IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

Development of Some Distribution's Test Statistic in Analogy to Kolmogorov-Smirnov Test

A. T. Soyinka [1*], E. O. Adeleke [2]

1. Department of Statistics, Federal University of Agriculture, Alabata Abeokuta, Nigeria.
2. Department of Mathematics, Federal University of Agriculture, Alabata Abeokuta, Nigeria.
* Corresponding author: soyinkaat@funaab.edu.ng

Article Info
Received: 06 June 2025 Revised: 03 December 2025
Accepted: 04 February 2026 Available online: 26 February 2026

Abstract

This study developed parametric and nonparametric test statistic, an analogue to Kolmogorov-Smirnov two sample test, for the testing of the equality between bivariate groups. We also established the performance of the developed test statistic in achieving accurate separation and classification. The concept layout model, which is based on Cartesian interaction between discrete random variables (rv's) 'x_m' and 'y_k' arranged in rows and columns respectively for $m, k \in \mathbb{N}$, has a behavioural pattern with bivariate cumulative distribution function (cdf) $F(x, y)$. We assumed that the content within the matrix '$m \times k$' frame followed log-logistic distribution (LLD) and is distribution free. The test statistic 't' is the absolute difference between two bivariate cdf, $|F_1(x, y) - F_2(x, y)|$, under the two distribution scenarios. We then optimize the test statistic which is a function of relationship matrices from the two groups and established its significance at optimal parameter value, from a newly introduced multivariate test significance, before investigating its performance analysis. This analysis enhances the understanding of the profiled content in the two groups, whether they are from the same group or not. In addition, we established the discrimination accuracy of the relationship matrix model towards perfect classification (diagnostic). Application to the discrimination and classification of Bumpus, cancer and mode of delivery data were established.

Keywords: Discrete distribution, Matrices equality and agreement, Test statistic, Discrimination function, Performance analysis.
MSC2010: 60E05, 60B20, 62H12, 62H15.

1 Introduction

Kolmogorov-Smirnov (KS) test in two samples is the most useful and general non-parametric methods for its sensitivity to the location and shape of the empirical cumulative distribution function (cdf). The introduction of histogram bins (Marcos, 2022) and its empirical cumulative behavioural function is a step forward to enhance the usefulness of KS test. However, the major challenge is the fact that the KS test was built around a point of maximum difference between two cdf. What happens when there are several maximum points. In situation of no unique maximum point, the KS

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

test is likely to yield an unreliable test and significance values. So, generalizing the KS test statistic, to accommodate several maximum points and pick the largest of several large ordered points is the only way forward to achieving reliable result from KS test. One of the several ways of helping in this regard, is to discretise the samples in the study before embarking on the KS test. That is by discretizing the intervals with no unique maximum points from its continuous function, we will achieve the slicing of the interval into discrete bins with several unique points and distinct frequencies. We thus discretize the continuous bivariate distribution, so as to get its discrete bivariate form over a matricial structure, with the sole aim of rounding up all several maximum points no matter how many, within the hierachical structure of a matrix. Hence, this study developed a better approach to testing two discrete bivariate groups. The developed test uses all the sample points in analysis, unlike KS-test which is restricted to single maximum point, which often times is not unique. The test statistic of differences between two relationship matrix from each groups, theoretically carved out of a bivariate data, was confirmed. Having developed the bivariate cdf $F(x, y)$ from Cartesian interaction of rv's 'x_m' and 'y_k' within the matrix frame '$m \times k$'. So, in this study, we developed the test statistic analogous to the KS test for the test of two samples from discrete bivariate log-logistic distribution and arbitrary (free) distribution. Owing that both logistic model and free distribution model have been used in small and large categorical analysis. Researchers that have worked on the test statistic tools to ascertain the equality between two matrices profile's includes: Ramos-Garcia et al., (2017) developed multivariate test statistic between experimental matrix and expected gamma matrix; Indahl et al., (2018) that constructed the significance of the statistical similarity index between two matrices frame work via its subspaces or factor subset combinations; Brusco and Steinley (2018) which uses quadratic assignment model to compute the measure of agreement between two proximity matrices; Guo and Modarres (2020) that derived the test statistic of matrices equality via likelihood ratio test, Frobenous norm and tringular test; Soyinka and Olosunde (2021, 2022) that generalized hotelling T^2 test statistic to accommodates matrix data structure whose content followed asymmetric multivariate exponential power distribution and Puritz et al. (2024) demonstrated the implementation of the 'fasano-franceschini.test' and confirmed its aptness in the absence of specific multivariate goodness of fit test (Justel et al., 1994) that is distribution free to test the equality of two random matrices. In all these studies, the approach of representing the content within the matrix layout with its supposed or assumed distribution has not been considered. Hence, this study proposed test statistic for contrasting and comparing bivariate structures in matrices layout whose content followed discretized bivariate log-logistic and arbitrary distributions. The performance rating of the test statistic models towards accurate discrimination and classification of the profiled content in each group, for each of the cases, were established using accuracy, precision, sensitivity, specificity and F1-score. The concept discussed in the study is only applicable to bivariate layout with matrices structure output. This is a possible limitation.

2 Cdf model of LLD Cartesian layout

In this section, the study developed the discrete bivariate cdf model '$0 < F(x, y) \leq 1$' for Cartesian layout, assuming that the interaction content '$H_{m\times k} = x \times y$' of the matrix frame encapsulated within the layout, formed a probability metric space, captured as collections of column probability vector matrix (CPVM), that followed LLD. This implies that we viewed the matrix frame as consisting of several cells, such that in each cell, the interacting varying values of x and y were assigned a probability value depending on the scenario under study to produce a column probability vector. The collection of all the cells probability vectors (in a row) across the kth dimensions (columns) is the cdf model '$F(x, y)$' required in this study. The literary representation of m and k differs depending on the kind of problem at hand. In repeated measurement, the data structure is such that m is the repeated measurement in each of the kth independent units, while in repeated sampling, m is the number of repeated sampling cycles in rows, while k is the number of units (columns) into which the entire population was initially partitioned (divided) for sampling convenience. The same data structure representation can be said of longitudinal data, studied over mth time (periodic or

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

non-periodic) across kth facets. For this study, the implication of m and k in the developed model is explained later with its limitation. Next we obtain the mathematical representation of the cdf model for each of the cases under study.

Proposition 2.1. The bivariate cdf '$F(x,y)$' model of the discrete Log Logistic Distribution (LLD) with parameter $p \in (0,1)$ can be derived as

$$\frac{xy\left(\ln p\right)^2}{1+(x+y)\ln p+xy\left(\ln p\right)^2}=\frac{H\left(\ln p\right)^2}{I+(\sqrt{xx'}+\sqrt{y'y})\ln p+H\left(\ln p\right)^2} \tag{2.1}$$

Proof. Discretizing the survival function of the univariate continuous LLD $S(x)=\left[1+\left(\frac{x}{\alpha}\right)^{\beta}\right]^{-1}$ (Xiaofang et al, 2020), via the re-parameterization function $p=e^{-\alpha^{\beta}}$, we obtained the discrete survival function of LLD as $S(x)=\left[1+x\ln p\right]^{-1}\ \forall\ x=0,1,2,.,..$ Using Cartesian product '$F(x,y)=([1-S(x)][1-S(y)])$', we have (2.1) as the cdf of discrete bivariate LLD.

3 Test statistic for equality of two matrices

The test statistic for the testing of equality between two matrices of the same dimension that followed the same cdf model and belong to the same probability metric space was derived in this section. Suppose we have the matrix $H_{m\times k}=x\times y$ with cdf model '$F(x,y)$' and matrix $Z_{m\times k}=w\times v$ with cdf model '$F(w,v)$', such that both $F(x,y)$ and $F(w,v)$ are from the same probability metric space, then the statistic $t=|F(x,y)-F(w,v)|$ is the tool required to ascertain the extent of the differences in the cdf model of the two matrices in line with KS test concept. In order to place the two cdf models under the same condition, we optimize the cdf difference model '$|F(x,y)-F(w,v)|$', to visualize the equality (differences) of the models under the same parameter values as against the usage of single maximum point used in KS test. Firstly, we ensured that the test statistic tool, '$t=|F(x,y)-F(w,v)|$', is simplified to reflect the practical realization of the matrices 'H' and 'Z' in its real-life bivariate data form. Also we capture single matrix function 'x' and 'w' from $H_{m\times k}=x\times y$ and $Z_{m\times k}=w\times v$ using '$\sqrt{xyy^Tx^T}$' and '$\sqrt{wvv^Tw^T}$' respectively. All these were to ensure that the test tool contains real life profiled data. Stating the null hypothesis (H_o), that the profile content in the two matrices are from the same population $f(x,y)=f(w,v)$, and that the alternative (H_a) implies otherwise, then we can obtain the statistic 't' for the testing of equality between two matrices under the different scenarios as follow:

3.1 Case 1. When the profiled interactive content followed LLD

Proposition 3.1. The test statistic 't' for the differences in cdf from (2.1), can be derived as

$$t=\left|\frac{f(w,v)-f(x,y)-Bf(x,y)f(w,v)}{A\frac{f(w,v)}{F(w,v)}\frac{f(x,y)}{F(x,y)}}\right| \tag{3.1}$$

where $A=\left([2HZ-HZ(H+Z)]\left(\ln p\right)^4+4HZ\left(\ln p\right)^3+3HZ\left(\ln p\right)^2+2HZ\ln p\right)$
and B is

$$\begin{pmatrix} HZ\left[Z(x+y)-H(w+v)\right]\left(\ln p\right)^3+\left[H(x+y)-Z(w+v)-4Z(x+y)+4H(w+v)\right]\left(\ln p\right)^2 \\ +\left[2(x+y)^2-2(w+v)^2+(x+y)-(w+v)+2H(x+y)-2Z(w+v)-5Z(x+y)\right]\ln p \\ -5H(w+v)\ln p+\left[3(x+y)^2-3(w+v)^2+3(x+y)-3(w+v)-2Z(x+y)+2H(w+v)\right] \\ +\left[2(x+y)^2-2(w+v)^2+2(x+y)-2(w+v)\right]\left(\ln p\right)^{-1} \end{pmatrix}.$$

Proof. Starting with the univariate probability mass function (pmf)

$$f(x)=F(x+1)-F(x)=\frac{\ln p}{(1+x\ln p+\ln p)(1+x\ln p)},$$

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

and

$$f(y) = \frac{\ln p}{(1 + y \ln p + \ln p)(1 + y \ln p)},$$

we can obtain the joint pmf $f(x, y)$ as the product $f(x).f(y)$ in (3.2)

$$f(x,y) = \frac{(\ln p)^2}{\left(\begin{array}{c} \left[H^2 + H(x+y) + H\right](\ln p)^4 + \\ \left[(x+y)^2 + 2H(x+y) + 2H + (x+y)\right](\ln p)^3 + \\ \left[(x+y)^2 + 3(x+y) + 2H\right](\ln p)^2 + 2(x+y)\ln p + \\ \left((\ln p)^2 + 2\ln p + 1\right) \end{array} \right)}. \tag{3.2}$$

Next we evaluate $\frac{1}{F(x,y)} - \frac{1}{F(w,v)}$ and $\frac{1}{f(x,y)} - \frac{1}{f(w,v)}$ from (2.1) and (3.2) respectively; and express its relationship in the form $\frac{1}{f(x,y)} - \frac{1}{f(w,v)} = A\left(\frac{1}{F(x,y)} - \frac{1}{F(w,v)}\right) + B$. Then multiplying the outcome of the relationship on both sides by $\frac{f(x,y)f(w,v)}{F(x,y)F(w,v)}$ and making '$|F(x,y) - F(w,v)|$' the subject of the resulting outcome, we have (3.1). Substituting $H_o : f(x,y) = f(w,v)$ in (3.1), then (3.1) reduces to (3.3). Note that (3.3) is a ratio, between two relationship matrix 'A' and 'B' earlier defined, that defines the level of equality between the two matrices H and Z provided A^{-1} exist.

$$t = \left| \frac{F(x,y) - F(w,v)}{F(x,y)F(w,v)} \right| = \left| \frac{B}{A} \right|. \tag{3.3}$$

Note: All preliminary matrices (H and Z) as well as its sub matrices are known. Likewise, owing that the test statistic 't' in (3.3), is itself a pdf, since it was obtained as a difference between two discretized cdf, we first estimate the parameters in the models (in order to place the model under a uniform parameter condition) via numerical optimization of its likelihood function, in view of the fact that its exact solution in each of the cases is either impossible or it is analytically intensive, before we proceed to test statistic estimation and significance.

3.2 Case 2. When the profiled content is distribution free

Proposition 3.2. Given the matrices $E(x,y)$ and $E(w,v)$ as weighted proportion of cells output in each column across all dimensions between the two matrices group. The test statistic 't' when the profile content has no specific distribution can be obtained as the matrices difference

$$t = |E(x,y) - E(w,v)|. \tag{3.4}$$

So, the level of equality between matrices H and Z, can be obtained in each of the cases by comparing relationship matrices A and B in (3.3) for the first case, and matrices $E(x,y)$ and $E(w,v)$ in (3.4) for the second case. Note that for all the derived expressions, the square matrix when '$m = k$' was used. This is another limitation to the study, as future study for similar problem, should consider the usage of rectangular matrix where '$m \neq k$', provided the matrices inverse exist. Such rectangular matrices structure is better captured under multivariate discrete model.

4 Significance of the Multivariate test statistic

In this section, we developed the measure of significance ($p_{value} = 1 - F(.)$) for the obtained multivariate test statistic ($3.3 - 3.4$) between two random relationship matrices. Owing that there are several random matrix distributions in literature, this study leveraged on the developed cdf of the multivariate exponential power random matrix distribution (Soyinka and Olosunde, 2021). The cdf which is developed by normalizing the multivariate exponential power random matrix

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

distribution with Wishart distribution was derived and written as beta of second kind given as

$$F(t) = \frac{I_{gb2}\left[\left|\frac{t}{I_k+t}\right|, \frac{1}{2\beta}, \left(\frac{2k-1}{2\beta} + \frac{1-k}{2}\right)\right]}{\Gamma_k\left(\frac{k}{2\beta}\right)\Gamma\left(\frac{k}{2\beta}\right)}, \tag{4.1}$$

where m and k are as earlier defined, $t_{m\times m}$ is the test statistic, $|.|$ is the determinant, β is the shape parameter of the combined relationship matrices, I_{gb2}, R_{gb2} and I^{-1}_{gb2} are the cdf of incomplete generalized beta of second kind, its regularized version and its inverse respectively. Owing that random squared matrix are often captured in categorical form as chi-square, whose generalized form is Wishart, in this study we required differences and ratios of two Wishart, whose asymptotic distribution is generalized beta distribution. The usage of beta of second kind is to accommodate the p_{value} of the test statistic models over a definite range $[0, 1]$, since both $t = \frac{B}{A}$ and $t = E(x, y) - E(w, v)$ are approximate Wishart ratio's and Wishart difference whose values varies over $(0, \infty)$ and $(0, 1)$ respectively.

Proposition 4.1. The $p_{value} = 1 - F(t)$ for the ratio of two relationship matrices, A and B, can be derived as

$$F(t) = R_{gb2}\left[\left|\frac{\frac{B}{A}}{I_k + \frac{B}{A}}\right|, \frac{1}{2\beta}, \left(\frac{2k-1}{2\beta} + \frac{1-k}{2}\right)\right]. \tag{4.2}$$

Proposition 4.2. The $p_{value} = 1 - F(t)$ for the differences between two relationship matrices can be derived as

$$F(t) = R_{gb2}\left[\left|\frac{E(x, y) - E(w, v)}{I_k + (E(x, y) - E(w, v))}\right|, \frac{1}{2\beta}, \left(\frac{2k-1}{2\beta} + \frac{1-k}{2}\right)\right]. \tag{4.3}$$

Proof. Substituting $t = \frac{B}{A}$ and $t = E(x, y) - E(w, v)$ in (4.1), we obtained (4.2) and (4.3). Note that two random matrices are said to be equal if the value of the suprema p_{value} across the entire matrix t is strictly $>$ than a significant bench mark (say 0.05) or the overall $p_{value} > 0.05$.

5 Performance of test statistic models

In this section, we established the performance of the obtained models for accurate discrimination and classification. So, having confirmed the equality (or otherwise) of the two relationship matrices from the two groups (H and Z), the performance of the model is dependent on the ability of the model to automatically becomes the mathematical function for the effective discrimination of the profiled content in the two groups (random matrices) to enhance good classification. That is, the performance of the model is centered on the ability of the model to uniquely identify and allocate each of the column probability vector matrix (CPVM) to their respective matrix profile. In order to investigate the model performance, we define the performance rating measure 'r' as

$$r = \ln\left[P'\left(\frac{B}{A}\right)P\right], \begin{Bmatrix} +r, & P \in H \\ -r, & P \in Z \end{Bmatrix}, \tag{5.1}$$

where $P_{m\times 1}$ is the CPVM in each of the kth dimension and $(A_{(.)m\times m}, B_{(.)m\times m})$ are the profiled data relationship matrix. $A_{(.)m\times m}$ and $B_{(.)m\times m}$ indicates profiled data relationship matrix from group one and group two respectively. To ascertain the model performance, each of the CPVM 'P' will be evaluated from (5.1) to determine if the model will rightly or wrongly allocate it to any of the two groups. That is if P actually belongs to profiled matrix H, and the model rating r is positive then P is also expected by the model (5.1) to have come from H (True positive). However, if r is negative, then P even though it is from group H, the model expectation revealed otherwise (false positive). Likewise, if P actually belongs to profiled matrix Z, and the model rating r is negative then P is also expected by the model (5.1) to have come from Z (True negative). If however, r is positive, then even though P comes from group Z, the model expectation revealed otherwise indicating

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

false negative. Having completed the model performance for each of the CPVM (P) allocation, the proportion of P that is rightly or wrongly classified (allocated) by the model is then used to confirm the model accuracy, precision, sensitivity and specificity. This performance analysis is crucial to the effectiveness of the models, at not only determining equality between two relationship matrices, but also at identifying factors (CPVM) that are needed to effect a reliable discrimination between two groups ($r = +/-$) and other factors whose contribution are unpredictable and indecisive ($r = NA, r = 0$) but their influence are significant to decision taking. Note that for the second model, the corresponding performance rating in line with the general measure (5.1) can be expressed as $r = P' \left| E(x, y) - E(w, v) \right| P$.

6 Application

In practice, we have two bivariate outputs to be compared. Each output has kth measurable parameters (in the columns), which are repeatedly measured m times to obtain a $m \times k$ matrix output. The probability metric space of the matrix output is first obtained by inscribing the output into the derived cdf models (2.1) and via the weighted cell outcome proportion in case of distribution free output. The test of equality between the two probability metric matrices (spaces) under each of the cases $(3.3 - 3.4)$, via its relationship matrices is then optimized and interpreted for equality significance using the newly introduced multivariate test significance. Having established the matrices equality (or otherwise) from the equality of its relationship matrices, the performance rating of the models to accurate discrimination and classification of the content in each of the two probability metric matrices (spaces) is then established. In order to demonstrate the application of the results obtained in this study, simulated data generated from r software, the packages mvtnorm (Genz, 2023) and NonNorMvtDist (Lun and Khattree, 2020) in CRAN, were modified to accommodate sampling LLD and distribution free data outputs. Samples of bivariate contingents LLD and arbitrary output (free distribution (FD)) were generated for bivariate population of 10^7 entries over the matrix $m \times k$. Two matrices layout of manageable sizes, matrix H and matrix Z, were randomly traced out from the population. The r codes, to generate the bivariate population for each case, trace out the study matrices, determine the optimal parameter values of t, decides the significance of t and establish the performance rating of t between two random relationship matrices, for the simulated and practical data, were given in the supplementary material. The discrimination and classification role of the ratio or difference relationship matrices between the two groups is dependent on the fitness of the bivariate data of the two groups to the assumed Cartesian layout cdf structure. In practice, the simulated or real life data is first inscribed into the assumed Cartesian layout. Owing that the layout via its cdf was already fragmented into two parts using its relationship matrices, we end-up having a plane sliced into three regions. The region where the content in the first group (H) is dominant, the region where the content in the second group (Z) is dominant and the region in between the dominating regions where clear cut dominance between the extreme regions can not be ascertained.

6.1 Simulation results

Using the LLD as a case study, the assumed Cartesian layout area is p^2 while the area within the layout occupied by the data is '$\hat{p}^2$'. The concept in-here is to inscribe a bivariate data for group one and group two with a total area $\hat{p}^2$, into a Cartesian layout plane with area p^2. The result obtained for simulation study is presented in table 1. Table 1, revealed the performance of the LLD model in discrimination and classification of simulated data, when the simulated matrices for groups H and Z have been confirmed to be the same. The table showed that the performance of the model at accurately discriminating and classifying profiled content in each group is dependent on the area of the initial Cartesian layout. The smaller the initial layout area p^2, the better the performance of the model. LLD at $p = 0.06$ performed best at 95.3%.
Table 1. Measure of model performance from simulated LLD data

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

	LLD	LLD	LLD	LLD	LLD	LLD	FD
p	0.06	0.1	0.2	0.3	0.4	0.5	
$\hat{p}$	0.1846	0.1225	0.1885	0.2765	0.2467	0.45	
p_v	0.2281	0.2454	0.3819	0.5277	0.2962	0.4038	0.5547
pre	0.953	0.82	0.939	0.739	0.708	0.51	0.84
npv	0.952	0.86	0.8	0.667	0.745	0.521	0.74
sen	0.953	0.854	0.793	0.694	0.739	0.521	0.764
spe	0.952	0.827	0.94	0.714	0.714	0.51	0.822
acc	0.953	0.84	0.861	0.703	0.726	0.515	0.79
F1	0.953	0.837	0.86	0.716	0.723	0.515	0.8

p_v-p value of relationship matrices difference, pre-precision, npv-negative predictive value, $p_v - p_{values}$, sen-sensitivity, spe-specificity, acc-accuracy, F1-score.

Similarly the performance of the difference model at discriminating and classifying profiled content of bivariate data sets from FD, has the following discrimination and classification properties: precision of 0.84, negative predictive value of 0.74, sensitivity of 0.764, specificity of 0.822, accuracy of 0.79 and F1-score of 0.8. The LLD performed better than the FD.

6.2 Real life data results

In the real life session, we validate the performance of the obtained model at accurately discriminating and classifying multivariate data. The mode of delivery data, the old generation Bumpus data and the cancer data were used to validate the model performance.

6.2.1 Model validation from Mode of delivery data

Firstly, we considered the performance of the obtained models at discriminating between the profile of pregnant women originally booked for normal vaginal delivery and those that were eventually delivered via unplanned emergency ceasarian section (Table 2). The performance of the various relationship matrix discriminant function at effecting accurate classification of the profiled content of pregnant women in the two categories showed a discrimination accuracy $F1$-scores of 1.0 at $p = 0.06$, 0.824 at $p = 0.1$, 0.776 at $p = 0.2$, 0.77 at $p = 0.3$ and 0.393 at $p = 0.4$ when the LLD relationship matrix discriminant function was used as the classifier. Besides the $F1$-scores under LLD, the score of 0.862 was recorded under the free distribution (FD) discriminant function. Hence mode of delivery data is better discriminated using LLD at $p = 0.06$ and the arbitrary relationship matrix discriminant models as classifier.

Table 2. Measure of discrimination accuracy performance on mode of delivery data

	LLD	LLD	LLD	LLD	LLD	FD
p	0.06	0.1	0.2	0.3	0.08	
p_{values}	0.096	0.117	0.2160	0.1444	0.1325	0.094
pre	1.0	0.9	0.839	0.83	0.393	0.9
npv	1.0	0.758	0.677	0.655	0.452	0.806
sen	1.0	0.758	0.722	0.714	0.393	0.824
spe	1.0	0.875	0.808	0.792	0.452	0.893
acc	1.0	0.803	0.758	0.746	0.424	0.855
F1	1.0	0.824	0.776	0.77	0.393	0.862

6.2.2 Model validation from Bumpus data

Secondly, we investigated the performance of our discrimination models on the old generation Bumpus data (Table 3). The result showed that the Bumpus multivariate data (Johnson and Wichern, 2006) did not in any way follow the LLD as its $F1$-scores across several parameters gave a score that is ≤ 0.6. The free distribution (FD) discriminant function gave a classification accuracy of 0.83 which is considered to be very good. The study revealed that the Bumpus data obeyed

IJMSO
International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

the difference discrimination model, and so the Bumpus data is better discriminated using the FD model.

Table 3. Measure of discrimination accuracy performance on Bumpus data

	LLD	LLD	LLD	FD
p	0.0001	0.001	0.1	
p_{values}	0.2403	0.302	0.8607	0.4675
pre	0.1	0.6	0.4	0.714
npv	0.2	0.556	0.4	1.0
sen	0.111	0.6	0.4	1.0
spe	0.182	0.556	0.4	0.6
acc	0.15	0.579	0.4	0.8
F1	0.105	0.6	0.4	0.83

6.2.3 Model validation from cancer data

Thirdly we considered the performance of the developed discrimination model to accurately approve the initial classification of cancer patients as done by screening test based on the consumption of some harmful substance. The result obtained from the model is presented in table 4. The LLD relationship matrix discriminant model gave a discrimination accuracy $F1$-score of 0.96 at $p = 0.06$ with 3.8% of indecisive variables. Further investigation at $p = 0.1$, though gave a higher $F1$-score of 1.0, but with a higher percentage of indecisive variables to about 26.92% of the entire variables.

Table 4. Measure of discrimination accuracy performance on cancer data

	LLD	LLD	LLD	FD
p	0.06	0.1	0.2	
p_{values}	0.4187	0.6053	0.5576	0.521
pre	1.0	1.0	0.083	0.846
npv	0.923	1.0	0.4	0.154
sen	0.923	1.0	0.143	0.5
spe	1.0	1.0	0.267	0.5
acc	0.96	1.0	0.227	0.5
F1	0.96	1.0	0.105	0.629

The $F1$-score of LLD at $p = 0.2$ is very poor (0.153) with 11.54% of indecisive variables. The result from the free distribution discriminant model has very low negative predictive value (0.154) with average sensitivity and specificity scores and an unreliable $F1$-score of 0.629. The LLD discriminant model at $p = 0.06$ behaved better than any other model in the discrimination of cancer patient.

7 Conclusion

In conclusion, this study established the procedure of generating practicable discrete bivariate matrix from the discretization of the cumulative distribution function of the bivariate data. This helped to facilitate the linkage between bivariate data output on excel or any other platform and theoretical practicable matrix that can be used in analysis. The study after then, established test of differences between two practicable probability matrices, from bivariate LLD and arbitrary bivariate output. The significance of the test statistic for the differences between the two matrices were established via a newly introduced multivariate significance test statistic. The performance of the study models at accurately discriminating and classifying profiled content between two groups, when the two groups were adjudged to be from the same population or not, were established.

8 Conflict of interest

No funding was received for the study. So no conflict of interest.

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
11(4), 2025, Pages 114 - 122
https://doi.org/10.5281/zenodo.19005219

9 Acknowledgement

I appreciate all the reviewers for their comments that has greatly improved the paper.

References

[1] Markos V. (2022). How to overlay empirical cunmulative distribution over histogram. *Statalist.org/forum*, com/doi/pdf/10.1177/1536867X211045583.

[2] Ramos-Garcia L.I., Perez-Azorin J.F., Redondo A.P-B, Moran-Velasco V. (2017). Improving gamma analysis comparison using binned multivariate test. *Journal of Physics in Medicine and Biology*, 62(18).

[3] Indahl U.G., Naes T., Liland K.H. (2018). A similarity index for comparing coupled matrices. *Journal of Chemometrics*, 32(10).

[4] Bulut O., Desjardins C.D. (2019). Profile analysis of multivariate data in r: A brief introduction to ProfileR package. *https:// Cran. R-project.org/package=profileR* (package version 0.3-5).

[5] Brusco, M.J., Steinley, D. (2018). Measuring and testing the agreement of matrices. *Journal of behaviour research and methods*, 50: 2256-2266.

[6] Guo L. and Modarres R. (2020). Testing the equality of matrix distribution. *Journal of statistical methods and applications*, 29: 289-307.

[7] Soyinka, A.T. and Olosunde, A.A (2021). Inferences from Asymmetric Multivariate Exponential Power Distribution. *Journal of Indian Statistical Institute Sankhya B*, 83(2):350-370.

[8] Soyinka A.T., Olosunde A.A. (2022). On Discretization of Continuous Random Variables for Contingency Tables: Discrete Johnson Systems of Distribution as a Case Study with Applications. *Journal of Statistical Theory and Practice*, 16(64). https://doi.org/10.1007/s42519-022-00290-8.

[9] Puritz C., Ness-Cohn E., Braun R., Weihs L. (2024). fasano.franceschini.test: An implementation of the two samples multivariate Kolmogorov-Smirnov test in r-software. *R Journal*, 15(3): 159-171.

[10] Justel A., Pena D., Zamar R. (1994). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Journal of statistics and econometrics*, 13:94-109.

[11] Xiaofang He, Wangxue Chen, Wenshu Qian (2020). Maximum likelihood estimators of the parameters of the log-logistic distribution, *Statistical Papers* 61:1875–1892.

[12] Genz A., Bretz F., Miwa T., Mi X., Leisch F., Scheipl F., Bornkamp B., Maechler M., Hothorn T. (2023). mvtnorm: Multivariate Normal and t Distributions. *R package version 1.2-4*. URL http://mvtnorm.R-forge.R-project.org.

[13] Lun Z., Khattree R. (2020). NonNorMvtDist: Multivariate Lomax (Pareto Type II) and Its Related Distributions. *R package version 1.0.2*. https://CRAN.R-project.org/package=NonNorMvtDist.

[14] Johnson, R.A. and Wichern, D.A. (2006). Applied multivariate statistical analysis, 3rd edition. Wiley, New Jersey, p. 140–237.

[15] Zheng X., Gogarten S.M., Lawrence M., Stilp A., Conomos M.P., Weir B.S., Laurie C., Levine D.(2017). SeqArray – A storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*; doi:10.1093/bioinformatics/btx145.