological processes also display surprisingly large variability, a good test would not focus on capturing the heterogeneity that does not result in disease. This creates a problem for the parametric ROC methods that have largely relied on using the same distribution for both diseased and non-diseased populations. The most popular of these parametric models, namely the binormal ROC curve (19), often falls short in capturing the heterogeneity of the diseased population. This has been long recognized in the ROC field. The most common approach is either an explicit or an implicit transformation to binormality. Examples of the former include the Box-Cox family of transformations (20, 21). Implicit transformations have been a focus of Metz's work, see for example (22). Despite these important advances on the use of transformations in ROC curves, the need to find (or presume the existence of) a single transformation that works for both the diseased and normal groups remains the achilles heel of this approach.

This article will propose an alternative that explicitly recognizes the heterogeneity in the diseased population and at the same time allows the use of familiar parametric models. It is made possible by an observation of the functional form of ROC curves and the use of mixture models. Mixture models have a rich history in statistics but it is likely that most radiologists will not be familiar with them. For this reason the next section includes a brief introduction to the common definitions and uses of mixture models.

MATERIALS AND METHODS

Mixture Models

To avoid any terminological confusion it is important to note that the kind of mixture models that are referred to here are the ones called "finite mixtures" in the statistical literature.

Perhaps the easiest way to motivate the ideas behind the use of mixture models is by example. The data set investigating the use of cerebrospinal fluid CK-BB isoenzyme in diagnosing severe head trauma, first published by Hans et al (23) and re-analyzed in detail (17) will be used as an example throughout this article. CK-BB was measured within 24 hours of admission to emergency care with a diagnosis of head trauma in a sample of 60 patients, 41 of whom did not recover. Death, vegetative state and or severe disability is considered lack of recovery (called poor outcome here) in this analysis. Although this is not a radiological study, it is among the most well-known and well-analyzed data sets in the field of continuous ROC curves and chosen for this purpose. All the issues discussed in this article applies equally well to imaging biomarkers such as the standardized uptake value from a positron emission tomography or Hounsfield units from computed tomography.

Figure 1 is a histogram of the CK-BB levels among the poor outcome group. The distribution is clearly non-normal with substantial right skew. Not surprisingly, a normal fit to this data (dashed line in Figure 1) is quite poor, missing the mode and substantial portion of the tails. There are many remedies to this and the one that will be explored here is to fit a mixture of two normal densities (24). Instead of assuming that Y, the random variable representing the CK-BB values in the poor outcome group, has a normal distribution it will instead be assumed that it is a mixture of two distributions, that is with probability λ, an observation from one component of the mixture ϕ_1 is obtained and with probability 1-λ, an observation from the other component, ϕ_2, is observed. The mathematical formulation for such a mixture density is given by

$$g(y) = \lambda\phi_1(y) + (1-\lambda)\phi_2(y)$$

where λ is known as the mixing proportion and ϕ_1 and ϕ_2 are known as the component densities; both normal in this case, with possibly different means and variances, i.e. ϕ_j has mean μ_j and standard deviation σ_j. It is possible to represent a variety of probability distributions with this formulation due to the flexibility afforded by the component densities as well as the mixing proportion.

It is, of course, natural to ask whether a mixture with more than two components can be considered. Mathematically this is not a problem: one can write

$$g(y) = \sum_{j=1}^{p} \lambda_j \phi_j (y)$$

where *p* is the number of components of the mixture. Estimation of p-component mixtures follow the same principles as estimation of two-component mixtures. In particular it is common to use maximum likelihood to estimate the parameters of this mixture density (25).

Mixture ROC Curves

First consider the formulation of an ROC curve in terms of the survivor functions of the diseased (*G*) and non-diseased (G_0) populations

$$ROC(t) = G\left(G_0^{-1}(t)\right)$$

where 0<t<1 and the survivor function is one minus the corresponding cumulative distribution:

$$G(x) = \int_{x}^{\infty} g(y)dy$$

with g(.) denoting the density function. If g(.) is a p-component mixture then so is G(.):

$$G(x) = \sum_{j=1}^{p} \lambda_j \mathrm{G}_j (x)$$

In this formulation G_j(.) are the survivor functions of the components of the mixture. This article focuses on mixture of normals, but there is no restriction on the choice of parametric family for G_j other than the practical consideration of being able to estimate them.

Now let G_0 be the survivor function of the observations in the non-diseased group. Only a normal form will be considered here but as above there is no restriction on the choice of G_0. The p-mixture ROC curve, denoted by $ROC^{(p)}$ is then defined in terms of the component survivor functions as

$$ROC^{(p)}(t) = \sum_{j=1}^{p} \lambda_j \mathrm{G}_j \left(\mathrm{G}_0^{-1}(\mathrm{t})\right)$$

Note that $G_j\left(\mathrm{G}_0^{-1}(\mathrm{t})\right)$ satisfies the definition of an ROC curve where the diseased and the non-diseased populations have survivor functions G_j and G_0. Therefore $ROC^{(p)}$ is indeed a mixture of

p ROC curves. In the case where it is assumed that all of the p+1 densities (including G_0) are normal, one can use the traditional parameterization where G_j has mean μ_j and standard deviation σ_j (j=0,1,…,p). Define the binormal parameters for j>0

$$a_j = \frac{\mu_j - \mu_0}{\sigma_j}$$

and

$$b_j = \frac{\sigma_0}{\sigma_j}$$

One can write

$$G_j\left(\mathrm{G}_0^{-1}(\mathrm{t})\right) = \Phi\left(\mathrm{a_j} + \mathrm{b_j}\Phi^{-1}(\mathrm{t})\right)$$

and the p-mixture ROC curve can be represented as

$$ROC^{(p)}(t) = \sum_{j=1}^{p} \lambda_j \Phi\left(\mathrm{a_j} + \mathrm{b_j}\Phi^{-1}(\mathrm{t})\right)$$

The biologic interpretation of this follows from the one given in the previous section: if the diseased population is heterogenous and made up of *p* different sub-populations with distinct distributions for the marker, then the component ROC curve given by $G_j\left(\mathrm{G}_0^{-1}(\mathrm{x})\right)$is the ROC curve evaluating the marker's ability to discriminate a non-diseased observation from one that is from the j^{th} diseased subpopulation. In practice these component ROC curves are not very useful since sub-population labels are not observed. In fact it is not even clear these sub-populations exist; all that can be said is that the data are consistent with a model that postulates these sub-populations.

Estimation of a p-mixture ROC curve is very simple: estimate G_0 using the non-diseased sample only and estimate $G_1,\ldots,G_p$ from the diseased population using maximum likelihood, as above (see appendix for computational issues). Then substitute these estimates to get an estimate of $ROC^{(p)}$.

Summary Measures of the Mixture ROC Curve

While plotting the ROC curve is traditionally the most popular and most informative presentation there are several reasons why one might want to summarize the entire curve by one or few carefully chosen summary indices. Commonly used summary measures are area under the curve (AUC), partial area under the curve (pAUC) and sensitivity at chosen thresholds or specificities. Among these AUC is arguably the most popular since it does not require

subjective choices like pAUC (area of interest) and sensitivity (choice of the specificities or thresholds at which it will be estimated) In addition, it admits a concordance interpretation which is useful in practice.

The AUC for a binormal model has the following closed form expression in terms of the parameters *a* and *b*:

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

Since AUC is an integral and the integral is a linear operator, the AUC of the mixture model turns out to be a linear combination of the component AUCs:

$$AUC^{(p)} = \sum_{j=1}^{p} \lambda_j \, AUC_j$$

Since each component is a binormal ROC curve AUC_j can be substituted with AUC of the binormal model resulting in

$$AUC^{(p)} = \sum_{j=1}^{p} \lambda_j \, AUC_j = \sum_{j=1}^{p} \lambda_j \, \Phi\left(\frac{a_j}{\sqrt{1+b_j^2}}\right)$$

Therefore estimation of the AUC for the p-mixture ROC curve requires only the estimation of the AUCs of its component ROC curves.

In principle, a similar expression holds for the pAUC:

$$pAUC^{(p)} = \sum_{j=1}^{p} \lambda_j \, pAUC_j$$

There is no closed form expression for the pAUC of a binormal curve (26), however, although it can be written as a function of the cumulative density function of the standard bivariate normal (27). It follows that there is no closed form expression for the pAUC of the mixture ROC curve either, but this is not a burden: an estimate of $pAUC^{(p)}$ can be obtained by numerically integrating the component ROC curves or using the bivariate formulation (27).

Estimates of sensitivities at given thresholds are very simple to obtain with the mixture ROC curves since $ROC^{(p)}(t)$ can be expressed in closed form. For a p-mixture ROC curve, an estimate of the sensitivity when the specificity is t_0 is given by

$$ROC^{(p)}(t) = \sum_{j=1}^{p} \lambda_j \mathrm{G}_j \left(\mathrm{G}_0^{-1}(\mathrm{t})\right)\Big|_{t=t_0}$$

and when the threshold is c is given by

$$ROC^{(p)}(t) = \sum_{j=1}^{p} \lambda_j \mathrm{G}_j \left(\mathrm{G}_0^{-1}(\mathrm{t})\right)\Big|_{t=G_0(c)}$$

RESULTS

Figure 1 is a histogram of the CK-BB levels among the poor outcome group. Skewness is estimated to be 1.42 (compared with 0 for normal distribution) and kurtosis, a measure of how thick the tails of the distribution are, is estimated to be 1.45 (again compared with 0 for normal). Not surprisingly, a normal fit to this data (dashed line in Figure 1) is quite poor, missing the mode and substantial portion of the tails. It is clear that a normal distribution is not a good candidate to model this data.

Table 1 presents parameter estimates for the two- and three-component mixture models along with the empirical estimates and those of the normal fit. Both the normal fit and mixtures of normal recover the mean and standard deviation successfully but normal fit is poor when it comes to skewness and kurtosis while the mixture model is much better, although still an underestimate when compared with the empirical estimates (see Appendix A for calculation of the moments for a mixture distribution).

In terms of choosing between the two- and three-component mixtures, Figure 1 provides some visual clues. The solid line is the two-component mixture and the dotted line is the three-component mixture. Visually the fit is an improvement, however slight it might be, especially in the right tail. Estimated mixing proportions suggest a small third component (10%) when compared with the other two (47% and 43%) but estimates of skewness and kurtosis are very close to the empirical ones. It may, thus, appear that a three-component mixture is the best fit. Unfortunately this informal process of matching moments and visually judging the goodness-of-fit is prone to over-fitting, i.e. choosing a mixture with more components than necessary. Most readers will be familiar with the type of over-fitting that involves adding independent variables

to a regression equation, which always results in an increase in R^2. The over-fitting encountered in mixture models is another instance of the same phenomenon: when more parameters are used to describe a given data set, a better fit is inevitably obtained. While we have several tools to minimize over-fitting in regression models (adjusted R^2, cross-validation, shrinkage estimation etc.) corresponding methods in mixture models are either lacking or not in common use. It is possible, however, to compare a *p*-component mixture to a *(p+1)*-component mixture using a likelihood ratio test. This turns out to be much more complicated than using a likelihood ratio test to compare two regression models due to technical issues that has to do with the null hypothesis being on the boundary of the parameter space (28). For the purposes of this article it is sufficient to note that this test is implemented in R (see Appendix for details). Performing this test suggests that a two-component mixture is a significant improvement over the single normal fit (p=0.01) but a three-component normal is not superior to a two-component normal (p=0.32) suggesting that the best model among the ones considered is the two-component mixture. A great deal of care must be exercised with this approach since these p-values are sensitive to sample size and it is quite possible that the true model has three components but our sample size did not afford us enough power to identify that. Instead of claiming to have found the right number of components it is more appropriate to speak in terms of the best model consistent with the data at hand.

Figure 2 presents ROC curves estimated and plotted using various methods for the head trauma example. Obvious competitors to the p-mixture ROC curve are the empirical ROC, direct binormal, ROCKIT (22, 29) and Box-Cox (20, 21). Direct binormal method is not recommended unless the underlying distribution is binormal (30), but included here for reference. Empirical ROC curve is obtained by plotting the pairs of sensitivity and one minus specificity fo reach observed threshold and connecting them with step functions. Direct binormal ROC curve results from using the estimates of the means and standard deviations from the observed data in the binormal ROC curve form. ROCKIT binormal curve is derived from the “truth runs” approach developed by Metz and his colleagues at the University of Chicago (22). Finally, the Box-Cox ROC curve estimates the Box-Cox power parameter that simultaneously transforms both the diseased and non-diseased observations to normality. The latter two are plotted here using the estimates reported in Chapter 4 of Zhou, Obuchowski and McClish (17).

Figure 2 is interpreted using the empirical ROC curve as the reference. The direct binormal fit is clearly a poor estimate, first overestimating and then underestimating the empirical curve with an irksome “hook” at high sensitivities where the estimated curve falls below the 45-degree line. While this hook is at the very-high sensitivity area which is not very relevant for applications (31), it is nevertheless an indication of model inadequacy. In fact such hooks are common with direct binormal fits and such ROC curves are called improper (32) since they do not satisfy the monotone likelihood ratio principle (33, 34). ROCKIT fit does the opposite of the

direct binormal fit: first underestimates, then overshoots the empirical curve, avoiding the hook in the process but offering minimal improvement over the direct binormal curve. The Box-Cox ROC curve is very similar to the ROCKIT one and similarly unsatisfactory. In contrast the 2-mixture ROC curve provides the best visual fit. While there is some overestimation for high-specificity and underestimation for high-sensitivity regions, it tracks the empirical ROC curve closely.

Table 2 presents the estimates of AUC and the sensitivity at three different operating points for the five ROC curves plotted in Figure 2. Poor fit of the direct binormal curve is once again evident in the underestimation of AUC. ROCKIT and Box-Cox estimates are very close to the empirical values. This may be surprising at first but Figure 2 makes it obvious that with an approximately equal amount of over- and under-estimation the errors cancel each other. This should be taken as less of a vote of confidence for these two methods and more of a shortcoming of AUC as a summary measure when the curves cross. In fact the poor performance of all the three binormal methods (direct, ROCKIT and Box-Cox) is evident from the estimates of sensitivity for three different values of specificity. The mixture ROC curve estimates of sensitivity are much closer to those obtained by empirical calculations.

DISCUSSION

Empirical estimates of the ROC curve are popular for two primary reasons: they are easy to compute and they are based on minimal assumptions. The latter reason begets a major drawback, namely that empirical ROC curves tend to be more variable than their parametric and semi-parametric counterparts (35). The flat regions of the empirical curve, which can be substantive in portions of the sample space not covered well by the sample at hand, create a transportability problem of estimates: they can differ greatly from one study to the other even if one study is a replicate of the other.

For these reasons a smooth estimate should be a good alternative to the empirical curves. Most common smoothing methods are parametric ROC curves, kernel-based ROC curves and semi-parametric ROC curves (36, 37). Parametric ROC curves are well-investigated, easy to use but highly sensitive to departures from the underlying assumptions. Kernel-based methods have slow rates of convergence, thus reliable only in large samples (38). Semi-parametric methods try to mediate between these two. While they could strike middle ground successfully, their computational complexity has limited widespread use.

This article proposes another parametric technique to produce smooth ROC curves based on the idea of mixture modeling of heterogenous populations. The fundamental argument is that diseased populations tend to be more diverse than non-diseased ones and might benefit from mixture modeling. The formulation of the ROC curve in terms of the survivor functions lends

itself to easy manipulation when the diseased population is represented by a finite mixture. It is shown here that under these conditions the ROC curve itself is a mixture of ROC curves each representing the discrimination of a particular component of the diseased population from the non-diseased population. Estimation of mixture ROC curves is greatly facilitated by the strides in fitting mixture models using the E-M algorithm (39); see appendix for various computing strategies in R and SAS.

It is possible that the non-diseased population is also heterogenous to the point of requiring a mixture model, however this fails to translate to a straightforward mixture of ROC curves. The reason has to do with the fact that the contribution of the non-diseased group to the entire curve is mediated through the inverse of their survivor function and the inverse of the mixture of survivor functions is not the same as the mixture of their inverse. This does not mean heterogeneity in the non-diseased population should be ignored; it just means that mixture modeling will not translate into a pragmatic solution as it did for the diseased population.

In the trauma example considered here a mixture model produced estimates closer to the empirical curve than several binormal methods: direct estimation, ROCKIT and Box-Cox. Only extensive simulations might shed some light as to whether this performance is generalizable to other scenarios. In the meantime the practitioner interested in mixture ROC curves is best guided by using the empirical curve as standard, as well as careful consideration of the modeling of diseased and non-diseased populations.

BIBLIOGRAPHY

1. Goodenough DJ, Rossmann K, Lusted LB. Radiographic applications of receiver operating characteristic (ROC) curves. Radiology. 1974; 110(1):89-95.
2. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. Radiology. 1975; 116(3):533-8.
3. Metz CE. ROC methodology in radiologic imaging. Investigative radiology. 1986; 21(9):720.
4. Gatsonis C, Begg C, Wieand S. Advances in statistical methods for diagnostic radiology: a symposium. Academic radiology. 1995; 2(suppl 1):S1-S84.
5. Wong R, Lin D, Schöder H, et al. Diagnostic and prognostic value of [18F] fluorodeoxyglucose positron emission tomography for recurrent head and neck squamous cell carcinoma. Journal of clinical oncology. 2002; 20(20):4199-208.
6. Yuan Y, Giger ML, Li H, Bhooshan N, Sennett CA. Multimodality computer-aided breast cancer diagnosis with FFDM and DCE-MRI. Academic radiology. 2010; 17(9):1158.
7. Tsushima Y, Takano A, Taketomi-Takahashi A, Endo K. Body Diffusion-Weighted MR Imaging Using High< i> b</i>-Value for Malignant Tumor Screening: Usefulness and Necessity of Referring to T2-Weighted Images and Creating Fusion Images. Academic radiology. 2007; 14(6):643-50.
8. Güttler A. Using a bootstrap approach to rate the raters. Financial Markets and Portfolio Management. 2005; 19(3):277-95.
9. Mailier P, Jollife I, Stephenson D. Quality of weather forecasts. Review and recommendations Royal Meteorological Society. 2006:1-89.
10. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010; 21(1):128.
11. Alemayehu D, Zou KH. Applications of ROC Analysis in Medical Research: Recent Developments and Future Directions. Academic radiology. 2012; 19(12):1457-64.
12. Skaron A, Li K, Zhou XH. Statistical Methods for MRMC ROC Studies. Academic radiology. 2012; 19(12):1499-507.
13. Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC Studies with Split-plot Designs: A Comparison of Statistical Methods. Academic radiology. 2012; 19(12):1508-17.
14. Gönen M. Analyzing receiver operating characteristic curves with sas: Sas Inst; 2007.
15. Krzanowski WJ. ROC curves for continuous data: Chapman & Hall/CRC; 2009.
16. Pepe MS. The statistical evaluation of medical tests for classification and prediction: Oxford University Press, USA; 2004.
17. Zhou XH, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine: Wiley-Interscience; 2009.
18. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011; 12(1):77.
19. Hillis SL. Simulation of Unequal-Variance Binormal Multireader ROC Decision Data: An Extension of the Roe and Metz Simulation Model. Academic radiology. 2012; 19(12):1518-28.
20. Zou KH, Tempany CM, Fielding JR, Silverman SG. Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. Academic radiology. 1998; 5(10):680-7.
21. Zou KH, Hall W. Two transformation models for estimating an ROC curve derived from continuous data. Journal of Applied Statistics. 2000; 27(5):621-31.
22. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic(ROC) curves from continuously-distributed data. Statistics in medicine. 1998; 17(9):1033-53.

23. Hans P, Albert A, Born J, Chapelle JP. Derivation of a bioclinical prognostic index in severe head injury. Intensive care medicine. 1985; 11(4):186-91.
24. McLachlan G, Peel D. Finite mixture models: Wiley-Interscience; 2000.
25. McLachlan GJ, Krishnan T. The EM algorithm and extensions: Wiley-Interscience; 2007.
26. McClish DK. Analyzing a portion of the ROC curve. Medical Decision Making. 1989; 9(3):190-5.
27. Hillis SL, Metz CE. An Analytic Expression for the Binormal Partial Area under the ROC Curve. Academic radiology. 2012; 19(12):1491-8.
28. McLachlan GJ. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. Applied Statistics. 1987:318-24.
29. Metz C. ROCKIT 0.9 B. Beta Version Department of Radiology, University of Chicago. 1998.
30. Goddard MJ, Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. Statistics in medicine. 1990; 9(3):325-37.
31. Hanley JA. The robustness of the" binormal" assumptions used in fitting ROC curves. Medical Decision Making. 1988; 8(3):197-203.
32. Metz CE, Kronman HB. Statistical significance tests for binormal ROC curves. Journal of Mathematical Psychology. 1980; 22(3):218-43.
33. Metz CE, Pan X. “Proper” binormal ROC curves: Theory and maximum-likelihood estimation. Journal of mathematical psychology. 1999; 43(1):1-33.
34. Pesce LL, Metz CE, Berbaum KS. On the convexity of ROC curves estimated from radiological test results. Academic radiology. 2010; 17(8):960-8. e4.
35. Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. Journal of the American Statistical Association. 1998; 93(444):1356-64.
36. Cai T. Semi-parametric ROC regression analysis with placement values. Biostatistics. 2004; 5(1):45-60.
37. Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. Biometrics. 2000; 56(2):352-9.
38. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Statistics in medicine. 1997; 16(19):2143-56.
39. Benaglia T, Chauveau D, Hunter DR, Young DS. mixtools: An R package for analyzing finite mixture models. Journal of Statistical Software. 2009; 32(6):1-29.
40. Frühwirth-Schnatter S. Finite Mixture and Markov Switching Models: Springer Science + Business Media, LLC; 2006.

APPENDICES

Appendix A

Formulae for the moments of mixture models (40):

$$\text{Mean}=\mu=\sum_{j=1}^{p}\lambda_j\mu_j$$

$$\text{Variance}=\sigma^2=\sum_{j=1}^{p}\lambda_j\left({\sigma_j}^2+{\mu_j}^2\right)-\mu^2$$

$$\text{Skewness}=\frac{1}{\sigma^3}\sum_{j=1}^{p}\lambda_j(\mu_j-\mu)\left[3\sigma_j^2+(\mu_j-\mu)^2\right]$$

$$\text{Kurtosis}=\frac{1}{\sigma^4}\sum_{j=1}^{p}\lambda_j\left[3\sigma_j^4+6\sigma_j^2(\mu_j-\mu)^2+(\mu_j-\mu)^4\right]-3$$

Appendix B

The R package `mixtools` provides a convenient and powerful way to fit various mixture models(39). The primary function for our purposes is `normalmixEM` which fits normal mixtures. For example, the following code is used to obtain the twp-component normal mixture for modeling the poor outcome group in the head trauma example:

```
normalmixEM(poor,lambda=0.5,mu=c(200,700),sigma=c(10,50))
```

Here `poor` is the vector of CK-BB values for the 41 patients with poor outcome. The rest of the inputs are initial guesses for the five parameters. This function infers the desired dimension of mixtures (*p*) from the length of the vectors `mu` and `sigma`.

The likelihood ratio test for comparing mixture models is implemented in `boot.comp` which generates the null distribution using bootstrap. In this case the relevant function call is

```
boot.comp(poor,mix.type="normalmix")
```

The output of `normalmixEM` contains everything that is required to work with the mixture ROC curves: estimated values for the component densities as well as the mixing proportions.

Some utility functions to evaluate and plot the mixture ROC curves are available from the author.

It is also possible to estimate finite mixture models using SAS. This is most simply done by using PROC FMM, a recently introduced functionality in version 9.3. For those who do not have access to version 9.3, a custom likelihood construction can be accommodated in PROC NLMIXED. Details of both approaches are available from the author.

FIGURE CAPTIONS

Figure 1: Histogram of the distribution of the CK-BB values for the poor outcome group overlaid with three different parametric density estimates: normal, 2-mixture and 3-mixture.

Figure 2: ROC curves for the head trauma example estimated using various parametric and non-parametric methods.

Table 1: Estimates of the Moments of the Diseased Distribution

	Mean	Standard Deviation	Skewness	Kurtosis
Empirical	427	373	1.42	1.44
Binormal	427	373	0	0
2-Mixture	427	368	1.07	0.60
3-Mixture	427	368	1.41	1.50

Table 2: Estimates of the AUC and the sensitivity at three different operating points for the five ROC curves plotted in Figure 2

	AUC	Sensitivity (specificity=0.60)	Sensitivity (specificity=0.75)	Sensitivity (specificity=0.90)
Empirical	0.828	0.818	0.683	0.561
Direct Binormal	0.790	0.777	0.745	0.693
ROCKIT	0.831	0.908	0.750	0.382
Box-Cox	0.831	0.887	0.742	0.427
Mixture	0.815	0.803	0.717	0.598