Communications for Statistical Applications and Methods
2013, Vol. 20, No. 3, 169–174

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.3.169

ROC Curve for Multivariate Random Variables

Chong Sun Hong[1,a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

The ROC curve is drawn with two conditional cumulative distribution functions (or survival functions) of the univariate random variable. In this work, we consider joint cumulative distribution functions of k random variables, and suggest a ROC curve for multivariate random variables. With regard to the values on the line, which passes through two mean vectors of dichotomous states, a joint cumulative distribution function can be regarded as a function of the univariate variable. After this function is modified to satisfy the properties of the cumulative distribution function, a ROC curve might be derived; moreover, some illustrative examples are demonstrated.

Keywords: Classification, default, discrimination, FPR, ROC curve, threshold, TPR.

1. Introduction

The ROC technique, developed in the signal detection theory, is used and studied to assign cases into dichotomous states in many applications. In this study, we consider some applications under the credit evaluation situation. The characteristics of the borrower are supposed to determine in terms of a continuous score random variable X and parameter space Θ. It is assumed that the borrower's population contains two sub-populations $\Theta = \{\theta_d, \theta_n\}$. The sub-populations consist of default(d) and non-default (n), which depend on the repayment ability of loans. The score variable X represents the credit information of the borrower and is used to expect the future state of the borrower.

The ROC curve is plotted with two cumulative distribution functions(CDF), $F(x;\theta_d)$ and $F(x;\theta_n)$, which are the true positive rate(TPR; sensitivity or hit rate) and the false positive rate(FPR; 1 − specificity or false alarm rate), respectively, for all threshold (cutoff point) x. $F(x;\theta_d)$ and $F(x;\theta_n)$ correspond to the Y and X axes of a unit square on a two dimensional plane (see Metz (1978), Zweig and Campbell (1992), Greiner *et al.* (2000), Gardner and Greiner (2006) and Tasche (2006) for further details).

For the ROC curve of a univariate score variable, there exists a paired value $(F(x;\theta_d), F(x;\theta_n))$ uniquely for any $X = x$. The ROC curve could be represented with these $(F(x;\theta_d), F(x;\theta_n))$ for all $X = x$. However for multivariate cases, there does not exist an unique value $(x_1, \ldots, x_k)$ that correspond to any paired value of CDFs $(F(\cdot, \ldots, \cdot;\theta_n), F(\cdot, \ldots, \cdot;\theta_d))$, so that a value of paired CDFs could not be defined uniquely for any $X_1 = x_1, \ldots, X_k = x_k$. Hence ROC curve for multivariate variables might not be drawn with paired CDFs $(F(\cdot, \ldots, \cdot;\theta_n), F(\cdot, \ldots, \cdot;\theta_d))$ for all $X_1 = x_1, \ldots, X_k = x_k$.

There exist significant literature reviews of ROC curve research that are mostly are based on the univariate random variable. The score random vector is extended to the k-dimension in this paper; subsequently, we will develop the ROC curve for a multivariate score random vector $X = (X_1, X_2, \ldots, X_k)'$. This ROC curve could be applied to discriminate into dichotomous states in many real multivariate data analysis. The ROC curve based on two conditional joint CDFs, $F(x_1, \ldots, x_k;\theta_d)$

[1] Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea. E-mail: cshong@skku.edu

and $F(x_1, \ldots, x_k; \theta_n)$, is suggested and explained in Section 2. In Section 3, with the assumption that the score random vector follows the multivariate normal distributions, ROC curves are obtained and discussed with a discriminant function. In Section 4, the empirical ROC curve is drawn based on random samples that are obtained from the multivariate normal distributions. Also the multivariate exponential distributions are considered among non-normal distributions and the ROC curve is explored with these random samples from the multivariate exponential distributions as illustrative examples. A conclusion is derived in Section 5.

2. ROC Curve for Multivariate Random Variables

The joint CDF, $F(x_1, x_2, \ldots, x_k)$, of the multivariate score random vector is supposed to be a convex combination of two conditional CDFs, $F(x_1, \ldots, x_k; \theta_d)$ and $F(x_1, \ldots, x_k; \theta_n)$, under the borrower's default and non-default states, such as

$$F(x_1, x_2, \ldots, x_k) = \lambda F(x_1, \ldots, x_k; \theta_d) + (1 - \lambda) F(x_1, \ldots, x_k; \theta_n), \tag{2.1}$$

where λ is the total probability of default $P(\Theta = \theta_d)$. It is assumed that the mean vectors of the multivariate random variable with default and non-default states are $(\mu_{1d}, \mu_{2d}, \ldots, \mu_{kd})'$ and $(\mu_{1n}, \mu_{2n}, \ldots, \mu_{kn})'$, respectively.

It cannot be determined unique $(x_1, \ldots, x_k)$ for any paired value of CDFs $(F(\cdot, \ldots, \cdot; \theta_n), F(\cdot, \ldots, \cdot; \theta_d))$. In addition, the ROC curve might not be plotted with paired CDFs $(F(\cdot, \ldots, \cdot; \theta_n), F(\cdot, \ldots, \cdot; \theta_d))$ for all $X_1 = x_1, \ldots, X_k = x_k$. If the random variables $X_2, \ldots, X_k$ might be represented as functions of the first variable X_1, say $x_i = g_i(x_1)$, $i = 2, \ldots, k$, then $(F(x_1, \ldots, x_k; \theta_n), F(x_1, \ldots, x_k; \theta_d))$ can be replaced with $(F(x_1, g_2(x_1), \ldots, g_k(x_1); \theta_d), F(x_1, g_2(x_1), \ldots, g_k(x_1); \theta_n))$. Moreover, there exists unique $(x_1, x_2 = g_2(x_1), \ldots, x_k = g_k(x_1))$ that correspond to any paired value of CDFs $(F(x_1, g_2(x_1), \ldots, g_k(x_1); \theta_n), F(x_1, g_2(x_1), \ldots, g_k(x_1); \theta_d))$. Since two CDFs $F(x_1, g_2(x_1), \ldots, g_k(x_1); \theta_n)$ and $F(x_1, g_2(x_1), \ldots, g_k(x_1)\ ; \theta_d)$ are not CDFs, those are adjusted to satisfy the properties of CDF. Then a ROC curve for multivariate random variables could be represented with a adjusted paired CDFs for all $x_1, x_2 = g_2(x_1), \ldots, x_k = g_k(x_1)$.

In order to define functions, $g_i(x_1)$, $i = 2, \ldots, k$, let us consider the following linear functions that pass through two mean vectors $(\mu_{1d}, \mu_{2d}, \ldots, \mu_{kd})'$ and $(\mu_{1n}, \mu_{2n}, \ldots, \mu_{kn})'$ of the borrower's default and non-default.

$$\frac{x_1 - \mu_{1d}}{\mu_{1n} - \mu_{1d}} = \frac{x_2 - \mu_{2d}}{\mu_{2n} - \mu_{2d}} = \cdots = \frac{x_k - \mu_{kd}}{\mu_{kn} - \mu_{kd}}. \tag{2.2}$$

Hence, $x_2, \ldots, x_k$ values of the coordinates $(X_1, X_2, \ldots, X_k)$ are expressed as functions of x_1, such as $g_2(x_1), \ldots, g_k(x_1)$, in equation (2.2). For example, $x_2 = g_2(x_1)$ function can be defined as

$$x_2 = g_2(x_1) \equiv \left(\frac{\mu_{2n} - \mu_{2d}}{\mu_{1n} - \mu_{1d}}\right) \times (x_1 - \mu_{1d}) + \mu_{2d}.$$

Then the coordinates $(x_1, x_2, \ldots, x_k)$ of the line pass through two mean vectors. The ROC curve for univariate random variable needs to assume that $F_d(x) \geq F_n(x)$ for all x. In addition, for multivariate random variables, it is also supposed that $F_d(x_1, \ldots, x_k) \geq F_n(x_1, \ldots, x_k)$ for any $x_1, \ldots, x_k$. This assumption implies that each μ_{id} is greater than or equal to μ_{in}, but some μ_{id} are greater than μ_{in} for $i = 1, \ldots, k$. Hence the slope, $(\mu_{in} - \mu_{id})/(\mu_{1n} - \mu_{1d})$, of $g_i(x_1)$ is non-negative for $i = 2, \ldots, k$, so that $g_i(x_1)$ must be non-decreasing in x_1.

Two conditional CDFs in (2.1) could be represented as $F(x_1,g_2(x_1),\ldots,g_k(x_1);\theta_d)$ and $F(x_1,g_2(x_1),\ldots,g_k(x_1);\theta_n)$, respectively, which are the functions of X_1 itself. Finally, the following $F^d(x_1)$ and $F^n(x_1)$ are defined to satisfy the properties of CDF of the univariate X_1, respectively,

$$F^d(x_1) = F\left(x_1, g_2(x_1),\ldots,g_k(x_1);\theta_d\right) \Big/ \int F(x_1, g_2(x_1),\ldots,g_k(x_1);\theta_d)dx_1, \tag{2.3}$$

$$F^n(x_1) = F(x_1, g_2(x_1),\ldots,g_k(x_1);\theta_n) \Big/ \int F(x_1, g_2(x_1),\ldots,g_k(x_1);\theta_n)dx_1.$$

Therefore, if the values of $F^n(x_1)$ and $F^d(x_1)$ for all x_1 correspond to the X and Y axes of a unit square, we could plot the ROC curve for two conditional CDFs of a k-dimensional score random vector.

3. ROC Curve for Multivariate Normal Distribution

It is assumed that two joint CDFs, $F(\underline{\mathbf{x}};\theta_d)$ and $F(\underline{\mathbf{x}};\theta_n)$, under the borrower's default and non-default states are the following multivariate normal distributions:

$$F\left(\underline{\mathbf{x}};\theta_d\right) \equiv \Phi\left(\underline{\mathbf{x}};\underline{\mu_d},\mathbf{\Sigma}_d\right), F\left(\underline{x};\theta_n\right) \equiv \Phi\left(\underline{\mathbf{x}};\underline{\mu_n},\mathbf{\Sigma}_n\right), \tag{3.1}$$

where $\underline{\mu_d}$ and $\underline{\mu_n}$ are the mean vectors; the variances of the random variable for the borrower's default and non-default states are supposed to be 1 and σ^2, respectively, and the covariance of the two random variables, X_i and X_j, are set as $\mathrm{Cov}(X_i,X_j)=\rho^{|i-j|}\sigma^2$. Then the covariance matrixes, $\mathbf{\Sigma}_d$ and $\mathbf{\Sigma}_n$, of the score random vectors for the borrower's default and non-default states are as follows:

$$\mathbf{\Sigma}_d = \begin{pmatrix} 1 & \cdots & \rho^{|1-k|} \\ \vdots & \ddots & \vdots \\ \rho^{|1-k|} & \cdots & 1 \end{pmatrix}, \qquad \mathbf{\Sigma}_n = \begin{pmatrix} \sigma^2 & \cdots & \rho^{|1-k|}\sigma^2 \\ \vdots & \ddots & \vdots \\ \rho^{|1-k|}\sigma^2 & \cdots & \sigma^2 \end{pmatrix}. \tag{3.2}$$

Hence, the models of the borrower's default and non-default states are $\underline{\mathbf{X}_d} \sim N_k(\underline{\mu_d},\mathbf{\Sigma}_d)$ and $\underline{\mathbf{X}_n} \sim N_k(\underline{\mu_n},\mathbf{\Sigma}_n)$, respectively. Various ROC curves will be obtained and compared with several values of r, σ^2 and ρ. Figure 1 represents two ROC curves: the left curve is for the trivariate and the right one is for four dimensional normal distributions. The mean vectors for the default state, $\underline{\mu_d}$, are null vectors; $\underline{\mu_n}$ is $(1, 1.5, 2)'$ on the left and $(1, 1.5, 2, 2.5)'$ on the right in Figure 1. The left graphs in Figure 1 are for $\sigma = 1.5$ and $\rho = -0.5, 0, 0.5$, and the right one is for $\rho = 0.5$ and $\sigma = 0.5, 1.0, 1.5$.

For the left ROC curves (Figure 1), it is found that as ρ increases, the ROC curves go towards the $(0, 1)$ point. Specifically, the discriminative ability of a diagnostic or prognostic test increases. As the standard deviation for the borrower's non-default state, σ, has increasing values from 0.5 to 1.5, while that for default state is fixed to be 1, the ROC curves run far away from the $(0, 1)$ point, such that the discriminative ability is reduced.

From each ROC curve (Figure 1), it is found that when $\mathbf{\Sigma}_d = \mathbf{\Sigma}_n$, the value (x_1, x_2, x_3) of the closest $(0, 1)$ point on the ROC curve corresponds to a zero value of the well-known discriminate function for multivariate normal distributions, such as

$$-2\ln\mathbf{\Lambda} = \ln|\mathbf{\Sigma}_d| + \underline{\mathbf{x}}'\mathbf{\Sigma}_d^{-1}\underline{\mathbf{x}} - \ln|\mathbf{\Sigma}_n| - \left(\underline{\mathbf{x}} - \underline{\mu_{\mathbf{n}}}\right)'\mathbf{\Sigma}_n^{-1}\left(\underline{\mathbf{x}} - \underline{\mu_{\mathbf{n}}}\right).$$

4. Illustrative Examples

First, let the random variables X_1, X_2, X_3 follow the trivariate normal distribution, such as

$$\Phi\left(x_1, x_2, x_3;\underline{\mu_d},\mathbf{\Sigma}_d\right), \qquad \Phi\left(x_1, x_2, x_3;\underline{\mu_n},\mathbf{\Sigma}_n\right),$$

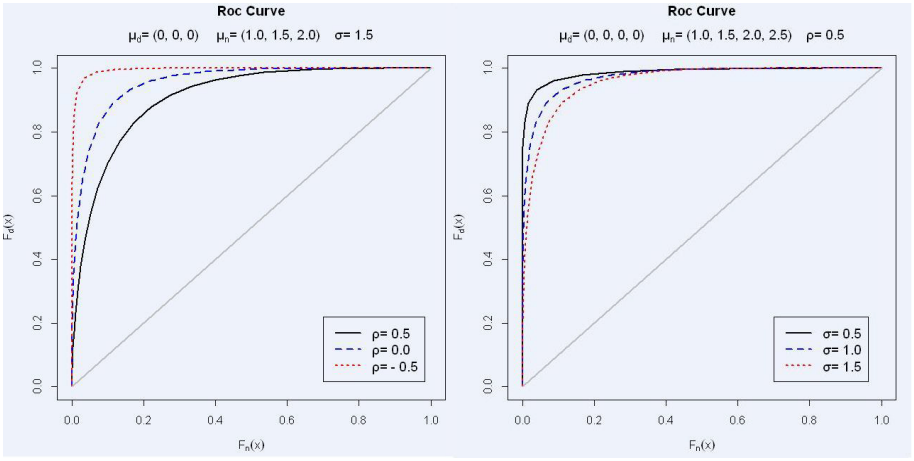

Figure 1: *ROC curves for multivariate normal distributions*

where $\underline{\mu_d} = (2, 3, 4)'$, $\underline{\mu_n} = (3, 4, 5)'$, and $\sigma = 1.5$ and $\rho = 0.5$ in the covariance matrix (3.1). Also, we obtain sample values of sizes 1,500 and 2,000 for each borrower's default and non-default states. Let the sample means and sample standard deviations of the three random variables for each state be $(\overline{X}_{id}, S_{id})$ and $(\overline{X}_{in}, S_{in})$, $i = 1, 2, 3$, respectively. By using equation (2.2), both values x_2 and x_3 in the following intervals for each state are selected for x_1 in the data:

$$x_2 \in \left(\frac{\overline{X}_{2n} - \overline{X}_{2d}}{\overline{X}_{1n} - \overline{X}_{1d}}\right) \times \left(x_1 - \overline{X}_{1d}\right) + \overline{X}_{2d} \pm \Delta_2, \tag{4.1}$$
$$x_3 \in \left(\frac{\overline{X}_{3n} - \overline{X}_{3d}}{\overline{X}_{1n} - \overline{X}_{1d}}\right) \times \left(x_1 - \overline{X}_{1d}\right) + \overline{X}_{3d} \pm \Delta_3,$$

where Δ_2 and Δ_3 may be a quarter of a minimum of two standard deviations for each state of random variables, X_2 and X_3, respectively. That is, $\Delta_2 = 0.25 \times \min\{S_{2d}, S_{2n}\}$ and $\Delta_3 = 0.25 \times \min\{S_{3d}, S_{3n}\}$. Then, the values (x_1, x_2, x_3) in the intervals (4.1) could be regarded as the points on the line that passes through two sample mean vectors $(\overline{X}_{1d}, \overline{X}_{2d}, \overline{X}_{3d})'$ and $(\overline{X}_{1n}, \overline{X}_{2n}, \overline{X}_{3n})'$.

In this example, 70 and 56 values among the data are selected from the borrower's default and non-default states, respectively, which are only 3.6% of the total data set. With this data, the empirical CDFs, $F^d(x_1)$ and $F^n(x_1)$, defined in (2.3) are obtained; hence, the empirical ROC curve could be explored in Figure 2.

Next, a random sample is obtained from the trivariate exponential distribution by using the Cholesky square root method among "Multivariate Computations in R", where $\underline{\mu_d} = (1, 1, 1)'$, $\underline{\mu_n} = (1.5, 1.5, 1.5)'$, and $\sigma = 1.5$, $\rho = 0.3$ in the covariance matrix (3.2). With similar arguments, two samples are generated of sizes 1,500 and 2,000 for each borrower's default and non-default states. In order to take the random values which pass through two mean vectors, the values (x_1, x_2, x_3) are selected in the following intervals:

$$x_2 \in x_1 \pm \Delta_2, \qquad x_3 \in x_1 \pm \Delta_3,$$

where Δ_2 and Δ_3 could set all 0.25.

These sample values are of sizes 105 and 69 from the borrower's default and non-default states,

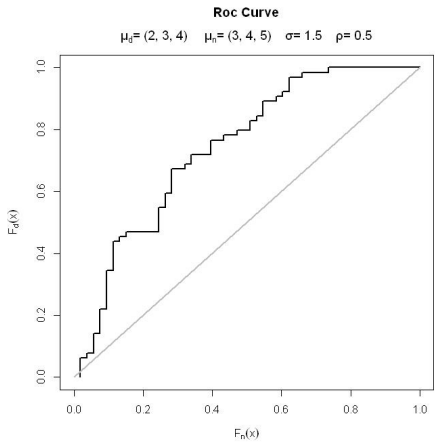

Figure 2: *Empirical ROC curve from multivariate normal distribution*

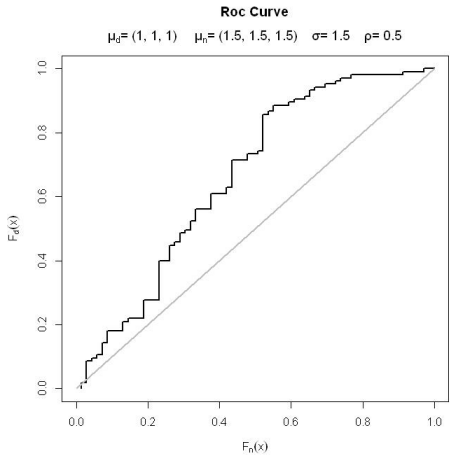

Figure 3: *Empirical ROC curve from multivariate exponential distribution*

respectively. These are only 4.9% of the total data set. With this illustrated sample, the empirical ROC curve could be represented in Figure 3.

5. Conclusion

The ROC curve is plotted with TPR and FPR for the univariate random variable. In this study, we extend to multivariate random vectors. The joint CDF of multivariate score random vector is supposed to be the convex combination of two conditional CDFs, $F(x_1, \ldots, x_k; \theta_d)$ and $F(x_1, \ldots, x_k; \theta_n)$, under the borrower's default and non-default states.

We consider only the values $(x_1, x_2, \ldots, x_k)$ of the line, that pass through two population mean vectors for two states. Then these $x_2, \ldots, x_k$ values of the coordinates $(X_1, X_2, \ldots, X_k)$ could be expressed as functions of x_1. With the values $(x_1, x_2, \ldots, x_k)$ of the line, which pass through two mean vectors, two conditional cumulative distribution functions could be represented as functions of X_1 itself. We modify this function of X_1 on the line, which passes through two mean vectors, in order to satisfy the

properties of CDF. With these transformed two CDFs under the borrower's default and non-default states, we suggest a multivariate ROC curve.

For multivariate normal distributions under the borrower's default and non-default states with various mean vectors and covariance matrixes, ROC curves are explored. Therefore, we could conclude that as ρ increases, ROC curves go towards the (0, 1) point in the unit square. As the standard deviation for the borrower's non-default state has larger values than in the default state, the ROC curves run far away from the (0, 1) point; hence, the discriminative ability decreases.

Two random samples of different sample sizes distributions for the two states are taken from the trivariate normal and exponential distributions. From these data sets, we could select two samples whose values (x_1, x_2, x_3) are regarded to be close to the straight line through two sample mean vectors. With the selected data, the ROC curves could be explored. Even though the selected data is small, these values are important since these are located around the mean vectors. Therefore, it is concluded that this ROC curve can be used and applied to discriminate into dichotomous states in many multivariate analysis whose distributions are normal as well as non Gaussian distributions.

References

Gardner, I. A. and Greiner, M. (2006). Receiver operating characteristic curves and likelihood ratios: Improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests, *American Society for Veterinary Clinical Pathology*, **35**, 8–17.

Greiner, M., Pfeiffer, D. and Smith, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests, *Preventive Veterinary Medicine*, **45**, 23–41.

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

Tasche, D. (2006). Validation of internal rating systems and PD estimates, arXiv.org, eprint arXiv: physics/0606071.

Zweig, M. H. and Campbell, G. (1993). Receiver operating characteristic(ROC) plots: A fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, **39**, 561–577.

Received February 1, 2013; Revised April 10, 2013; Accepted May 14, 2013