Stat Papers
DOI 10.1007/s00362-015-0715-x

REGULAR ARTICLE

Multivariate tests of uniformity

Mengta Yang[1] · **Reza Modarres**[1]

Received: 21 September 2014 / Revised: 1 June 2015

Abstract We present tests of multivariate uniformity using data depth, the normal quantiles and the interpoint distances between the observations. We investigate the properties of the interpoint distances among uniform random vectors. We compare the performance of the proposed tests with two existing statistics under the hypothesis of uniformity and obtain their empirical power under various alternatives in a Monte Carlo study.

Keywords Data depth · Multivariate uniformity · Uniformity test · Interpoint distance

Mathematics Subject Classification 62H15 · 60E05

1 Introduction

Given n vectors of observations $\mathbf{x} \in \mathbb{R}^d$, a central question in several scientific fields, including ecology, physics, epidemiology, computer science and statistics, is determining whether the observations are generated from a uniform distribution with compact support $S \subseteq \mathbb{R}^d$. For example, one may examine whether a certain species is distributed uniformly for an ecological study (Wiegand and Moloney 2004). Uniformity tests provide indispensable tools for verifying the quality of pseudo-random number generators. The sets of k consecutive points from a good uniform generator should

✉ Reza Modarres
reza@gwu.edu

Mengta Yang
danto.yang@gmail.com

[1] Department of Statistics, George Washington University, Washington, DC, USA

Published online: 22 September 2015

not display any pattern in $\mathbb{R}^d$ (Marsaglia 1968). Physicists are interested in knowing whether the galaxies in the universe display any particular pattern (Barrow et al. 1985). In this article we focus on the unit hyper-cube $S = [0, 1]^d$ support.

A fundamental property in univariate statistics is the probability integral transformation, which states that $W = F^{-1}(X)$ is uniformly distributed on $(0, 1)$ if and only if X has distribution function (c.d.f.) F. It follows from probability integral transformation that testing $H_0 : X \sim F$ is equivalent to testing $H_0 : F^{-1}(X) \sim U(0, 1)$. The Rosenblatt transformation (Rosenblatt 1952) is a multivariate generalization of probability integral transformation. Stephens (1986) provides a comprehensive review of univariate tests of uniformity. More recently, Krumbholz and Schmid (1996) propose a χ^2 statistic for testing univariate uniformity, Schellhaas (1999) suggests a modified Kolmogorov–Smirnov test for a rectangular distribution with unknown parameters, and Pardo (2003) uses spacings to test univariate uniformity.

The notion of statistical depth functions provides a methodology for ordering multivariate observations and extending many of the univariate nonparametric methods to multivariate settings. A data depth function $D(\mathbf{t}; F)$ or $D(\mathbf{t}; F_n)$ measures the closeness or centrality of a point $\mathbf{t} \in \mathbb{R}^d$ with respect to the center of a distribution F or the empirical distribution F_n associated with a data cloud. It provides a center-outward ordering of multivariate observations in $\mathbb{R}^d$ with respect to the center of F or F_n. The larger the value of $D(\mathbf{t}; F)$ or $D(\mathbf{t}; F_n)$ is, the deeper or more central $\mathbf{t}$ is. Liu (1990) and Zuo and Serfling (2000) provide a framework for statistical depth functions. Liu et al. (1999) and Zuo and Serfling (2000) discuss recent developments of statistical depth functions for multivariate nonparametric analysis.

In the next section, we extend the uniformity tests of Hegazy and Green (1975) to the multivariate case by ranking the multivariate observations using data depth function. We propose a new test of multivariate uniformity based on the normal quantiles. We also consider the interpoint distances of observations that are uniformly distributed over the hyper-cube $[0, 1]^d$ and propose tests based on the first two moments of the interpoint distances. We compare the performance of the proposed test statistics with two existing statistics (distance to boundary test and the total edge weight of the minimal spanning tree) in a Monte Carlo study when $n > d$ and $n < d$. The last section is devoted to discussion and summary. The Appendix contains the proof for Theorem 1.

2 New multivariate uniformity tests

2.1 Depth-based Hegazy–Green test

A natural approach for testing multivariate uniformity is generalizing univariate tests based on quantiles and spacings using depth functions. We assume that the distribution function F is continuous so that there are no ties among the observations and their depth values. Consider n i.i.d. random vectors $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ in $\mathbb{R}^d$ with c.d.f F and let F_n be the associated empirical distribution function (e.d.f.). For some depth function $D(\cdot; \cdot)$ that is bounded above, let $Z_i = D(\mathbf{X}_i; F_n)$ be the depth of $\mathbf{X}_i$ with respect to the e.d.f. and $Z_{(i)}$ be the order statistics of Z_i's. The ordered sequence $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \ldots, \mathbf{X}_{(n)}$,

where $Z_{(i)} = D(\mathbf{X}_{(i)}; F_n)$, is referred to as depth induced order statistics and the p-th quantile of the sample induced by the depth function is $\mathbf{X}_{(\lfloor np \rfloor + 1)}$. We note that it is possible to have observations with tied sample depth values. Hence, there can be ties within the depth induced order statistics in an observed sample. Random tie-breaking schemes are employed in such cases. Let $T_i = P\big(\{\mathbf{x} : D(\mathbf{x}; F) \leq D(\mathbf{X}_i; F)\}\big)$. One can verify that $T_i \sim U(0, 1)$ (Li and Liu 2008).

Given n univariate observations $X_1, X_2, \ldots, X_n$, the univariate Hegazy–Green procedure regresses the order statistics $X_{(i)}$ against equidistant numbers t_i. We perform a simple linear regression with model $X_{(i)} = \alpha + \beta t_i + \epsilon_i$. Consider the case of testing $H_0 : X \sim U(0, 1)$ against $H_1 : X \nsim U(0, 1)$. Let $t_i = E[X_{(i)}] = \frac{i}{n+1}$. One may verify that under H_0, $\alpha = 0$ and $\beta = 1$. The residuals $e_i = X_{(i)} - \frac{i}{n+1}$ are statistics based on the characteristics of the fitted line and the univariate Hegazy–Green test with statistic $M^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$ is a test based on regression.

For testing multivariate uniformity, $H_0 : \mathbf{X} \sim U[0, 1]^d$ against $H_1 : \mathbf{X} \nsim U[0, 1]^d$, we replace X_i with

$$T_i^0 = P\left(\{\mathbf{x} : \mathrm{D}(\mathbf{x}; F) \leq \mathrm{D}(\mathbf{X}_i; F)\} \,|\, F = U[0, 1]^d\right)$$

and use the regression test coupled with depth functions. The residuals $e_i = T_{(i)}^0 - \frac{i}{n+1}$ lead to the generalized Hegazy–Green statistic

$$M^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{1}$$

Since T_i^0's are i.i.d. univariate $U(0, 1)$ random variables under H_0, the sampling distribution of M^2 under H_0 is exactly the same as the sampling distribution of the univariate version. Green and Hegazy (1976) fit the first four moments to those of Pearson distribution and a three parameters log-normal distribution. Both distributions provide good approximations of the distribution of M^2.

2.2 Uniformity test using normal quantiles

We propose a new uniformity test based on the relationship between multivariate uniforms and normals. Suppose we take n i.i.d. random vectors $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ in $\mathbb{R}^d$ from c.d.f. F_X. Let $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n\}$ where

$$\mathbf{Z}_i = (\Phi^{-1}(X_{i1}), \Phi^{-1}(X_{i2}), \ldots, \Phi^{-1}(X_{id}))'$$

and $\Phi(\cdot)$ is the c.d.f. of standard normal distribution. The test is carried out using the statistic

$$C_{\mathrm{N}} = n\bar{\mathbf{Z}}'\bar{\mathbf{Z}}. \tag{2}$$

Under $H_0 : F_X \sim U[0, 1]^d$ and $\mathbf{Z}_i$'s are $N(\mathbf{0}, \mathbf{I})$ random vectors. Hence, testing $H_0 : F_X \sim U[0, 1]^d$ against $H_1 : F_X \nsim U[0, 1]^d$ is equivalent to testing $H_0 : F_Z \sim N(\mathbf{0}, \mathbf{I})$

against $H_1 : F_Z \nsim N(\mathbf{0}, \mathbf{I})$. It follows from that $\bar{\mathbf{Z}} \sim N(\mathbf{0}, \frac{1}{n}\mathbf{I})$ and $n\bar{\mathbf{Z}}'\bar{\mathbf{Z}} \sim \chi^2_d$. We reject the uniformity hypothesis if $C_{\mathbf{N}}$ is large.

2.3 Interpoint distances of uniforms

The interpoint distances (IDs) constitute the engine of many multivariate techniques and find applications in comparisons of point patterns, correspondence analysis, multidimensional scaling, spatial clustering, clustering techniques, and testing the equality of distribution functions. Jammalamadaka and Janson (1986) consider the asymptotic properties of small IDs observed in a sample. Modarres (2014) investigates the moments and distribution of the interpoint distances between multivariate Bernoulli random vectors.

We will use the first two moments of the distribution of IDs to test the uniformity of n i.i.d. random vectors $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \in \mathbb{R}^d$. Our tests are based on the asymptotic distribution of the sample mean and variance of all interpoint distances,

$$\bar{\delta} = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} \|\mathbf{X}_i - \mathbf{X}_j\|^2 \tag{3}$$

$$\bar{S} = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} \left(\|\mathbf{X}_i - \mathbf{X}_j\|^2 - d/6\right)^2. \tag{4}$$

Theorem 1 *Under* $H_0 : \mathbf{X} \sim U[0,1]^d$, *one can show* $\mathrm{Var}(\bar{\delta}) = \frac{d(2n+3)}{90n(n-1)}$ *and*

$$\mathrm{Var}(\bar{S}) = \begin{cases} \frac{1}{\binom{n}{2}} \left(\frac{1}{56700}\left[989 + 202(n-2)\right]\right), & d = 2, \\ \frac{1}{\binom{n}{2}} \left(\frac{1}{1050}\left[37 + 6(n-2)\right]\right), & d = 3, \\ \frac{1}{\binom{n}{2}} \left(\frac{49d^2}{16200} + \frac{101d}{37800} + 2(n-2)\left[\frac{d^2}{16200} + \frac{29d}{37800}\right]\right), & d \geq 4. \end{cases} \tag{5}$$

Let $\sigma_{\bar{\delta}} = \sqrt{\mathrm{Var}(\bar{\delta})}$ and $\sigma_{\bar{S}} = \sqrt{\mathrm{Var}(\bar{S})}$. It follows from the Central Limit Theorem for U-processes (Arcones and Giné 1993) that $\frac{\bar{\delta} - d/6}{\sigma_{\bar{\delta}}} \xrightarrow{\mathcal{L}} N(0,1)$ and $\frac{\bar{S} - 7d/180}{\sigma_{\bar{S}}} \xrightarrow{\mathcal{L}} N(0,1)$ as $n \to \infty$. Let

$$Q_1 = \left(\bar{\delta} - d/6\right)^2 / \mathrm{Var}(\bar{\delta}), \tag{6}$$

$$Q_2 = \left(\bar{S} - 7d/180\right)^2 / \mathrm{Var}(\bar{S}). \tag{7}$$

One observes that $Q_1 \xrightarrow{\mathcal{L}} \chi^2_1$, $Q_2 \xrightarrow{\mathcal{L}} \chi^2_1$. Furthermore, the sampling covariance between the mean and any central moments of even order is zero to the order of $1/n$ provided the sample is drawn from a symmetrical population (Stuart and Ord 1994). Furthermore, the mean $\bar{\delta}$ and the variance $\bar{S}$ of asymototically normal variables are independent (Anderson 2003). Hence, $Q_3 = Q_1 + Q_2 \xrightarrow{\mathcal{L}} \chi^2_2$. We propose testing

multivariate uniformity using Q_1, Q_2 and Q_3 and reject the uniformity hypothesis for large values.

3 Monte Carlo study

In this section, we study the empirical power of the generalized Hegazy–Green test M^2, normal quantile test (C_N) and ID tests (Q_1, Q_2 and Q_3). For M^2, the depth induced order statistics are computed with respect to spherical, lens and random Tukey (1975) depth functions. Further details on these depth functions are in Sect. 3.3. The quantiles of the depth functions under $H_0 : \mathbf{X} \sim U[0,1]^d$ are obtained via parametric bootstrap. We compare the proposed uniformity tests against two existing tests in the literature.

3.1 Distance to boundary test

Let $\mathbf{X}$ be n i.i.d. random vectors in $\mathbb{R}^d$ with c.d.f. F and compact support S. Denote the boundary of S by ∂S. The distance to boundary of a point $\mathbf{u} \in S$ is defined to be $\Delta(\mathbf{u}, \partial S) = \min_{\mathbf{t}\in\partial S}(\|\mathbf{u}-\mathbf{t}\|)$, where $\|\mathbf{u}-\mathbf{t}\|$ denotes the Euclidean distance between the vectors $\mathbf{u}$ and $\mathbf{t}$. Berrendero et al. (2006) propose a test based on the statistics $Y_i = \frac{\Delta(\mathbf{X}_i, \partial S)}{R}$, where $R = \sup_{\mathbf{t}\in S}\big(\Delta(\mathbf{t}, \partial S)\big)$. They show that if $\{\mathbf{t} : B(\mathbf{t}, r) \subset S\} = \lambda_r S, \quad \forall \quad r \in (0, R)$, where λ_r is a constant depending only on r, then $Y_i \sim \text{Beta}(1, d)$ under $H_0 : \mathbf{X}$ is uniformly distributed over S. Berrendero et al. (2006) also prove that $dY \xrightarrow{\mathcal{L}} Exp(1)$ as $d \to \infty$ under H_0.

Berrendero et al. (2006) also consider several classes of supports, including the d-dimensional hyper-cubes centered at the origin $[-M, M]^d$ and the d-dimensional balls centered at the origin $B(0, M)$. In particular, they show that if $S = \prod_{j=1}^{d}[0, 2R_j]$, then Y_i are i.i.d. with c.d.f. $F_Y = 1 - \prod_{j=1}^{d}(1 - k_j y), \quad 0 \le y \le 1$ under H_0. Hence, testing $H_0 : \mathbf{X} \sim U[0,1]^d$ is equivalent to testing $H_0 : Y_i \sim \text{Beta}(1, d)$. Berrendero et al. (2006) suggest carrying out the test using Kolmogorov–Smirnov statistic. We denote this test by KS_{BCV} in our simulation studies.

3.2 Uniformity test based on minimum spanning tree

Consider a set of n random vectors in $\mathbb{R}^d$ and its minimal spanning tree (MST) where the edge weights are the Euclidean interpoint distances. Let L_{MST} be the total edge weight of the MST. Under certain regularity conditions including absolute continuity of density and compactness of support, L_{MST} is asymptotically normally distributed (Lee 1999). Petrie and Willemain (2013) suggest a test based on a result by Steele (1988) stating that if the vertices of the MST are uniformly distributed over some compact support S, then $\frac{1}{n^{(d-1)/d}} E[L_{MST}]$ converges to $\beta_d V_s^{1/d}$, where β_d is some constant that depends only on the dimension d and V_s is the volume of the support S, as $n \to \infty$.

For testing $H_0 : \mathbf{X} \sim U[0,1]^d$, the support is the unit hyper-cube and $V_s = 1$. Furthermore, Petrie and Willemain (2013) note that although Avram and Bertsimas (1992) have given the analytical expression for β_d and the variance of the asymptotic distribution can be derived, the computation is impractical for $d > 2$. Hence, they

propose using parametric bootstrap to obtain the sampling distribution of L_{MST} under H_0 and rejecting the uniformity hypothesis for extreme values of L_{MST}.

3.3 Comparisons

The Tukey depth satisfies many statistical properties as shown in Zuo and Serfling (2000). However, it is extremely computationally intensive and considered infeasible for high dimensional data. Cuesta-Albertos and Nieto-Reyes (2008) propose a stochastic approximation of the Tukey depth known as the random Tukey depth. The random Tukey depth of a point $\mathbf{t} \in \mathbb{R}^d$ with respect to a sample $\mathbf{x}$ of size n is obtained by projecting $\{\mathbf{t}, \mathbf{x}\}$ on the d-dimensional vectors $\mathbf{V}_i, i = 1, 2, \ldots, k$. The minimum of the k univariate Tukey depths is the random Tukey depth. It is shown that the random Tukey depth converges to the Tukey depth as $k \to \infty$. In our Monte Carlo study, the random Tukey depth is computed using the R package DD-procedure by Mozharovskyi et al. (2014) with $k = 3000$. The random projections are uniformly distributed on the (d-1) sphere, which is the surface of a d-dimensional ball.

The spherical depth $\mathrm{SPD}(\mathbf{t}; F)$ (Elmore et al. 2006) of a vector $\mathbf{t}$ in $\mathbb{R}^d$ with respect to distribution F is defined to be the probability that $\mathbf{t}$ is contained within the closed hypersphere $B(\frac{\mathbf{X}_1+\mathbf{X}_2}{2}, \frac{\|\mathbf{X}_1-\mathbf{X}_2\|}{2})$ where $\mathbf{X}_1$ and $\mathbf{X}_2$ are i.i.d. random vectors with c.d.f. F. The sample version of the spherical depth of a point $\mathbf{t} \in \mathbb{R}^d$ is defined as $\mathrm{SPD}(\mathbf{t}; F_n) = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} I[\mathbf{t} \in \mathrm{Sph}(\mathbf{X}_i, \mathbf{X}_j)]$, where $\mathrm{Sph}(\mathbf{X}_i, \mathbf{X}_j) = B(\frac{1}{2}(\mathbf{X}_i + \mathbf{X}_j), \frac{1}{2}\|\mathbf{X}_i - \mathbf{X}_j\|)$ and F_n is the e.d.f.

The lens depth $LD(\mathbf{t}; F)$ (Liu and Modarres 2011) of a vector $\mathbf{t}$ in $\mathbb{R}^d$ with respect to distribution F is defined to be the probability that $\mathbf{t}$ is contained within the closed hyper-lens $L(\mathbf{X}_1, \mathbf{X}_2) = B(\mathbf{X}_1, \|\mathbf{X}_1 - \mathbf{X}_2\|) \cap B(\mathbf{X}_2, \|\mathbf{X}_1 - \mathbf{X}_2\|)$ where $\mathbf{X}_1$ and $\mathbf{X}_2$ are i.i.d. random vectors with c.d.f. F. The sample lens depth of a point $\mathbf{t} \in \mathbb{R}^d$ is defined as $LD(\mathbf{t}; F_n) = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} I[\mathbf{t} \in L(\mathbf{X}_i, \mathbf{X}_j)]$.

We denote the generalized Hegazy–Green statistic based on the spherical depth, the lens depth and the random Tukey depth by M_S^d, M_L^d and M_T^d, respectively. We first investigate the α-levels of these tests and consider samples of sizes n=10, 20, 50 in $\mathbb{R}^d$, $d = 2$, 15 and 25. Table 1 displays the observed significance levels of the tests of uniformity at $\alpha = 0.05$.

We note that when the dimension is high, the sample random Tukey depth resulted in many ties such that the critical regions for the test statistics were not reliable even with 3000 repetitions. One main reason to this phenomenon is that all points exist on the surface of a d-dimensional simplex when $n \geq d$ and have depth valule of $1/n$. Hence, one should exert extra care when choosing the depth functions. To amend this issue, one may also compute the depth quantiles under H_0 using parametric bootstrap. We were able to get satisfactory results by computing the depth quantiles against a bootstrap sample of 10,000 draws. Similar behavior is also observed for the spherical depth and is amended in the same fashion. The remaining table entries are all close to $\alpha = 0.05$. While one can theoretically approximate the depth quantiles under H_0 arbitrarily well, we maintain that the extra computational cost incurred is still of critical importance and there are other well-behaved depth functions that can be computed fairly efficiently in high dimensional space, e.g. the zonoid depth (Mosler 2002).

Table 1 Empirical α-level of tests of uniformity: Multivariate Hegazy–Green tests using spherical (M_S^2), lens (M_L^2), and random Tukey (M_T^2) depth functions, normal quantiles (C_N), distance to boundary (KS_{BCV}), total MST weight (L_{MST}), and interpoint distances of uniforms (Q_1, Q_2, and Q_3)

n	d	M_S^2	M_L^2	M_T^2	$C_{\mathbf{N}}$	KS_{BCV}	L_{MST}	Q_1	Q_2	Q_3
10	2	0.056	0.058	0.069	0.060	0.040	0.059	0.048	0.054	0.067
	15	0.040	0.046	0.046	0.062	0.060	0.054	0.051	0.046	0.057
	25	0.040	0.048	0.040	0.064	0.072	0.052	0.048	0.042	0.052
25	2	0.052	0.062	0.048	0.064	0.046	0.052	0.046	0.040	0.069
	15	0.064	0.046	0.054	0.058	0.054	0.045	0.046	0.050	0.058
	25	0.024	0.040	0.054	0.066	0.056	0.054	0.047	0.048	0.051
50	2	0.056	0.054	0.046	0.074	0.054	0.035	0.045	0.045	0.074
	15	0.052	0.052	0.058	0.080	0.064	0.055	0.049	0.045	0.051
	25	0.038	0.086	0.046	0.072	0.044	0.041	0.053	0.044	0.056

Clearly, the necessary number of directions to approximate the Tukey depth with some given precision increases rapidly with the dimension d. Hence, it is recommended to use different number of directions for the dimensions. Mozharovskyi et al. (2014) and Lange et al. (2014) discuss the question of selecting directions in detail.

We consider two different classes of alternatives for our power study, dependence and shape. We consider dependence using copulas. A copula is a multivariate distribution with all univariate margins being $U(0, 1)$ (Joe 1997). For detecting dependence, we will examine the uniformity tests under the following bivariate copulas: the FGM (Farlie–Gumbel–Morgenstern) copula $C_\theta(u, v) = uv + \theta uv(1-u)(1-v)$ for $\theta \in [-1, 1]$, the Clayton copula $C(u, v)_\theta = \max[(u^{-\theta} + v^{-\theta} - 1), 0]^{-1/\theta}$ for $\theta \in [-1, \infty)$ and $\theta \neq 0$, the AMH (Ali–Mikhail–Haq) copula $C(u, v)_\theta = \frac{uv}{1-\theta(1-u)(1-v)}$, for $-1 \leq \theta < 1$, and the Plackett copula $C(u, v)_\theta = uv$ if $\theta = 0$ and $C(u, v)_\theta = \frac{1}{2}(\theta-1)A - \sqrt{A^2 - 4uv(\theta-1)}$, otherwise, where $A = 1+(u+v)(\theta-1)$. For more information on these copulas, we refer the readers to Nelson (2006). The performance of our tests for detecting dependencies are documented in Table 2. The samples are drawn from the $\mathrm{AMH}_{\theta=0.9}$ copula, $\mathrm{Plackett}_{\theta=5}$ copula, $\mathrm{FGM}_{\theta=1}$ copula and $\mathrm{Clayton}_{\theta=2}$ copula for several sample sizes $n = 10$, 25 and 50.

One observes that $C_{\mathbf{N}}$, Q_1 and KS_{BCV} all perform poorly against all copula alternatives while M_S^2 is very powerful against copula alternatives in all our experiments. On the other hand, M_L^2 and L_{MST} show excellent power against the $\mathrm{Clayton}_{\theta=2}$ alternative and moderate power against the $\mathrm{Plackett}_{\theta=5}$ alternative. One also notes that M_S^2 and M_L^2 outperforms L_{MST} against copula alternatives. Finally, the performances of Q_2 and Q_3 are comparable to those of M_L^2 and L_{MST} when the samples are drawn from $\mathrm{AMH}_{\theta=0.9}$, $\mathrm{Plackett}_{\theta=5}$ and $\mathrm{FGM}_{\theta=1}$. While they show slightly lower power against the $\mathrm{Clayton}_{\theta=2}$ alternative, Q_2 and Q_3 have the best small sample performance ($n = 10$) other than M_S^2.

In Table 3, we evaluate our tests against shape alternatives. Samples of size 50 are drawn from several different bivariate beta distributions the components of which are independent ($a = 0.5, 1, 2, 3$ and $b = 0.5, 1, 2, 3$). We note that $Beta(a, b) \stackrel{d}{=}$

Table 2 Empirical power of uniformity tests against copula alternatives

	n	M_S^2	M_L^2	M_T^2	$C_{\mathbf{N}}$	KS_{BCV}	L_{MST}	Q_1	Q_2	Q_3
$\mathrm{AMH}_{\theta=0.9}$	10	0.376	0.038	0.062	0.056	0.056	0.066	0.065	0.121	0.127
	25	0.328	0.118	0.054	0.068	0.062	0.112	0.066	0.170	0.164
	50	0.504	0.166	0.060	0.078	0.072	0.154	0.063	0.233	0.204
$\mathrm{FGM}_{\theta=1}$	10	0.672	0.046	0.096	0.060	0.044	0.044	0.055	0.090	0.094
	25	0.590	0.076	0.060	0.072	0.040	0.052	0.052	0.104	0.115
	50	0.390	0.072	0.05	0.062	0.040	0.094	0.049	0.126	0.127
$\mathrm{Clayton}_{\theta=2}$	10	0.384	0.016	0.078	0.088	0.078	0.164	0.097	0.257	0.237
	25	0.638	0.472	0.076	0.074	0.136	0.592	0.101	0.427	0.370
	50	0.984	0.850	0.060	0.090	0.194	0.894	0.098	0.640	0.566
$\mathrm{Plackett}_{\theta=5}$	10	0.572	0.026	0.064	0.078	0.051	0.082	0.078	0.162	0.153
	25	0.414	0.170	0.046	0.072	0.078	0.152	0.076	0.234	0.210
	50	0.632	0.356	0.038	0.086	0.082	0.270	0.071	0.349	0.295

Table 3 Empirical power of uniformity tests against bivariate Beta alternatives

α	β	M_S^2	M_L^2	M_T^2	$C_{\mathbf{N}}$	KS_{BCV}	L_{MST}	Q_1	Q_2	Q_3
0.5	0.5	0.140	0.356	0.472	0.268	0.998	0.106	0.997	0.999	0.999
	1	0.330	0.242	0.182	1.000	0.976	0.254	0.184	0.415	0.386
	2	0.950	0.698	0.09	1.000	1.000	0.998	0.998	0.951	0.991
	3	0.996	0.776	0.086	1.000	1.000	1.000	1.000	1.000	1.000
1	1	0.056	0.054	0.044	0.056	0.056	0.030	0.048	0.042	0.074
	2	0.124	0.254	0.018	1.000	0.066	0.856	0.971	0.495	0.965
	3	0.374	0.456	0.07	1.000	0.426	1.000	1.000	0.221	1.000
2	2	0.262	0.222	0.07	0.030	0.992	0.880	1.000	0.949	1.000
	3	0.172	0.314	0.096	0.806	0.998	0.998	1.000	1.000	1.000
3	3	0.166	0.426	0.15	0.030	1.000	1.000	1.000	0.544	1.000

$1 - Beta(b, a)$ and only show the result for $a \leq b$. One observes that $C_{\mathbf{N}}$ is weak against symmetric alternatives while being extremely powerful against asymmetric alternatives. On the other hand, M_S^2 and M_L^2 are reasonably powerful against all shape alternatives. Finally, both Q_1 and Q_3 show excellent power while Q_2 performs poorly against skewed alternatives. Overall, L_{MST}, KS_{BCV}, Q_1 and Q_3 seem to be the best choices against shape alternatives.

4 Discussion and summary

In this paper, we propose several new tests of multivariate uniformity. We study their performance under the null and against several alternatives. We note that the sample size $n = 10$ is in general too small for all tests in our study to detect the dependency

structure of the copula alternatives with the exceptions of M_S^2, Q_2 and Q_3. Overall, we suggest M^2, $C_{\mathbf{N}}$, Q_3 and L_{MST} for testing multivariate uniformity. We also note that the power of $C_{\mathbf{N}}$ against alternatives that are symmetric could be further improved by using a two-tailed test while sacrificing power against asymmetric alternatives. Finally, one observes that the same test statistic can display different behaviors when different depth functions are used. In particular, we see in our experiments that spherical depth function often leads to test statistics of higher power while lens depth leads to test statistics that are better behaved for high dimensional data.

Acknowledgments The authors would like to thank two anonymous referees and the Editor for comments that led to an improved manuscript.

Appendix

Proof for Theorem 1 Let $e_{i,j}^2 = \|\mathbf{u}_i - \mathbf{u}_j\|^2$. Consider

$$\begin{aligned} \bar{\delta} &= \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} e_{i,j}^2 \\ \bar{S} &= \frac{1}{\binom{n}{2}} \sum_{i<j} \left(e_{i,j}^2 - \frac{d}{6}\right)^2 \end{aligned} \tag{8}$$

and

$$\begin{aligned} \mathrm{Var}(\bar{\delta}) &= \frac{1}{\binom{n}{2}} \left[\mathrm{Var}\left(e_{i,j}^2\right) + 2(n-2)\mathrm{Cov}\left(e_{i,j}^2, e_{i,k}^2\right)\right] \\ \mathrm{Var}(\bar{S}) &= \frac{1}{\binom{n}{2}} \left[\mathrm{Var}\left(\left(e_{i,j}^2 - \frac{d}{6}\right)^2\right) + 2(n-2)\mathrm{Cov}\left(\left(e_{i,j}^2 - \frac{d}{6}\right)^2, \left(e_{i,k}^2 - \frac{d}{6}\right)^2\right)\right]. \end{aligned} \tag{9}$$

From Eq. 9, we have

$$\mathrm{Var}\left(\left(e_{i,j}^2 - \frac{d}{6}\right)^2\right) = \mathrm{Var}\left(e_{i,j}^4 - \frac{d}{3}e_{i,j}^2\right) = E^2\left[e_{i,j}^4 - \frac{d}{3}e_{i,j}^2\right] - E[e_{i,j}^4 - \frac{d}{3}e_{i,j}^2]^2. \tag{10}$$

Let $\epsilon_{ijkh} = (u_{i,j} - u_{k,h})$. One observes,

$$\mathrm{Var}(\bar{\delta}) = \frac{1}{\binom{n}{2}} \left[d\mathrm{Var}(\epsilon_{1121}^2) + 2d(n-2)\mathrm{Cov}\left(\epsilon_{1121}^2, \epsilon_{1131}^2\right)\right] = \frac{d(2n+3)}{90n(n-1)}. \tag{11}$$

Furthermore,

$$E[e_{i,j}^4 - \frac{d}{3}e_{i,j}^2] = E\left[\left(\sum_{i=1}^{d}\epsilon_{1i2i}^2\right)^2\right] - \frac{d^2}{18} = E\left[\sum_{i=1}^{d}\epsilon_{1i2i}^4 + 2\sum_{i<j}^{d}\epsilon_{1i2i}^2\epsilon_{1j2j}^2\right] - \frac{d^2}{18}$$
$$= dE\left[\epsilon_{1i2i}^4\right] + d(d-1)E\left[\epsilon_{1i2i}^2\right]^2 - \frac{d^2}{18} = \frac{d}{15} + \frac{d(d-1)}{36} - \frac{d^2}{18}$$
$$= \frac{7d}{180} - \frac{d^2}{36} \tag{12}$$

and

$$E^2\left[e_{i,j}^4 - \frac{d}{3}e_{i,j}^2\right] = E\left[e_{i,j}^8 - \frac{2d}{3}e_{i,j}^6\right] + \frac{d^2}{9}\left(\frac{d}{15} + \frac{d(d-1)}{36}\right). \tag{13}$$

One can verify that

$$E\left[e_{i,j}^8\right] = \begin{cases} \frac{1}{1296}d^4 + \frac{7}{1080}d^3 + \frac{2789}{226800}d^2 + \frac{101}{37800}d, & d \geq 4, \\ \frac{2}{15} + \frac{1}{7} + \frac{2}{25}, & d = 3, \\ \frac{2}{45} + \frac{2}{75} + \frac{1}{21}, & d = 2. \end{cases} \tag{14}$$

Similarly,

$$E\left[e_{i,j}^6\right] = \begin{cases} \frac{1}{216}d^3 + \frac{7}{360}d^2 + \frac{11}{945}d, & d \geq 3, \\ \frac{1}{14} + \frac{1}{15}, & d = 2. \end{cases} \tag{15}$$

It follows from Eqs. 10 to 15,

$$\mathrm{Var}\left(\left(e_{i,j}^2 - \frac{d}{6}\right)^2\right) = \begin{cases} \frac{989}{56700}, & d = 2, \\ \frac{37}{1050}, & d = 3, \\ \frac{49}{16200}d^2 + \frac{101}{37800}d, & d \geq 4. \end{cases} \tag{16}$$

The covariance term in Eq. 9 can be expanded as

$$\mathrm{Cov}\left(\left(e_{i,j}^2 - \frac{d}{6}\right)^2, \left(e_{i,k}^2 - \frac{d}{6}\right)^2\right) = \mathrm{Cov}\left(e_{i,j}^4 e_{i,k}^4\right) - \frac{2d}{3}\mathrm{Cov}\left(e_{i,j}^4, e_{i,k}^2\right) + \frac{d^3}{1620}. \tag{17}$$

One can verify

$$E\left[e_{i,j}^4 e_{i,k}^2\right] = E\left[\left(\sum_{i=1}^{d}\epsilon_{1i2i}^2\right)^2 \sum_{j=1}^{d}\epsilon_{1j3j}^2\right]$$
$$= dE\left[\epsilon_{1131}^2 \sum_{i=1}^{d}\epsilon_{1i2i}^4\right] + 2dE\left[\epsilon_{1131}^2 \sum_{i<k}^{d}\epsilon_{1i2i}^2\epsilon_{1k2k}^2\right]. \tag{18}$$

One observes,

$$dE\left[\epsilon_{1131}^2\sum_{i=1}^{d}\epsilon_{1i2i}^4\right]=d\left(E\left[\epsilon_{1131}^2\epsilon_{1121}^4\right]+(d-1)E\left[\epsilon_{1131}^2\right]E\left[\epsilon_{1222}^4\right]\right), \quad (19)$$

where $E[\epsilon_{1131}^2\epsilon_{1121}^4]=\frac{19}{1260}$. Similarly,

$$2dE\left[\epsilon_{1131}^2\sum_{i<k}^{d}\epsilon_{1i2i}^2\epsilon_{1k2k}^2\right]=2d\left((d-1)E[\epsilon_{1131}^2\epsilon_{1121}^2]E[\epsilon_{1121}^2]+\binom{d-1}{2}E[\epsilon_{1121}^2]^3\right), \quad (20)$$

where $E[\epsilon_{1131}^2\epsilon_{1121}^2]=\frac{1}{30}$. It follows from Eqs. 18 to 20 that

$$\mathrm{Cov}\left(e_{i,j}^4,e_{i,k}^2\right)=E\left[e_{i,j}^4e_{i,k}^2\right]-E\left[e_{i,j}^4\right]E\left[e_{i,k}^2\right]=\frac{d^2}{540}+\frac{2d}{945}. \quad (21)$$

One can also verify

$$E\left[e_{i,j}^4e_{i,k}^4\right]=E\left[\sum_{i=1}^{d}\epsilon_{1i2i}^4\sum_{i=1}^{d}\epsilon_{1i3i}^4+2\sum_{i=1}^{d}\epsilon_{1i3i}^4\sum_{i<j}^{d}\epsilon_{1i2i}^2\epsilon_{1j2j}^2 +2\sum_{i=1}^{d}\epsilon_{1i2i}^4\sum_{i<j}^{d}\epsilon_{1i3i}^2\epsilon_{1j3j}^2+4\sum_{i<j}^{d}\epsilon_{1i2i}^2\epsilon_{1j2j}^2\sum_{i<j}^{d}\epsilon_{1i3i}^2\epsilon_{1j3j}^2\right]. \quad (22)$$

Consider the first term in Eq. 22. One observes

$$E\left[\sum_{i=1}^{d}\epsilon_{1i2i}^4\sum_{i=1}^{d}\epsilon_{1i3i}^4\right]=dE\left[\epsilon_{1121}^4\epsilon_{1131}^4\right]+(d-1)E\left[\epsilon_{1121}^4\right]^2=\frac{23d}{3150}+\frac{d(d-1)}{225}. \quad (23)$$

Similarly,

$$E\left[\sum_{i=1}^{d}\epsilon_{1i3i}^4\sum_{i<j}^{d}\epsilon_{1i2i}^2\epsilon_{1j2j}^2\right]=E\left[\sum_{i=1}^{d}\epsilon_{1i2i}^4\sum_{i<j}^{d}\epsilon_{1i3i}^2\epsilon_{1j3j}^2\right] =\frac{19d(d-1)}{7560}+\frac{d(d-1)(d-2)}{1080} \quad (24)$$

by symmetry. Consider the last term in Eq. 22. One observes

$$E\left[\sum_{i<j}^{d}\epsilon_{1i2i}^2\epsilon_{1j2j}^2\sum_{i<j}^{d}\epsilon_{1i3i}^2\epsilon_{1j3j}^2\right]=\binom{d}{2}E\left[\epsilon_{1121}^2\epsilon_{1222}^2\sum_{i<j}^{d}\epsilon_{1i3i}^2\epsilon_{1j3j}^2\right]$$
$$=\binom{d}{2}E\left[\epsilon_{1121}^2\epsilon_{1222}^2\left\{\epsilon_{1131}^2\epsilon_{1232}^2+\epsilon_{1131}^2\sum_{2<j}\epsilon_{1j3j}^2\right.\right.$$
$$\left.\left.+\epsilon_{1232}^2\sum_{2<j}\epsilon_{1j3j}^2+\sum_{2<i<j}\epsilon_{1i3i}^2\epsilon_{1j3j}^2\right\}\right]. \quad (25)$$

From Eqs. 22 to 25, we have

$$\mathrm{Cov}\left(e_{i,j}^4,e_{i,k}^4\right)=E\left[e_{i,j}^4e_{i,k}^4\right]-E\left[e_{i,j}^4\right]E\left[e_{i,k}^4\right]$$
$$=\begin{cases}\frac{701}{56700}, & d=2,\\ \frac{29}{900}, & d=3,\\ \frac{d^3}{1620}+\frac{167d^2}{113400}+\frac{29d}{37800}, & d\geq 4,\end{cases} \quad (26)$$

Equation 17 becomes

$$\mathrm{Cov}\left(\left(e_{i,j}^2-\frac{d}{6}\right)^2,\left(e_{i,k}^2-\frac{d}{6}\right)^2\right)=\begin{cases}\frac{101}{56700}, & d=2,\\ \frac{1}{350}, & d=3,\\ \frac{d^2}{16200}+\frac{29d}{37800}, & d\geq 4.\end{cases} \quad (27)$$

and

$$\mathrm{Var}(\bar{S})=\begin{cases}\frac{1}{\binom{n}{2}}\left(\frac{1}{56700}\left[989+202(n-2)\right]\right), & d=2,\\ \frac{1}{\binom{n}{2}}\left(\frac{1}{1050}\left[37+6(n-2)\right]\right), & d=3,\\ \frac{1}{\binom{n}{2}}\left(\frac{49d^2}{16200}+\frac{101d}{37800}+2(n-2)\left[\frac{d^2}{16200}+\frac{29d}{37800}\right]\right), & d\geq 4.\end{cases} \quad (28)$$

References

Anderson TW (2003) An introduction to multivariate statistical analysis. Wiley-Interscience, Hoboken

Arcones MA, Giné E (1993) Limit theorems for u-processes. Ann Probab 21:1494–1542

Avram F, Bertsimas D (1992) The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach. Ann Appl Probab 2:113–130

Barrow J, Bhavsar S, Sonoda D (1985) Minimal spanning trees, filaments and galaxy clustering. R Astron Soc Mon Not 216:17–35

Berrendero JR, Cuevas A, Vázquez-Grande F (2006) Testing multivariate uniformity: the distance-to-boundary method. Can J Stat 34:693–707

Cuesta-Albertos JA, Nieto-Reyes A (2008) The random Tukey depth. J Comput Stat Data Anal 52:4979–4988

Elmore RT, Hettmansperger TP, Xuan F (2006) Spherical data depth and a multivariate median. In: Liu RY, R Serfling, DL Souvaine (eds) Proceedings of data depth: robust multivariate analysis, computational geometry and applications. American Mathematical Society, Rhode Island, pp 87–101

Green JR, Hegazy YA (1976) Powerful modified-EDF goodness-of-fit tests. J Am Stat Assoc 71:204–209

Hegazy YAS, Green JR (1975) Some new goodness-of-fit tests using order statistics. J R Stat Soc 24:299–308

Jammalamadaka SR, Janson S (1986) Limit theorems for a triangular scheme of U-statistics with applications to inter-point distances. Ann Probab 14:1347–1358

Joe H (1997) Multivariate models and dependence concepts. Chapman & Hall, New York

Krumbholz W, Schmid F (1996) A non standard χ^2 test of fit for testing uniformity with unknown limits. Stat Pap 37(4):365–373

Lange T, Mosler K, Mozharovskyi P (2014) DD-classification of asymmetric and fat-tailed data. In: Spiliopoulou M, Schmidt-Thieme L, Janning R (eds) Data analysis. Machine learning and knowledge discovery. Springer, Berlin, pp 71–78

Lee S (1999) The central limit theorem for Euclidean minimal spanning trees ii. Adv Appl Probab 31:969–984

Li J, Liu RY (2008) Multivariate spacings based on data depth. I. Construction of nonparametric multivariate tolerance regions. Ann Stat 36:1299–1323

Liu RY (1990) On a notion of data depth based on random simplices. Ann Stat 18:405–414

Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann Stat 27:783–858

Liu Z, Modarres R (2011) Lens data depth and median. J Nonparametr Stat 23:1063–1074

Marsaglia G (1968) Random numbers fall mainly in the planes. Proc Natl Acad Sci 61:25–28

Modarres R (2014) On the interpoint distances of Bernoulli vectors. Stat Probab Lett 84:215–222

Mosler K (2002) Multivariate dispersion, central regions and depth: the lift zonoid approach. Springer, New York

Mozharovskyi P, Mosler K, Lange T (2014) Classifying real-world data with the DD-procedure. Adv Data Anal Classif 9:287–314 (Springer online-first)

Nelson RB (2006) An introduction to copulas, 2nd edn. Springer, New York

Pardo MC (2003) A test for uniformity based on informational energy. Stat Pap 44(4):521–534

Petrie A, Willemain TR (2013) An empirical study of tests for uniformity in multidimensional data. Comput Stat Data Anal 64:253–268

Rosenblatt M (1952) Remarks on multivariate transformation. Ann Math Stat 23:470–472

Schellhaas H (1999) A modified Kolmogorov–Smirnov test for a rectangular distribution with unknown parameters: computation of the distribution of the test statistic. Stat Pap 40(3):343–349

Steele J (1988) Growth rates of Euclidean minimal spanning trees with power-weighted edges. Ann Probab 16:1767–1787

Stephens MA (1986) Test for the uniform distribution. In: D'Agostino RB, Stephens MA (eds) Goodness-of-fit techniques. Marcel Dekker, Inc, New York, pp 331–336

Stuart A, Ord K (1994) Kendall's advanced theory of statistics, distribution theory, vol 1, 6th edn. Oxford University Press, New York

Tukey JW (1975) Mathematics and picturing data. In: James RD (ed) Proceedings of the international congress on mathematics. Canadian Mathematical Congress, Montreal, pp 523–531

Wiegand T, Moloney K (2004) Rings, circles, and null-models for point pattern analysis in ecology. OIKOS 104:209–229

Zuo Y, Serfling R (2000) General notions of statistical depth function. Ann Stat 28:461–482