11-10-2011

Statistical Test for Multi-dimensional Uniformity

Tieyong Hu
Florida International University, thu001@fiu.edu

DOI: 10.25148/etd.FI11120504

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

STATISTICAL TEST FOR MULTI-DIMENSIONAL UNIFORMITY

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

by

Tieyong Hu

2011

To: Dean Kenneth Furton
College of Arts and Sciences

This thesis, written by Tieyong Hu and entitled Statistical Test for Multi-dimensional Uniformity, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

Sneh Gulati

Gauri L. Ghai

Zhenmin Chen, Major Professor

Date of Defense: November 10, 2011

The thesis of Tieyong Hu is approved.

Dean Kenneth Furton
College of Arts and Sciences

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2011

ACKNOWLEDGMENTS

I would like to express my sincere thanks to Dr. Zhenmin Chen for giving me the opportunity to pursue a Master's degree and for his support, guidance and patience throughout this research.

I would like to thank all those who gave me help to complete this thesis. In particular, I would like to thank Dr. Sneh Gulati and Dr. Guari L. Ghai for their time to serve on my committee.

Finally, I am eternally grateful to my wife, Ting, for her constant love and strength throughout the years.

ABSTRACT OF THE THESIS

STATISTICAL TEST FOR MULTI-DIMENSIONAL UNIFORMITY

by

Tieyong Hu

Florida International University, 2011

Miami, Florida

Professor Zhenmin Chen, Major Professor

Testing uniformity in the univariate case has been studied by many researchers. Many papers have been published on this issue, whereas the multi-dimensional uniformity test seems to have received less attention in the literature. A new test statistic for the multi-dimensional uniformity is proposed in this thesis. The proposed test statistic can be used to test whether an underlying multivariate probability distribution differs from a multi-dimensional uniform distribution. Some important properties of the proposed test statistic are discussed. As a special case, the bivariate test statistic is discussed in detail and the critical values of test statistic are obtained. By performing Monte Carlo simulation, the power of the new test is compared with the Distance to Boundary test, which was a recently proposed statistical test for multi-dimensional uniformity by Berrendero, Cuevas and Vazquez-Grande (2006). It has been shown that the test proposed in this thesis is more powerful than the Distance to Boundary test in some cases.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Testing uniformity in the univariate case has been studied by many researchers. Many papers have been published on this issue, whereas the multi-dimensional uniformity test seems to have received less attention in the literature. The problem of testing whether the pattern of points in the multi-dimensional space is distributed uniformly has applications in biology, astronomy, computer science and some other fields.

A commonly used goodness-of-fit test for uniformity is the chi-square test proposed by Pearson (1900). Theoretically, the chi-square test can be applied for any multivariate distribution test. However, the problem for the chi-square test is how to determine the cell limits. Two other well-known methods are the Kolmogorov-Smirnov test proposed by Kolmogorov (1933) and Smirnov (1939) and the Cramer-von Mises test proposed by Cramer (1952) and von Mises. Those two methods are derived from the empirical distribution and are widely used for univariate case. However, Berrendero, Cuevas and Vazquez-Grande (2006) pointed out that the probability distributions of those multivariate statistics don not have distribution free property. Therefore, those two methods are not applied for high dimensional distributions.

Justel, Pena and Zamar (1997) proposed a multivariate goodness-of-fit test based on the Kolmogorov-Smirnov test. By using Rosenblatt's transformation, they reduce the Kolmogorov-Smirnov test from multivariate case to univariate case. The test has distribution free property and can be applied to any dimensional case. However, the computation of the test is complicated especially for over two dimensions. In their paper, the authors presented bivariate statistic test (d=2) and performed power study. It showed that the test is powerful when sample size is moderately large.

In the paper by Liang, Fang, Hichernell and Li (2001), several statistical tests are proposed for testing uniformity in multivariate case. Those tests are use number-theoretic and quasi-Monte Carlo methods for measuring the discrepancy of the point in multidimensional unit. They proved that the test statistics can be approximated by the standard normal distribution $N(0,1)$ or the chi-squared distribution $\chi^2(2)$ under the null hypothesis that the underlying distribution is uniformly distributed. The Monte Carlo simulation is performed for three special cases: the symmetric discrepancy, the centered discrepancy and the star discrepancy. They take the sample size as n=25, 50, 100, 200. Since the sample size is moderately large, the approximation is quite reasonable. It is shown that the power will increase as the dimension increases.

Berrendero, Cuevas and Vazquez-Grande (2006) proposed a test based on the distance to the boundary. They define the distance from a point X to the boundary as $D(X,\partial S) = \min\{x \in \partial S : \|x - X\|\}$. They also define R as the maximum distance to the boundary which can be obtained in support S. Then, they get the relative distance from X to the boundary which is $Y = D(X,\partial S)/R$. In the paper, they prove that the relative depth will follow the beta distribution with parameters 1 and p (p is dimensional size) if the data are uniformly distributed on compact support. In the two dimensional case, the relative depth Y will follow the beta distribution $Beta(1,2)$ and it also makes sense in univariate case. In univariate case, the relative depth Y will follow the beta distribution $Beta(1,1)$ which is also the uniform distribution. Then, the test will be reduced from p dimensional case to one dimensional case. Finally they used Kolmogorov-Smirnov method to test whether the distribution of relative depth Y, which is from underlying distribution sample, follows the corresponding beta distribution. By performing the Monte Carlo simulation, under the same situations (same sample size and same dimension), the Distance to Boundary test is more powerful in the case of mixture model than the test proposed by Liang, Fang, Hichernell and Li (2001). The other advantage of the Distance to Boundary test is easy to compute and easy to explain.

Chen and Ye (2009) developed an alternative test for uniformity in univariate case. In that paper, the authors proposed a new test statistic based on the order statistics in support set [0, 1]. The test statistic proposed in the paper is

$$G\left(X_1, X_2, \cdots, X_n\right) = \frac{(n+1)}{n} \sum_{i=1}^{n+1} (X_{(i)} - X_{(i-1)} - \frac{1}{n+1})^2 .$$

By performing the Monte Carlo simulation, it has been shown that the test is more powerful when the alternative distribution is V-shape distribution and when the sample size is small compared with the Kolmogorov-Smirnov test. By using the probability integral transformation, the uniformity test can be used to check whether the underlying distribution follows any specified distribution. The idea is adopted in this research to develop a test for the multi-dimensional case.

The main purpose of this research is to propose a new test statistic for testing multi-dimensional uniformity. It has been shown that the newly proposed test improves the power of the multi-dimensional uniformity tests. Since the Distance to Boundary test is a recently published paper in multivariate uniformity test, the power of the proposed test is compared with the power of the Distance to Boundary test when different alternative distributions are used.

The thesis is organized as follow. In Section 2, we propose a new test statistic and discuss some important properties of the proposed test statistic. As a special case, the bivariate statistic test will be discussed in detail. In Section 3, the performance of the proposed test is evaluated using Monte Carlo simulation. Critical values of the new test statistic are obtained and tabulated. Power study is conducted for evaluating the test statistic. Several different alternative distributions are used in the power study. The final conclusions are given in Section 4.

Monte Carlo simulation is performed to find the critical values of the new test statistic and to conduct power study. The SAS/IML and SAS/Base are used for statistical simulation.

2. NEW TEST STATISTIC

2.1 General Case

In this research, a new test statistic is proposed. The basic idea of the proposed test statistic follows the paper of Chen and Ye (2009). The univariate uniformity test is discussed and evaluation of the test statistic is performed by power comparison in that paper. The result is extended to the multi-dimensional case.

Suppose

$$\mathbf{X}_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1k} \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2k} \end{bmatrix}, \ldots, \mathbf{X}_n = \begin{bmatrix} X_{n1} \\ X_{n2} \\ \vdots \\ X_{nk} \end{bmatrix}$$

is a random sample from a *k*-dimensional population distribution with support set $[0,1]^{(k)}$. Here $[0,1]^{(k)}$ is the *k*-dimensional unit cube which the set is defined as

$$\left\{ \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{bmatrix} : 0 \le t_i \le 1 \ \ (i = 1, 2, \cdots, k) \right\}.$$

Let $X_{(1)i}, X_{(2)i}, \cdots, X_{(n)i}$ be the ordered values of $X_{1i}, X_{2i}, \cdots, X_{ni} \quad (i = 1, 2, \cdots, k)$.

The purpose of the multi-dimensional uniformity test is to test

H_0 : The population distribution is a uniform distribution on $[0,1]^{(k)}$,

H_a : The population distribution is not uniform distribution on $[0,1]^{(k)}$.

The test statistics proposed in this research is

$$G_k\left(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\right) = \frac{(n+1)^k}{(n+1)^k - 1} \sum_{i_1=1}^{n+1} \sum_{i_2=1}^{n+1} \cdots \sum_{i_k=1}^{n+1} \left[\prod_{j=1}^{k} \left(X_{(i_j)j} - X_{(i_j-1)j} \right) - \frac{1}{(n+1)^k} \right]^2.$$

Here it is assumed that $X_{(0)j} = 0$ and $X_{(n+1)j} = 1$ $(j = 1, 2, \cdots k)$. It can be seen that if the underlying distribution is a uniform distribution on $[0,1]^{(k)}$, then

$$E\left[\prod_{j=1}^{k}\left(X_{(i_j)j} - X_{(i_j-1)j}\right)\right] = \frac{1}{(n+1)^k} \quad \left(i_j = 1, 2, \cdots n+1 \quad (j = 1, 2, \cdots k)\right).$$

Therefore, if the value of $G_k(\mathbf{X}_1, \mathbf{X}_2, \cdots \mathbf{X}_n)$ is too far away from zero, it could be an indication that the underlying distribution is not uniform distribution on $[0,1]^{(k)}$. This motivates the following test procedure. Under H_0, let $G_{k,1-\alpha}$ be a number such that

$$P\left(G_k(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n) > G_{k,1-\alpha}\right) = \alpha \quad (0 < \alpha < 1).$$

Then H_0 should be rejected at significant level α if $G_k(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n) > G_{k,1-\alpha}$. It can be shown that $G_k(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$ is always between 0 and 1.

2.2 Bivariate Case

The bivariate test statistic is discussed in detail as a special case of the multi-dimensional G test. The critical values of the test statistic are obtained and tabulated. Power comparison is performed in Section 3.

The purpose of the bivariate uniformity test is to test

H_0: The population distribution is a uniform distribution on $[0,1]\times[0,1]$,

H_a: The population distribution is not uniform distribution on $[0,1]\times[0,1]$.

In order to raise the power of the proposed test statistic, the definition of $G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$ is modified by adopting the idea of the Kendall's τ statistic (1938).

Suppose $\mathbf{X}_1 = \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix}, \cdots, \mathbf{X}_n = \begin{bmatrix} X_{n1} \\ X_{n2} \end{bmatrix}$ form a random sample from a bivariate population distribution with support set $[0,1]\times[0,1]$.

$\begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix}$ and $\begin{bmatrix} X_{j1} \\ X_{j2} \end{bmatrix}$ are said to be concordant if $\dfrac{X_{j2}-X_{j1}}{X_{i2}-X_{i1}} > 0$.

$\begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix}$ and $\begin{bmatrix} X_{j1} \\ X_{j2} \end{bmatrix}$ are said to be discordant if $\dfrac{X_{j2}-X_{j1}}{X_{i2}-X_{i1}} < 0$.

$\begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix}$ and $\begin{bmatrix} X_{j1} \\ X_{j2} \end{bmatrix}$ are said to be half concordant and half discordant if $X_{j2}-X_{j1} = 0\left(X_{i2}-X_{i1} \neq 0\right)$.

No comparison is made if $X_{i2}-X_{i1}=0$.

Let n_c be the total number of concordant pairs, and let n_d be the total number of discordant pairs.

Suppose also that $X_{(1)1}, X_{(2)1}, \cdots, X_{(n)1}$ are the ordered values of $X_{11}, X_{21}, \cdots, X_{n1}$, and $X_{(1)2}, X_{(2)2}, \cdots, X_{(n)2}$ are the ordered values of $X_{12}, X_{22}, \cdots, X_{n2}$.

Define

$$G_2\left(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\right) = \frac{(n+1)^2\left(\left|n_c - n_d\right|+1\right)}{n(n+2)\left(n_c+n_d+1\right)} \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} \left[\left(X_{(i)1}-X_{(i-1)1}\right)\left(X_{(j)2}-X_{(j-1)2}\right) - \frac{1}{(n+1)^2}\right]^2.$$

Here it is assumed that $X_{(0)1} = X_{(0)2} = 0$ and $X_{(n+1)1} = X_{(n+1)2} = 1$.

It can be seen that if the underlying distribution is a uniform distribution on $[0,1]\times[0,1]$, then

$$E\left[\left(X_{(i)1}-X_{(i-1)1}\right)\left(X_{(j)2}-X_{(j-1)2}\right)\right] = \frac{1}{(n+1)^2} \quad \left(i = 1,2,\cdots,n+1;\ j = 1,2,\cdots,n+1\right).$$

Therefore, if the value of $G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$ is too far away from zero, it could be an indication that the underlying distribution is not uniform distribution on $[0,1]\times[0,1]$. This motivates the following test procedure. Under H_0, let $G_{1-\alpha}$ be a number such that

$$P\left(G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n) > G_{1-\alpha}\right) = \alpha \quad (0 < \alpha < 1).$$

Then H_0 should be rejected at significant level α if $G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n) > G_{1-\alpha}$.

The purpose of including the term $\dfrac{|n_c - n_d| + 1}{n_c + n_d + 1}$ is to raise the power of the test when the two variables of alternative bivariate distribution are correlated.

It can be shown that $0 \le G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n) \le 1$ for any $\mathbf{X}_1, \mathbf{X}_2, \cdots \mathbf{X}_n \in [0,1]\times[0,1]$. The lower and upper bounds of the inequality $0 \le G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n) \le 1$ cannot be improved. It means that one may construct bivariate data sets such that the values of $G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$ will reach 0 and 1, respectively.

3. CRITCAL VALUES AND POWER STUDY

3.1 Critical Values

In this research, we focus on the bivariate case of the $G(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$ test. Critical values of the G test in bivariate case are obtained by applying the Monte Carlo method in SAS/IML.

Firstly, a pseudo-random sample $\mathbf{X}_1 = \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix}, \cdots, \mathbf{X}_n = \begin{bmatrix} X_{n1} \\ X_{n2} \end{bmatrix}$ is drawn from the uniform distribution on the support set $[0,1]\times[0,1]$. Since the marginal distributions of the bivariate uniform distribution are independent and uniform on $[0,1]$, $X_{11}, X_{21,} \cdots, X_{n1}$ and $X_{12}, X_{22,} \cdots, X_{n2}$ can be generated by the univariate uniform distribution. Secondly, the coefficient of the G test, $\dfrac{(n+1)^2 \left(|n_c - n_d| + 1\right)}{n(n+2)(n_c + n_d + 1)}$, is calculated by SAS. Then we sort the values of $X_{11}, X_{21}, \cdots, X_{n1}$ to obtain $X_{(1)1}, X_{(2)1}, \cdots, X_{(n)1}$ and sort the values of $X_{12}, X_{22}, \cdots, X_{n2}$ to obtain $X_{(1)2}, X_{(2)2}, \cdots, X_{(n)2}$. Finally, G value can be calculated by using the formula:

$$G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n) = \frac{(n+1)^2 \left(|n_c - n_d| + 1\right)}{n(n+2)(n_c + n_d + 1)} \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} \left[\left(X_{(i)1} - X_{(i-1)1}\right)\left(X_{(j)2} - X_{(j-1)2}\right) - \frac{1}{(n+1)^2} \right]^2.$$

In this research, 100,000 replications are used. Then the 100,000 calculated G values are sorted in ascending order. Theoretically, the value of the 95,000th is treated as the 95th percentile. For more accuracy, the following formula is used to calculate the critical values at 95% confidence level:

$$G_{1-\alpha} = \frac{\left(G_{(0.95*rep)} + G_{(0.95*rep)+1}\right)}{2}$$

Here $\alpha = 0.05$ is the significant level and rep is the number of replications.

The critical values of the bivariate G test are displayed in the Table 1. The first column in the table is the sample size (from 5 to 50). Columns 2 to 5 are the critical values for significant levels $\alpha = 0.10, 0.05, 0.025, 0.01$. In the table, the values are the G test critical value corresponding to the sample size and significant level. The critical values in this research are moderately small, so we keep five decimal places for all the values.

Table1: Critical values of G_2 test statistic

	α=0.10	α=0.05	α=0.025	α=0.01
n	$G_{0.900}$	$G_{0.950}$	$G_{0.975}$	$G_{0.990}$
5	0.05194	0.06722	0.083	0.10584
6	0.03355	0.04323	0.05385	0.06894
7	0.02321	0.02997	0.03695	0.04726
8	0.01676	0.02154	0.02678	0.0343
9	0.01256	0.0161	0.0199	0.02562
10	0.00973	0.0125	0.01525	0.01955
11	0.00773	0.00984	0.0121	0.01509
12	0.00618	0.0079	0.00972	0.01212
13	0.00515	0.00657	0.00802	0.01002
14	0.00426	0.0054	0.00657	0.00829
15	0.00359	0.00455	0.00552	0.00698
16	0.00308	0.00388	0.00468	0.00585
17	0.00265	0.00334	0.00404	0.00502
18	0.00231	0.00292	0.00356	0.00439
19	0.00201	0.00254	0.00308	0.00379
20	0.00177	0.00222	0.0027	0.00333
21	0.00158	0.002	0.00241	0.00298
22	0.0014	0.00176	0.00212	0.00262
23	0.00126	0.00158	0.00191	0.00236
24	0.00113	0.00141	0.00169	0.00211
25	0.00102	0.00127	0.00153	0.00188
26	0.00093	0.00116	0.00139	0.00171
27	0.00085	0.00106	0.00127	0.00156
28	0.00077	0.00096	0.00114	0.00139
29	0.00071	0.00088	0.00106	0.00129
30	0.00065	0.00081	0.00096	0.00117
31	0.0006	0.00074	0.00089	0.00108
32	0.00055	0.00069	0.00082	0.00099
33	0.00052	0.00064	0.00077	0.00094
34	0.00048	0.00059	0.0007	0.00085
35	0.00044	0.00055	0.00065	0.00079
36	0.00041	0.00051	0.00061	0.00075
37	0.00039	0.00048	0.00057	0.0007
38	0.00036	0.00045	0.00053	0.00064
39	0.00034	0.00042	0.0005	0.0006
40	0.00032	0.0004	0.00047	0.00057
41	0.0003	0.00037	0.00044	0.00053
42	0.00028	0.00035	0.00041	0.0005
43	0.00027	0.00033	0.00039	0.00047
44	0.00025	0.00031	0.00037	0.00044
45	0.00024	0.00029	0.00035	0.00041
46	0.00023	0.00028	0.00033	0.0004
47	0.00021	0.00026	0.00031	0.00037
48	0.0002	0.00025	0.00029	0.00035
49	0.00019	0.00024	0.00028	0.00034
50	0.00018	0.00022	0.00026	0.00032

3.2 Power Study

The power of a test is the probability of rejecting the null hypothesis when the null hypothesis is false. It means the power is the probability of making correct decision, so the power is desired to be as large as possible. The power of the test statistic mentioned in this research is

$$P\left(G_2\left(\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_n\right) > G_{1-\alpha} \middle| H_a\right),$$

where α is the significant level and H_a is alternative hypothesis.

Theoretically, the power of a statistical test depends on the sample size, the significant level and the sensitivity of the data. In this research we focus on the case that the sample size is less than or equal to 50. Significant level $\alpha = 0.10, 0.05, 0.01$ and several alternative distributions are used in this power study.

In this section, the performance of the $G_2\left(\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_n\right)$ test is evaluated by the power study and the bivariate case of the $G_2\left(\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_n\right)$ test is compared with other existing test statistics. Several alternative distributions are used to evaluate the performance of the $G_2\left(\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_n\right)$ test proposed in this research. As mentioned before, the recent research results in the multivariate uniformity test found in the literature are the “discrepancy measures” tests proposed by Liang, Fang, Hichernell and Li (2001) and the Distance to Boundary test proposed by Berrendero, Cuevas and Vazquez-Grande (2006). Since the Distance to Boundary test is better than the “discrepancy measures” tests in many cases, the power of the $G_2\left(\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_n\right)$ test is compared with the Distance to Boundary test proposed by Berrendero, Cuevas and Vazquez-Grande (2006).

Let $\mathbf{X}_1 = \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix}, \cdots, \mathbf{X}_n = \begin{bmatrix} X_{n1} \\ X_{n2} \end{bmatrix}$ be the observed sample points from the underlying population distribution in the support set $[0,1]\times[0,1]$. Let $X_{(1)1}, X_{(2)1}, \cdots, X_{(n)1}$ be the ordered values of $X_{11}, X_{21}, \cdots, X_{n1}$, and let $X_{(1)2}, X_{(2)2}, \cdots, X_{(n)2}$ be the ordered values of $X_{12}, X_{22}, \cdots, X_{n2}$. Then, the value of the $G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$ can be calculated by the test statistic formula. To test whether or not the underlying population distribution is a uniform distribution in the support set $[0,1]\times[0,1]$, the null and alternative hypotheses are

H_0 : The population distribution is a uniform distribution on $[0,1]\times[0,1]$,

H_a : The population distribution is not uniform distribution on $[0,1]\times[0,1]$.

If the value of test statistic $G_2(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$ is greater than $G_{1-\alpha}$, the null hypothesis should be reject at the significant level α. Otherwise, the null hypothesis is not rejected. Here, $G_{1-\alpha}$ is the critical value of the G test and can be found from the Table 1 corresponding to the different significant level.

The Distance to Boundary test is proposed by Berrendero, Cuevas and Vazquez-Grande (2006). They define the distance from a point X to the boundary as $D(X, \partial S) = \min\{x \in \partial S : \|x - X\|$. They also define R as the maximum distance to the boundary which can be obtained in support S. Then, they get the relative distance from X to the boundary which is $Y = D(X, \partial S)/R$. The test is based on the real variables $Y_i = D(X_i, \partial S)/R,\ i = 1, \cdots n$. It is proved in the paper that if X is from the uniform distribution, the relative distance Y would follow the beta distribution with parameters $\alpha = 1$ and $\beta = p$ (p is the number of dimension). Let $G(y)$ be the distribution function of Y and $G_n(y)$ be the empirical distribution function associated with the sample. The test statistic is

$$\sup_y \sqrt{n}\left|G_n(y) - H(y)\right|,$$

where H(y) is the distribution of Y under the null hypothesis.

The null hypothesis assumes that the underlying distribution is the bivariate uniform distribution and the alternative hypothesis is that the underlying distribution is not the bivariate uniform distribution. As for the power study, two different models of the alternative distributions including beta distribution and meta-type uniform distribution are used.

Monte Carlo simulation is used to find the power of the G test and the Distance to Boundary test. Sample sizes $\mathrm{n}=5,10,15,20,25,30,35,40,45,50$ are selected to compare the power and the selected significant levels $\alpha = 0.10, 0.05, 0.01$. Based on the principal of efficiency and accuracy, 100,000 replications are performed in simulation.

The procedure for the power calculation is summarized as follows:

1. Generate $\mathbf{X}_1 = \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix}, \cdots, \mathbf{X}_n = \begin{bmatrix} X_{n1} \\ X_{n2} \end{bmatrix}$ samples in the support set $[0,1]\times[0,1]$ follow the alternative distribution.
2. Let $X_{(1)1}, X_{(2)1}, \cdots, X_{(n)1}$ be the ordered values of $X_{11}, X_{21}, \cdots, X_{n1}$, and let $X_{(1)2}, X_{(2)2}, \cdots, X_{(n)2}$ be the ordered values of $X_{12}, X_{22}, \cdots, X_{n2}$.
3. Calculate the value of the G test statistics.
4. Compare the value of the test statistic with the critical value for specific significant level and determine whether the null hypothesis is rejected.

5. Repeat steps 1 to 4 for 100,000 times.

6. Calculate the rejection rate which is the power for the G test.

The power of the Distance to Boundary test can be obtained by performing the same procedure.

In this research, three different models are chosen as the choice of the alternative distribution in power study. In the following subsections the alternative distributions are described and the performance of the two tests under different alternative distributions is compared.

3.2.1 Beta Distribution

The Beta distribution family is used by authors as an alternative distribution to conduct power study for univariate uniformity test. By choosing different shape parameter and scale parameter, it gives rich and flexible result in power comparison. The probability density function of the Beta distribution is displayed as follow:

$$f(x)=\begin{cases}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} & 0<x<1\\ 0 & \text{elsewhere}\end{cases}\quad(\alpha>0,\beta>0).$$

To derive an alternative distribution for k-dimensional case, let Y_i be random variable following $Beta(\alpha_i,\beta_i)$ distribution $(i=1,2,\cdots,k)$. It is assumed that $Y_1,Y_2,\cdots,Y_k$ are independent. Then, the distribution of $\mathbf{Y}=(Y_1,Y_2,\cdots,Y_k)'$ can be treated as a k-dimensional alternative distribution. For convenience, such a distribution is called the k-dimensional distribution derived from $Beta(\alpha_1,\beta_1)$, $Beta(\alpha_2,\beta_2)$,…,

$Beta(\alpha_k, \beta_k)$ distributions.

Alternative Distribution 1

The bivariate distribution derived from Beta (5, 2) and Beta (5, 2). The variance of the marginal distribution is 0.629.

Alternative Distribution 2

The bivariate distribution derived from Beta (5, 1) and Beta (5, 1). The marginal distribution is a left-skewed Beta distribution and the variance of marginal distribution is 0.34.

Alternative Distribution 3

The bivariate distribution derived from Beta (0.5, 0.5) and Beta (0.5, 0.5). The marginal distribution is a symmetric Beta distribution and the variance of marginal distribution is 0.1667.

Figures 1-3 show the power comparisons for the G test and the Distance to Boundary test in the condition that the alternative distributions are derived from $Beta(\alpha_1, \beta_1)$ and $Beta(\alpha_2, \beta_2)$. Figure 1 shows that the G test is more powerful than the Distance to Boundary test if the marginal distribution is not symmetric and the variance is large. Figure 2 shows that the G test is more powerful when sample size is less than 25 and with the sample size increasing the power of two tests close to each other and almost close to 1. In this case, the marginal distribution is left–skewed and the variance is not large. Figure 3 shows that the power of the Distance to Boundary test is significantly large than the G test when the alternative distribution 3 is used which is marginal distribution is symmetric and the variance is small.

The symmetric situation is presented in the paper of Berrendero, Cuevas and Vazquez-Grande (2006). It has shown the power of Distance to Boundary is much high in this case. After changing the symmetric condition, the result seems different.

3.2.2 Meta-type Uniform Distribution

Meta-type uniform distribution is mentioned in the papers of Liang, Fang, Hichernell and Li (2001) and Berrendero, Cuevas and Vazquez-Grande (2006). They introduce this distribution for the power comparison.

The basic idea for creating meta-type multivariate distribution is as follows. Let the random vector $\mathbf{X}=\left(X_1,X_2\right)'$ have a distribution function $F\left(\mathbf{X}\right)$ .It is defined that $F_1\left(X_1\right)$ is the marginal distribution function of X_1 and $F_2\left(X_2\right)$ is the marginal distribution function of X_2. Then random vector $\mathbf{Y}$ as $\mathbf{Y}=\left(F_1\left(X_1\right),F_2\left(X_2\right)\right)'$ can be obtained. The new random vector $\mathbf{Y}$ has those properties: $F_1\left(X_1\right)$ and $F_2\left(X_2\right)$ are uniform distributed in the support set [0, 1] and the joint distribution is different from the bivariate uniform distribution since $F_1\left(X_1\right)$ and $F_2\left(X_2\right)$ are not independent. This kind of multivariate distribution is easily generated by statistical software and is useful to check the multivariate uniform distribution. Specifically, two kinds of the meta-type uniform distributions are considered in power study.

Alternative Distribution 4

MNU (Meta Normal Uniform): obtained from bivariate normal distribution with $\boldsymbol{\mu}=\left[0,0\right]'$ and $\boldsymbol{\Sigma}=\begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}$. For the consistence of comparison, the same parameters are chosen as in the paper of Berrendero, Cuevas and Vazquez-Grande (2006).

Alternative Distribution 5

MTU (Meta T distribution Uniform): obtained from bivariate Student's-t distribution with $\boldsymbol{\mu} = [0,0]'$, $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ and 5 degree of freedom.

The power comparison results under meta-type uniform distribution are shown in the Figure 4 to 5. The G test is more powerful than Distance to Boundary test in each case. With the sample size increasing, the power of G test increase and the power of distance to boundary test still don't change too much. Based on the paper of Liang, Fang, Hichernell and Li (2001), the G test is also powerful compare with the "discrepancy measures" tests.

3.3.3 Dirichlet Distribution

Dirichlet distribution is a family of continuous multivariate probability distributions with parameter vector $\boldsymbol{\alpha}$. It is the multivariate generalization of the beta distribution. In this research, the two dimensional Dirichlet distribution is used in power comparison. The probability density function of the two dimensional Dirichlet distribution is displayed as follow:

$$f(x_1, x_2; \alpha_1, \alpha_2, \alpha_3) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)} (x_1)^{\alpha_1 - 1} (x_2)^{\alpha_2 - 1} (1 - x_1 - x_2)^{\alpha_3 - 1}.$$

for $x_1 > 0, x_2 > 0, 1 - x_1 - x_2 > 0$ and $\alpha_1, \alpha_2, \alpha_3 > 0$

Alternative Distribution 6

Choose $\boldsymbol{\alpha} = (2, 2, 2)'$. This is the symmetrical Dirichlet distribution.

Alternative Distribution 7

Choose $\boldsymbol{\alpha} = (3, 4, 9)'$. This is the nonsymmetrical Dirichlet distribution.

Figures 6-7 show the power comparisons for the G test and the Distance to Boundary test in the condition that the alternative distribution is Drichlet distribution. The G test is more powerful than Distance to Boundary test in each case. One is symmetrical Dirichlet distribution and the other is nonsymmetrical Dirichlet distribution.

3.2.4 Summary of the Power Comparison

1. In i.i.d case which is the marginal distribution are independent and identical distributed, if the marginal of the Beta distribution is not symmetric and variance is high, the G test performs better than the Distance to Boundary test.
2. In i.i.d case, if the marginal of the Beta distribution is left-skewed and variance is not high, the G test is more powerful than the Distance to Boundary test under the sample size of 25. When sample size goes over 25, two tests perform in same way.
3. In i.i.d case, if the marginal of the Beta distribution is symmetric and the variance is low, the Distance to Boundary test performs much better than the G test.
4. For the meta-type uniform distribution, the G test has more power than the

Distance to Boundary test both in MNU and MTU cases.

5. For the Dirichlet distribution, the G test has more power than the Distance to Boundary test especially in nonsymmetrical case.

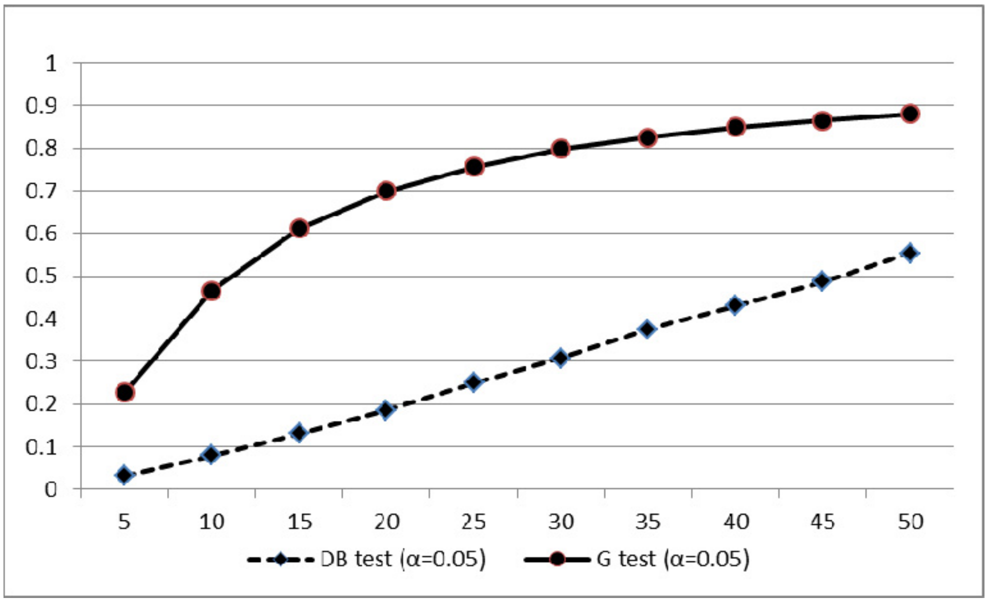

Figure 1. Power Comparison: Alternative Distribution 1
The bivariate distribution derived from Beta (5, 2) and Beta (5, 2)

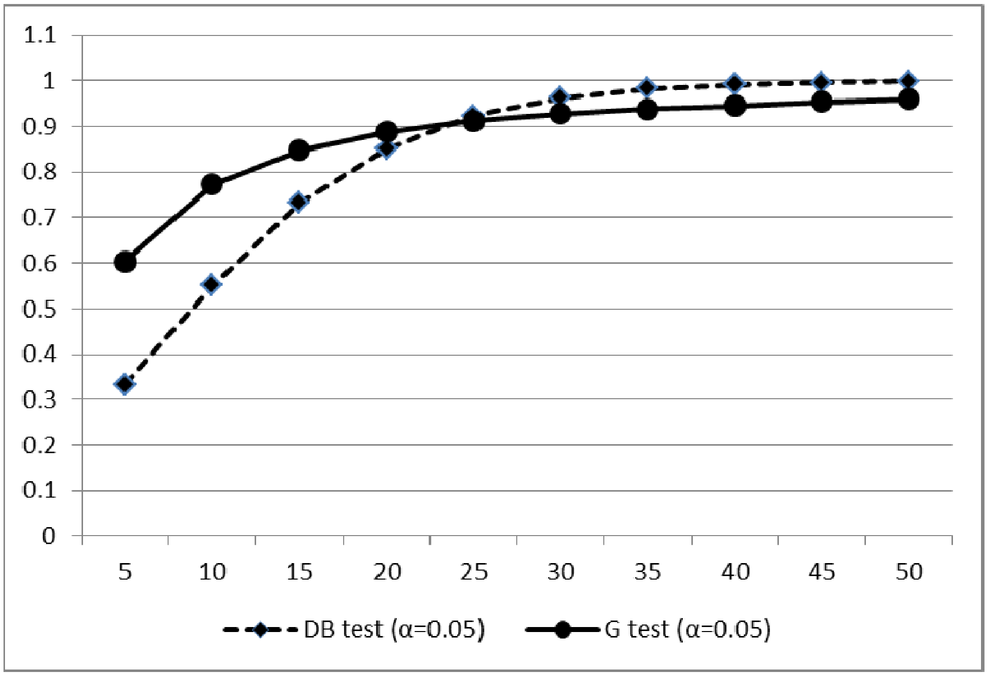

Figure 2. Power Comparison: Alternative Distribution 2
The bivariate distribution derived from Beta (5, 1) and Beta (5, 1)

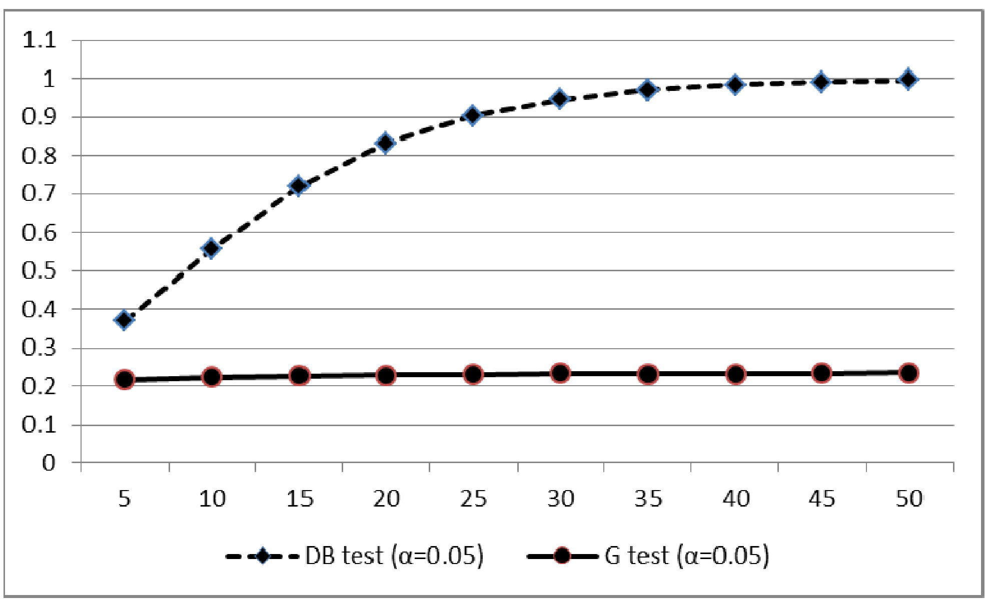

Figure 3. Power Comparison: Alternative Distribution 3
The bivariate distribution derived from Beta (0.5, 0.5) and Beta (0.5, 0.5)

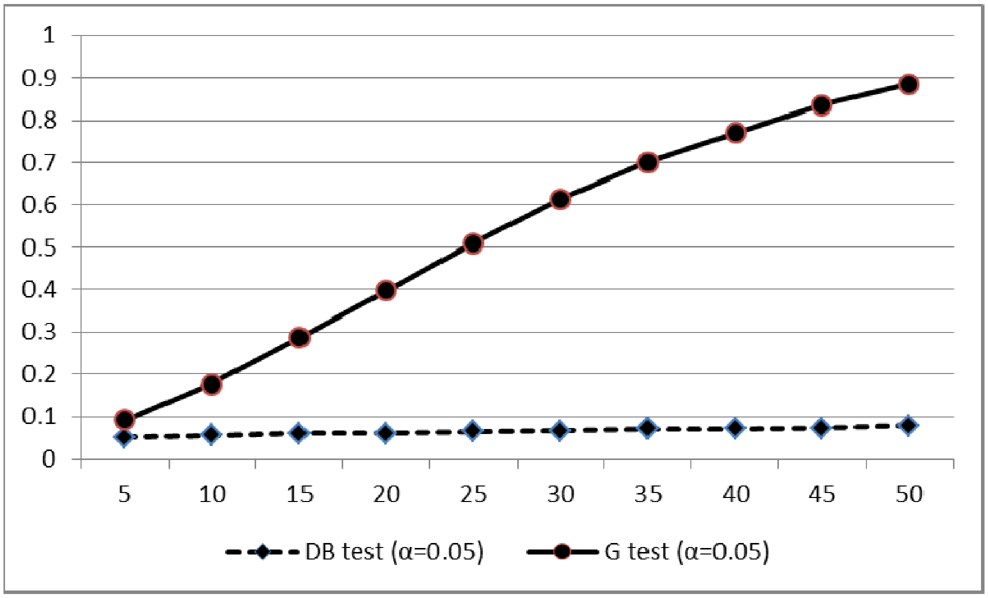

Figure 4. Power Comparison: Alternative Distribution 4
Meta-type Uniform Distribution: MNU

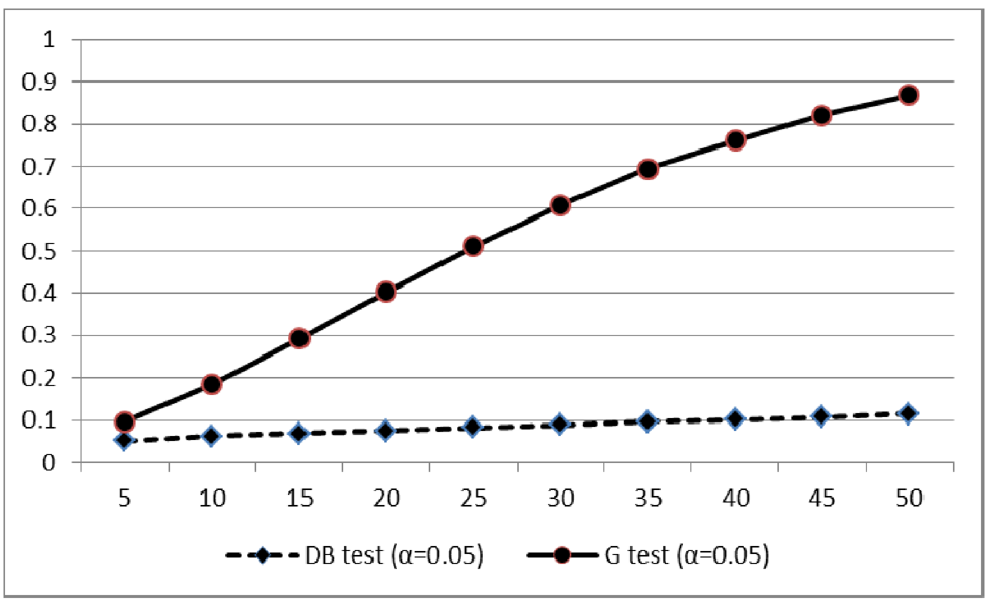

Figure 5. Power Comparison: Alternative Distribution 5
Meta-type Uniform Distribution: MTU

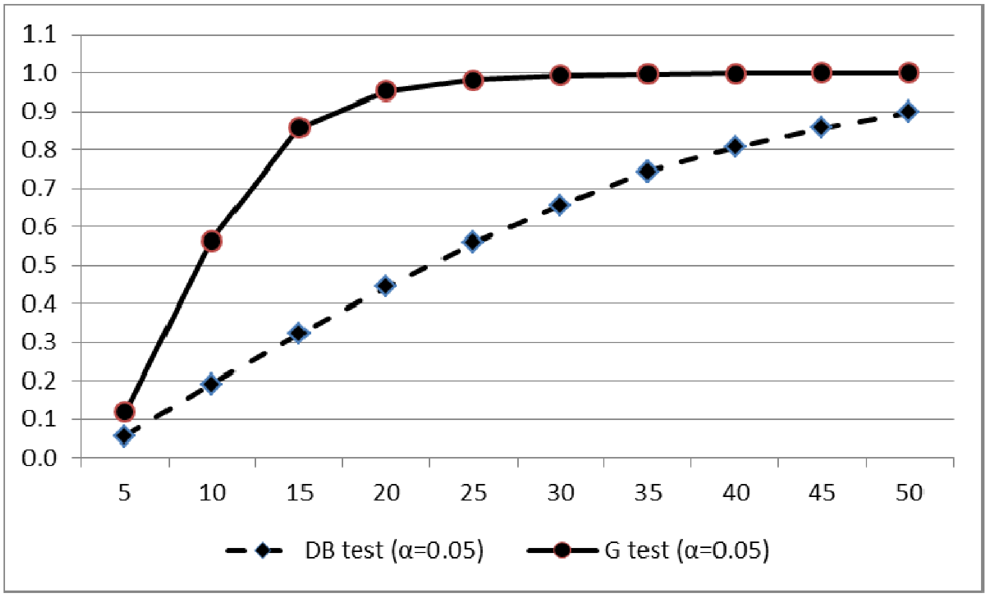

Figure 6. Power Comparison: Alternative Distribution 6
Dirichlet distribution: Dir (2, 2, 2)

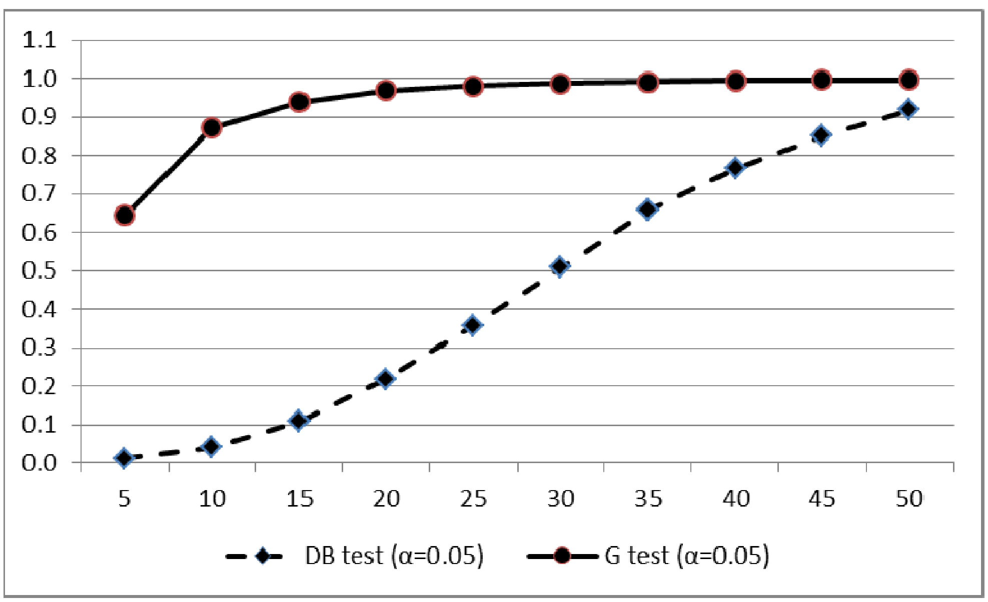

Figure 7. Power Comparison: Alternative Distribution 7
Dirichlet distribution: Dir (3, 4, 9)

4. CONCLUSION AND DISCUSSION

In this research, a new multi-dimensional uniformity test is proposed. The basic idea is from the univariate uniformity test proposed by Chen and Ye (2009). The method is extended to the multi-dimensional case and the bivariate case is discussed in detail. The new test can be used to test whether an underlying multivariate probability distribution differs from a uniform distribution. The critical values of bivariate uniformity test are obtained and the power study is performed by comparing with the recently published multivariate uniformity test.

On the basis of the central limit theorem, if the sample size is sufficiently large with finite mean and variance, the distribution of the sample points will be approximately normally distributed. In this research, we focus on the small sample study which is the case that the sample size is less than or equal to 50.

The Distance to Boundary test is a recently published multivariate uniformity test by Berrendero, Cuevas and Vazquez-Grande (2006). The result of the power comparison shows that the test proposed in this research is more powerful than the Distance to Boundary test in some cases. Especially, when the marginal distribution of alternative distribution is not symmetric and all the marginal distributions are independent and identical distributed. The meta-type uniform distribution is introduced in this research. We generate the meta-type uniform distribution from normal distribution and student's t distribution. The power study shows that new test is more powerful than Distance to the Boundary test in this case. The two dimensional Dirichlet Distribution is used as an alternative distribution. The G test has more power than the

Distance to Boundary test both in symmetrical case and nonsymmetrical case.

Theoretically, we can generate the multivariate distribution under the condition that all the marginal distributions are independent and identical distributed. However, when random samples are drawn from a multivariate population distribution, the marginal distributions are usually dependent. Because of the inclusion of the term $\frac{\left(\left|n_c - n_d\right| + 1\right)}{\left(n_c + n_d + 1\right)}$ in the bivariate case, the power of the modified test statistic will be raised significantly when the marginal distributions are not independent.

REFERENCES

K. Pearson (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably Supposed to Have Arisen from Random Sampling, *Philosophical Magazine,* 5, 157-175.

A. N. Komogorov (1933) Sulla Determinazione Empirica di Una Legge di Distribuzione, Giornale dell' Istituto degli Attuari, 4, 83-91.

N. V. Smirnov (1939) Estimate of Deviation Between Empirical Distributions (Russian), *Bulletin Moscow University*, 2, 3-16.

H. Cramer (1928) On the composition of elementary errors. *Skandinavisk Aktuarietidskrift*, 11, 13-74, 141-180.

A. Justel, D. Pena and R. Zamar (1997) A multivariate Kolmogorov-Smirnov test of goodness-of-fit, *Statistics & Probability Letters*, 35, 251-259.

J. J. Liang, K. T. Fang, F. J. Hickernell and R. Li (2001) Testing multivariate uniformity and its application. *Mathematics of Computation*, 70, 337-355.

J. R. Berrendero, A. Cuevas and F. Vazquez-Grande (2006) Testing multivariate uniformity: The distance-to-boundary method. *The Canadian Journal of Statistics*, 34, 693-707.

Zhenmin Chen and Chunmiao Ye (2009) An alternative test for uniformity. *International Journal of Reliability, Quality and Safety Engineering*, 14, 343-356.

Kendall's, M. (1938) A New Measure of Rank Correlation, *Biomtrika*, 30(1-2), 81-98.

APPENDIX

Tables 2-7: Power of G test and Distance to Boundary test

The bivariate distribution derived from Beta (5, 2) and Beta (5, 2)

sample Size	DB test (α=0.10)	G test (α=0.10)	DB test (α=0.05)	G test (α=0.05)	DB test (α=0.01)	G test (α=0.01)
5	0.4615	0.7024	0.3336	0.6036	0.132	0.3748
10	0.6867	0.837	0.5505	0.7726	0.2806	0.6296
15	0.8377	0.8867	0.7325	0.8472	0.4645	0.7519
20	0.92	0.9174	0.8505	0.8876	0.6323	0.8224
25	0.9628	0.9329	0.9239	0.9122	0.7609	0.864
30	0.9837	0.9445	0.9622	0.9289	0.8544	0.8895
35	0.993	0.954	0.9829	0.9388	0.9149	0.9086
40	0.9968	0.9592	0.9919	0.9461	0.9516	0.9208
45	0.9988	0.9643	0.9965	0.9542	0.973	0.9327
50	0.9994	0.9682	0.9984	0.96	0.9874	0.9403

The bivariate distribution derived from Beta (5, 1) and Beta (5, 1)

sample Size	DB test (α=0.10)	G test (α=0.10)	DB test (α=0.05)	G test (α=0.05)	DB test (α=0.01)	G test (α=0.01)
5	0.0742	0.3429	0.0314	0.2259	0.0037	0.0766
10	0.1553	0.5747	0.08	0.4657	0.0158	0.2659
15	0.2347	0.6946	0.1307	0.6114	0.0319	0.437
20	0.3088	0.7635	0.1847	0.6992	0.0505	0.5601
25	0.3857	0.8059	0.2494	0.7552	0.0757	0.6437
30	0.4602	0.8377	0.3069	0.7976	0.1025	0.7055
35	0.5276	0.8592	0.3754	0.8255	0.1337	0.7467
40	0.5892	0.8773	0.4299	0.8487	0.1664	0.7789
45	0.6458	0.8915	0.4872	0.8649	0.2056	0.8052
50	0.6996	0.9024	0.5527	0.8796	0.2506	0.8302

The bivariate distribution derived from Beta (0.5, 0.5) and Beta (0.5, 0.5)

sample Size	DB test (α=0.10)	G test (α=0.10)	DB test (α=0.05)	G test (α=0.05)	DB test (α=0.01)	G test (α=0.01)
5	0.488	0.3203	0.3688	0.2158	0.1841	0.0865
10	0.674	0.3223	0.5555	0.2226	0.3242	0.0926
15	0.8152	0.3248	0.7177	0.2267	0.4904	0.0944
20	0.898	0.3228	0.831	0.2289	0.6367	0.0978
25	0.9458	0.3204	0.9036	0.2306	0.7535	0.1022
30	0.9728	0.3219	0.9461	0.2333	0.8359	0.105
35	0.9862	0.3253	0.9714	0.2314	0.8966	0.1052
40	0.9935	0.3253	0.9843	0.2326	0.9359	0.1056
45	0.997	0.3233	0.9919	0.2338	0.9621	0.1095
50	0.9985	0.326	0.9963	0.2357	0.978	0.1085

Meta-type uniform distribution: MNU

sample Size	DB test (α=0.10)	G test (α=0.10)	DB test (α=0.05)	G test (α=0.05)	DB test (α=0.01)	G test (α=0.01)
5	0.0997	0.16196	0.05027	0.0914	0.00868	0.02374
10	0.1065	0.29389	0.05476	0.17659	0.01098	0.05111
15	0.1139	0.42999	0.05936	0.28619	0.01316	0.09358
20	0.1162	0.56441	0.06017	0.39789	0.01388	0.1497
25	0.119	0.67094	0.06461	0.50956	0.01455	0.21915
30	0.1236	0.76164	0.06599	0.61473	0.01437	0.29748
35	0.1284	0.83164	0.07084	0.70064	0.0162	0.37723
40	0.1318	0.88216	0.07053	0.77079	0.01589	0.45411
45	0.1327	0.91995	0.07232	0.83584	0.01747	0.54819
50	0.1397	0.94617	0.07774	0.88545	0.01866	0.62383

Meta-type uniform distribution: MTU

sample Size	DB test (α=0.10)	G test (α=0.10)	DB test (α=0.05)	G test (α=0.05)	DB test (α=0.01)	G test (α=0.01)
5	0.1038	0.17149	0.05187	0.09686	0.00937	0.02696
10	0.1185	0.29977	0.06338	0.18476	0.0139	0.05605
15	0.1304	0.43545	0.06915	0.29283	0.01674	0.09927
20	0.1382	0.56048	0.07509	0.40396	0.02015	0.15653
25	0.1447	0.66351	0.0825	0.51075	0.02078	0.22316
30	0.1586	0.7505	0.08887	0.60855	0.02398	0.30418
35	0.1648	0.81941	0.09691	0.69431	0.02582	0.38331
40	0.1727	0.86871	0.10195	0.76063	0.02868	0.45699
45	0.1811	0.9065	0.10786	0.8213	0.02977	0.54708
50	0.1929	0.93458	0.11566	0.86891	0.03342	0.61896

Dirichlet distribution: Dir (2, 2, 2)

sample Size	DB test (α=0.10)	G test (α=0.10)	DB test (α=0.05)	G test (α=0.05)	DB test (α=0.01)	G test (α=0.01)
5	0.1228	0.2325	0.0561	0.1185	0.0061	0.0212
10	0.3109	0.7070	0.1914	0.5644	0.0536	0.2553
15	0.4717	0.9079	0.3216	0.8573	0.1143	0.6622
20	0.5989	0.9676	0.4430	0.9514	0.1893	0.8866
25	0.7026	0.9865	0.5592	0.9815	0.2761	0.9608
30	0.7901	0.9944	0.6555	0.9923	0.3614	0.9862
35	0.8510	0.9975	0.7439	0.9965	0.4487	0.9938
40	0.8973	0.9989	0.8059	0.9986	0.5333	0.9977
45	0.9277	0.9995	0.8570	0.9993	0.6096	0.9989
50	0.9532	0.9999	0.8987	0.9997	0.6851	0.9994

Dirichlet distribution: Dir (3, 4, 9)

sample Size	DB test (α=0.10)	G test (α=0.10)	DB test (α=0.05)	G test (α=0.05)	DB test (α=0.01)	G test (α=0.01)
5	0.040	0.762	0.012	0.643	0.001	0.372
10	0.121	0.907	0.040	0.871	0.002	0.771
15	0.273	0.956	0.108	0.939	0.009	0.896
20	0.446	0.978	0.217	0.969	0.027	0.944
25	0.618	0.987	0.357	0.980	0.057	0.966
30	0.764	0.990	0.508	0.987	0.103	0.979
35	0.865	0.993	0.657	0.991	0.176	0.987
40	0.927	0.996	0.766	0.995	0.263	0.991
45	0.965	0.997	0.850	0.996	0.364	0.994
50	0.983	0.998	0.919	0.997	0.482	0.996