Int. J. Nonlinear Anal. Appl. 13 (2022) 1, 2703-2711
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.5992

Distribution-free two-sample homogeneity test for circular data based on geodesic distance

Ahmed Jebur Ali[a], Samira Faisal Abushilah[a,*]

[a]Department of Mathematics, College of Education for Girls, University of Kufa, Najaf, Iraq

Abstract

A new distance-based homogeneity test for circular data is considered in this paper. This test statistic could be used with a permutation test to detect the homogeneity between the distributions of two angular groups. A comparison between the proposed test against the Randomisation Watson test (RWT) and Wheeler Watson test (WWT) is addressed as well. The power of the new test has been computed and compared against RWT and WWT powers based on the angular simulated data which have been generated from the von Mises distributions. The simulation study based on the power demonstrates that the proposed statistical test outperforms classical tests.

Keywords: Circular Data, Geodesic Distance, Von Mises Distribution, Randomisation Watson Test, Wheeler Watson Test, Energy Statistic.

1. Introduction

Two-sample nonparametric test to detect the homogeneity between the distributions of two groups of circular data is presented in this paper. Circular data differs from traditional linear data and statistical methods to deal with this type of data are still under development. Angular data can be represented by an angle $\theta \in [0, 2\pi)$ and can be seen as a point on the unit circle [7].

There are a few contributions on a data lying on the circle or on the sphere. Watson proposed a nonparametric test to test the null hypothesis that two circular samples are drawn from the same population. Suppose $S_1 = \{\phi_1, ..., \phi_n\}$ and $S_2 = \{\psi_1, ..., \psi_m\}$ be two samples of sizes n and m respectively, which are collected from circular populations. The Watson test statistic is defined by the following form:

$$\boldsymbol{U}^2{}_{n,m} = \frac{nm}{N^2}\Big[\sum_{k=1}^{N} d_k^2 - \frac{(\sum_{k=1}^{N} d_k)^2}{N}\Big], \tag{1.1}$$

Email addresses: ahmedj.alkhuzai@student.uokufa.edu.iq (Ahmed Jebur Ali), sameerah.hathoot@uokufa.edu.iq (Samira Faisal Abushilah)

Received: September 2021 *Accepted:* December 2021

where $N = n + m$, and d_k represent the differences between the two cumulative relative frequency distributions [1].

Wheeler and Watson [10] presented a statistical test to test the null hypothesis there is no difference between two circular distributions (independently developed by Mardia, 1967) [5, 6]. The Wheeler Watson test statistic is defined as:

$$WWT = \frac{2(N-1)\left(S_i^2 + C_i^2\right)}{nm}, \tag{1.2}$$

where

$$C_i = \sum_{j=1}^{n} \cos(d_j), \qquad S_i = \sum_{j=1}^{m} \sin(d_j),$$

and

$$d_j = \frac{(360^o) \text{ rank of } \theta_j}{N}.$$

Please note that, the index i in Equation (1.2) refers to either sample S_1 or S_2; and in the calculation it does not matter which one of the two samples is used [11].

In 2018, Landler et al. [4] presented a comparison study to the statistical power of five widely used tests (Watson's test, Rayleigh test, Kuiper's test, V-texst, and Rao's spacing test). They showed that the V-test had more power for symmetrical distributions, that Rao's spacing performed poorly for all investigated unimodal distributions, and that the remaining three tests performed similarly.

In metric spaces, the energy statistic ($\mathcal{E}$-statistic), which has been introduced in 1984-1985 by Gábor Székely in a series of lectures [8], is a function of distances between observations. The statistic is very useful, more common and powerful than classical statistic (non-energy type) such as F-statistic and correlation [9]. The energy statistic is defined by

$$\mathcal{E}(X,Y) = \frac{nm}{n+m}\Big(\frac{2}{nm}A_1 - \frac{1}{n^2}A_2 - \frac{1}{m^2}A_3\Big), \tag{1.3}$$

where

$$A_1 = \sum_{i=1}^{n}\sum_{j=1}^{m} D(x_i, y_j), \quad A_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} D(x_i, x_j), \quad A_3 = \sum_{i=1}^{m}\sum_{j=1}^{m} D(y_i, y_j).$$

However, since the distribution of the energy statistic is unknown, then this statistic could be used with a permutation test (which is very time-consuming) to test the homogeneity between the distributions of two groups. Moreover, energy statistic has been considered for traditional linear data, and in order to use energy statistic for data lying on a torus we need to make some modifications to achieve the required behaviour for circular data.

Therefore, the following questions arise: Can we reduce the computational time for permutation test by considering energy-like statistic? Can we create statistics which have the ability to deal with not only linear data but also with circular data?

In this paper, we answer these questions by first considering energy-like statistic which could be used to detect the homogeneity between the distributions of two circular groups. Moreover, a comparison between the proposed test against Randomisation Watson test (RWT) and Wheeler Watson test (WWT) is addressed as well. The power of the new test has been computed and compared against RWT and WWT powers based on the circular simulated data which have been generated from the von Mises distributions. The simulation study based on the power demonstrates that the proposed statistical test outperform classical tests.

2. Notation and setting

It is important that we set out the notation as well as the context that we use in this manuscript before we discuss the test. With regard to the motivating dataset, let $\phi_1, \phi_2, ..., \phi_n$ be n angles from the first group and $\psi_1, \psi_2, ..., \psi_m$ be m angles. Given this, our setting is as follows:

- We assume that the $\phi_i, i = 1, ..., n$ and $\psi_j, j = 1, ..., m$ are samples of random variables Φ and Ψ with cumulative distribution function G_Φ and G_Ψ, respectively.
- The test of equality of the distribution of angles between two groups is a test on two independent samples with n and m observations. These two samples are not paired, i.e. ϕ_1 is not paired with ψ_1.
- We are interested in testing the null hypothesis $H_0 : G_\Phi = G_\Psi$ vs $H_a : G_\Phi \neq G_\Psi$. The distance between two probability distribution is considered as the Cramer distance [2]

  $$\int_0^{2\pi} \{G_\Phi(\theta) - G_\Psi(\theta)\}^2 d\theta,$$

  which is equal to zero if $G_\Phi = G_\Psi$.

In the next section, we will discuss the test of equality of distribution of angles between two groups of circular data.

3. Angular Randomisation Test (ART)

Let $S_1 = \{\phi_1, \phi_2, ..., \phi_n\}$ and $S_2 = \{\psi_1, \psi_2, ..., \psi_m\}$ be two groups of circular data with probability distributions G_Φ and G_Ψ, respectively. We would like to test the null hypothesis

$$H_0 : G_\Phi = G_\Psi \quad \text{vs} \quad H_a : G_\Phi \neq G_\Psi.$$

The proposed statistic is defined by the following form:

$$\mathcal{T}_d(\phi_i, \psi_j) = \left(\frac{n+m}{nm}\right)^{\frac{-1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{m} D(\phi_i, \psi_j) \tag{3.1}$$

where $D(\phi_i, \psi_j)$ is a distance measure between the angles $\phi_i, i = 1, ..., n$ and $\psi_j, j = 1, ..., m$, and in this paper we use a geodesic distance measure which is given by

$$D(\phi_i, \psi_j) = \pi - |\pi - |\phi_i - \psi_j||.$$

The powerful of statistic depends on Geodesic distance which give the exact distance between observation on the circumference of the unit circle which makes it sensitive to change in values of populations. The factor $\left(\frac{n+m}{nm}\right)^{\frac{-1}{2}}$ showed stability in the results on the tests among several factors that were previously tested.

4. Simulation Study

In this section, a simulation study is performed to understand the working characteristics of the test of equality of distribution of two groups, which distinguishes whether the two groups shall have the same distributions or not. The purpose of this simulation study is two-fold, which are described below:

First, we are interested to understand whether the suggested test (ART) has an appropriate control of Type-I error rate (false positive). To do this, two groups have been generated under the null hypothesis that they are equal to identify the control of false positive rate. This simulation is critical to identify that when we say that the significance of the test is 0.05, then the test actually controls the probability of Type-I error rate at 0.05. In other words, we wish to confirm that the estimated p-value is accurate, as it is going to be the basis for a measure of dissimilarity between the distributions of groups.

Second, we are interested to identify that the test is able to distinguish two circular groups when they are truly from different distributions, and at what amount of differences that the test is able to distinguish the two. We perform a simulation study under the alternative hypothesis, where there is a difference in the distribution between two circular groups either in terms of mean or concentration parameter. Both purposes of the simulation can be constructed in a single simulation framework as described below:

1. Generate two samples $S_1 = \{\phi_1, \phi_2, ..., \phi_n\}$ and $S_2 = \{\psi_1, \psi_2, ..., \psi_m\}$ from von Mises distributions vM(μ_1, κ_1) and vM(μ_2, κ_2), respectively. The von Mises distribution is the most popular model for unimodal samples of circular data, which is a symmetric unimodal distribution (see Figure **??**). The probability density function (PDF) is defined as follows:

$$f(\vartheta \mid \mu, \kappa) = \frac{e^{\kappa \cos(\vartheta - \mu)}}{2\pi I_0(\kappa)}, \tag{4.1}$$

where

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} e^{k \cos(\vartheta)} d\vartheta$$

is the modified Bessel function of the first kind and order one. The concentration parameter is κ, and the mean direction is μ [3].

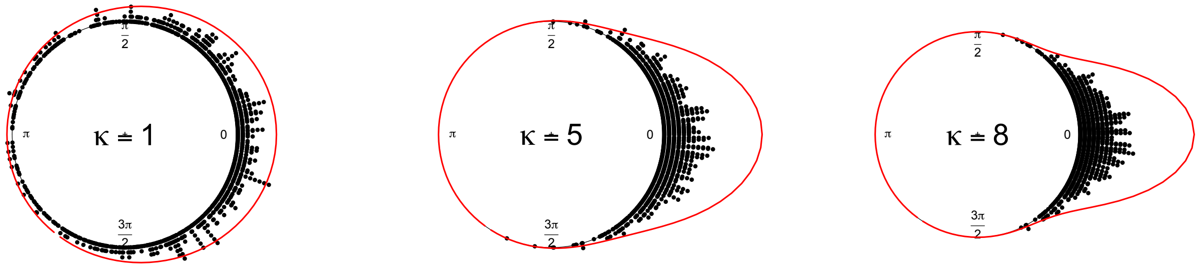

Figure 1: Examples for the distribution used in this analysis. The von Mises distribution over a range of concentration parameters (values given in the plots, $n = 1000$ for each).

2. For the original data calculate the value of the test statistic, the observed value ($\mathcal{T}_d^*$), using the proposed test statistic which is given by

$$\mathcal{T}_d(\phi_i, \psi_j) = \Big(\frac{n+m}{nm}\Big)^{\frac{-1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{m} D(\phi_i, \psi_j) \tag{4.2}$$

3. Shuffle and separate the data $\{\phi_1, \phi_2, ..., \phi_n, \psi_1, \psi_2, ..., \psi_m\}$ into two groups SS_1 and SS_2 with sizes n and m, respectively.

4. Recompute the test statistic (4.2) for the permuted sample and repeat the process until the distribution of the test statistic for many permutations (n_{rand}) is computed $\mathcal{T}_1^*, \mathcal{T}_2^*, ..., \mathcal{T}_{n_{rand}}^*$.

5. Test the null hypothesis of equal distributions $H_0 : F_\Phi = F_\Psi$ between the two groups and record the p-value of the test using the following formula

$$\frac{\sum_{t=1}^{n_{rand}} I(\mathcal{T}_t^* \geq \mathcal{T}_0^*)}{N_{rand}}, \tag{4.3}$$

   where $I(\cdot)$ is the function which is equal to one if the statement in the bracket is true and zero otherwise, and N_{rand} is the number of all possible permutations including the original sample.

6. When the mean is varied, m data points were generated from vM(μ, κ) for the second group and Steps 1–5 were repeated 1000 for each of $\mu \in \{0, 0.2, 0.4, 0.5, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.5\}$ so that there were 1000 p-values under each setting.

7. When the variance is varied, m data points were generated from vM(μ, κ) for the second group and Steps 1–5 were repeated 1000 for each of $\kappa \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3, 4\}$ so that we will have 1000 p-values under each setting.

5. Results

The results of the simulation study are presented in Figures 2 – 5, and Tables 1 and 2. The results indicates that the proposed ART test has good control of the type-I error rate (0.05) for the nominal 0.05 significance level. Figure 2 indicates that, as the concentration parameter increases (and the mean remains the same), the test is able to identify that the two distributions of circular groups are different. Figure 3 indicates that, as the mean difference increases (and concentration parameter remains the same), the test is also able to detect that the two distributions of circular groups are different.

The power of the ART test, WWT test, and WRT test have been computed and presented in Figures 2 – 5, and Tables 1 and 2. A comparison investigation was performed to see how well the suggested statistic versus Watson and Wheeler Watson statistics performed as shown in Figures 4 and 5.

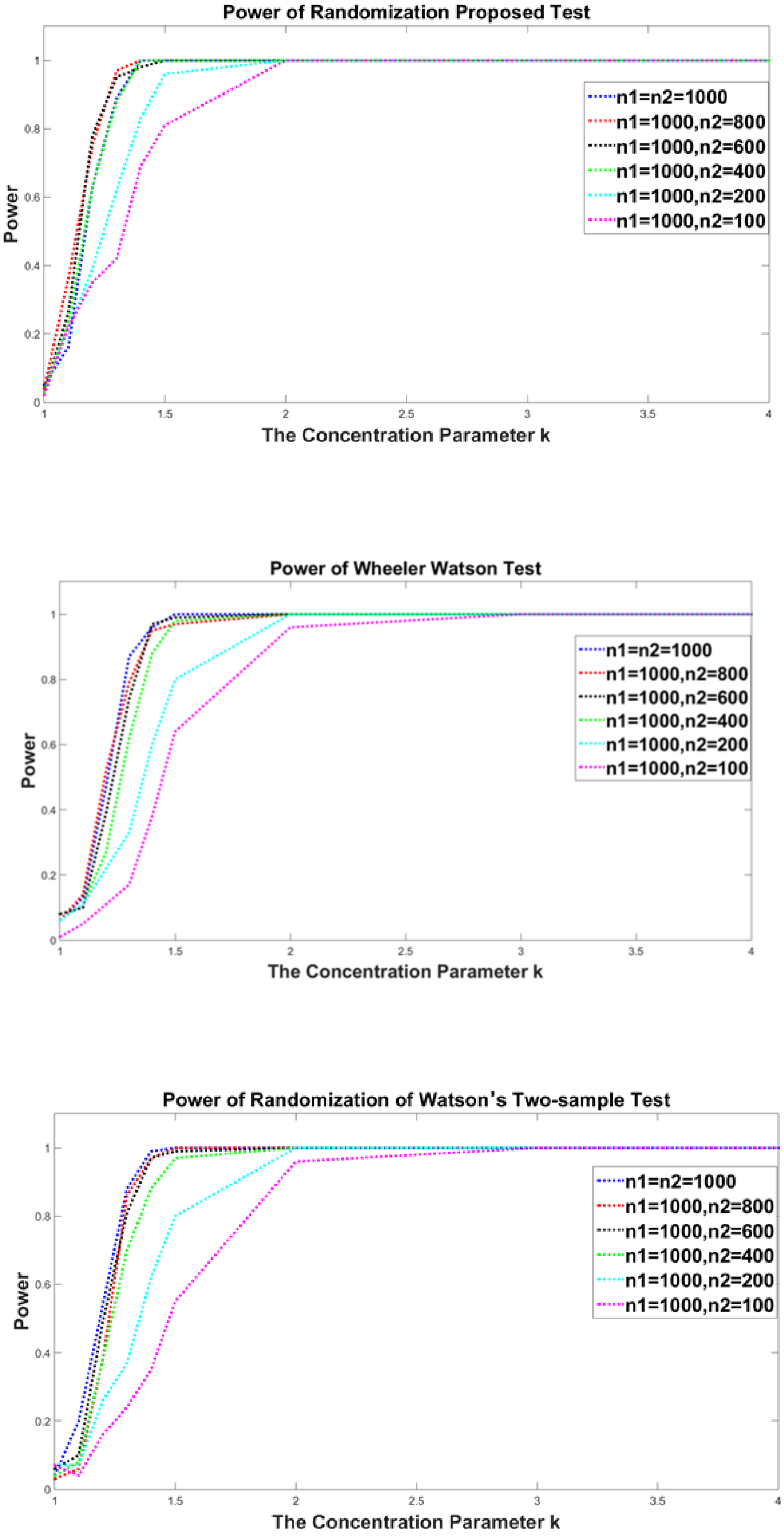

Figure 2: Sensitivity and Type-I error rate of the ART test (top-panel), WWT test (middle panel), RWT test (bottom-panel) between two simulated groups of angles as a function of concentration ratio for different number of observations in the second group: 100, 200, 400, 600, 800, 1000. The number of observations in the first group remains the same (1000). The value for concentration parameter one corresponds to type-I error rate, while the other values correspond to sensitivity.

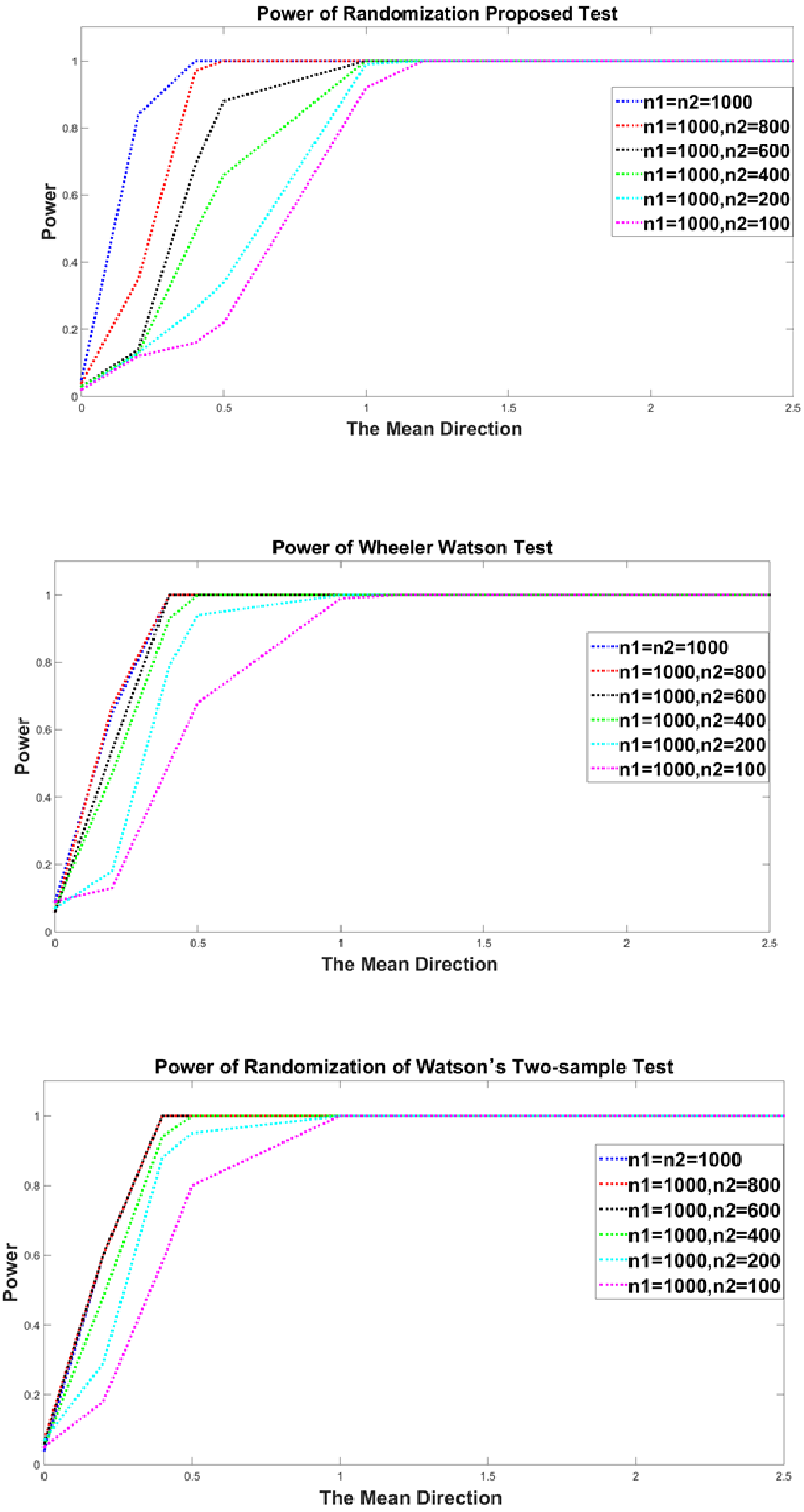

Figure 3: Sensitivity and Type-I error rate of the ART test (top-panel), WWT test (middle panel), RWT test (bottom-panel) between two simulated groups of angles as a function of mean difference and for different number of observations in the second group: 100, 200, 400, 600, 800, 1000, while the number of observations in the first group remains the same (1000). The value for mean difference zero corresponds to type-I error rate, while the other values correspond to sensitivity.

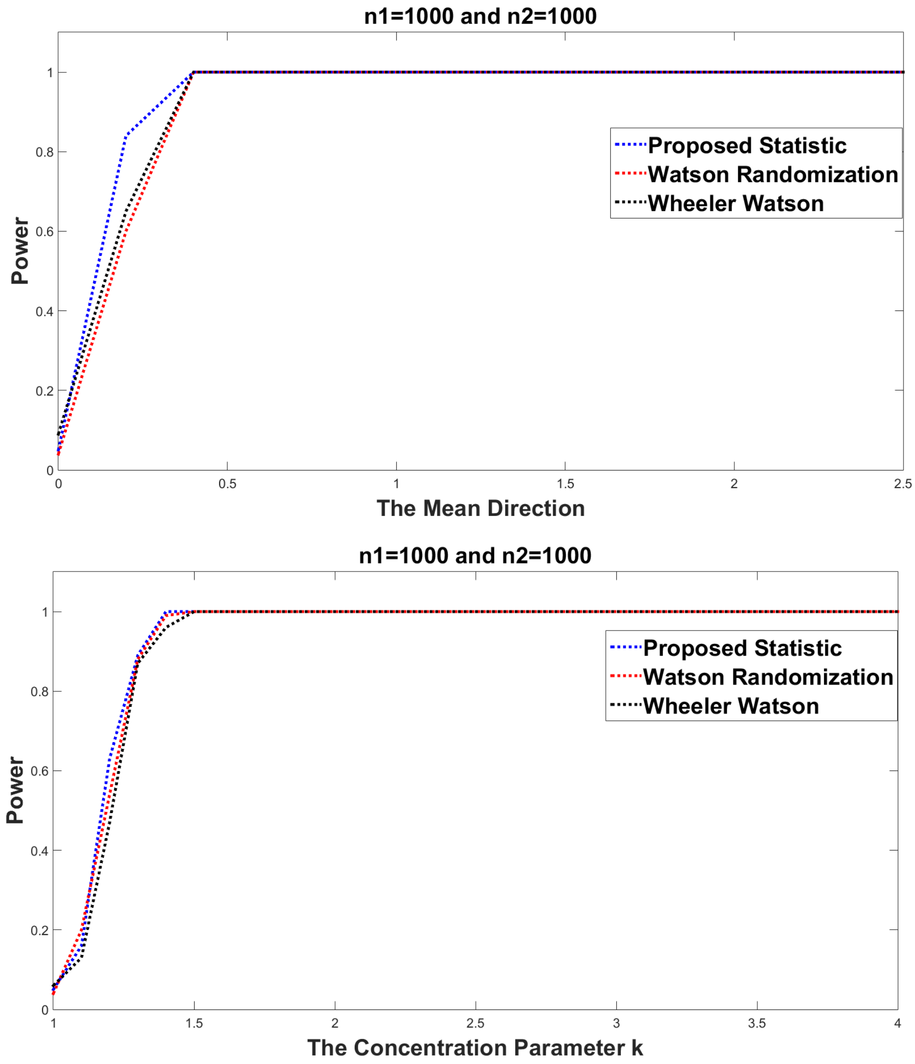

Figure 4: Comparison between the powers of the homogeneity tests ART, WWT, and WRT tests when the mean is varied $\mu \in \{0, 0.2, 0.4, 0.5, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.5\}$ (top-panel) and when the concentration parameters is varied $\kappa_2 \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3, 4\}$.

Table 1: Comparison between the power of the two-sample homogeneity tests: ART, WRT, and WWT for angular data which have been generated from v$M(\mu_1, \kappa_1)$ and v$M(\mu_2, \kappa_2)$, when $\mu_2 \in \{0, 0.2, 0.4, 0.5, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.5\}$ and with different sample sizes.

μ_1, μ_2	ART						WRT						WWT					
	m						m						m					
	100	200	400	600	800	1000	100	200	400	600	800	1000	100	200	400	600	800	1000
0, 0	0.02	0.01	0.03	0.03	0.04	0.05	0.05	0.07	0.05	0.06	0.07	0.04	0.09	0.07	0.07	0.06	0.06	0.09
0, 0.2	0.12	0.13	0.13	0.14	0.35	0.84	0.18	0.29	0.48	0.60	0.60	0.60	0.13	0.18	0.47	0.54	0.67	0.65
0, 0.4	0.16	0.26	0.49	0.69	0.97	1.00	0.58	0.88	0.94	1.00	1.00	1.00	0.50	0.79	0.93	1.00	1.00	1.00
0, 0.5	0.22	0.34	0.66	0.88	1.00	1.00	0.80	0.95	1.00	1.00	1.00	1.00	0.68	0.94	1.00	1.00	1.00	1.00
0, 1	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
0, 1.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0, 1.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0, 1.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0, 1.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0, 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0, 2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Comparison between the power of the two-sample homogeneity tests: ART, WRT, and WWT for angular data which have been generated from v$M(\mu_1, \kappa_1)$ and v$M(\mu_2, \kappa_2)$, when $\kappa_2 \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3, 4\}$ and with different sample sizes.

κ_1, κ_2	ART						WRT						WWT					
	m						m						m					
	100	200	400	600	800	1000	100	200	400	600	800	1000	100	200	400	600	800	1000
1, 1	0.02	0.02	0.03	0.03	0.04	0.05	0.07	0.07	0.04	0.06	0.03	0.04	0.01	0.06	0.06	0.08	0.06	0.06
1, 1.1	0.22	0.21	0.23	0.26	0.36	0.16	0.04	0.07	0.08	0.1	0.06	0.20	0.05	0.11	0.11	0.10	0.14	0.13
1, 1.2	0.35	0.39	0.63	0.78	0.75	0.63	0.16	0.26	0.38	0.49	0.38	0.54	0.11	0.22	0.27	0.39	0.52	0.47
1, 1.3	0.42	0.62	0.88	0.95	0.97	0.89	0.24	0.37	0.70	0.81	0.86	0.88	0.17	0.33	0.62	0.74	0.79	0.87
1, 1.4	0.69	0.83	1.00	0.98	1.00	1.00	0.35	0.62	0.88	0.97	0.97	0.99	0.38	0.60	0.88	0.97	0.95	0.96
1, 1.5	0.81	0.96	1.00	1.00	1.00	1.00	0.55	0.80	0.97	0.99	1.00	1.00	0.64	0.80	0.98	0.99	0.97	1.00
1, 2	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00
1, 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1, 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

6. Conclusion

This study proposes a novel free-distribution two-sample test which has the ability to detect the homogeneity between the distributions of two groups of circular data. A comparison between the proposed test (ART) against RWT and WWT is addressed as well. The power of the new homogeneity test has been computed under geodesic distance and compared against RWT and WWT powers based on the circular simulated data which have been generated from the von Mises distributions. The simulation study based on the power demonstrates that the proposed two-sample test has a good sensitivity and a proper control of type-I error.

References

[1] F.H. Clarke, Y.S. Ledyaev, R.J. Stern and P.R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer Science and Business Media, 2008.

[2] H.Cramér, *On the composition of elementary errors*, Scand. Actuarial J. 1928(1) (2011) 13—74.

[3] S.R. Jammalamadaka and A. Sengupta, *Topics in Circular Statistics*, Volume 5, World Scientific, 2001.

[4] L. Landler, G.D. Ruxton and E.P. Malkemper, *Circular data in biology: advice for effectively implementing statistical procedures*, Behav. Eco.Sociobio. 72 (2018) 1—10.

[5] K. Mardia, *A non-parametric test for the bivariate two-sample location problem*, J. Royal Stat. Soc. Ser. B 29 (1967) 320-–342.

[6] K. Mardia, *On Wheeler and Watson's two-sample test on a circle*, Indian J. Stat. Ser. A 31(2) (1969) 177-–190.

[7] K.V. Mardia and P.E. Jupp, *Directional Statistics*, volume 494, John Wiley & Sons, 2009.

[8] M.L. Rizzo and G.J. Székely, *Energy distance*, Wiley Interdisciplinary Reviews: Comput. Stat. 8 (2016) 27—38.

[9] G.J. Székely and M.L. Rizzo, *Energy statistics: A class of statistics based on distances*, J. Stat. Plan. Infer. 143 (2013) 1249—1272.

[10] S. Wheeler and G.S. Watson, *A distribution-free two-sample test on a circle*, Biometrika 51(1-2) (1964) 256–257.

[11] J.H. Zar, *Biostatistical Analysis*, Pearson, 2014.