Hindawi
Advances in Astronomy
Volume 2023, Article ID 8682054, 11 pages
https://doi.org/10.1155/2023/8682054

Research Article

A New Approach to Astronomical Data Analysis Based on Multiple Variables

Prasenjit Banerjee,[1] **Asis Kumar Chattopadhyay**,[1] **and Soumita Modak**[2]

[1]*Department of Statistics, University of Calcutta, 35 Ballygunge Circular Road, Kolkata 700019, India*
[2]*Faculty of Statistics, University of Calcutta, Basanti Devi College, 147B Rash Behari Avenue, Kolkata 700029, West Bengal, India*

Correspondence should be addressed to Prasenjit Banerjee; b.prasenjit1994@gmail.com

Received 4 April 2023; Revised 29 May 2023; Accepted 16 June 2023; Published 6 July 2023

Academic Editor: Yu Liu

Data analysis for a sample of celestial bodies generally is preceded by the completeness test in order to verify whether the sample objects are proper representatives of the corresponding part of the universe. A data set following a multivariate, continuous, uniform distribution is said to be "complete in space." This paper introduces a new approach to check for this completeness for any astronomical data set under a multivariate setup. Our proposed procedure, using the multiple tests of hypotheses based on nonparametric statistics, and consequently, combining their p values, outperforms others from the literature.

1. Introduction

In astronomy, different catalogs from various sources are generally combined to create a master data set, where it is supposed to be "complete in space." Under the univariate setup, a sample related to a particular astronomical parameter (variable), original or transformed [1], is uniform in distribution (continuous), then it is referred to as "complete in space." In this context, the popular $V/V_{\max}$ test [1] has been proposed. However, it is restricted to univariate analysis; therefore, it cannot take into account the multivariate structure of the samples, and it provides only a point estimation of the concerned statistics. On the other hand, the other implemented statistical tests in astronomy, to check for uniformity of a multidimensional sample, involve comparison of each individual variable with a univariate uniform distribution, irrespective of the important dependence structure underlying the multivariate sample under analysis [2, 3]. Therefore, in this paper, we propose a new approach to investigate the completeness of a multivariate data set in space. To our very knowledge, it is the first multivariate test of hypothesis to check completeness for an astronomical sample. We discuss two nonparametric tests [4, 5] to check whether the data set follows a multivariate uniform distribution over the range $[0,1]^d$ (denoted by $U_d[0,1]$) or not, where d is the dimension of data set for $d \geq 2$ and $[0,1]^d$ is the set of the d-times Cartesian product of the closed interval $[0, 1]$. Any deviation of the given sample from $U_d[0,1]$ will lead to the rejection of the fact that the data set is "complete in space."

Establishment of a test to check $U_d[0,1]$ becomes difficult for higher values of d, whereas the existing tests are either not well defined or not feasible for big d [6, 7]. In the literature, the multiple tests for goodness-of-fit are presented for checking the null hypothesis if a sample is from a specific multivariate distribution. However, only few of them are put forward for the multivariate uniformity of the data set. Two popular tests, among them, are (i) the multivariate Kolmogorov–Smirnov test [8] and (ii) the test based on empirical characteristic function [9, 10]. The empirical distribution function has jumps and discontinuity at various points apart from the sample observations, which makes it quite challenging to be computed for large d. Therefore, the algorithm for the concerned test statistic is yet unavailable for $d > 2$ [6, 8]. Any distribution is characterized by its characteristic function, which is consistently estimated by an empirical version. However, computation of the test statistics and the critical value for the test, based on the empirical characteristic function, is very difficult for high-dimensional sample as well as for big data, which induce greater testing error [9–11]. Avoiding all these concerns, we propose a novel approach as follows.

In this work of Astrostatistic, we suggest a new testing procedure based on multiple nonparametric tests of hypotheses, where we check whether the individual marginal of the data set is from a univariate, continuous, uniform distribution over the range [0, 1] (denoted by $U[0,1]$) or not. Here, we use the fact that if the given multivariate sample follows U_d [0,1], then all the d marginals of the data set will be from $U[0,1]$ and vice versa. Our final decision is taken uniquely by properly combining the dependent multiple tests or their corresponding p values. With advanced fashion of data collection, we focus on the high-dimensional big data from astronomical field (see, [5, 12, 13], and references therein), where our data study shows that the proposed technique is effective and superior compared to its competitors.

This paper is organized as follows. Two proposed methods are described in Section 2. The simulation is carried out in Section 3. Section 4 holds application of our proposed tests to an astronomical data set. Finally, Section 5 concludes the paper.

2. Proposed Method

Our main objective is to investigate the completeness of a multivariate sample in space, which is done in terms of hypothesis testing. Suppose $\mathbf{X} = (X_1, X_2, \ldots, X_d)'$ is a real-valued$d$-variate observation vector and we want to test whether it follows $U_d[0,1]$ or not, that is, we test the null.

$$H_0\colon \mathbf{X} \sim U_d[0,1] \text{ against the alternative } H_1\colon \mathbf{X} \nsim U_d[0,1], \tag{1}$$

where '$\sim$' is used to mean following and '$\nsim$' not following. We perform our test using the given sample:

$\{\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})', i = 1, 2, \ldots, n\}$ with size $n(\gg d)$.

2.1. Multiple Tests. The abovementioned proposed hypothesis testing can be equivalently performed in terms of the following d number of multiple tests, which are carried out in a univariate setup for each variable. Here, we implement the fact that if the given multivariate sample follows $U_d[0,1]$, then all the d marginals of the data set will be from $U[0,1]$ and vice versa. The dependent multiple tests are formulated as follows:

$$H_{0j}\colon X_j \sim U[0,1] \text{ agianst } H_{1j}\colon X_j \nsim U[0,1]\ \forall j = 1, 2, \ldots, d. \tag{2}$$

Then, each of the univariate multiple tests is done with the help of the popular nonparametric one-sample tests: (t1) Kolmogorov–Smirnov test [14] and (t2) Anderson–Darling test [15], to check whether the given sample for each dimension follows $U[0,1]$ or not. Acceptance of all $H_{0j}s$ for $j = 1, 2, \ldots, d$ concludes with acceptance of H_0, whereas rejection of any H_{0j} for at least one $j = 1, 2, \ldots, d$ causes rejection of H_0.

(t1) We implement the univariate, nonparametric, distribution-free, one-sample Kolmogorov–Smirnov test of hypothesis to check whether the unknown continuous distribution function $F(X)$ of a random variable X is equal to a completely specified reference distribution $F_0(X)$. This is done in terms of the test statistic:

$$\underset{X_i, i=1(1)n}{\text{Sup}} \left|F_n(X_i) - F_0(X_i)\right|, \tag{3}$$

which involves a distance between the empirical distribution function $F_n(X)$ computed using a random sample $X_1, X_2, \ldots, X_n$ on X and the cumulative distribution function $F_0(X)$ of the reference distribution. The null hypothesis is accepted if the computed test statistic is smaller than or equal to the upper α point of the distribution for the test statistic (equation (3)) under the null.

(t2) Then, we suggest the nonparametric, distribution-free Anderson–Darling test, which is a modification of the Kolmogorov–Smirnov test, assigning more weight to the tails of the distribution for the given sample. It tests whether a univariate sample $\{X_{(1)}, X_{(2)}, \ldots, X_{(n)}\}$ comes from a population with a specific continuous distribution function F. When it is true, we can assume that $F \sim U(0,1)$ and the sample $F(X_i), i = 1, 2, \ldots, n$ are then tested for uniformity [16]. The test statistic $\widetilde{T}$ is defined as follows:

$$\begin{aligned} \widetilde{T}^2 &= -n - S \quad \text{with} \\ S &= \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln\{F(X_i)\} + \ln\{1 - F(X_{n+1-i})\}\right], \end{aligned} \tag{4}$$

where values, greater than its upper α point under the null hypothesis, reject the null of uniformity against the both-sided alternative.

The influence of ties on (t2) varies depending on the characteristics and frequency of ties present in the data. Ties can have a noticeable impact on the precision of the test and potentially affect the test results. Presence of ties disrupts the estimation of the distribution function, particularly in the tails of the distribution, which will lead to inaccurate calculations of the test statistic and p value. If the numbers of ties are less or if they are evenly distributed across the data set, their impact on (t2) will be minimal. However, when there are numerous ties or if they cluster around specific values, the precision of the test can be compromised.

2.2. Test Statistics. Suppose the statistics for testing H_{0j} against H_{1j}, carried out in terms of (t1) or (t2), is T_j for $j = 1, 2, \ldots, d$. Then, the critical region for the right-tailed alternative in the j-th one among the multiple tests is given by

$$\{\text{observed values of } T_j > C_\alpha\}, \tag{5}$$

where C_α is the required upper α point of the sampling distribution for our proposed test statistic $T_j \ni$.

$$P_{H_{0j}}\left(T_j > C_\alpha\right) = \alpha \,\text{holds}. \tag{6}$$

Here, α is the nominal level of significance for each of our marginal tests which we perform using the (t1) Kolmogorov–Smirnov test, wherein the asymptotic $C_\alpha = 1.36/\sqrt{n}$ [17] and (t2) Anderson–Darling test with the asymptotic value for $C_\alpha = 2.4986$ [18, 19]. We denote the statistics for testing H_0 against H_1 by T, where the test statistics from the multiple tests corresponding to the marginals are combined together with equal weights which defines the following:

$$P(T > C) = \frac{1}{d}\sum_{j=1}^{d} P\left(T_j > C\right) \,\text{for any}\, C, \tag{7}$$

and subsequently we obtain:

$$P_{H_0}(T > C_\alpha) = \alpha. \tag{8}$$

Thus, it is a right-tailed test, so the null hypothesis is rejected at α% level of significance if the observed value of T based on the given sample is greater than C_α. Being a data-driven test, the distribution of T and the corresponding p value (discussed in the following section) are determined empirically.

2.3. p Value Computation. To obtain the p value of our proposed test, we have calculated the p value for the j-th marginal test as p_j, for $j = 1, 2, \ldots, d$. Since the multiple tests are interdependent, so are their p values. There are various ways to combine these p values among themselves [20–22]. We consider the following:

$$p = \sum_{j=1}^{d} p_j. \tag{9}$$

The null hypothesis H_0 is rejected if the p value 'p' computed from the data set is less than its upper α point, say p_α, which is estimated as $\widehat{p}_\alpha$, by applying the bootstrap technique to the given sample. Thus, H_0 is rejected in favor of H_1 at α% level of significance if the computed value of $p < \widehat{p}_\alpha$.

3. Simulation

The performance of our proposed technique of testing is demonstrated through an extensive simulation study, in this section, where we implement both (t1) and (t2) tests separately. The scenarios from which the samples are drawn are (a) $U_d[0,1]$ under independence and (b) $U_d[0,1]$ under dependence structure. Case (a) has the $d \times d$ correlation matrix given by $\rho = (\rho_{ij})$, where $\rho_{ij} = 0$ for $i \neq j$ and $\rho_{ii} = 1 \forall i = 1, 2, \ldots, d$. Thus,

$$\rho = \begin{pmatrix} 1 & 0 & . & . & . & 0 \\ 0 & 1 & . & . & . & 0 \\ & & & . & & \\ & & & . & & \\ 0 & 0 & . & . & . & 1 \end{pmatrix}. \tag{10}$$

On the other hand, the dependence structure in (b) is induced in two distinct ways.

(b1) A nonidentity correlation matrix is considered as $\rho = \mathbf{J}\mathbf{J}' - kA$, where $k = 2/d(d-1)$, $\mathbf{J} = (1, 1, \ldots, 1)'$ is a $d \times 1$ vector, and $A = (a_{ij})$ is a $d \times d$ symmetric matrix $\ni a_{ij} = i + (j-1)(j-2)/2$ for $i < j$, $a_{ij} = a_{ji} \forall i \neq j$, and $a_{ii} = 0 \forall i = 1, 2, \ldots, d$. Thus, ρ explicitly looks like:

$$\rho = \begin{pmatrix} 1 & \dfrac{d^2-d-2}{d^2-d} & . & . & \dfrac{2\times(d-2)}{d^2-d} \\ \dfrac{d^2-d-2}{d^2-d} & 1 & . & . & \dfrac{2\times(d-3)}{d^2-d} \\ . & . & . & . & . \\ . & . & . & 1-\dfrac{2\times i+(j-1)\times(j-2)}{d\times(d-1)} & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & 1 & . \\ \dfrac{2\times(d-2)}{d^2-d} & \dfrac{2\times(d-3)}{d^2-d} & . & . & 1 \end{pmatrix}^{d\times d}. \tag{11}$$

(b2)The later way of generating random samples from $U_d[0,1]$ under the dependent setup is carried out through the Clayton copula modeling [23, 24] by implementing the multivariate uniform distribution from Cook and Johnson [25], where the scalar parameter involved in the distribution is taken to be 2.

We compute the size and power under (a) and (b) with $d = 5, 10, 15, \dots, 50$ and $n = 100, 150, 200, \dots, 1000$, as we focus on the multivariate large astronomical data sets. Both size and power are estimated by the Monte Carlo simulation with the number of replications equals 10,000. The size is estimated as the proportion (out of 10,000) with H_0 rejected when the simulated samples are originally drawn from $U_d[0,1]$.

Analogously, the powers are computed when the simulated samples are not coming from $U_d[0,1]$, where we consider the following setups.

a1 The multivariate beta (Dirichlet) distribution over the range $[0,1]^d$ with the shape parameter vector as $\mu = (1, 2, 3, \dots.., d)'$ and the scale parameter (beta) taken as 3 [26, 27].

a2 The truncated multivariate normal distribution, over the range $[0,1]^d$, with the mean vector $0.5\mathbf{J}$ where $\mathbf{J} = (1, 1, \dots., 1)'$, and the correlation matrix: $\mathbf{J}\mathbf{J}' - kA$ where $k = 2/d(d-1)$ and $A = (a_{ij})$ with $a_{ij} = i + (j-1)(j-2)/2$ for $i < j, a_{ij} = a_{ji} \forall i \neq j, a_{ii} = 0 \forall i = 1, 2, \dots, d$ [28, 29].

It is to be noted that the samples are drawn through a Gibbs sampler technique [21, 30] with a thinning of 10 (that is, every 10th observation is selected) to get rid of the autocorrelation present in the synthetic data.

a3 Multivariate normal distribution with the same mean vector and the correlation matrix as mentioned in (a2) [31, 32].

a4 $U_d[0,2]$ under independent structure.

a5 $U_d[-1,1]$ under independent structure [33, 34].

3.1. Competitor Tests.

Several goodness-of-fit tests checking for multivariate uniformity, from Yang and Modarres [35], are considered as competitors: (i) the test based on normal quantiles and (ii) a set of tests based on interpoint distances, as discussed below.

3.1.1. Uniformity Test Based on Normal Quantiles.

Suppose the random vectors, $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{id})' \in R^d$ for $i = 1, 2, \dots, n$, constitute a random sample of size n from a population of the random vector $\mathbf{X}$ characterized by a continuous multivariate distribution function F_X. We consider the following transformation from $\mathbf{X}$ to $\mathbf{Z}$:

$$\begin{aligned}\mathbf{Z}_i &= (Z_{i1}, Z_{i2}, \dots, Z_{id})' \\ &= \left(\Phi^{-1}(X_{i1}), \Phi^{-1}(X_{i2}), \dots., \Phi^{-1}(X_{id})\right)',\end{aligned} \tag{12}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. The test statistics under study is given by

$$\chi^2 = n\overline{\mathbf{Z}}'\overline{\mathbf{Z}}, \tag{13}$$

where $\overline{\mathbf{Z}} = (1/n\sum_{i=1}^{n} Z_{i1}, 1/n\sum_{i=1}^{n} Z_{i2}, \dots.., 1/n\sum_{i=1}^{n} Z_{id})'$.

Under the null setup:

$$F_X \sim U_d[0,1], \text{and } \mathbf{Z}_i \sim N_d[\mathbf{0}, I_d] \text{ for } i = 1, 2, \dots, n, \tag{14}$$

where I_d is the identity matrix of order d and $N_d[\mathbf{0}, I_d]$ denotes a d-variate normal distribution with the null vector as the mean and the dispersion matrix I_d. It implies $\overline{\mathbf{Z}} \sim N_d[\mathbf{0}, I_d/n]$ and $\chi^2 \sim \chi^2_d$ (a central chi-square distribution with degrees of freedom = d). Then, testing

$$\text{the null: } F_X \sim U_d[0,1] \text{ vs the alternative: } F_X \nsim U_d[0,1], \tag{15}$$

is equivalent to testing

$$\text{the null: } \mathbf{Z} \sim N_d[\mathbf{0}, I_d] \text{ vs the alternative: } \mathbf{Z} \nsim N_d[\mathbf{0}, I_d]. \tag{16}$$

The null hypothesis is rejected at α% level of significance if the calculated $\chi^2 > \chi^2_{d,\alpha} \ni P(\chi^2 > \chi^2_{d,\alpha}) = \alpha$ holds under the null.

3.1.2. Uniformity Test Based on Interpoint Distances.

For a given sample $\mathbf{X}_1, \mathbf{X}_2, \dots., \mathbf{X}_n$ of $n$real-valued vectors on $\mathbf{X}$, we use a test based on the first two moments of the interpoint distances [36, 37]. The moments and the distribution of the interpoint distances between the multivariate Bernoulli random vectors are investigated by Modaeres in [38], whereas the asymptotic properties of the small interpoint distances in a sample are introduced by Jammalamadaka and Janson [39]. The test, discussed in this section, uses the asymptotic distribution of the sample mean and the sample variance of all interpoint distances.

The sample mean (m_1) and the sample variance (m_2) of the interpoint distances are respectively expressed as follows:

$$m_1 = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} \left\|\mathbf{X}_i - \mathbf{X}_j\right\|^2 \text{ and } m_2 = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} \left(\left\|\mathbf{X}_i - \mathbf{X}_j\right\| - \frac{d}{6}\right)^2, \tag{17}$$

where their corresponding expectations are as follows:

$$E(m_1) = \frac{d}{6} \text{ and } E(m_2) = \frac{7d}{180}. \tag{18}$$

Under the null $\mathbf{X} \sim U_d[0, 1]$, the respective variances would be described by Yang and Modarres [35]:

$$V(m_1) = \frac{d(2n+3)}{90n(n-1)} \text{and} \tag{19}$$

$$V(m_2) = \begin{cases} \dfrac{1}{\binom{n}{2}}\left[\dfrac{1}{56700}\{989 + 202(n-2)\}\right], \text{if } d = 2, \\ \dfrac{1}{\binom{n}{2}}\left[\dfrac{1}{1050}\{37 + 6(n-2)\}\right], \text{if } d = 3, \\ \dfrac{1}{\binom{n}{2}}\left[\dfrac{49d^2}{16200} + \dfrac{101d}{37800} + 2(n-2)\left\{\dfrac{d^2}{16200} + \dfrac{29d}{37800}\right\}\right], \text{if } d \geq 4. \end{cases} \tag{20}$$

The central limit theorem for U-process says that under the null, the followings hold (Arcones and Giné [40]):

$$Q_1 = \frac{m_1 - (d/6)}{\sqrt{V(m_1)}},$$
$$Q_2 = \frac{m_2 - (7d/180)}{\sqrt{V(m_2)}} \sim N(0,1) \text{ as } n \longrightarrow \infty, \tag{21}$$

$$\Longrightarrow Q_1^2, Q_2^2 \sim \chi_1^2 \text{ as } n \longrightarrow \infty, \tag{22}$$

$$\Longrightarrow Q_3 = Q_1^2 + Q_2^2 \sim \chi_2^2 \text{ as } n \longrightarrow \infty, \tag{23}$$

as the first two order moments are independent of each other [41, 42]. Any of the statistics Q_1^2, Q_2^2, or Q_3 (see equations (22) and (23)) may be regarded as our test statistics. The null hypothesis is rejected in favor of the two-sided alternative for large values of the statistic, which is done at α% level of significance if the calculated value of the test statistics is larger than its upper α point under the null.

3.2. Results. In the simulation study, we choose $\alpha = 0.05$. Tables 1–3 show the estimated sizes, for samples from the null distributions under (a), (b1) and (b2), are all coming out close to the nominal level of significance, with both the proposed tests (t2) and (t2), for all considered values of n and d.

To address ties in (t2), the averaging technique has been used. The averaging technique is a tie-breaking method, which involves assigning distinct values to tied observations by taking the average of the tied values. Moreover, as our simulated data set is from $U_d(0,1)^d$ setup, it contains an insignificant number of ties for each of the marginal $U_d(0,1)$. Hence, the original data set with ties and the modified data set where ties are resolved using tie-breaking techniques are almost alike. By averaging the tied observations, we ensure that each tied value is distinct, allowing (t2) to provide more accurate results and better estimate of the distribution function.

As competitors, we consider the four tests discussed in Section 3.1. They are referred to as their respective test statistics: χ^2, Q_1^2, Q_2^2, and Q_3, where we first investigate their empirical sizes under all the conditions as considered for our proposed tests. Tables for competitor tests show that, among all the competitors, only the χ^2 test for (a) independent $U_d(0,1)$ samples attains its nominal α, whereas it also fails under the more sophisticated multivariate structures such as (b1) and (b2). However, it can be deemed as a rival to compare the performance of our proposed tests.

Just like Tables 1–3, we have also computed the size values of the test (t2), competitor tests χ^2, and competitor tests based on the statistics Q_1^2, Q_2^2, and Q_3.

Both the first and second proposed tests (t1) and (t2), for the samples from a non-null distribution (a2), exhibit an increasing power computed with the increase in n and/or d. A maximum of powers for (t1) comes out to be 0.542114 with $n = 1000$ and $d = 50$ (Table 4), whereas (t2) has its highest power calculated as 0.586738 which is attained for $n = 1000$ and $d = 50$ (Table 5). For every choice of n, the powers of the tests are optimally good with a value 1 for samples from each of the non-null distributions (a1, a3–a5) under consideration.

The power estimated for the first competitor test χ^2 comes out to be very low under the non-null distribution (a2). However, it gradually increases with an increase in n as well as d (Table 6), with a largest value 0.2285. The empirical

Table 1: Size computed with the test (t1) for (a) independent $U_d[0, 1]$.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.04476	0.04568	0.045147	0.045385	0.045576	0.045577	0.04546	0.045563	0.045482	0.04553
150	0.04714	0.04587	0.046027	0.046145	0.045932	0.045893	0.045883	0.04597	0.045978	0.046068
200	0.04788	0.04728	0.04744	0.047095	0.04696	0.046913	0.046903	0.04683	0.046716	0.046638
250	0.0467	0.04715	0.046013	0.046265	0.046556	0.046613	0.04656	0.046395	0.04646	0.04663
300	0.04632	0.0463	0.046327	0.046095	0.046168	0.046303	0.046303	0.046563	0.046593	0.046528
350	0.0464	0.04698	0.046713	0.04663	0.046784	0.0467	0.046823	0.046855	0.046822	0.046826
400	0.0474	0.04748	0.047033	0.047285	0.046984	0.047137	0.047	0.047045	0.046996	0.047054
450	0.04892	0.04776	0.047247	0.04692	0.046948	0.047123	0.04706	0.04709	0.047158	0.047164
500	0.04818	0.04804	0.048407	0.04798	0.048	0.04788	0.047811	0.047883	0.048058	0.048078
550	0.04858	0.04704	0.04738	0.04753	0.047352	0.047337	0.04756	0.047825	0.04788	0.04787
600	0.04852	0.04723	0.046887	0.047605	0.047224	0.047247	0.047209	0.047403	0.047496	0.047476
650	0.04826	0.04823	0.047587	0.047695	0.047664	0.047677	0.047823	0.047873	0.047858	0.047778
700	0.04796	0.04829	0.048053	0.04808	0.048052	0.047867	0.047886	0.047968	0.047898	0.04788
750	0.04756	0.0481	0.04788	0.04794	0.047804	0.047983	0.047894	0.047818	0.047673	0.047766
800	0.04784	0.0479	0.048	0.047755	0.047892	0.047893	0.047777	0.047743	0.047724	0.04761
850	0.04776	0.04752	0.04752	0.047635	0.047708	0.048067	0.048094	0.048015	0.048064	0.047994
900	0.04762	0.04759	0.047953	0.047725	0.04792	0.04811	0.047897	0.048135	0.04802	0.047934
950	0.0486	0.04802	0.048753	0.04883	0.048656	0.048593	0.048554	0.048598	0.048516	0.048478
1000	0.04814	0.04818	0.048327	0.04859	0.048816	0.048657	0.048537	0.04835	0.048224	0.04802

Table 2: Size computed with the test (t1) for (b1) dependent $U_d[0, 1]$.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.04642	0.04398	0.04432	0.044025	0.044548	0.044683	0.044754	0.043635	0.044711	0.044448
150	0.04834	0.04571	0.047713	0.046325	0.045624	0.04587	0.047157	0.04685	0.047051	0.04562
200	0.04718	0.046	0.04684	0.046055	0.045404	0.046487	0.046197	0.04694	0.045722	0.046548
250	0.04684	0.04657	0.045787	0.045895	0.045884	0.045353	0.047289	0.046028	0.045836	0.046752
300	0.04712	0.04635	0.04592	0.04672	0.045996	0.045867	0.047111	0.046793	0.046869	0.04623
350	0.04594	0.04706	0.048153	0.04742	0.049148	0.04718	0.047723	0.047748	0.046487	0.047896
400	0.04798	0.04601	0.047893	0.047045	0.04834	0.046547	0.04804	0.046998	0.046582	0.047054
450	0.04966	0.04838	0.04804	0.04747	0.046208	0.04798	0.047246	0.04695	0.047924	0.04724
500	0.04862	0.04638	0.04816	0.047765	0.04816	0.047253	0.047346	0.047208	0.048238	0.048348
550	0.0485	0.04834	0.047727	0.0484	0.046908	0.04776	0.047723	0.048365	0.047713	0.048418
600	0.04736	0.0484	0.046653	0.046105	0.047184	0.047773	0.04634	0.046143	0.048751	0.047692
650	0.04816	0.0487	0.04798	0.04716	0.049032	0.048283	0.048503	0.047708	0.047502	0.047254
700	0.04866	0.04808	0.04746	0.048675	0.046612	0.047553	0.04916	0.048445	0.046327	0.047046
750	0.04756	0.04763	0.04758	0.04776	0.048612	0.04886	0.048117	0.049095	0.046978	0.048064
800	0.04886	0.04786	0.047487	0.048035	0.04752	0.047707	0.048743	0.048758	0.046916	0.047998
850	0.04776	0.0478	0.04838	0.047455	0.047004	0.04901	0.049209	0.047393	0.048542	0.046976
900	0.04732	0.04721	0.04874	0.04841	0.047944	0.049267	0.047766	0.04765	0.049489	0.04741
950	0.04688	0.04814	0.049513	0.04876	0.0484	0.04968	0.047291	0.049613	0.049542	0.048952
1000	0.0477	0.04829	0.04746	0.04806	0.05	0.049123	0.048729	0.048663	0.04708	0.048332

power takes a value 1 $\forall n, d$ under the distribution (a1). Thus, we comment that, in this situation, our proposed technique with both the tests is competitive with this competitor. Here, the test statistic involved in the rival is based on $\Phi^{-1}(z)$ for $z \in [0, 1]$; therefore, among all non-null distributions (a1–a5), only the Dirichlet distribution (a2) and the truncated multivariate normal distribution (3) are considered, for power calculation of test χ^2, as those sample values lie in $[0, 1]^d$.

For the later set of competitors based on the measures Q_1^2, Q_2^2, and Q_3, under the non-null distributions (a1) and (a2), the empirical power is increasing in n and d and reaches 1 for most values of the pair (n, d) (see, Tables 7–10). For the samples from (a3–a5), the powers all attain 1. In spite of this optimal power execution, the use of these tests in identifying "completeness in space" is highly questionable due to the drastic failure in satisfying the size condition, even for the multivariate uniform distribution under independence.

4. Application

We apply our proposed technique to the observed data set in space obtained from NEWFIRM Medium Band Survey (NMBS). Data set from the NMBS catalog consists of two

Table 3: Size computed with the test (t1) for (b2) dependent $U_d[0, 1]$.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.04482	0.04488	0.04526	0.044785	0.045064	0.044063	0.044466	0.045533	0.045369	0.044342
150	0.04644	0.04595	0.045153	0.046565	0.044804	0.046363	0.045477	0.04545	0.045731	0.045248
200	0.04536	0.04559	0.046087	0.045125	0.047116	0.046203	0.045634	0.046443	0.046829	0.046406
250	0.04574	0.04647	0.046147	0.04666	0.046544	0.04705	0.046746	0.046873	0.046429	0.046656
300	0.04564	0.04658	0.047673	0.0483	0.047328	0.047113	0.046423	0.046583	0.046547	0.047424
350	0.04746	0.04664	0.047193	0.04823	0.04708	0.047227	0.046654	0.046745	0.047051	0.047642
400	0.04804	0.04743	0.046933	0.04654	0.0472	0.047663	0.047466	0.046628	0.047238	0.047142
450	0.04826	0.04727	0.04706	0.04807	0.04758	0.04791	0.046794	0.046745	0.047556	0.046676
500	0.04636	0.04769	0.048173	0.04759	0.04798	0.047843	0.047134	0.047108	0.047944	0.0472
550	0.04698	0.0469	0.047033	0.04746	0.04798	0.047573	0.047474	0.04816	0.04816	0.0471
600	0.0478	0.04661	0.047153	0.04782	0.048496	0.04844	0.047897	0.04801	0.047404	0.047924
650	0.0478	0.04765	0.04714	0.048255	0.04698	0.046607	0.048374	0.04804	0.047489	0.047934
700	0.0469	0.04863	0.04838	0.04807	0.04688	0.04736	0.04832	0.048365	0.047949	0.04772
750	0.0489	0.04981	0.048447	0.047125	0.048088	0.04792	0.047906	0.04807	0.047907	0.047648
800	0.04812	0.04774	0.048173	0.04861	0.048056	0.048317	0.047751	0.047795	0.048313	0.048362
850	0.04812	0.0476	0.04714	0.048245	0.047696	0.047473	0.047606	0.04802	0.047887	0.04789
900	0.04924	0.04847	0.048613	0.04756	0.047636	0.0471	0.048357	0.048623	0.048202	0.047826
950	0.0482	0.04921	0.048593	0.04726	0.047392	0.048187	0.048569	0.04869	0.048273	0.047908
1000	0.05	0.04907	0.04754	0.04724	0.0483	0.04824	0.04848	0.047168	0.047164	0.048144

Table 4: Power estimated with the first proposed test (t1) for (a1) multivariate truncated normal sample.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.07478	0.162	0.18902	0.21238	0.233184	0.246853	0.256389	0.263498	0.268558	0.274604
150	0.1026	0.20196	0.242847	0.269775	0.288792	0.298597	0.305523	0.312215	0.31536	0.31962
200	0.13706	0.23465	0.284347	0.309	0.32422	0.332747	0.338497	0.34391	0.347029	0.34996
250	0.17176	0.26479	0.316347	0.338605	0.351276	0.358063	0.364303	0.367873	0.371764	0.374508
300	0.2027	0.29381	0.340387	0.36097	0.372216	0.379727	0.384577	0.388495	0.390798	0.394482
350	0.22804	0.31744	0.361527	0.379335	0.390308	0.397063	0.402117	0.40634	0.409131	0.412306
400	0.25244	0.33986	0.378747	0.39601	0.405784	0.41297	0.416834	0.422085	0.424927	0.42763
450	0.2712	0.35754	0.3955	0.410765	0.420452	0.42645	0.432037	0.435383	0.438587	0.44146
500	0.28732	0.37477	0.409047	0.42406	0.432736	0.43908	0.444149	0.448043	0.451011	0.454104
550	0.3016	0.38936	0.420853	0.435605	0.444908	0.451113	0.455557	0.459428	0.462836	0.465366
600	0.31362	0.40175	0.43216	0.44653	0.455652	0.462357	0.465709	0.47019	0.473744	0.476916
650	0.3267	0.41432	0.442807	0.456925	0.465796	0.47208	0.476157	0.479868	0.483411	0.485982
700	0.33468	0.42642	0.45248	0.46601	0.47554	0.48069	0.485497	0.489483	0.492962	0.495248
750	0.34502	0.43565	0.462727	0.47478	0.484152	0.490083	0.494371	0.49859	0.501191	0.504396
800	0.35666	0.44593	0.470493	0.48322	0.492392	0.49834	0.50218	0.506488	0.510542	0.512518
850	0.36584	0.4561	0.47894	0.491695	0.500476	0.506613	0.510571	0.514313	0.517993	0.520508
900	0.37444	0.46257	0.48522	0.499045	0.508176	0.51426	0.518591	0.522615	0.52564	0.528576
950	0.3843	0.47354	0.49362	0.5067	0.515516	0.520923	0.526234	0.529033	0.53254	0.535292
1000	0.39536	0.48021	0.500273	0.513705	0.522092	0.528797	0.53306	0.536483	0.540473	0.542114

versions for the photometric samples as the original SExtractor output and a catalog with additional deblending. We consider the first version that contains the photometric redshifts and rest-frame colors from EAZY, and the stellar population synthesis (SPS) variables from FAST using the Bruzual and Charlot [43] models. Here, we study the early type galaxies (ETGs) [44] from the AEGIS 1 catalog, whose redshift ranges from 0.5 to 4. As our interest is to study the intrinsic properties of the galaxies, we consider the following parameters (variables) that remain invariant with the change in distance: (i) K_{ellip} is the K-band ellipticity, (ii) K_{R50} is the K-band half-light radius, (iii) z is the redshift of the galaxies, (iv) l_{age} is the log (age/year), (v) l_{mass} is the log (mass/$M_{\odot}$), and (vi) l_{ssfr} is the log (specific star formation rate × year).

Our data set consists of the abovementioned variables on 6,661 ETGs. We apply our technique, in terms of the proposed two tests, to investigate whether the data set is "complete in space." Here, for 'x' as an observed variable, we consider the following transformation "y" as follows:

$$y = \frac{\log(|x|) - \log(|x|_{\min})}{\log(|x|_{\max}) - \log(|x|_{\min})}, \tag{24}$$

Table 5: Power estimated with the second proposed test (t2) for (a1) multivariate truncated normal sample.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.08706	0.18502	0.226047	0.252485	0.271076	0.282643	0.291117	0.296515	0.300667	0.304674
150	0.13546	0.23222	0.285373	0.309035	0.323816	0.331347	0.338	0.342938	0.34548	0.349036
200	0.18192	0.2784	0.325107	0.34561	0.358504	0.365677	0.37036	0.374968	0.377864	0.380338
250	0.22294	0.31424	0.35638	0.3745	0.385236	0.390967	0.39684	0.400335	0.403776	0.406614
300	0.25302	0.3428	0.380947	0.396105	0.406404	0.413237	0.417974	0.42171	0.424169	0.427572
350	0.27758	0.36682	0.401273	0.414785	0.426112	0.432093	0.436494	0.44055	0.443536	0.446374
400	0.29494	0.38808	0.418153	0.432625	0.442024	0.448197	0.453223	0.456693	0.459744	0.46292
450	0.31186	0.40588	0.433913	0.447665	0.45702	0.462753	0.468123	0.471495	0.47482	0.477348
500	0.32654	0.4217	0.447933	0.46099	0.470884	0.47675	0.481574	0.484995	0.488813	0.491202
550	0.34116	0.43654	0.460153	0.473895	0.483248	0.489407	0.493343	0.49755	0.50118	0.503004
600	0.35716	0.449	0.472707	0.485105	0.49486	0.500793	0.505191	0.509313	0.512644	0.515478
650	0.3721	0.46175	0.483113	0.496515	0.5059	0.512023	0.516177	0.5202	0.523047	0.525594
700	0.38478	0.47148	0.493733	0.50638	0.515872	0.522043	0.526271	0.53045	0.533	0.53509
750	0.39934	0.48141	0.503773	0.516015	0.524692	0.53087	0.535954	0.539903	0.542709	0.54521
800	0.41464	0.49258	0.512833	0.525215	0.533808	0.54015	0.545054	0.54851	0.552324	0.554448
850	0.42922	0.50198	0.521573	0.534	0.543068	0.5493	0.553251	0.557185	0.561124	0.562808
900	0.44046	0.50967	0.52952	0.54209	0.550968	0.5568	0.561949	0.56581	0.568938	0.571622
950	0.4541	0.51873	0.536893	0.549885	0.558672	0.565363	0.569234	0.572825	0.576489	0.5789
1000	0.46576	0.52686	0.54374	0.557295	0.565496	0.572923	0.577389	0.58106	0.584838	0.586738

Table 6: Power estimated with the competitor test χ^2 for (a1) multivariate truncated normal sample.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.03	0.0198	0.0169	0.0265	0.0404	0.0753	0.1102	0.1467	0.1875	0.2277
150	0.0289	0.0193	0.017	0.0225	0.0421	0.0702	0.103	0.1434	0.1879	0.23
200	0.0296	0.0181	0.0168	0.0251	0.0401	0.0718	0.1024	0.143	0.1866	0.2212
250	0.0285	0.0187	0.0195	0.0211	0.0426	0.0777	0.102	0.1476	0.184	0.2149
300	0.0265	0.0198	0.0177	0.0259	0.0455	0.0711	0.1054	0.1489	0.1831	0.2263
350	0.0294	0.0195	0.0191	0.0263	0.0475	0.0707	0.1089	0.1475	0.1811	0.2228
400	0.029	0.0195	0.0168	0.0281	0.0422	0.0712	0.1063	0.1435	0.1832	0.221
450	0.0288	0.0172	0.0172	0.0248	0.0429	0.0678	0.1082	0.1444	0.1891	0.2268
500	0.0291	0.018	0.0157	0.0238	0.0432	0.0685	0.1018	0.1415	0.1888	0.2225
550	0.029	0.019	0.0165	0.0252	0.0424	0.0693	0.1086	0.1517	0.1824	0.2237
600	0.0293	0.0184	0.0163	0.0276	0.0463	0.0721	0.1077	0.1453	0.18	0.2227
650	0.0264	0.0198	0.0168	0.0229	0.0437	0.0693	0.1082	0.147	0.1806	0.2174
700	0.0264	0.0185	0.019	0.0244	0.0428	0.0706	0.1065	0.1446	0.1853	0.2257
750	0.03	0.0184	0.0182	0.0237	0.0441	0.073	0.1038	0.1399	0.1885	0.2285
800	0.0308	0.0197	0.0167	0.0236	0.0416	0.0689	0.1045	0.1411	0.1896	0.2274
850	0.029	0.0165	0.0184	0.0278	0.0427	0.0664	0.1054	0.1425	0.1833	0.2277
900	0.0294	0.0176	0.0184	0.0222	0.0419	0.0735	0.1064	0.1474	0.1865	0.2192
950	0.0286	0.0186	0.0173	0.027	0.0444	0.0708	0.1065	0.1407	0.1814	0.2281
1000	0.0276	0.0197	0.0168	0.0252	0.0459	0.0714	0.1053	0.1444	0.179	0.2217

where $|x|$ is the absolute value of x, $\log(\cdot)$ is the natural logarithmic function with base e, $|x|_{\max}$ is the maximum value of $|x|$, and $|x|_{\min}$ is the minimum value of $|x|$. This transformation is done on each of the 6 original variables in such a way that the ranges, under the null hypothesis, remain the same in the transformed space.

We now implement our tests in terms of the p value (see, (9)), where we obtain $\widehat{p}_\alpha$ (see Section 2.3) through the nonparametric bootstrap technique (Modak and Bandyopadhyay (2018)) as follows:

(i) For each of the multiple tests, we perform bootstrapping individually

(ia) $B = 10,000$ bootstrap samples are drawn from the given data set and used to compute B number of bootstrap p values for the j-th marginal test as $p_{j,b}$ for $b = 1, 2, \ldots, B$

(ib) Estimate the upper α point $\widehat{p}_{j,\alpha}$ for p_j using the sampling distribution of the computed bootstrap p values from step (ia) as

$$\frac{\left\{number\ of\ p_{j,b} s \forall b = 1(1)B \leq \widehat{p}_{j,\alpha}\right\}}{B} = \alpha. \quad (25)$$

(ii) Redo steps (ia)-(ib) for $j = 1, 2 \ldots, d$

Table 7: Power computed with the competitor test Q_2^2 for (a2) multivariate truncated normal sample.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.0212	0.0137	0.6514	0.9962	1	1	1	1	1	1
150	0.056	0.0179	0.8607	1	1	1	1	1	1	1
200	0.0936	0.0227	0.9537	1	1	1	1	1	1	1
250	0.1424	0.029	0.987	1	1	1	1	1	1	1
300	0.1992	0.0309	0.9967	1	1	1	1	1	1	1
350	0.253	0.0357	0.9993	1	1	1	1	1	1	1
400	0.3122	0.0465	0.9996	1	1	1	1	1	1	1
450	0.3668	0.0553	1	1	1	1	1	1	1	1
500	0.4244	0.0611	1	1	1	1	1	1	1	1
550	0.4812	0.0715	1	1	1	1	1	1	1	1
600	0.5285	0.08	1	1	1	1	1	1	1	1
650	0.5816	0.0886	1	1	1	1	1	1	1	1
700	0.6229	0.0991	1	1	1	1	1	1	1	1
750	0.6631	0.107	1	1	1	1	1	1	1	1
800	0.7031	0.1202	1	1	1	1	1	1	1	1
850	0.7397	0.1351	1	1	1	1	1	1	1	1
900	0.7634	0.1418	1	1	1	1	1	1	1	1
950	0.7978	0.1564	1	1	1	1	1	1	1	1
1000	0.8245	0.1712	1	1	1	1	1	1	1	1

Table 8: Power computed with the competitor test Q_1^2 for (a2) multivariate truncated normal sample.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.6256	0.9973	1	1	1	1	1	1	1	1
150	0.8145	0.9999	1	1	1	1	1	1	1	1
200	0.9196	1	1	1	1	1	1	1	1	1
250	0.9673	1	1	1	1	1	1	1	1	1
300	0.9876	1	1	1	1	1	1	1	1	1
350	0.9946	1	1	1	1	1	1	1	1	1
400	0.9977	1	1	1	1	1	1	1	1	1
450	0.9994	1	1	1	1	1	1	1	1	1
500	0.9997	1	1	1	1	1	1	1	1	1
550	1	1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	1	1	1	1	1	1	1	1	1	1

(iii) For a given data set, the null hypothesis of multivariate uniformity is rejected at α% level of significance if $p = \sum_{j=1}^{d} p_j < \widehat{p}_\alpha = \sum_{j=1}^{d} \widehat{p}_{j,\alpha}$

Based on our procedure, the tests (t1) and (t2) both produce a p value zero with the cutoff values 2.608102×10^{-7} and 2.571811×10^{-7}, respectively. Therefore, in the light of the sample given, we reject the null hypothesis at $\alpha = 5\%$ level of significance and conclude that the sample does not come from a $U_d(0,1)$ distribution and hence is not "complete in space." Moreover, the data set under consideration does not have any ties present in the marginal; hence, no tie-breaking (such as, averaging technique 3.2) technique is required to eradicate the ties from the data set before the application of (t2) in the marginal.

Here, we cross-check our results by the popular and very classical of its kind $V/V_{\max}$ test [1] from astronomy fraternity. It calls a univariate data set "complete in space," if $< V/V_{\max} > = 0.5$, where $\langle \cdot \rangle$ denotes the mean of the study variable V with its maximum $V_{\max}$. However, it is not a statistical test for an appropriate hypothesis rather provides only a point estimate. Moreover, for multivariate data, only the marginal means are determined independently by this procedure. Anyway, the computed values for $\langle V/V_{\max} \rangle$ corresponding to the 6 study variables are 0.6136369, 0.3906980, 0.4305935, 0.4915570, 0.5898555, and 0.1268829, respectively. It shows the mean values for only 1 amongst 6 marginals are close to 0.5, whereas for the others they are less than 0.5 and for 2 variables it is greater. Therefore, the outcome of rejecting the null distribution resulted in our method is supported by the well-known$V/V_{\max}$ test.

Table 9: Power computed with the competitor test Q_2^2 for (a1) Dirichlet sample.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.6397	1	1	1	1	1	1	1	1	1
150	0.8801	1	1	1	1	1	1	1	1	1
200	0.9723	1	1	1	1	1	1	1	1	1
250	0.9953	1	1	1	1	1	1	1	1	1
300	0.9988	1	1	1	1	1	1	1	1	1
350	0.9999	1	1	1	1	1	1	1	1	1
400	1	1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	1	1	1	1	1	1	1	1	1	1

Table 10: Power computed with the competitor test Q_3 for (a2) multivariate truncated normal sample.

n	d									
	5	10	15	20	25	30	35	40	45	50
100	0.5197	0.9935	1	1	1	1	1	1	1	1
150	0.7272	0.9997	1	1	1	1	1	1	1	1
200	0.8642	1	1	1	1	1	1	1	1	1
250	0.9331	1	1	1	1	1	1	1	1	1
300	0.9683	1	1	1	1	1	1	1	1	1
350	0.9851	1	1	1	1	1	1	1	1	1
400	0.994	1	1	1	1	1	1	1	1	1
450	0.9981	1	1	1	1	1	1	1	1	1
500	0.999	1	1	1	1	1	1	1	1	1
550	0.9999	1	1	1	1	1	1	1	1	1
600	1	1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	1	1	1	1	1	1	1	1	1	1

5. Conclusion

This paper checks the completeness for the multivariate astronomical samples, implementing our novel approach. The advised procedure, using two tests (t1) and (t2), has been shown to perform well with the help of multiple tests of hypotheses and then combining the results of the dependent marginal tests. A few characteristics of our technique are listed below.

(1) If an astronomical data set is from a continuous multivariate uniform setup, then it is said to be

"complete in space" and vice versa. Our test is the first, to our best knowledge, to check for completeness of an astronomical sample in the multivariate setup.

(2) Our approach, although proposed and analyzed for checking multivariate uniformity, can be used for any other arbitrary, continuous, multivariate distribution.

(3) We have used two univariate, nonparametric, one-sample tests: (t1) Kolmogorov–Smirnov and (t2) Anderson–Darling, to check for uniformity of the data set, corresponding to each of the marginals. However, any other test, appropriate for use on the multiple tests individually, can be implemented analogously. All the shortcomings of the (t1) and (t2) tests have been taken into consideration before their application.

(4) The proposed tests' efficiency, supremacy, and wide applicability for high-dimensional, big data sets are demonstrated through extensive data study.

(5) Our proposed test is established as an efficient method in astronomy for the objective under analysis.

In the near future, we are planning to develop a new test based on the regression analysis to check for completeness of astronomical samples.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The author Prasenjit Banerjee acknowledges the research fellowship provided by the University Grants Commission, India, with UGC Id. NOV2017-422665 and UGC reference no. 1141/(CSIR-UGC Net June 2017).

References

[1] M. Schmidt, "Space distribution and luminosity functions of quasi-stellar radio sources," *The Astrophysical Journal*, vol. 151, p. 393, 1968.

[2] T. De, T. Chattopadhyay, and A. K. Chattopadhyay, "Use of Cross-Correlation Function to Study Formation Mechanism of Massive Elliptical Galaxies," *Pub©lications of the Astronomical Society of Australia*, vol. 31, 2014.

[3] S. Modak, T. Chattopadhyay, and A. K. Chattopadhyay, "Two phase formation of massive elliptical galaxies: study through cross-correlation including spatial effect," *Astrophysics and Space Science*, vol. 362, no. 11, p. 206, 2017.

[4] M. L. Puri and P. K. Sen, "Nonparametric Methods in Multivariate Analysis," *Biometrische Zeitschrift*, vol. 16, 1971.

[5] S. Modak, *Uncovering Astrophysical Phenomena Related to Galaxies and Other Objects through Statistical Analysis*, Ph.D. thesis, Handle System, San Francisco, CA, USA, 2019.

[6] R. Singh, S. Dutta, and N. Misra, "Some multivariate goodness of fit tests based on data depth," *Journal of Nonparametric Statistics*, vol. 34, no. 2, pp. 428–447, 2022.

[7] Z. Chen and T. Hu, "Statistical test for bivariate uniformity," *Advances in Statistics*, vol. 2014, Article ID 740831, 6 pages, 2014.

[8] A. Justel, D. Peña, and R. Zamar, "A multivariate Kolmogorov-Smirnov test of goodness of fit," *Statistics & Probability Letters*, vol. 35, no. 3, pp. 251–259, 1997.

[9] Y. Fan, "Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function," *Journal of Multivariate Analysis*, vol. 62, no. 1, pp. 36–63, 1997.

[10] M.-D. Jiménez-Gamero, V. Alba-Fernández, J. Muñoz-García, and Y. Chalco-Cano, "Goodness-of-fit tests based on empirical characteristic functions," *Computational Statistics & Data Analysis*, vol. 53, no. 12, pp. 3957–3971, 2009.

[11] A. Feuerverger and R. A. Mureika, "The Empirical Characteristic Function and its Applications," *The annals of Statistics*, vol. 5, pp. 88–97, 1977.

[12] A. K. Chattopadhyay and T. Chattopadhyay, *Statistical methods for astronomical data analysis*, Springer, Berlin, Germany, 2014.

[13] P. Banerjee, T. Chattopadhyay, and A. K. Chattopadhyay, "Investigation of the effect of bars on the properties of spiral galaxies: a multivariate statistical study," *Communications in Statistics-Simulation and Computation*, pp. 1–31, 2022.

[14] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[15] F. W. Scholz and M. A. Stephens, "K-Sample anderson–darling tests," *Journal of the American Statistical Association*, vol. 82, no. 399, pp. 918–924, 1987.

[16] S. S. Shapiro, "How to test normality and other distributional assumptions," ASQ, Milwaukee, WI, USA, Tech. rep, 1990.

[17] E. S. Pearson, *Biometrika Tables for Statisticians*, USP, Sao Paulo, SP Brasil, 1800.

[18] M. Rahman, L. M. Pearson, and H. C. Heien, "A modified Anderson-Darling test for uniformity," *Bulletin of the Malaysian Mathematical Sciences Society*, vol. 29, no. 1, 2006.

[19] L. Jäntschi and S. D. Bolboacă, "Computation of probability associated with Anderson–Darling statistic," *Mathematics*, vol. 6, no. 6, p. 88, 2018.

[20] M. B. Brown, "400: A Method for Combining Non-independent, One-Sided Tests of Significance," *Biometrics*, vol. 31, pp. 987–992, 1975.

[21] W. Poole, D. L. Gibbs, I. Shmulevich, B. Bernard, and T. A. Knijnenburg, "Combining dependent P-values with an empirical adaptation of Brown's method," *Bioinformatics*, vol. 32, no. 17, pp. i430–i436, 2016.

[22] S. Modak and U. Bandyopadhyay, "A New Nonparametric Test for Two Sample Multivariate Location Problem with Application to Astronomy," 2018, https://arxiv.org/abs/1801.06809.

[23] R. B. Nelsen, "Methods of Constructing Copulas," *An Introduction to Copulas*, vol. 139, pp. 51–108, 2006.

[24] T. K. Nayak, "Multivariate Lomax distribution: properties and usefulness in reliability theory," *Journal of Applied Probability*, vol. 24, no. 1, pp. 170–177, 1987.

[25] R. D. Cook and M. E. Johnson, "A family of distributions for modelling non-elliptically symmetric multivariate data,"

Journal of the Royal Statistical Society: Series B, vol. 43, no. 2, pp. 210–218, 1981.

[26] B. Sivazlian, "On a multivariate extension of the gamma and beta distributions," *SIAM Journal on Applied Mathematics*, vol. 41, no. 2, pp. 205–209, 1981.

[27] C. Troskie, "Noncentral multivariate Dirichlet distributions," *South African Statistical Journal*, vol. 1, no. 1, pp. 21–32, 1967.

[28] S. Wilhelm and B. Manjunath, "tmvtnorm: a package for the truncated multivariate normal distribution," *Sigma*, vol. 2, no. 2, pp. 1–25, 2010.

[29] J. A. Breslaw, "Random sampling from a truncated multivariate normal distribution," *Applied Mathematics Letters*, vol. 7, no. 1, pp. 1–6, 1994.

[30] G. Casella and E. I. George, "Explaining the Gibbs sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.

[31] Y. L. Tong, *The Multivariate normal Distribution*, Springer Science & Business Media, Berlin, Germany, 2012.

[32] S. Ghurye and I. Olkin, "A characterization of the multivariate normal distribution," *The Annals of Mathematical Statistics*, vol. 33, no. 2, pp. 533–541, 1962.

[33] P. L'Ecuyer, "Random Number Generation," *Handbook Of Computational Statistics*, pp. 35–71, Springer, Berlin, Germany, 2012.

[34] J.-J. Liang, K.-T. Fang, F. Hickernell, and R. Li, "Testing multivariate uniformity and its applications," *Mathematics of Computation*, vol. 70, no. 233, pp. 337–355, 2000.

[35] M. Yang and R. Modarres, "Multivariate tests of uniformity," *Statistical Papers*, vol. 58, no. 3, pp. 627–639, 2017.

[36] S. Modak, "A new measure for assessment of clustering based on kernel density estimation," *Communications in Statistics-Theory and Methods*, pp. 1–10, 2022.

[37] S. Modak, "A new nonparametric interpoint distance-based measure for assessment of clustering," *Journal of Statistical Computation and Simulation*, vol. 92, no. 5, pp. 1062–1077, 2022.

[38] R. Modarres, "On the interpoint distances of Bernoulli vectors," *Statistics & Probability Letters*, vol. 84, pp. 215–222, 2014.

[39] S. R. Jammalamadaka and S. Janson, "Limit Theorems for a Triangular Scheme of U-Statistics with Applications to Inter-point Distances," *The Annals of Probability*, vol. 14, pp. 1347–1358, 1986.

[40] M. A. Arcones and E. Giné, "Limit Theorems for U-Processes," *The Annals of Probability*, vol. 21, pp. 1494–1542, 1993.

[41] A. Stuart, "Kendalls advanced theory of statistics," *Distribution theory*, vol. 1, 1994.

[42] T. W. Anderson and T. Anderson, *An introduction to multivariate statistical analysis*, John Wiley & Sons, New York, NY, USA, 1958.

[43] G. Bruzual and S. Charlot, "Stellar population synthesis at the resolution of 2003," *Monthly Notices of the Royal Astronomical Society*, vol. 344, no. 4, pp. 1000–1028, 2003.

[44] K. E. Whitaker, I. Labbé, P. G. Van Dokkum et al., "The NEWFIRM Medium-Band Survey: photometric catalogs, redshifts, and the bimodal color distribution of galaxies out to z 3," *The Astrophysical Journal*, vol. 735, no. 2, p. 86, 2011.