LogitROC: An R Package for Making Inferences on ROC Curves and Surfaces Using Nonparametric and Semiparametric Approaches

Dong Zhang [*] Biao Zhang [†] Kuo-Hao Lee [‡]

Abstract
LogitROC is an R package for simulation, visualization and estimation of receiver operating characteristic (ROC) curves and surfaces using nonparametric and semiparametric approaches. Empirical likelihood estimators and semiparametric estimators considering the density ratio model of ROC curves and surfaces, and comparison between correlated ROC curves are generated. Plots of estimated ROC curves or surfaces, and estimated distribution functions are produced for different parameter values. In addition, empirical likelihood and semiparametric confidence intervals are proposed. Especially, on the Windows system, there is an interactive feature for the user to find the specific sensitivity and specificity by clicking the ROC curves or surfaces.

Key Words: Density ratio model, Empirical likelihood, Receiver operating characteristic (ROC), Area under the ROC curve (AUC)

1. Introduction

Receiver operating characteristic (ROC) curves are commonly used in clinical trials to describe and compare the performance of diagnostic technology and diagnostic algorithms. The accuracy of a diagnostic test with binary test results is typically summarized by its sensitivity (SE) and specificity (SP) defined, respectively, as the probability that a truly diseased subject has a positive test result and the probability that a truly non-diseased subject has a negative test result. When the test results are ordinal or continuous, the methodology of the ROC curve is widely used for assessing the accuracy of a continuous or ordinal-valued diagnostic test; the ROC curve is defined as a plot of SE against one minus SP across all possible choices of threshold values. For a diagnostic test with multi-level results, the ROC surface can be constructed. Similar to the ROC curve, the ROC surface is constructed based on the distribution functions of test results from each disease class.

Logistic regression models are commonly used in analyzing binary data which arise in studying relationships between diseases and environment or genetic characteristics; see for example Breslow and Day (1980), Prentice and Pyke (1979) and Farewell (1979). In the set-up of a generalized linear model, Pregibon (1980) considered examining the adequacy of the hypothesized link for a given prospective sampling dataset. A nonparametric regression method was proposed by Azzalini et al. (1989) to test the validity of the logistic regression assumption. Some graphical methods for assessing logistic regression models were studied by Landwehr, Landwehr et al. (1984). Let D denote the class indicator, and x be the associated $p \times 1$ covariate vector. Suppose the diagnostic test results can be classified as two classes, 0 and 1, denoting nondiseased and diseased groups respectively. Let $f_0(t)$ and $f_1(t)$ denote density functions of the two classes. The standard logistic regression

[*] Department of Mathematics, Computer Science and Statistics, Bloomsburg University, 400 E 2nd Street, Bloomsburg, PA, 17815

[†] Department of Mathematics and Statistics, University of Toledo, OH, 43606

[‡] Department of Finance, Bloomsburg University, 400 E 2nd Street, Bloomsburg, PA, 17815

model is

$$\begin{aligned} P(D=1|x) &= \frac{\exp\{\alpha^* + \beta^T r(x)\}}{1+\exp\{\alpha^* + \beta^T r(x)\}}, \\ P(D=0|x) &= \frac{1}{1+\exp\{\alpha^* + \beta^T r(x)\}}, \end{aligned} \tag{1}$$

where α^* is a scalar, β is a $p \times 1$ vector, $r(x)$ is a p-dimensional vector function, such as $r(x) = x$ or $r(x) = (x, x^2)^T$.

The density ratio model discussed in Qin and Zhang (1997) describes the relationship among the two density functions as

$$f_1(t) = f_0(t) \exp\{\alpha + \beta^T r(t)\}, \tag{2}$$

where $\alpha = \alpha^* - \log \rho$, $n_1/n_0 \to \rho$, n_d is the sample size of class d, $d = 0, 1$. The most important advantage of density ratio models is that the semiparametric inference for the ROC curve is more efficient than the nonparametric inference. Moreover, the estimated ROC curves are smoother under density ratio models.

This paper describes an R package for statistical analysis of ROC curves for diagnostic tests with binary classification and ROC surfaces for diagnostic tests with a three-level classification. Empirical likelihood estimation and semiparametric estimation under density ratio models are proposed. Also, confidence intervals using both nonparametric and semiparametric approaches are constructed for comparison between two correlated ROC curves or surfaces.

2. Nonparametric and semiparametric estimation

For nonparametric estimation of the ROC curve or surface, we employ empirical distribution functions for the nondiseased and diseased populations to construct empirical estimators of the ROC curve or surface. For a diagnostic test with binary results, let F_0 and F_1 denote, respectively, the distribution functions of the nondiseased and diseased groups, and let f_0 and f_1 denote the corresponding density functions. Furthermore, let $X_1, \ldots, X_{n_0} \overset{iid}{\sim} F_0$ and $Y_1, \ldots, Y_{n_1} \overset{iid}{\sim} F_1$. The empirical distribution functions of the nondiseased and diseased groups are given by $\hat{F}_0(t) = 1/n_0 \sum_{i=1}^{n_0} I(X_i \le t)$ and $\hat{F}_1(t) = 1/n_1 \sum_{j=1}^{n_1} I(Y_j \le t)$. Consequently, the empirical estimator of the ROC curve is $\widehat{\mathrm{ROC}}(q) = 1 - \hat{F}_1(\hat{F}_0(1-q))$ for $q \in [0, 1]$.

In the semiparametric approach under density ratio models, we need to find maximum likelihood estimators of α, β, and $F_0(t)$, and then we construct the semiparametric estimator of the ROC curve. Let $\{T_1, \ldots, T_n\} = \{X_1, \ldots, X_{n_0}, Y_1, \ldots, Y_{n_1}\}$ be the pooled sample and $n_1/n_0 \to \rho$. The semiparametric estimators of F_0 and F_1 are given by

$$\tilde{F}_0(t) = \frac{1}{n_0} \sum_{l=1}^{n} \frac{I(T_l \le t)}{1 + \rho \exp\{\tilde{\alpha} + \tilde{\beta}^T r(t)\}}, \tilde{F}_1(t) = \frac{1}{n_0} \sum_{l=1}^{n} \frac{\exp\{\tilde{\alpha} + \tilde{\beta}^T r(t)\} I(T_l \le t)}{1 + \rho \exp\{\tilde{\alpha} + \tilde{\beta}^T r(t)\}}$$

so that the semiparametric estimator of the ROC curve is $\widetilde{\mathrm{ROC}}(q) = 1 - \tilde{F}_1(\tilde{F}_0(1-q))$. Details can be found in Zhang and Zhang (2014) and Zhang and Zhang (2014).

3. The LogitROC package

3.1 Making inference on a single diagnostic test

There are two functions in this package named as *LogitROC* and *LogitROC3d* for estimating ROC curve and surface. LogitROC is the function used to estimate and make inferences for a single ROC curve and corresponding area under the ROC curve (AUC), while LogitROC3d is used for a single ROC surface with a three-level diagnostic test. Both of these functions can generate point estimators, variances, and plots of estimated distribution functions and estimated ROC curves or surfaces. The structure of parameters required in those two functions are similar.

The function *LogitROC* can be used by passing values of the corresponding parameters as follows.

```
logitROC(dataset, plotmethod, interactive, r_function, ROCplot,
Distplot,QQplot)
```

- *dataset* is a two-column matrix. The first column contains test results from the non-diseased group, and the second column contains test results from the diseased group.
- *plotmethod* shows the style of plots.
  - "p" for points,
  - "l" for lines,
  - "b" for both,
  - "c" for the lines part alone of "b",
  - "o" for both over-plotted,
  - "h" for histogram like (or high-density) vertical lines,
  - "s" for stair steps (default value),
  - "S" for other steps,
  - "n" for no plotting.

  All other types give a warning or an error; using, e.g., type = "punkte" as the equivalent to type ="p" for S compatibility. Note that some methods, e.g. plot.factor, do not accept this.
- *interactive* is a logic value to control the feature of interactive operation. This function can allow one to find corresponding sensitivity and specificity by clicking points on ROC plots. This feature is only available on the Window system. Default is activated.
- *r-function* is a list of functions of covariates under a density ratio model. One can pass values in the form of r_function=list(f1,f2,....) The default value is r_function=list(f1=x), i.e. $f(x) = x$.
- *ROCplot, Distplot, QQplot* are all logic values to control whether or not plots of estimated ROC curved and distribution functions and estimated QQ plots (between estimated distribution functions) are produced. The default setup is just to pop-up the estimated ROC plots.

To demonstrate the package, we run the following command in R console. In the example, we use the melanoma data set, which is proposed in Venkatraman and Begg (1996). There are two diagnostic tests, biopsy and dermoscope, to determine the disease stage of melanoma. In the data set named dermoscope, column1 is the test scores of those nondiseased subjects; while column2 contains those scores from diseased patients.

```
>> # Use demo dataset, generate ROC, distributions, qq plots.
>> logitROC(dermoscope, Distplot=TRUE, QQplot=TRUE)
```

This command will produce the following plots and some numerical results in R console. Fig 1 shows the estimated ROC curves usingthe non-parametric and semi-parametric approaches with different colors. Fig 3 and 4 show QQ-plots to compare semiparametrically and nonparametrically estimated distribution functions for the non-diseased and the diseased groups. Fig 5 and 6 show the estimated distribution functions of the non-diseased and diseased groups using the nonparametric and semiparametric approaches. The usage and results of *LogitROC3d* are similar to those of *LogitROC*, but the parameter *data* should contain three columns of test results from three classifications, i.e. the first column contains nondiseased test scores; the second column contains moderate diseased test scores; and the third column contains serious test scores. Following Fig 2 shows the estimated ROC surface of a three-level diagnostic test.

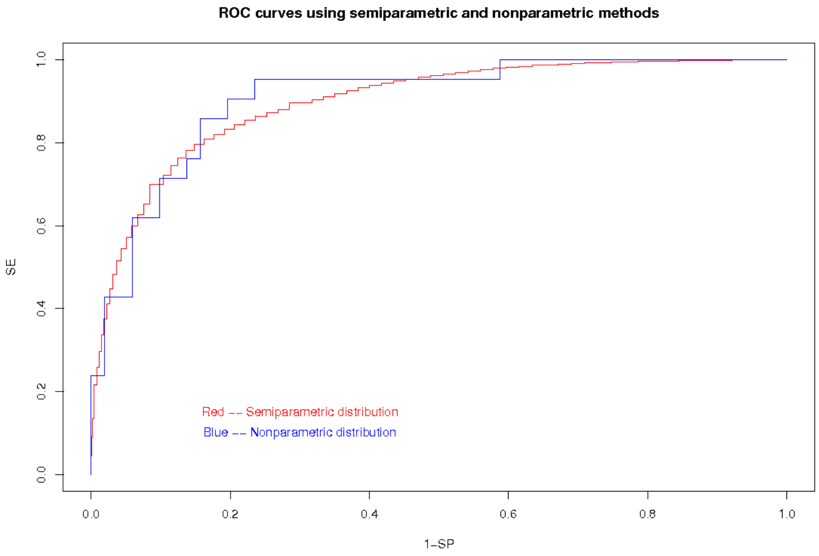


Figure 1: Estimated ROC curve

Numerical results in R console include the maximum semiparametric likelihood estimator of (α, β) under a density ratio model, estimated probabilities for each classification of a diagnostic test, and estimated values of AUCs using both nonparametric and semiparametric approaches. All numerical results are saved in a list. Here is a demonstration.

```
>> A=logitROC(dermoscope, Distplot=TRUE, QQplot=TRUE)
>> # The estimation of parameters in density ratio model
>> A$result
$coefficients
[1] 0.8872473 1.0001523

$stderror
[1] 0.3701204 0.2376495

>> A$NP_AUC
[1] 0.9056956
>> A$SP_AUC
[1] 0.8892279
```

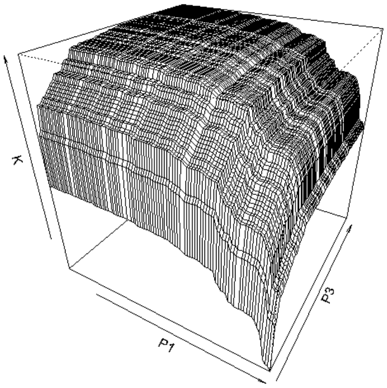


Figure 2: Estimated ROC surface

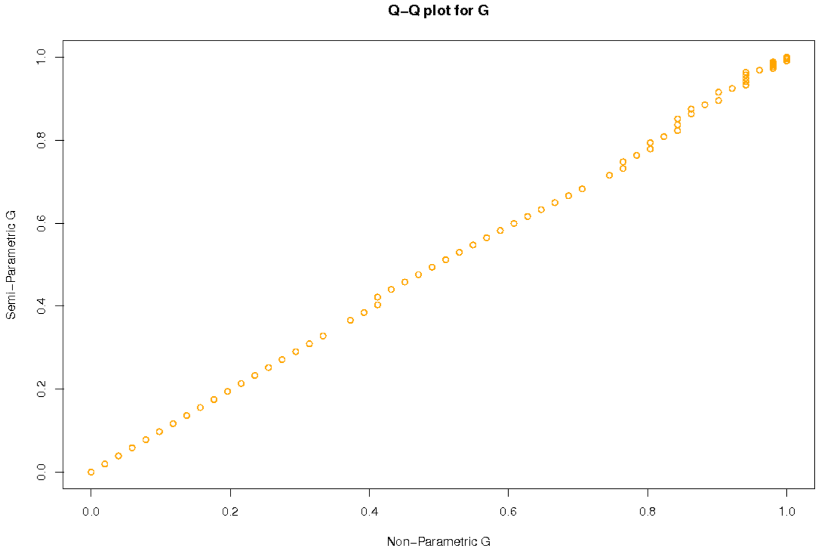


Figure 3: QQ-plots of distribution G

From the above results, the maximum semiparametric likelihood estimator of α and β are 0.8872473 and 1.0001523 with the corresponding standard errors of 0.3701204 and 0.2376495. The nonparametric estimator of AUC is 0.9056956, while the semiparametric estimator of AUC is 0.8892279.

In this package, the parametric ROC curves and surfaces can also be drawn and corresponding AUCs and VUSs can be calculated for normal, exponential and χ^2 distributions. Functions named P_Norm_ROC, P_Exp_ROC and P_Chisq_ROC compute the ROC curves and AUCs based on normal distribution, exponential distribution and χ^2 distribution, respectively. Meanwhile, the functions named P_Norm_ROC3d, P_Exp_ROC3d and P_Chisq_ROC3d are prepared for ROC surfaces and VUSs, which require *rgl* package to be installed. The surface can be rotated using mouse in the graphic device. The usage of these functions are listed below.

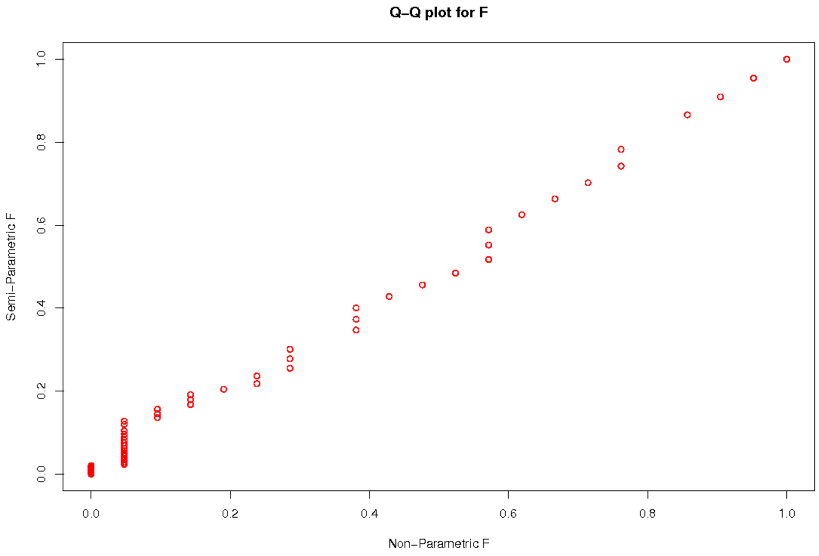


Figure 4: QQ-plots of distribution F

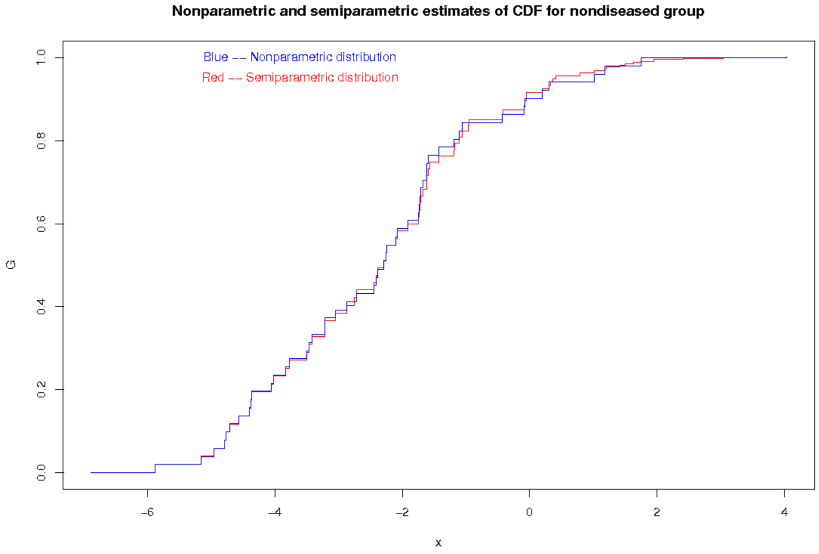


Figure 5: Estimated distribution function G

```
P_Norm_ROC(mu0,mu1,std0,std1,dense=1000)
P_Exp_ROC(lambda1,lambda2, dense=1000)
P_Chisq_ROC(df1,df2,dense=1000)
P_Norm_ROC3d(mu0,mu1,mu2,std0,std1,std2,dense=1000)
P_Exp_ROC3d(lambda0,lambda1,lambda2,dense=1000)
P_Chi_ROC3d(df0,df1,df2,dense=1000)
```

- *mu0,mu1,mu2* are means of normal distributions, and *std0, std1, std2* are standard deviations.

- *lambda0, lambda1, lambda2* are rates of exponential distributions with the form of $\lambda \exp(-\lambda x)$.

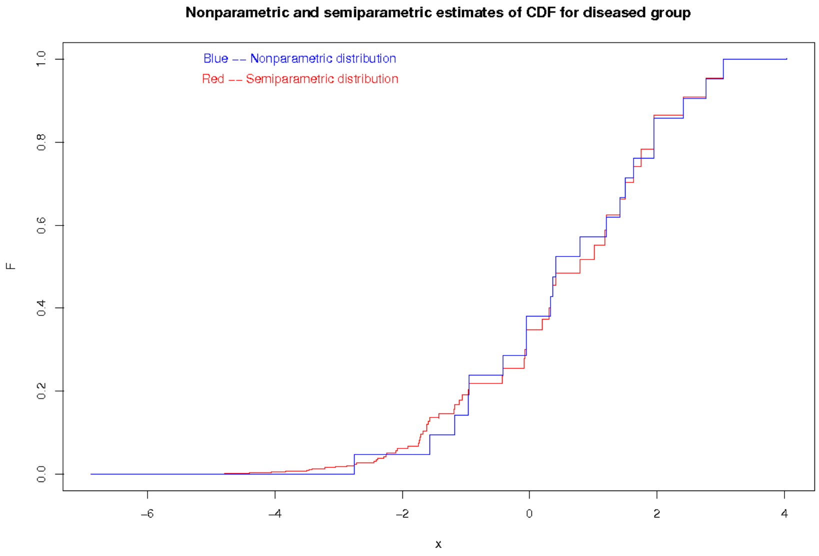
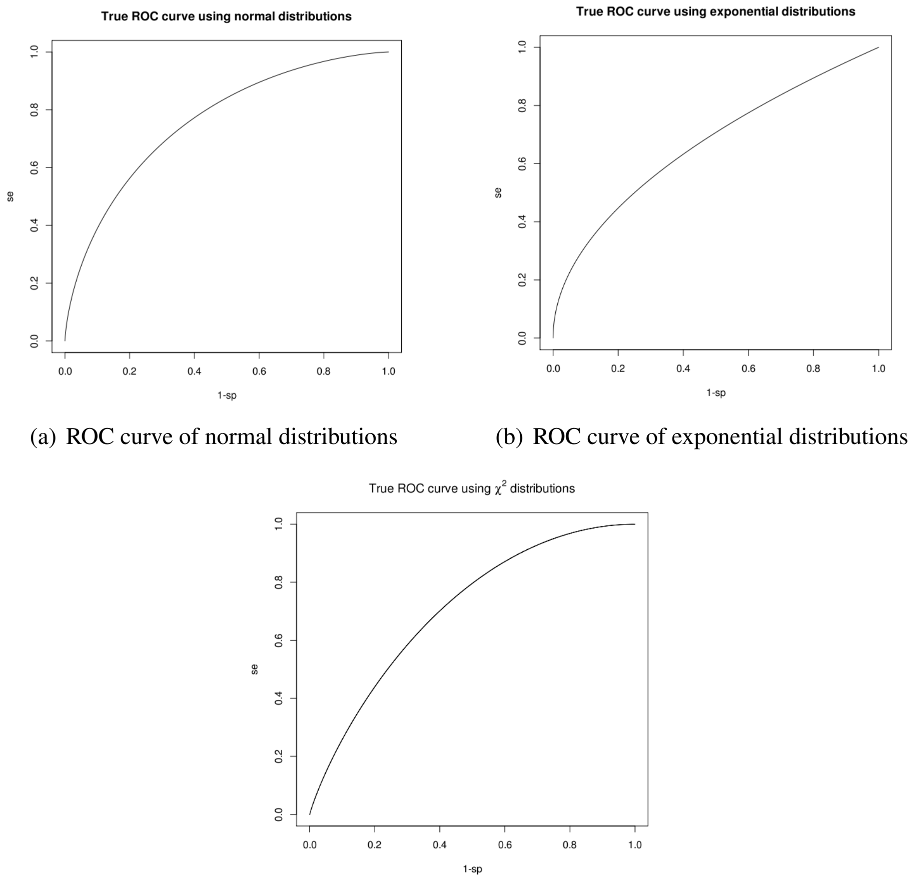


Figure 6: Estimated distribution function F

- *df0,df1,df2* are degree of freedoms of χ^2 distributions.
- *dense* control the accuracy of the plot.

To demonstrate these functions, we can run following examples.

```
>>P_Norm_ROC(0,1,1,1)
$AUC
[1] 0.7602499

>>P_Exp_ROC(2,1)
$AUC
[1] 0.6666667

>>P_Chisq_ROC(1,2)
$AUC
[1] 0.7071068


>>P_Norm_ROC3d(0,1,2,1,1,1)
$VUS
[1] 0.5361515

>>P_Exp_ROC3d(2,1,0.5)
$VUS
[1] 0.3809525

>>P_Chisq_ROC3d(1,2,3)
$VUS
[1] 0.3961623
```



(a) ROC curve of normal distributions

(b) ROC curve of exponential distributions

(c) ROC curve of chi-square distributions

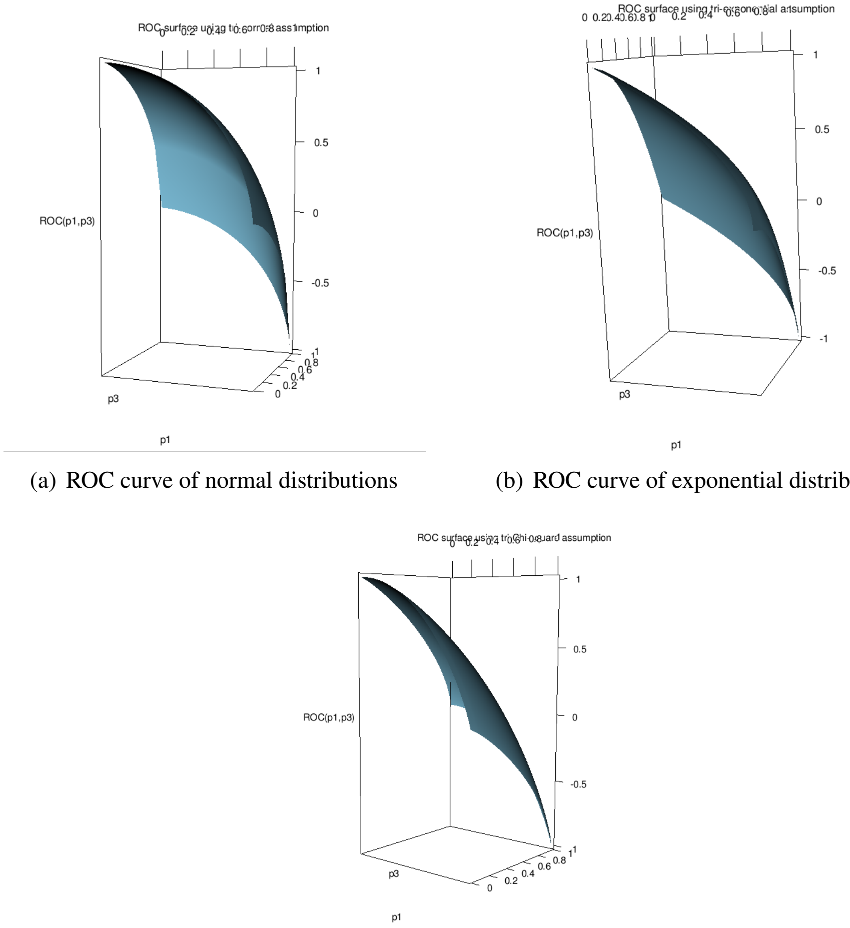
Figure 7: parametric ROC curves

3.2 Comparison between two correlated diagnostic tests

In practice, inferences only on ROC curves and AUCs of a single diagnostic test are not enough. It is a common issue for doctors that they need to choose an alternative diagnostic test for some patients with the consideration of convenience and safety. This issue involves the comparison of two diagnostic tests. There are two function in this package for comparing two correlated diagnostic tests with binary test scores. In this part, we employ the AUC for comparison, which is a summary statistic of a ROC curve. The functions will construct $(1-\alpha)$ confidence interval for the difference between two AUCs. The empirical likelihood approaches are employed to construct confidence intervals. Details can be found in the revised manuscript of Zhang and Zhang (2011) for *Statistics in Medicine*.

The program will give numerical confidence intervals in R console, and the plots used to demonstrate the corresponding empirical likelihood confidence intervals using both nonparametric and semiparametric settings are shown. The function named NP_AUC_Com constructs nonparametric empirical likelihood confidence intervals using two different scalar estimates, and the function named SP_AUC_Com is used to construct semiparametric empirical likelihood confidence intervals under a density ratio model. The syntax of parameters is shown below.

```
NP_AUC_Com(dataX, dataY, level=0.95)
SP_AUC_Com(dataX, dataY, level=0.95)
```



(a) ROC curve of normal distributions

(b) ROC curve of exponential distributions

(c) ROC curve of chi-square distributions

Figure 8: parametric ROC curves

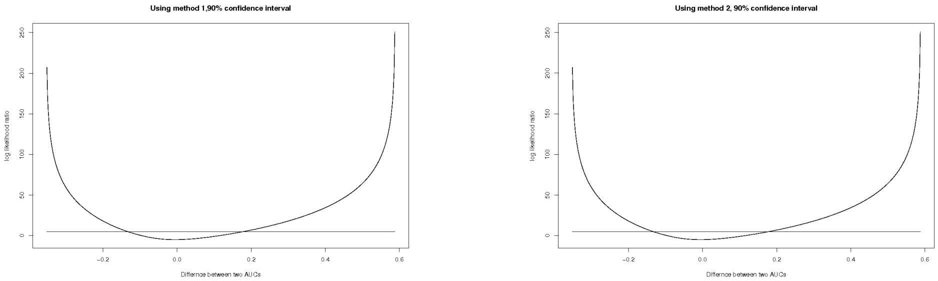
- *dataX* is a two-column matrix containing test results of two correlated diagnostic tests from the nondiseased group.
- *dataY* is similar to dataX, but contains test results from the diseased group. We need to make sure that dataX and dataY have the SAME number of diagnostic test records located in the same column.
- *level* is a vector of confidence levels.

In the example we use both dermoscope test scores and biopsy test scores. The data set named biop_dermo has totally 4 columns. The first two columns are test scores of dermoscope and biopsy, respectively, from nondiseased subjects; while other two columns contains test scores of dermoscope and biopsy, respectively, from diseased subjects The function can be run in R console as follows.

```
>> NP_AUC_Com(biop_dermo[,1:2], biop_dermo[,3:4], level=c(0.9,0.95,0.99))
$scalar1
[1] 0.5470283

$scalar2
[1] 0.5398363
```

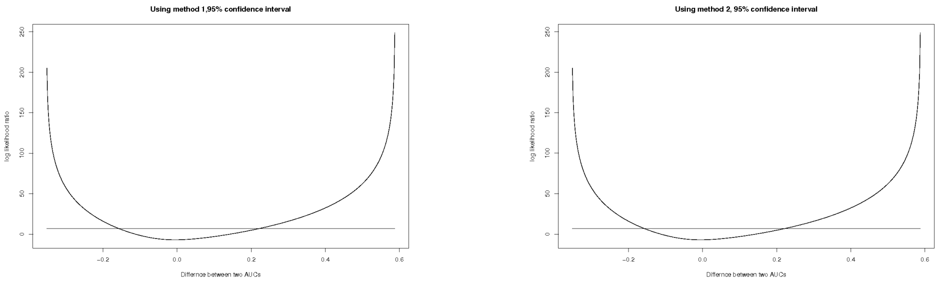
```
$Empirical_CI_r1
          Lower Bound Upper Bound
EL 90% CI -0.09541627   0.1136878
EL 95% CI -0.11272053   0.1425562
EL 99% CI -0.14702807   0.2026727

$Empirical_CI_r2
          Lower Bound Upper Bound
EL 90% CI -0.09601486   0.1146588
EL 95% CI -0.11344110   0.1437833
EL 99% CI -0.14798637   0.2043871
```

Meanwhile, the following plots will be produced and shown.



(a) Using r_1 (b) Using r_2

Figure 9: Plots for solving 90% CI

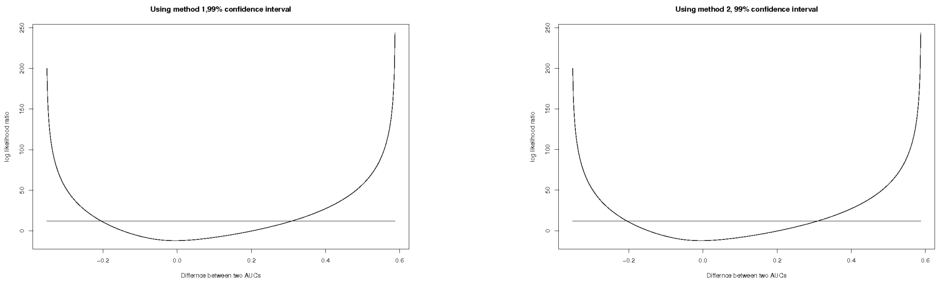


(a) Using r_1 (b) Using r_2

Figure 10: Plots for solving 95% CI

In these figures, the curves represent the log likelihood ratio, and the lines show the locations of the quantiles of a χ_1^2 random variable. The two intersections are limits of the corresponding confidence intervals for the difference between the two AUCs. The results of function SP_AUC_Com are similar, we only show numerical results in the following.

```
>> SP_AUC_Com(biop_dermo[,1:2], biop_dermo[,3:4], level=c(0.9,0.95,0.99))
$scalar1
[1] 0.5398363

$Empirical_CI
          Lower Bound Upper Bound
```



(a) Using r_1

(b) Using r_2

Figure 11: Plots for solving 99% CI

```
EL 90% CI -0.09601486   0.1146588
EL 95% CI -0.11344110   0.1437833
EL 99% CI -0.14798637   0.2043871
```

4. Discussion

In practice, some diagnostic tests are multi-levels, which involve inferences on hyper ROC surfaces and comparison of two ROC surfaces. The package now only has functions for estimating ROC curve for binary output diagnostic test or ROC surface for three-level diagnostic test, and calculating areas under the ROC (AUCs) curve or volumes under ROC surfaces (VUSs). In the future, the comparison between two correlated VUSs will be developed and attached to the package. In addition to the aforementioned nonparametric and semiparametric methods, smoothing methods can also be employed to estimate distribution functions, ROC curves, AUCs, and VUSs.

References

Breslow, N.E., Day, N.E. (1980),*Statistical Methods in Cancer Research*, Vol 1, The Analysis of Case-Control Studies, International Agency for Research on Cancer, Lyon.

Prentice, R.L., Pyke, R. (1979), “Logistic disease incidence models and case-control studies”. *Biometrika*, 66,3,403–411.

Farewell, V. (1979), “Some results on the estimation of logistic models based on retrospective data”. *Biometrika*, 66, 27–32.

Pregibon, D. (1980), “Goodness of link tests for generalized linear models”. *Journal of Applied Statistics*, 29, 15–24.

Azzalini, A., Bowman, A., Hardle, W. (1989), “On the use of nonparametric regression for model checking”. *Biometrika*, 76, 1–11.

Landwehr, J.M., Pregibon, D., Showmaker, A.C. (1984), “Graphical methods for assessing logistic regression models (with Discussion)”. *Journal of the American Statistical Association*, 79, 61–83.

Qin, J., Zhang, B. (1997), “A goodness-of-fit test for logistic regression models based on case-control data”. *Biometrika*, 84, 3, 609–618.

Venkatraman, E.S., Begg, C.B. (1996), "A distribution free procedure for comparing receiver operating characteristic curves from a paired experiment". *Biometrika*, 83, 835–848.

Zhang, D., Zhang, B. (2014), "Empirical Likelihood Confidence Intervals for the Difference of Areas Under Two Correlated ROC Curves". *Journal of Statistical Theory and Practice*, 8, 3, 482–508.

Zhang, D., Zhang, B. (2014), "Semiparametric empirical likelihood confidence intervals for the difference of areas under two correlated ROC curves under density ratio model". *Biometrical Journal*, 56, 4, 678–698.