Journal of the Korean Statistical Society (2021) 50:233–271
https://doi.org/10.1007/s42952-020-00071-8

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Two-sample tests for interval-valued data

Hyejeong Choi[1] · Johan Lim[1] · Donghyeon Yu[2] · Minjung Kwak[3]

Received: 15 October 2019 / Accepted: 22 April 2020 / Published online: 26 May 2020

Abstract

Methods to compare two samples of interval-valued data are discussed. In the interval-valued data, observations have the form of intervals and are often read as two-dimensional vector of lower and upper bounds. We consider four methods, two of which are the applications of existing methods for bivariate data, and the other two are based on the marginalization (univariate distributional representation) of the interval-valued data. We conduct a comprehensive numerical study and analysis of real data to understand the performance of four methods.

Keywords Combined test · Hotelling's test · Interval-valued data · Two-sample test · Univariate marginalization

1 Introduction

Statistical procedures for testing equality of two populations of interval valued data are developed in this paper. In interval-valued data, the variable of interest is a univariate random object on a probability space of intervals, and has the form of an interval $(L, U]$, with lower and upper bounds L, U satisfying $L < U$. More precisely, the lower and upper bounds L and U of the interval are real-valued random variables on the probability space of intervals.

Here, we denote an interval-valued data as a form of a half-open interval but the interval could be open, closed, or other form of half-open intervals. In general, the interval-valued data can be categorized into two types; min–max (MM) and measurement error (ME) types (Blanco-Fernández and Winker 2016). The MM type assumes that lower and upper bounds of each interval-valued observation are the minimum and maximum values of object of interest, respectively. In practice, the

✉ Donghyeon Yu
dyu@inha.ac.kr

[1] Department of Statistics, Seoul National University, Seoul, Korea

[2] Department of Statistics, Inha University, Inchoen, Korea

[3] Department of Statistics, Yeungnam University, Gyeongsan, Korea

MM type data is generated when aggregating large datasets to the minimum and maximum values or focusing on the range of variation of the variables. The typical example of the MM type data is the blood pressure data, where the blood pressure usually recorded in minimum and maximum during a heartbeat cycle. On the other hand, the ME type assumes that there exists a true value and the true value is not observable directly, but only observable as an interval that contains the true value. The ME type data is occurred when the exact value is not available due to the confidentiality issues or the use of non-sufficiently accurate measurement device. The typical example of the ME type data is the interval-censored data that commonly encountered in clinical trials.

While the same notation is used for both the MM-type and ME-type interval-valued data, the analysis and inference in both types should be different (Blanco-Fernández and Winker 2016; Grzegorzewski 2018). In the ME type, we deal with usual real-valued random variables, but the problem is that the realization is not precise and obtained as an interval. Thus, the statistical analysis is based on this imprecise information about the point data. On the other hand, in the MM type, we focus on the random interval itself, not the point value. Thus, the statistical analysis aims at developing probabilistic models for the interval itself by considering the models of the lower and upper bounds (or the center and half-range) of the interval.

In this paper, we assume that the observed interval-valued data is of MM-type and develop statistical procedures to test equality of two populations of the interval-valued data. Among many statistical procedures, the comparison of two populations is one of the most fundamental statistical questions. There are several literatures for testing equality of two populations for the ME-type interval-valued data which are related to the interval censored data. Most of the existing methods are developed by the nonparametric test procedures such as Wilcoxon test (Perolat et al. 2015; Grzegorzewski and Śpiewak 2017), U-statistic (Choi et al. 2019) and the sign test (Grzegorzewski and Śpiewak 2019). However, little research has been done in the context of the MM-type interval-valued data. The only method we are aware of is the combined test (CB) proposed by Grzegorzewski (2018). To develop a more powerful testing procedure, we consider three additional testing procedures. One is, by considering the bivariate nature (or bivariate real-valued representation) of interval-valued data, the Hotelling's T^2 (HT) test. The forementioned two are direct applications of the existing methods for bivariate data. The other two newly suggested are based on the univariate marginalization (or univariate distributional representation) of interval-valued data that depends on kernalization. To this end, the uniform kernel method (UK) and Gaussian kernel method (GK) by Jeon et al. (2015) are used to estimate the marginal distribution. We suggest using the Kolmogorov–Smirnov (KS) distance between the kernel marginal distributions to test the equality of two populations. The null distribution of the KS distance is approximated by a permutation procedure (Præstgaard 1995).

The remainder of the paper is organized as follows. In Sect. 2, we precisely describe four methods to compare two-sample interval-valued data. In Sect. 3, we compare performance of the four methods in various settings through a comprehensive simulation study. In Sect. 4, we apply the methods to the blood pressure data of female students in the US. In Sect. 5, we conclude the paper with a summary.

2 Methods

To verify if there is a significant difference between two populations where samples are observed by intervals, we study four methods: the CB, HT, UK, and GK tests. For the CB and HT test, we transform interval-valued data into a bivariate real-valued vector of center C and (logarithm of) half-range R ($\log R$), where $C = (L + U)/2$ and $R = (U - L)/2$, in order to remove the constraint in L and U.

2.1 Combined (CB) test

Let F_{C} and F_{R} be the cumulative distribution function (c.d.f.) of the center and half-range (or log-transformed half-range), respectively, from one population. We define G_{C} and G_{R} similarly for another population. Assume that m and n random samples are observed from each population and independent of each other, i.e., $\big\{(C_{1j}, R_{1j}), j = 1, \ldots, m\big\}$ and $\big\{(C_{2j}, R_{2j}), j = 1, \ldots, n\big\}$. Grzegorzewski (2018) suggests verifying the equivalence of the two populations by testing the overall hypothesis below

$$\mathcal{H}_0 \,:\, F_{\mathrm{C}} = G_{\mathrm{C}} \;\text{ and }\; F_{\mathrm{R}} = G_{\mathrm{R}}.$$

Grzegorzewski (2018) proposes the KS goodness-of-fit test that individually applies the usual KS test to the center and half-range, and combines the respective results.

The KS statistics for each hypothesis $\mathcal{H}_{0,C} \,:\, F_{\mathrm{C}} = G_{\mathrm{C}}$ and $\mathcal{H}_{0,R} \,:\, F_{\mathrm{R}} = G_{\mathrm{R}}$ are

$$\begin{aligned} T_C =& D_{m,n}(\widehat{F}_{m,\mathrm{C}}, \widehat{G}_{n,\mathrm{C}}) = \left(\frac{mn}{m+n}\right)^{1/2} \sup_{t\in\mathbb{R}} |\widehat{F}_{m,\mathrm{C}}(t) - \widehat{G}_{n,\mathrm{C}}(t)|, \\ T_R =& D_{m,n}(\widehat{F}_{m,\mathrm{R}}, \widehat{G}_{n,\mathrm{R}}) = \left(\frac{mn}{m+n}\right)^{1/2} \sup_{t\in\mathbb{R}} |\widehat{F}_{m,\mathrm{R}}(t) - \widehat{G}_{n,\mathrm{R}}(t)|, \end{aligned}$$

where $\widehat{F}_{m,\mathrm{C}}(t) = (1/m)\sum_{j=1}^{m} \mathrm{I}(C_{1j} \le t)$, $\widehat{F}_{m,\mathrm{R}}(t) = (1/m)\sum_{j=1}^{m} \mathrm{I}(R_{1j} \le t)$, $\widehat{G}_{n,\mathrm{C}}(t) = (1/n)\sum_{j=1}^{n} \mathrm{I}(C_{2j} \le t)$, and $\widehat{G}_{n,\mathrm{R}}(t) = (1/n)\sum_{j=1}^{n} \mathrm{I}(R_{2j} \le t)$. The asymptotic null distribution of T_C (or T_R) is known as *Kolomogorov–Smirnov distribution* (Feller 1948), where, for every fixed $z \ge 0$,

$$\mathrm{P}\big\{T_C \le z\big\} \to \; L(z) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2 z},$$

as $m \to \infty$, $n \to \infty$ so that $m/n \to a \in (0, \infty)$. In the numerical study and data example, we use the permutation method to estimate the distribution of the test statistic T_C (or T_R) due to finiteness of the sample sizes.

To test the overall hypothesis $\mathcal{H}_0$, Grzegorzewski (2018) exploits the Bonferroni procedure when combining p values of $\mathcal{H}_{0,C}$ and $\mathcal{H}_{0,R}$. To be specific, let p_C and p_R be the p values related to T_C and T_R, respectively. Then, the overall p value is set as $p = 2\min(p_C, p_R)$, and we reject $\mathcal{H}_0$ if p is small enough, such as $p < \alpha$, where $\alpha \in (0, 1)$ is the significance level.

2.2 Hotelling's T^2 (HT) test

Two-sample HT test is one of the most popular procedures to test the equality of two mean vectors of the populations. We apply this method to center and log-transformed half range by transforming the interval data, which is a two-dimensional problem. We assume that the real-valued random variables $\mathbf{X}_i = (C_{1i}, \log R_{1i}), i = 1, \dots, m$ ($\mathbf{Y}_j = (C_{2j}, \log R_{2j}), j = 1, \dots, n$, respectively) are independently from the population with $N_2(\boldsymbol{\mu}_\mathbf{x}, \Sigma_\mathbf{x})$ ($N_2(\boldsymbol{\mu}_\mathbf{y}, \Sigma_\mathbf{y})$, respectively), where $N_2(\boldsymbol{\mu}, \Sigma)$ denotes the bivariate normal distribution with mean vector μ and covariance matrix Σ.

2.2.1 Equal covariance case

We assume that the covariances of the two populations are equal, $\Sigma_\mathbf{x} = \Sigma_\mathbf{y}$. Then, the null hypothesis $\mathcal{H}_0 : \boldsymbol{\mu}_\mathbf{x} = \boldsymbol{\mu}_\mathbf{y}$ can be tested using HT_{eq}:

$$\mathrm{HT}_{\text{eq}} = \frac{mn}{m+n}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^\top S_p^{-1}(\overline{\mathbf{X}} - \overline{\mathbf{Y}}),$$

where $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ are the sample mean vectors of two samples, respectively, and S_p is the pooled covariance matrix calculated by

$$S_p = \frac{(m-1)S_\mathbf{x} + (n-1)S_\mathbf{y}}{m+n-2},$$

where $S_\mathbf{x}$ and $S_\mathbf{y}$ are the sample covariance matrices from $\mathbf{X}_i$s and $\mathbf{Y}_j$s, respectively. Under the null hypothesis, we know that

$$\frac{m+n-3}{2(m+n-2)}\mathrm{HT}_{\text{eq}} \sim F(2, m+n-3),$$

where $F(2, m+n-3)$ is the F-distribution with parameters 2 and $m+n-3$.

2.2.2 Unequal covariance case

If $\Sigma_\mathbf{x} \neq \Sigma_\mathbf{y}$, the HT statistic given by

$$\mathrm{HT}_{\text{un}} = (\overline{\mathbf{X}} - \overline{\mathbf{Y}})^\top \left(\frac{S_\mathbf{x}}{m} + \frac{S_\mathbf{y}}{n}\right)^{-1} (\overline{\mathbf{X}} - \overline{\mathbf{Y}})$$

follows the null distribution below:

$$\frac{m+n-3}{2(m+n-2)}\mathrm{HT}_{\text{un}} \sim F(2, \nu),$$

where ν is an appropriately defined degrees of freedom.

2.3 Marginalization-based test

In this section, we propose two-step marginalization-based approaches to test the equality of two interval-valued samples. First, we find a univariate distributional representation, which we named as *marginalization*, that attempts to summarize the interval-valued sample with single real-valued variables. Then, we adopt a procedure to compare two univariate distributions. Here, we should remark that the *marginalization* above is a univariate real-valued representation of an interval, not the typical marginalization of the bivariate real-valued representation of the interval, e.g. (L, U) or (C, R).

2.3.1 Two marginalizations

We first introduce two popular marginalization methods: an empirical histogram estimator (also known as a marginal histogram estimator or a kernel estimator with the uniform kernel) and a Gaussian kernel estimator. Suppose we observe n independent intervals $\left\{I_i = (\ell_i, u_i], i = 1, \dots, n\right\}$. The estimator with the uniform kernel (Bertrand and Goupil 2000) for a univariate density of interval-valued data is

$$f_n^{UK}(g) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{u_i - \ell_i} \mathrm{I}(\ell_i < g \le u_i). \tag{1}$$

The rationale behind (1) is that the value of a univariate representation of I_i is uniformly distributed in the interval $(\ell_i, u_i]$. Thus, the marginalization is represented as the uniform mixture of n uniform distributions. We refer to this estimator as the uniform kernel estimator (UK).

Jeon et al. (2015) improve the uniform kernel estimator by imposing some structures on the distribution of data. The proposed estimator is a mixture of n univariate normal densities. That is,

$$f_n^{GK}(g;h) = \frac{1}{n} \sum_{k=1}^{n} \phi(g | \hat{\mu}_k(h), \hat{\sigma}_k(h)), \tag{2}$$

where h is a bandwidth, $\phi(\cdot | \hat{\mu}_k(h), \hat{\sigma}_k(h))$ is the univariate normal density with mean $\hat{\mu}_k(h)$ and standard deviation $\hat{\sigma}_k(h)$ computed by

$$\hat{\mu}_k(h) = \frac{1}{n} \sum_{i=1}^{n} w_{ki}(h) m_i, \qquad \hat{\sigma}_k^2(h) = \frac{1}{n} \sum_{i=1}^{n} w_{ki}(h) v_i,$$
$$m_i = (\ell_i + u_i)/2, \quad v_i = (u_i - \ell_i)^2/12.$$

For the given bandwidth h, the local weights $w_{ki}(h)$ are determined as follows.

Using the center of intervals, we calculate Euclidean distances between kth and ith intervals, say $d_{ki} (= d_{ik})$. Let R_{ki} be the rank of the d_{ki} (in increasing order) among $\left\{d_{k1}, d_{k2}, \dots, d_{kn}\right\}$ with $R_{kk} = 1$. The weights are determined such that

$$w_{ki}(h) \propto \frac{1}{h} K\left(\frac{R_{ki}-1}{h}\right) \text{ and } \sum_{i=1}^{n} w_{ki}(h) = 1,$$

where K is the standard normal density. Jeon et al. (2015) propose to select h that minimizes the Kullback–Leibler loss between the uniform kernel estimator in (1) and the Gaussian kernel estimator in (2). Details can be found in Jeon et al. (2015). We refer to this estimator (2) as the Gaussian kernel estimator (GK).

2.3.2 Test statistic

Let us consider two independent random intervals: first sample $\mathbf{X}_1, \ldots, \mathbf{X}_m$ is drawn from the population with c.d.f. $F(\ell_1, u_1)$ where ℓ_1 and u_1 indicate the lower and upper bound of the interval $\mathbf{X}$, respectively. The second sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ comes from the population with c.d.f. $G(\ell_2, u_2)$ where ℓ_2 and u_2 are defined similarly for $\mathbf{Y}$. We check the equality of distributions $\mathcal{H}_0 : F = G$ by using the univariate marginal estimators introduced previously. In other words, we compare F_{M} and G_{M}, where F_{M} and G_{M} are the marginal distributions of $F(\ell_1, u_1)$ and $G(\ell_2, u_2)$, respectively.

We consider two types UK and GK of test statistics based on the UK and GK estimators, respectively. For the UK type, the test statistic T_{M}^{UK} is similar to the KS statistic and defined as follows:

$$T_{\text{M}}^{UK} = D_{m,n}(\widehat{F}_{\text{M},m}^{UK}, \widehat{G}_{\text{M},n}^{UK}) = \left(\frac{mn}{m+n}\right)^{1/2} \sup_{t\in\mathbb{R}} |\widehat{F}_{\text{M},m}^{UK}(t) - \widehat{G}_{\text{M},n}^{UK}(t)|, \tag{3}$$

where $\hat{F}_{M,m}^{UK}$ and $\hat{G}_{M,m}^{UK}$ are the UK estimators of the marginal cumulative distribution functions F_M and G_M based on ($\mathbf{X}_1$,..., $\mathbf{X}_m$), and ($\mathbf{Y}_1, \ldots, \mathbf{Y}_n$), respectively. The estimator $\hat{F}_{M,m}^{UK}$ of F_M is defined by the estimated density functions $\hat{f}_m^{UK}$ as follows:

$$\begin{aligned}\hat{F}_{M,m}^{UK}(t) &= \int_{-\infty}^{t} \hat{f}_m^{UK}(x)\, dx \\ &= \frac{1}{m}\sum_{i=1}^{m}\{I(U_{1i} < t) + \frac{t - L_{1i}}{U_{1i} - L_{1i}} I(L_{1i} < t \le U_{1i})\},\end{aligned}$$

where L_{1i} and U_{1i} are the lower and upper bounds of the ith observed interval of $\mathbf{X}_i$.

To develop the GK type test statistic, we only adopt the structure of the Gaussian kernel estimator for the interval-valued data proposed in Jeon et al. (2015) and define the test statistic T_M^{GK} based on the GK estimator as the maximal distance between $\hat{F}_{M,m}^{GK}(t;h)$ and $\hat{G}_{M,n}^{GK}(t;h)$ with respect to t for the given common bandwidth h that maximizes $\sup_t |\hat{F}_{M,m}^{GK}(t;h) - \hat{G}_{M,n}^{GK}(t;h)|$, where $\hat{F}_{M,m}^{GK}(\cdot;h)$ and $\hat{G}_{M,n}^{UK}(\cdot;h)$ are the GK estimators of the marginal cumulative distribution functions F_M and G_M for a given common bandwidth h based on ($\mathbf{X}_1$,..., $\mathbf{X}_m$), and ($\mathbf{Y}_1, \ldots, \mathbf{Y}_n$), respectively. The estimator $\hat{F}_{M,m}^{GK}(\cdot;h)$ of F_M is obtained by the estimated density function $\hat{f}_m^{GK}(\cdot;h)$ as follows:

$$\hat{F}^{GK}_{M,m}(t;h) = \int_{-\infty}^{t} \hat{f}^{GK}_{m}(x;h)\,dx = \frac{1}{m}\sum_{i=1}^{n} \Phi(t|\hat{\mu}_i(h), \hat{\sigma}_i(h)),$$

where $\Phi(t|\hat{\mu}_i(h), \hat{\sigma}_i(h))$ is the cumulative distribution function of the normal distribution with mean $\hat{\mu}_i(h)$ and variance $\hat{\sigma}^2_i(h)$. To be specific, we first choose the common bandwidth $h_{\max}$ such that

$$h_{\max} = \mathrm{argmax}_h \sup_{t\in\mathcal{R}} |\hat{F}^h_{\mathrm{M},m}(t) - \hat{G}^h_{\mathrm{M},n}(t)|.$$

Therefore, the test statistic T^{GK}_{M} for the GK type is defined as follows:

$$T^{GK}_{\mathrm{M}} = \left(\frac{mn}{m+n}\right)^{1/2} \sup_{t\in\mathbb{R}} |\hat{F}^{GK}_{\mathrm{M},m}(t;h_{\max}) - \hat{G}^{GK}_{\mathrm{M},n}(t;h_{\max})|. \tag{4}$$

Note that we propose the GK type test with the bandwidth $h_{\max}$ since the GK type test with the proposed bandwidth $h_{\max}$ has similar powers for center change and larger powers for range change than the GK type test with the bandwidth selection by Jeon et al. (2015) in our numerical study (see Appendix 2). However, it is worth noting that the proposed bandwidth $h_{\max}$ does not guarantee the better performance for density estimation compared to the bandwidth selection by Jeon et al. (2015). In addition, the common bandwidth selection for the GK type test statistic does not need the calculation of the cross-validated Kulback–Leibler loss as applied in Jeon et al. (2015) and hence we can considerably reduce the computational cost in the evaluation of p value by the permutation procedure while we need to choose the optimal bandwidth h for every permutation if the test statistic is defined by the optimal bandwidth chosen by the cross-validation.

2.3.3 Permutation procedure to approximate the null distribution

We use the permutation method to estimate the sampling distribution of the test statistic (3) under the null $\mathcal{H}_0$. The permutation procedure is straightforward and briefly described as follows. For the b-th permutation, we combine all the $m+n$ observations from both groups together, and then randomly take m observations without replacement. This sample constitutes the first group and the remaining n observations are set as the second group. We compute the test statistic $t_{\mathrm{M},b}$ as in (3) using these permuted samples and repeat this procedure B many times. The permutation distribution for the test statistic T_{M} is given by the empirical distribution of $t_{\mathrm{M},1}, \ldots, t_{\mathrm{M,B}}$. Now, let $t^{\mathrm{obs}}_{\mathrm{M}}$ be the observed test statistic from the original two samples. The p value for hypothesis $\mathcal{H}_0$ based on permutation is

$$p = \frac{\sum_{b=1}^{B} \mathrm{I}(t_{\mathrm{M},b} \geq t^{\mathrm{obs}}_{\mathrm{M}})}{B}.$$

In the numerical study, since we know the underlying distribution such as normal or t distribution, the reference distribution can be better approximated by generating random samples from the known distribution under the null rather than permuting observed samples.

3 Numerical study

In this section, we compare the finite-sample performance of the four methods described in the previous section. We generate interval variable by generating a bivariate real-valued random variable $(C, \log R)$ under various situations. Each situation depends on different factor(s) to induce difference between two populations, where the magnitude of difference is controlled by $\delta = 0, 0.5, 1, 1.5$. By the setting, the null hypothesis is expressed as $\mathcal{H}_0 : \delta = 0$ for all four tests. Thus, when $\delta = 0$, we examine the size of each test, while for $\delta > 0$, we assess the power of competing tests. For the sample size, we consider following 4 cases: $(m, n) = (30, 30),\ (30, 120),\ (50, 50),\ (50, 200)$. To investigate the effect of correlation between the center and range, we use three values for a correlation parameter $\rho = (0,\ 0.4,\ 0.8)$. All other settings we consider for the study are summarized in Table 1. Generative models of each simulation are given in the beginning of each subsection.

Table 1 Summary of the settings. At the first column, the left character of the hyphen (-) denotes the distribution of $(C, \log R)$: N is for "normal", T for "T with df 5", and SN for "skew-normal". The right character represents a source of difference between the two populations: C is for "mean of center", R for "mean of range", C.S for "mean and skewness of center", COV for "covariance", C.V for "mean and variance of center", and R.V for "mean and variance of range". Each population ($i = 1, 2$) is denote by $\varPi_i$ with parameters μ_i (mean), $\varSigma_i$ (covariance matrix), and γ_i (skewness). We define $\varSigma = (1\ \ \rho\ ;\rho\ \ 1)$

Case	Distribution of $(C, \log R)$	$\varPi_1$			$\varPi_2$		
		μ_1	$\varSigma_1$	γ_1	μ_2	$\varSigma_2$	γ_2
(N-C)	Normal	$(0, 0)$	$\varSigma$	$(0, 0)$	$(\delta, 0)$	$\varSigma$	$(0, 0)$
(N-R)	Normal	$(0, 0)$	$\varSigma$	$(0, 0)$	$(0, \delta)$	$\varSigma$	$(0, 0)$
(T-C)	T with df 5	$(0, 0)$	$\varSigma$	$(0, 0)$	$(\delta, 0)$	$\varSigma$	$(0, 0)$
(T-R)	T with df 5	$(0, 0)$	$\varSigma$	$(0, 0)$	$(0, \delta)$	$\varSigma$	$(0, 0)$
(SN-C)	Skew normal	$(0, 0)$	$\varSigma$	$(-0.6, -0.1)$	$(\delta, 0)$	$\varSigma$	$(-0.6, -0.1)$
(SN-C.S)	Skew normal	$(0, 0)$	$\varSigma$	$(0, -0.1)$	$(\delta, 0)$	$\varSigma$	$(-0.4\delta, -0.1)$
(N-COV)	Normal	$(0, 0)$	$\varSigma$	$(0, 0)$	$(0, 0)$	$(1+\delta)\varSigma$	$(0, 0)$
(N-C.V1)	Normal	$(0, 0)$	$\varSigma$	$(0, 0)$	$(\delta, 0)$	$\begin{pmatrix} 1+2\delta & \sqrt{1+2\delta}\rho \\ \sqrt{1+2\delta}\rho & 1 \end{pmatrix}$	$(0, 0)$
(N-C.V2)	Normal	$(0, 0)$	$\begin{pmatrix} 4 & 2\rho \\ 2\rho & 1 \end{pmatrix}$	$(0, 0)$	$(\delta, 0)$	$\begin{pmatrix} 4-2\delta & \sqrt{4-2\delta}\rho \\ \sqrt{4-2\delta}\rho & 1 \end{pmatrix}$	$(0, 0)$
(N-R.V)	Normal	$(0, 0)$	$\varSigma$	$(0, 0)$	$(0, \delta)$	$\begin{pmatrix} 1+2\delta & \sqrt{1+2\delta}\rho \\ \sqrt{1+2\delta}\rho & 1 \end{pmatrix}$	$(0, 0)$

For test statistics T_C, T_R, and T_{M}, we numerically approximate its null distribution by generating m and n samples under the null and calculating corresponding test statistics. We repeat this procedure 20, 000 times to get their reference distributions. For $\mathrm{HT}_{\mathrm{eq}}$ and $\mathrm{HT}_{\mathrm{un}}$, the simulated distribution is similarly obtained if a setting does not meet underlying assumptions of the HT test.

The significance level α is set as 5%. The size and power of each test are evaluated as the rejection rate through 2000 repetitions.

3.1 Normal distribution with equal covariances

We set a bivariate normal distribution for the center and log-transformed half-range. We compare the rejection power of four tests by varying the mean vector value of the second population, assuming that the covariances of two populations are equal. By denoting the first population as Π_1 and the second as Π_2, the setting is expressed as follows:

$$\Pi_1 : \begin{pmatrix} C_1 \\ \log R_1 \end{pmatrix} \sim N_2(\mu_1, \Sigma_1), \quad \Pi_2 : \begin{pmatrix} C_2 \\ \log R_2 \end{pmatrix} \sim N_2(\mu_2, \Sigma_2),$$

where mean and variance parameters are

$$\mu_1 = (0,0)^{\top}, \ \ \Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$
$$\text{(N-C) } \mu_2 = (\delta, 0)^{\top}, \text{ or (N-R) } \mu_2 = (0, \delta)^{\top}.$$

Note that the mean vector in the second population (Π_2) is set to either $(\delta, 0)$ or $(0, \delta)$. The reason for varying mean of center and half-range separately is that they differently affect the rejection power, which will be explained later.

We first explain a general trend across methods. When we look at the null case where $\delta = 0$ in Table 2, the size of each test is well controlled since the rejection rate is close to the significance level $\alpha = 0.05$ in all cases. Under the alternative hypothesis ($\delta > 0$), it can be seen for every setting that the larger δ is, the greater probability of rejection is. Similarly, each test becomes more powerful as more samples are available.

To summarize the winners based on the case where $\rho = 0$, the HT test shows the highest power among the four tests in both cases (N-C) and (N-R). This consequence is natural when considering that other methods test the equality of distributions, while the HT test only compares mean vectors between two populations. In addition, the data generation setting (a bivariate normal distribution with equal covariances) satisfies the underlying assumptions of the HT test. Note that in case (N-C), where two distributions differ in mean of the center, two marginal tests are comparable to the HT, but perform better than the CB. However, in case (N-R), where mean vectors are different at the range, the result is reversed, i.e., the CB performs better than the marginal tests.

Looking closely at the properties of each test, in the CB and HT tests, the power in case (N-C) is almost the same to the power in (N-R) under the same simulation

Table 2 Size and power of each test in case of the bivariate normal distribution with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-C)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.293	0.381	**0.402**	0.387	0.296	**0.442**	0.420	0.412	0.242	**0.803**	0.520	0.498
		1.0	0.844	**0.931**	0.928	0.916	0.847	**0.966**	0.940	0.931	0.832	**1.000**	0.982	0.979
		1.5	0.997	**1.000**	0.998	0.998	0.997	**1.000**	0.999	0.999	0.995	**1.000**	**1.000**	**1.000**
	(30, 120)	0.0	0.044	0.053	0.046	0.043	0.052	0.053	0.048	0.045	0.048	0.053	0.047	0.043
		0.5	0.451	**0.577**	0.576	0.562	0.467	**0.659**	0.603	0.581	0.460	**0.958**	0.716	0.689
		1.0	0.974	**0.996**	0.993	0.992	0.974	**0.999**	0.995	0.993	0.977	**1.000**	0.999	0.999
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 50)	0.0	0.042	0.051	0.048	0.046	0.041	0.051	0.048	0.051	0.042	0.051	0.049	0.050
		0.5	0.466	**0.597**	0.594	0.575	0.454	**0.677**	0.626	0.606	0.446	**0.974**	0.752	0.745
		1.0	0.986	**0.996**	0.995	0.995	0.985	**1.000**	0.997	0.996	0.984	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 200)	0.0	0.040	0.047	0.047	0.048	0.046	0.047	0.046	0.047	0.040	0.047	0.053	0.051
		0.5	0.678	**0.810**	0.792	0.778	0.684	**0.888**	0.825	0.813	0.686	**1.000**	0.909	0.901
		1.0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(N-R)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.275	**0.365**	0.043	0.071	0.251	**0.430**	0.047	0.065	0.244	**0.795**	0.089	0.083
		1.0	0.840	**0.931**	0.109	0.315	0.840	**0.965**	0.159	0.347	0.840	**1.000**	0.409	0.433
		1.5	0.996	**0.999**	0.537	0.822	0.995	**1.000**	0.648	0.874	0.997	**1.000**	0.952	0.978
	(30, 120)	0.0	0.044	0.053	0.046	0.043	0.052	0.053	0.048	0.045	0.048	0.053	0.047	0.043
		0.5	0.470	**0.573**	0.060	0.123	0.462	**0.644**	0.075	0.146	0.465	**0.957**	0.155	0.217
		1.0	0.979	**0.995**	0.291	0.654	0.980	**0.998**	0.367	0.723	0.985	**1.000**	0.738	0.869

Table 2 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
		1.5	**1.000**	**1.000**	0.924	0.986	**1.000**	**1.000**	0.965	0.992	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 50)	0.0	0.042	0.051	0.048	0.046	0.041	0.051	0.048	0.051	0.042	0.051	0.049	0.050
		0.5	0.479	**0.602**	0.050	0.091	0.456	**0.680**	0.066	0.109	0.478	**0.971**	0.154	0.155
		1.0	0.981	**0.996**	0.281	0.607	0.983	**0.999**	0.394	0.680	0.987	**1.000**	0.752	0.791
		1.5	**1.000**	**1.000**	0.920	0.984	**1.000**	**1.000**	0.965	0.997	**1.000**	**1.000**	0.999	**1.000**
	(50, 200)	0.0	0.040	0.047	0.047	0.048	0.046	0.047	0.046	0.047	0.040	0.047	0.053	0.051
		0.5	0.694	**0.830**	0.093	0.201	0.699	**0.893**	0.111	0.238	0.713	**0.997**	0.264	0.334
		1.0	0.999	**1.000**	0.609	0.889	**1.000**	**1.000**	0.724	0.942	**1.000**	**1.000**	0.949	0.985
		1.5	**1.000**	**1.000**	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

parameters. This result is also natural because both tests are designed with the same priority for the center and range. On the other hand, in the marginal tests, the power in case (N-C) is much higher than the power in case (N-R), especially when δ is small. This implies that the two marginalization methods, the UK and GK, are more sensitive to the change of the center rather than range. In addition, the marginalization-based tests show much less powers than those of the CB and HT tests. Thus, two-dimensional test procedure is preferable to the marginalization-based tests when the difference of two distributions is caused by the difference in the range of the interval. However, it is worth noting the performance of the marginalization-based tests in case (N-R) with $\rho = 0$. That is, even if the range and center are independent, the power of the GK and UK is close to 1 as δ grows. It should also be noted that the performance of the GK and UK is similar in case (N-C), but the GK performs much better than the UK in case (N-R).

Now, we examine the effect of correlation on the power of each test. In general, larger correlation results in higher power of each test. This phenomenon can be explained using the Mahalanobis distance between the two mean vectors from Π_1 and Π_2. In case (N-C), for instance, the distance is $(\delta, 0)\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1}(\delta, 0) = \delta^2/(1-\rho^2)$, which increases as ρ gets larger. Specifically, when ρ is 0, 0.4, and 0.8, the corresponding distance is δ^2, $1.2\delta^2$ and $2.8\delta^2$, respectively. Thus, it is evident to see that two population distributions are easily distinguished from each other, especially when $\rho = 0.8$. However, the effect size of ρ in power differs from each test. The HT test shows the most significant increment in power among the four tests as ρ increases, which could be reasonable considering that the HT statistic is in the form of the Mahalanobis distance between two mean vectors. The followings are the UK and GK tests showing a similar increase. On the other hand, the power of the CB test hardly changes. We, hereafter, would avoid a discussion on ρ since interpretation of its effect is almost same in most of the following settings. Thus, the case of $\rho = 0$ will be mainly discussed.

3.2 Non-normal cases

We examine the size and power in terms of tail thickness and skewness of an underlying bivariate distribution for the center and log-transformed half-range.

3.2.1 Thickness of the tail

We use a bivariate t-distribution with the degrees of freedom 5 denoted by t_5, which has a thicker tail than the normal distribution. We assume two populations have equal covariance matrices. Other details regarding the setup are identical to the normal case. That is,

$$\Pi_1 : \begin{pmatrix} C_1 \\ \log R_1 \end{pmatrix} \sim t_5(\mu_1, \Sigma_1), \quad \Pi_2 : \begin{pmatrix} C_2 \\ \log R_2 \end{pmatrix} \sim t_5(\mu_2, \Sigma_2),$$

where mean and variance parameters are

$$\mu_1 = (0,0)^\top,\ \ \Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$
$$\text{(T-C) } \mu_2 = (\delta,0)^\top,\ \text{ or (T-R) } \mu_2 = (0,\delta)^\top.$$

Since the Gaussian assumption is broken, the null distribution of $\mathrm{HT}_{\mathrm{eq}}$ is calculated by the permutation method as mentioned earlier.

First of all, it is noticeable in Table 3 that the testing power decreases overall compared to that of the normal distribution. Next, based on the case where ρ is 0, the UK test outperforms the other three tests in case (T-C), while in case (T-R), the CB test is most powerful, which is different from the normal case where the HT test shows the highest power. Performance degradation of the HT is obvious since the Gaussian assumption is not satisfied. Third, in case (T-C), the power of the UK test uniformly dominates that of the GK, contrary to their similar performance in the normal case (N-C). The less better performance of the GK test is attributed to its dependency on the Gaussian kernel. Finally, as in the previous results, in case (T-R), the performance of the marginal tests is much worse than that of the two other tests except the case with large $\delta = 1.5$. Meanwhile, when center and range are highly correlated ($\rho = 0.8$), the HT test shows better performance than the others. This is because as ρ gets larger, the increase of power in the HT test is more substantial than the other tests, as explained before.

3.2.2 Skewness

We generate the center and log-transformed half-range from the following bivariate skew-normal distribution. We use a centered parameterization to fix the marginal parameters at prescribed values (Azzalini and Capitanio 1999). That is,

$$\begin{pmatrix} C \\ \log R \end{pmatrix} \sim SN\left[\boldsymbol{\mu} = \begin{pmatrix} \mu_{\mathrm{C}} \\ \mu_{\mathrm{R}} \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \boldsymbol{\gamma} = \begin{pmatrix} \gamma_{\mathrm{C}} \\ \gamma_{\mathrm{R}} \end{pmatrix}\right],$$

where $(\gamma_{\mathrm{C}}, \gamma_{\mathrm{R}})^\top$ represents skewness of the marginal distribution of the center and log-transformed half-range, respectively. For the sake of simplicity, we only consider two cases for sample size $(m, n) = (30, 30), (30, 120)$, and the case of different mean at center. We additionally include the case where skewness and mean of the center are varying together, which is motivated from the real data example described in the next section.

(SN-C) Mean of the center is different while covariance and skewness are the same in two populations:

$$\Pi_1 : \mu_1 = (0,0)^\top,\ \ \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},\ \ \gamma_1 = (-0.6, -0.1)^\top$$
$$\Pi_2 : \mu_2 = (\delta,0)^\top,\ \ \Sigma_2 = \Sigma_1,\ \ \gamma_2 = \gamma_1.$$

(SN-C.S) Skewness of the center as well as mean of the center are different in two populations, and two covariances are equal:

Table 3 Size and power of each test in case of the bivariate t-distribution with df 5 with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(T-C)	(30, 30)	0.0	0.049	0.058	0.059	0.049	0.043	0.058	0.060	0.053	0.044	0.058	0.057	0.055
		0.5	0.262	0.253	**0.356**	0.256	0.214	0.304	**0.351**	0.298	0.245	**0.618**	0.428	0.398
		1.0	0.777	0.778	**0.870**	0.775	0.749	0.845	**0.886**	0.822	0.791	**0.993**	0.940	0.925
		1.5	0.983	0.977	**0.994**	0.979	0.977	0.991	**0.996**	0.988	0.983	**1.000**	0.998	0.998
	(30, 120)	0.0	0.042	0.053	0.047	0.053	0.044	0.053	0.054	0.055	0.041	0.053	0.048	0.050
		0.5	0.354	0.378	**0.486**	0.388	0.386	0.442	**0.507**	0.413	0.381	**0.807**	0.610	0.523
		1.0	0.937	0.924	**0.972**	0.947	0.950	0.956	**0.978**	0.961	0.951	**1.000**	0.995	0.988
		1.5	**1.000**	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 50)	0.0	0.047	0.049	0.058	0.054	0.042	0.049	0.052	0.047	0.046	0.049	0.047	0.048
		0.5	0.385	0.396	**0.508**	0.393	0.390	0.460	**0.534**	0.444	0.388	**0.809**	0.641	0.603
		1.0	0.952	0.931	**0.976**	0.945	0.952	0.966	**0.985**	0.966	0.959	**1.000**	0.996	0.995
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 200)	0.0	0.043	0.043	0.047	0.050	0.044	0.043	0.050	0.057	0.040	0.043	0.052	0.048
		0.5	0.610	0.568	**0.713**	0.601	0.611	0.662	**0.736**	0.631	0.610	**0.958**	0.827	0.788
		1.0	0.997	0.997	**0.999**	0.996	0.995	**0.999**	**0.999**	0.997	0.998	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(T-R)	(30, 30)	0.0	0.049	0.058	0.059	0.049	0.043	0.058	0.060	0.053	0.044	0.058	0.057	0.055
		0.5	**0.257**	0.251	0.044	0.067	0.218	**0.291**	0.051	0.071	0.206	**0.611**	0.083	0.085
		1.0	**0.802**	0.777	0.119	0.214	0.763	**0.845**	0.152	0.225	0.773	**0.988**	0.329	0.316
		1.5	**0.989**	0.977	0.436	0.531	**0.986**	**0.986**	0.523	0.606	0.989	**1.000**	0.809	0.798
	(30, 120)	0.0	0.042	0.053	0.047	0.053	0.044	0.053	0.054	0.055	0.041	0.053	0.048	0.050
		0.5	**0.409**	0.388	0.058	0.100	0.392	**0.454**	0.076	0.118	0.396	**0.796**	0.133	0.174
		1.0	**0.959**	0.925	0.244	0.450	**0.957**	0.955	0.301	0.504	0.958	**0.999**	0.612	0.682

Table 3 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
		1.5	**1.000**	0.998	0.804	0.830	**0.999**	**0.999**	0.870	0.894	**1.000**	**1.000**	0.980	0.976
	(50, 50)	0.0	0.047	0.049	0.058	0.054	0.042	0.049	0.052	0.047	0.046	0.049	0.047	0.048
		0.5	0.407	**0.418**	0.066	0.102	0.419	**0.481**	0.080	0.105	0.400	**0.830**	0.134	0.129
		1.0	**0.961**	0.941	0.250	0.369	0.963	**0.964**	0.321	0.410	0.961	**1.000**	0.594	0.564
		1.5	**1.000**	0.999	0.803	0.808	**1.000**	**1.000**	0.863	0.860	**1.000**	**1.000**	0.980	0.966
	(50, 200)	0.0	0.043	0.043	0.047	0.050	0.044	0.043	0.050	0.057	0.040	0.043	0.052	0.048
		0.5	**0.619**	0.570	0.087	0.158	0.605	**0.655**	0.102	0.180	0.610	**0.968**	0.208	0.246
		1.0	**0.998**	0.994	0.499	0.658	0.996	**0.997**	0.596	0.730	0.998	**1.000**	0.858	0.870
		1.5	**1.000**	**1.000**	0.981	0.965	**1.000**	**1.000**	0.995	0.979	**1.000**	**1.000**	0.999	**1.000**

$$\Pi_1 : \mu_1 = (0,0)^\top,\ \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},\ \gamma_1 = (0,-0.1)^\top$$
$$\Pi_2 : \mu_2 = (\delta,0)^\top,\ \Sigma_2 = \Sigma_1,\ \gamma_2 = (-2\delta/5,-0.1)^\top.$$

It is shown in Table 4 that the case (SN-C) is similar to the normal case in that the HT test shows the best performance and the power of two marginal tests is better than that of the CB. In the case (SN-C.S), we control the skewness of the second population to gradually increase so that its marginal distribution is more left-skewed. We find that when correlation is small ($\rho = 0$), the UK and GK tests are superior to the other two tests, unlike the previous case (SN-C), but under the highly correlated structure ($\rho = 0.8$), the HT test is the most powerful, as before.

3.3 Normal distribution with unequal covariances

We also set a bivariate normal distribution for the center and log-transformed half-range, but this time we assume that covariances of two populations are not equal. We consider the following four cases, one of which represents characteristics of the real data example. We use two cases for sample size for simplicity: $(m,n) = (30,30), (30,120)$.

(N-COV) The covariance matrices are unequal while the mean vectors are equal:

$$\Pi_1 : \mu_1 = (0,0)^\top,\ \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$
$$\Pi_2 : \mu_2 = (0,0)^\top,\ \Sigma_2 = (1+\delta)\Sigma_1.$$

(N-C.V1) The mean and variance of the center are different in two populations. In the second population, both the mean and variance of the center increase:

$$\Pi_1 : \mu_1 = (0,0)^\top,\ \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$
$$\Pi_2 : \mu_2 = (\delta,0)^\top,\ \Sigma_2 = \begin{pmatrix} 1+2\delta & \sqrt{1+2\delta}\rho \\ \sqrt{1+2\delta}\rho & 1 \end{pmatrix}.$$

(N-C.V2) In the second population, the mean of center increases while the variance of center decreases:

$$\Pi_1 : \mu_1 = (0,0)^\top,\ \Sigma_1 = \begin{pmatrix} 4 & 2\rho \\ 2\rho & 1 \end{pmatrix}$$
$$\Pi_2 : \mu_2 = (\delta,0)^\top,\ \Sigma_2 = \begin{pmatrix} 4-2\delta & \sqrt{4-2\delta}\rho \\ \sqrt{4-2\delta}\rho & 1 \end{pmatrix}.$$

Table 4 Size and Power of each test in case of the bivariate skew-normal distribution with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(SN-C)	(30, 30)	0.0	0.045	0.047	0.042	0.040	0.044	0.051	0.045	0.046	0.040	0.047	0.057	0.053
		0.5	0.303	**0.387**	0.345	0.341	0.306	**0.419**	0.373	0.359	0.300	**0.801**	0.487	0.474
		1.0	0.889	**0.927**	0.908	0.907	0.890	**0.962**	0.924	0.916	0.894	**1.000**	0.972	0.965
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(30, 120)	0.0	0.046	0.052	0.051	0.052	0.043	0.056	0.052	0.052	0.041	0.053	0.049	0.049
		0.5	0.498	**0.566**	0.515	0.523	0.498	**0.648**	0.556	0.529	0.499	**0.958**	0.669	0.634
		1.0	0.993	**0.999**	0.993	0.996	0.993	**0.999**	0.992	0.991	0.994	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(SN-C.S)	(30, 30)	0.0	0.045	0.050	0.044	0.043	0.042	0.052	0.044	0.049	0.041	0.050	0.049	0.052
		0.5	0.317	0.366	**0.424**	0.409	0.315	0.423	**0.437**	0.434	0.310	**0.803**	0.536	0.522
		1.0	0.886	0.922	**0.937**	0.923	0.888	**0.962**	0.952	0.944	0.891	**1.000**	0.983	0.982
		1.5	0.998	**1.000**	**1.000**	**1.000**	0.998	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**	**1.000**
	(30, 120)	0.0	0.048	0.051	0.058	0.056	0.046	0.051	0.056	0.055	0.043	0.056	0.046	0.046
		0.5	0.532	0.582	**0.634**	0.625	0.529	**0.656**	0.652	0.629	0.531	**0.962**	0.764	0.731
		1.0	0.990	**0.995**	**0.995**	**0.995**	0.990	**0.999**	0.997	0.996	0.992	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

(N-R.V) The mean and variance of the range differ in two populations. In the second population, both mean and variance of the range increase:

$$\Pi_1 : \mu_1 = (0,0)^\top, \ \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

$$\Pi_2 : \mu_2 = (0,\delta)^\top, \ \Sigma_2 = \begin{pmatrix} 1 & \sqrt{1+2\delta}\rho \\ \sqrt{1+2\delta}\rho & 1+2\delta \end{pmatrix}.$$

As mentioned earlier, we give an interpretation to the cases where $\rho = 0$ based on Table 5. The most interesting result is the case (N-COV), where the marginal tests have much higher power than two other tests, and the GK outperforms the UK. This result means that the marginal tests, especially the GK test, effectively detect the difference in covariance over the other tests. On the contrary, the HT test, which tests the difference between two mean vectors, is incapable of detecting covariance differences between two populations, as it shows the power same to the size. In cases of (N-C.V1) and (N-C.V2), where variance of the center in the second population varies (increases or decreases) together with the mean change, the marginal tests performed best, compared to the case (N-C) where the HT test is the best. Finally, in case (N-R.V), where both mean and variance of the range are controlled, the GK test shows much higher power than the other tests for $\rho = 0, 0.4$, unlike the poor performance in case (N-R). For $\rho = 0.8$, the HT still has the highest powers than the others as shown in the case of (N-R). When $\rho = 0.8$, the HT test has the highest power in all cases but (N-COV), where there is no difference in the two mean vectors.

We summarize the numerical study in Table 6 that shows the best and worst methods in each case. Two major findings we make are as follows. First, when the center and range are highly correlated, the HT performs best among all. Second, the marginal tests, the UK and GK tests, show higher power than other methods if two distributions differ by more than one factor (mean, covariance, and skewness, etc). In addition, the marginal tests tend to detect the difference in center better than in range. Note that the results of numerical study for the significance levels 1% and 10% are reported in Tables 10, 11, 12, 13, 14, 15, 16 and 17 in Appendix 1.

4 Data example

We conduct a real data analysis using the methods discussed in this paper. We use the data from National Heart, Lung, and Blood Institute Growth and Health Study (NGHS), which is a cohort study to investigate temporal trends of cardiovascular risk factors, such as systolic and diastolic blood pressures (SBP, DBP) through up to ten annual visits of 2379 African–American and Caucasian girls. The blood pressure (BP) measured at two levels can be an example of the MM-type interval-valued data. The goal of our real data analysis is to find the difference in BP distributions between African–American and Caucasian girls at the initial points of the study.

Table 5 Size and power of each test in case of the bivariate normal distribution with unequal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-COV)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.061	0.050	0.080	**0.104**	0.053	0.050	0.082	**0.112**	0.047	0.050	0.098	**0.115**
		1.0	0.077	0.046	0.155	**0.227**	0.086	0.046	0.157	**0.256**	0.071	0.046	0.217	**0.280**
		1.5	0.122	0.052	0.239	**0.408**	0.115	0.052	0.263	**0.431**	0.097	0.052	0.384	**0.496**
	(30, 120)	0.0	0.044	0.054	0.046	0.043	0.052	0.054	0.048	0.045	0.048	0.054	0.047	0.043
		0.5	0.051	0.045	0.072	**0.117**	0.047	0.045	0.078	**0.142**	0.040	0.045	0.111	**0.178**
		1.0	0.092	0.047	0.161	**0.325**	0.085	0.047	0.176	**0.407**	0.077	0.047	0.284	**0.483**
		1.5	0.138	0.044	0.306	**0.592**	0.144	0.044	0.345	**0.689**	0.127	0.044	0.546	**0.784**
(N-C.V1)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.262	0.258	**0.420**	0.401	0.265	0.305	**0.385**	0.343	0.241	**0.601**	0.402	0.374
		1.0	0.700	0.664	**0.865**	0.837	0.716	0.734	**0.834**	0.818	0.715	**0.971**	0.865	0.841
		1.5	0.944	0.897	**0.988**	0.981	0.946	0.938	**0.982**	0.977	0.938	**1.000**	0.990	0.987
	(30, 120)	0.0	0.044	0.054	0.046	0.043	0.052	0.054	0.048	0.045	0.048	0.054	0.047	0.043
		0.5	0.436	0.489	**0.611**	0.594	0.448	0.553	**0.564**	0.536	0.450	**0.899**	0.573	0.556
		1.0	0.949	0.956	**0.986**	**0.986**	0.951	**0.984**	0.979	0.975	0.946	**1.000**	0.981	0.980
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(N-C.V2)	(30, 30)	0.0	0.046	0.052	0.043	0.045	0.041	0.052	0.041	0.046	0.040	0.052	0.049	0.050
		0.5	0.098	0.142	**0.152**	0.149	0.105	0.154	**0.174**	0.161	0.098	**0.298**	0.202	0.198
		1.0	0.438	0.470	**0.557**	0.528	0.430	0.551	**0.613**	0.610	0.430	**0.896**	0.710	0.719
		1.5	0.934	0.898	**0.956**	0.948	0.938	0.939	0.977	**0.978**	0.935	**0.999**	0.996	0.997
	(30, 120)	0.0	0.043	0.054	0.045	0.047	0.048	0.054	0.045	0.048	0.042	0.054	0.045	0.049
		0.5	0.180	0.185	**0.256**	0.243	0.181	0.209	**0.268**	0.253	0.157	**0.422**	0.299	0.281
		1.0	0.673	0.612	**0.759**	0.745	0.676	0.687	**0.802**	0.784	0.666	**0.968**	0.870	0.854

Table 5 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
		1.5	0.989	0.940	**0.994**	0.993	0.992	0.973	**0.998**	**0.998**	0.992	**1.000**	**1.000**	**1.000**
(N-R.V)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.285	0.264	0.067	**0.351**	0.271	0.303	0.084	**0.334**	0.260	**0.597**	0.166	0.256
		1.0	0.732	0.653	0.195	**0.886**	0.721	0.736	0.265	**0.894**	0.724	**0.969**	0.552	0.859
		1.5	0.947	0.907	0.517	**0.993**	0.950	0.938	0.635	**0.994**	0.948	**1.000**	0.905	0.995
	(30, 120)	0.0	0.044	0.054	0.046	0.043	0.052	0.054	0.048	0.045	0.048	0.054	0.047	0.043
		0.5	0.466	0.492	0.079	**0.614**	0.473	0.558	0.120	**0.620**	0.458	**0.892**	0.250	0.602
		1.0	0.963	0.958	0.425	**0.994**	0.962	0.979	0.535	**0.992**	0.962	**0.998**	0.852	0.994
		1.5	**1.000**	**1.000**	0.932	**1.000**	**1.000**	**1.000**	0.965	**1.000**	**1.000**	**1.000**	0.999	**1.000**

Table 6 Summary of the results for the significance level $1\%, 5\%$ and 10%. The best and worst tests are represented for each case. At the second column, the left character of the hyphen (-) denotes the distribution of (C,log$R$) and the right represents the difference between the two populations

	Case	$\rho = 0$		$\rho = 0.4$		$\rho = 0.8$	
		Best	Worst	Best	Worst	Best	Worst
Equal covariances	(N-C)	HT ($\approx$ UK, GK)	CB	HT	CB	HT	CB
	(N-R)	HT	UK	HT	UK	HT	UK
	(T-C)[a]	UK	HT($\approx$ CB)	UK	CB	HT	CB
	(T-R)	CB($\approx$ HT)	UK	HT($\approx$ CB)	UK	HT	UK($\approx$ GK)
	(SN-C)	HT	CB	HT	CB	HT	CB
	(SN-C.S)	UK($\approx$GK)	CB	HT($\approx$UK)	CB	HT	CB
Unequal covariances	(N-COV)	GK	HT	GK	HT	GK	HT
	(N-C.V1)	UK($\approx$GK)	CB($\approx$ HT)	UK	CB	HT	CB
	(N-C.V2)	UK($\approx$GK)	CB($\approx$ HT)	UK($\approx$GK)	CB	HT	CB
	(N-R.V)	GK	UK	GK	UK	HT	UK

After we removing subjects with missing measurement, the total number of subjects remaining is $N = 2256$ ($m = 1112$ Caucasians and $n = 1144$ African–Americans). Table 7 shows descriptive statistics of the BP data by race and results of univariate t tests on whether the BP of African–Americans is higher than that of Caucasians. Mean value of SBP, DBP, and their center from African–American girls is significantly larger than that from Caucasians, but the range shows no significant difference between the two groups. The distributions of the center and log-transformed half-range of African–American are more skewed to the upper-left than those of Caucasians (see Fig. 1). Correlation coefficients between the center and log-transformed half-range for the two groups are as low as −0.26 and −0.27, respectively. Thus, the data are roughly matched with the simulation setting (SN-C.S) or (N-C.V2) with small ρ.

Table 8 shows the results when two-sample comparison methods are applied to the BP data. In all tests, the p values are smaller than 0.001, confirming the significant difference between the two groups.

Table 7 Descriptive statistics of the BP data by race. The p-value is from a univariate t-test on the alternative hypothesis that the BP of the African-American is higher than that of Caucasian

	Caucasian	African–American	p value
Center	78.67 (9.09)	80.13 (8.03)	< 0.001
DBP	56.72 (12.19)	58.03 (11.72)	0.005
SBP	100.62 (9.28)	102.23 (8.65)	< 0.001
Half-range	21.95 (5.89)	22.10 (6.44)	0.279

Table 8 Two-sample tests for the whole BP data

	CB	HT	UK	GK
p-value	< 0.001	< 0.001	< 0.001	< 0.001

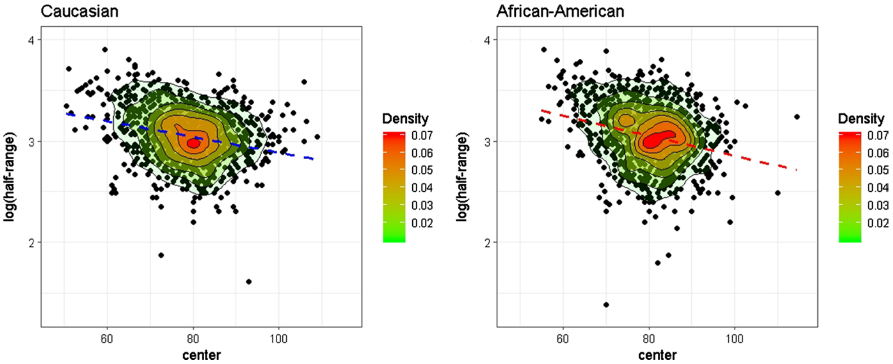

Fig. 1 Contour plots of the two groups of BP data

Table 9 Powers of four two-sample testing methods for different sub-sample sizes with significance levels 1%, 5%, and 10%. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting

m′	n′	$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 10\%$			
		CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
30	30	0.020	0.027	**0.033**	0.030	0.074	0.082	0.097	**0.098**	0.114	0.146	0.164	**0.166**
30	120	0.025	0.038	0.040	**0.046**	0.084	0.116	0.130	**0.132**	0.157	0.190	0.217	**0.226**
50	50	0.025	0.030	**0.040**	**0.040**	0.100	0.110	0.143	**0.150**	0.132	0.170	0.218	**0.226**
50	200	0.036	0.058	**0.074**	0.073	0.125	0.162	**0.194**	0.188	0.208	0.251	0.296	**0.300**
100	100	0.039	0.050	0.075	**0.077**	0.146	0.176	0.232	**0.238**	0.230	0.253	0.332	**0.344**
100	400	0.084	0.112	0.140	**0.148**	0.215	0.264	0.320	**0.323**	0.332	0.389	**0.472**	0.460
300	300	0.182	0.182	**0.308**	0.274	0.404	0.436	**0.593**	0.554	0.574	0.588	**0.719**	0.660

4.1 Sub-sampling

Since the sizes of two samples ($m = 1112$, $n = 1144$) are very large compared to the typical sample size, the p value of each test is close to 0 and it is difficult to compare the performance of four tests. Therefore, considering the original sample as a population, we sub-sample m' and n' and examine corresponding powers with the significance levels 1%, 5% and 10%.

Table 9 summarizes the rejection power depending on the size of sub-samples among 2000 replicates. The two marginal tests perform best with similar power, followed by the HT and CB tests. This result is consistent with our findings from the numerical study, especially for (SN-C.S) and (N-C.V2) with small ρ.

5 Conclusion

In this paper, we test equality of two population of MM-type interval samples by testing their real-valued representations. We first consider the Hotelling's T^2 test to examine the equality of mean vectors of the center and range of interval-valued data. We then propose marginalization-based test statistics, T_M^{UK} and T_M^{GK}, which are based on two univariate distributional representation (named as *marginalization* in this paper) of the interval-valued data.

Numerical study and real data analysis show that the marginalization-based tests perform better than the existing methods when two population distributions are different due to more than one factor, such as mean, covariance, skewness, and so on. This implies that the marginal tests can be more suitable for testing real problems of interval-valued data. Further, the power of the GK test is much higher than that of the UK when the two populations differ in range and covariance.

However, we need to be cautious when we apply the marginalization-based tests since they are only valid for the case that the null hypothesis is rejected. That is, the rejection of the equality test using the marginalization implies that two bivariate distributions are unequal. On the other hand, the acceptance of the null hypothesis does not imply the equality of two bivariate distributions.

Finally, it is worth remarking that both the marginalization (or univariate real-valued representation) and bivariate real-valued representation (e.g. (L, U) or $(C, \log R)$) are induced by the probability measure on intervals, but the converse is not. To be specific, the interval-valued data is a univariate random object on an appropriately defined probability space of intervals. For example, suppose we consider a sample space of intervals, say Ω, equipped with a metric $d(\omega_1, \omega_2)$ for $\omega_1, \omega_2 \in \Omega$. The metric introduces the Borel σ–field, say $\mathcal{F}$, and the probability measure $\mathcal{P}$ is defined on $\mathcal{F}$. In this paper, we write the interval-valued data as the form of $(L(\omega), U(\omega)]$, where $L(\omega)$ and $U(\omega)$ are real-valued random variables on the above probability space and their joint distribution, say $F(\ell, u)$, is induced by the probability measure $\mathcal{P}$. Thus, in this paper, we test equality of two probability measures by testing the equality of "their real-valued representations".

Acknowledgements Johan Lim and Minjung Kwak was supported by the National Research Foundation of Korea (NRF-2017R1A2B2012264, NRF-2016R1D1A1B03931050). Donghyeon Yu's research was supported by INHA UNIVERSITY Research Grant.

Appendix 1: Additional results of numerical study with $\alpha = 0.01, 0.1$

See Tables 10, 11, 12, 13, 14, 15, 16 and 17.

Table 10 Simulation results for the significance level 1%. Power of each test in case of the bivariate normal distribution with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-C)	(30, 30)	0.0	0.010	0.012	0.009	0.008	0.009	0.012	0.010	0.009	0.007	0.012	0.008	0.008
		0.5	0.119	0.175	**0.198**	0.173	0.105	0.210	**0.211**	0.197	0.108	**0.597**	0.279	0.277
		1.0	0.663	**0.798**	0.795	0.775	0.645	**0.878**	0.820	0.805	0.661	**1.000**	0.923	0.915
		1.5	0.978	**0.998**	0.993	0.991	0.981	**0.999**	0.996	0.995	0.980	**1.000**	0.999	**1.000**
	(30, 120)	0.0	0.009	0.010	0.008	0.008	0.008	0.010	0.009	0.007	0.012	0.010	0.011	0.011
		0.5	0.235	**0.351**	0.330	0.330	0.231	**0.422**	0.354	0.331	0.219	**0.868**	0.485	0.445
		1.0	0.906	**0.972**	0.958	0.952	0.902	**0.992**	0.972	0.965	0.904	**1.000**	0.995	0.993
		1.5	**1.000**	**1.000**	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**
	(50, 50)	0.0	0.007	0.011	0.007	0.007	0.007	0.011	0.007	0.007	0.008	0.011	0.010	0.009
		0.5	0.237	0.344	**0.364**	0.354	0.234	**0.426**	0.380	0.367	0.214	**0.881**	0.516	0.505
		1.0	0.922	**0.977**	0.970	0.964	0.922	**0.992**	0.974	0.970	0.908	**1.000**	0.994	0.994
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 200)	0.0	0.008	0.009	0.010	0.009	0.007	0.009	0.008	0.011	0.007	0.009	0.011	0.010
		0.5	0.435	**0.594**	0.565	0.546	0.448	**0.693**	0.581	0.588	0.412	**0.992**	0.760	0.744
		1.0	0.995	**1.000**	0.998	0.997	0.997	**1.000**	0.999	0.999	0.992	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(N-R)	(30, 30)	0.0	0.010	0.012	0.009	0.008	0.009	0.012	0.010	0.009	0.007	0.012	0.008	0.008
		0.5	0.110	**0.180**	0.008	0.013	0.112	**0.222**	0.009	0.014	0.105	**0.589**	0.016	0.021
		1.0	0.638	**0.803**	0.016	0.085	0.632	**0.883**	0.027	0.101	0.644	**1.000**	0.123	0.124
		1.5	0.978	**0.997**	0.135	0.503	0.981	**1.000**	0.225	0.560	0.980	**1.000**	0.661	0.755
	(30, 120)	0.0	0.009	0.010	0.008	0.008	0.008	0.010	0.009	0.007	0.012	0.010	0.011	0.011
		0.5	0.244	**0.346**	0.011	0.022	0.242	**0.415**	0.012	0.027	0.257	**0.865**	0.037	0.064
		1.0	0.929	**0.976**	0.039	0.291	0.925	**0.992**	0.078	0.360	0.929	**1.000**	0.390	0.561

Table 10 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
		1.5	**1.000**	**1.000**	0.468	0.887	**1.000**	**1.000**	0.643	0.928	**1.000**	**1.000**	0.976	0.997
	(50, 50)	0.0	0.007	0.011	0.007	0.007	0.007	0.011	0.007	0.007	0.008	0.011	0.010	0.009
		0.5	0.241	**0.327**	0.005	0.014	0.235	**0.416**	0.009	0.017	0.230	**0.884**	0.035	0.039
		1.0	0.932	**0.979**	0.043	0.263	0.933	**0.993**	0.083	0.315	0.931	**1.000**	0.383	0.422
		1.5	**1.000**	**1.000**	0.519	0.896	**1.000**	**1.000**	0.668	0.939	**1.000**	**1.000**	0.982	0.993
	(50, 200)	0.0	0.008	0.009	0.010	0.009	0.007	0.009	0.008	0.011	0.007	0.009	0.011	0.010
		0.5	0.468	**0.604**	0.012	0.041	0.497	**0.695**	0.016	0.060	0.476	**0.989**	0.075	0.113
		1.0	0.996	**1.000**	0.161	0.625	0.999	**1.000**	0.254	0.744	0.997	**1.000**	0.772	0.879
		1.5	**1.000**	**1.000**	0.957	0.995	**1.000**	**1.000**	0.983	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**0

Table 11 Simulation results for the significance level 10%. Power of each test in case of the bivariate normal distribution with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-C)	(30, 30)	0.0	0.094	0.102	0.101	0.101	0.094	0.102	0.103	0.098	0.087	0.102	0.095	0.095
		0.5	0.416	0.513	**0.529**	0.502	0.405	**0.581**	0.550	0.524	0.353	**0.883**	0.646	0.633
		1.0	0.918	**0.968**	0.966	0.959	0.913	**0.986**	0.976	0.965	0.905	**1.000**	0.991	0.991
		1.5	**1.000**	**1.000**	**1.000**	0.999	0.999	**1.000**	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**
	(30, 120)	0.0	0.087	0.100	0.100	0.098	0.090	0.100	0.098	0.092	0.091	0.100	0.100	0.097
		0.5	0.550	0.700	**0.703**	0.689	0.555	**0.756**	0.722	0.702	0.559	**0.982**	0.813	0.786
		1.0	0.987	0.998	**0.999**	0.997	0.985	**1.000**	0.998	0.998	0.990	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 50)	0.0	0.089	0.102	0.090	0.102	0.086	0.102	0.091	0.095	0.085	0.102	0.096	0.096
		0.5	0.576	0.708	**0.714**	0.684	0.574	**0.785**	0.740	0.724	0.592	**0.987**	0.843	0.843
		1.0	0.993	**0.999**	0.998	0.999	0.993	**1.000**	0.998	0.998	0.995	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 200)	0.0	0.088	0.089	0.098	0.093	0.086	0.089	0.100	0.103	0.088	0.089	0.101	0.100
		0.5	0.782	**0.896**	0.887	0.866	0.790	**0.933**	0.901	0.894	0.794	**1.000**	0.953	0.947
		1.0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(N-R)	(30, 30)	0.0	0.094	0.102	0.101	0.101	0.094	0.102	0.103	0.098	0.087	0.102	0.095	0.095
		0.5	0.382	**0.500**	0.104	0.146	0.370	**0.555**	0.114	0.152	0.360	**0.865**	0.182	0.169
		1.0	0.905	**0.966**	0.252	0.502	0.907	**0.983**	0.325	0.537	0.916	**1.000**	0.615	0.647
		1.5	0.999	**1.000**	0.786	0.917	0.999	**1.000**	0.870	0.949	0.999	**1.000**	0.990	0.997
	(30, 120)	0.0	0.087	0.100	0.100	0.098	0.090	0.100	0.098	0.092	0.091	0.100	0.100	0.097
		0.5	0.566	**0.702**	0.138	0.236	0.574	**0.767**	0.156	0.275	0.568	**0.980**	0.252	0.351
		1.0	0.987	**0.997**	0.517	0.795	0.990	**0.998**	0.616	0.852	0.992	**1.000**	0.876	0.952

Table 11 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
		1.5	**1.000**	**1.000**	0.985	0.995	**1.000**	**1.000**	0.994	0.999	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 50)	0.0	0.089	0.102	0.090	0.102	0.086	0.102	0.091	0.095	0.085	0.102	0.096	0.096
		0.5	0.588	**0.718**	0.118	0.188	0.593	**0.793**	0.137	0.213	0.594	**0.986**	0.263	0.261
		1.0	0.991	**0.999**	0.507	0.771	0.992	**1.000**	0.595	0.825	0.995	**1.000**	0.886	0.906
		1.5	**1.000**	**1.000**	0.978	0.995	**1.000**	**1.000**	0.993	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 200)	0.0	0.088	0.089	0.098	0.093	0.086	0.089	0.100	0.103	0.088	0.089	0.101	0.100
		0.5	0.808	**0.901**	0.181	0.334	0.801	**0.941**	0.232	0.406	0.825	**0.999**	0.397	0.500
		1.0	**1.000**	**1.000**	0.824	0.954	**1.000**	**1.000**	0.884	0.980	**1.000**	**1.000**	0.985	0.997
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 12 Simulation results for the significance level 1%. Power of each test in case of the bivariate t-distribution with df 5 with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(T-C)	(30, 30)	0.0	0.007	0.012	0.010	0.010	0.010	0.012	0.008	0.011	0.010	0.012	0.009	0.009
		0.5	0.099	0.103	**0.151**	0.069	0.092	0.128	**0.161**	0.089	0.090	**0.390**	0.203	0.163
		1.0	0.554	0.557	**0.685**	0.527	0.554	0.651	**0.725**	0.571	0.556	**0.967**	0.838	0.777
		1.5	0.934	0.928	**0.968**	0.933	0.937	0.958	**0.977**	0.944	0.935	**1.000**	0.994	0.991
	(30, 120)	0.0	0.008	0.010	0.010	0.008	0.007	0.010	0.012	0.008	0.007	0.010	0.014	0.010
		0.5	0.172	0.199	**0.258**	0.159	0.197	0.244	**0.276**	0.172	0.195	**0.618**	0.366	0.304
		1.0	0.815	0.818	**0.902**	0.797	0.845	0.881	**0.922**	0.839	0.848	**1.000**	0.971	0.954
		1.5	0.998	0.997	**1.000**	0.997	0.997	0.999	**1.000**	0.998	0.999	**1.000**	**1.000**	**1.000**
	(50, 50)	0.0	0.010	0.010	0.008	0.013	0.008	0.010	0.013	0.008	0.007	0.010	0.011	0.011
		0.5	0.191	0.193	**0.284**	0.177	0.192	0.240	**0.318**	0.187	0.184	**0.616**	0.412	0.362
		1.0	0.856	0.820	**0.918**	0.836	0.864	0.897	**0.935**	0.867	0.858	**1.000**	0.983	0.973
		1.5	0.998	0.998	**1.000**	0.996	0.998	**1.000**	**1.000**	0.999	0.999	**1.000**	**1.000**	**1.000**
	(50, 200)	0.0	0.008	0.007	0.009	0.009	0.009	0.007	0.012	0.009	0.009	0.007	0.013	0.012
		0.5	0.369	0.331	**0.476**	0.306	0.369	0.409	**0.500**	0.362	0.371	**0.872**	0.643	0.554
		1.0	0.982	0.972	**0.990**	0.972	0.985	0.993	**0.996**	0.982	0.982	**1.000**	0.999	0.998
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(T-R)	(30, 30)	0.0	0.007	0.012	0.010	0.010	0.010	0.012	0.008	0.011	0.010	0.012	0.009	0.009
		0.5	**0.099**	0.096	0.005	0.011	0.096	**0.124**	0.009	0.012	0.098	**0.366**	0.016	0.018
		1.0	**0.567**	0.550	0.015	0.051	0.565	**0.655**	0.027	0.054	0.605	**0.963**	0.099	0.083
		1.5	**0.948**	0.919	0.082	0.203	0.952	**0.961**	0.142	0.235	0.950	**0.999**	0.450	0.390
	(30, 120)	0.0	0.008	0.010	0.010	0.008	0.007	0.010	0.012	0.008	0.007	0.010	0.014	0.010
		0.5	0.182	**0.194**	0.012	0.016	0.201	**0.239**	0.014	0.023	0.208	**0.629**	0.028	0.047
		1.0	**0.849**	0.811	0.044	0.153	0.859	**0.887**	0.082	0.196	0.874	**0.997**	0.277	0.381

Table 12 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
		1.5	**0.997**	0.994	0.356	0.574	**0.998**	0.996	0.473	0.658	0.998	**1.000**	0.843	0.886
	(50, 50)	0.0	0.010	0.010	0.008	0.013	0.008	0.010	0.013	0.008	0.007	0.010	0.011	0.011
		0.5	0.182	**0.200**	0.013	0.022	0.174	**0.255**	0.015	0.018	0.183	**0.644**	0.038	0.034
		1.0	**0.860**	0.821	0.044	0.135	0.860	**0.892**	0.084	0.143	0.858	**0.998**	0.264	0.239
		1.5	**0.999**	0.995	0.379	0.511	0.998	**0.999**	0.506	0.566	0.999	**1.000**	0.866	0.800
	(50, 200)	0.0	0.008	0.007	0.009	0.009	0.009	0.007	0.012	0.009	0.009	0.007	0.013	0.012
		0.5	**0.380**	0.332	0.018	0.032	0.384	**0.401**	0.024	0.043	0.376	**0.873**	0.062	0.083
		1.0	**0.985**	0.972	0.120	0.319	0.984	**0.990**	0.228	0.441	0.988	**1.000**	0.590	0.643
		1.5	**1.000**	**1.000**	0.819	0.846	**1.000**	**1.000**	0.894	0.905	**1.000**	**1.000**	0.992	0.986

Table 13 Simulation results for the significance level 10%. Power of each test in case of the bivariate t-distribution with df 5 with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(T-C)	(30, 30)	0.0	0.088	0.114	0.113	0.105	0.085	0.115	0.114	0.107	0.085	0.115	0.114	0.111
		0.5	0.339	0.384	**0.459**	0.381	0.310	0.430	**0.474**	0.413	0.303	**0.739**	0.557	0.524
		1.0	0.835	0.848	**0.921**	0.865	0.831	0.899	**0.937**	0.890	0.847	**0.999**	0.965	0.961
		1.5	0.989	0.989	**0.998**	0.993	0.988	0.996	**0.998**	0.997	0.992	**1.000**	**1.000**	0.999
	(30, 120)	0.0	0.089	0.099	0.103	0.101	0.091	0.099	0.097	0.103	0.086	0.099	0.100	0.102
		0.5	0.494	0.500	**0.619**	0.517	0.519	0.570	**0.634**	0.538	0.508	**0.878**	0.720	0.654
		1.0	0.969	0.953	**0.986**	0.974	0.974	0.979	**0.989**	0.977	0.980	**1.000**	0.999	0.996
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 50)	0.0	0.090	0.097	0.107	0.108	0.085	0.097	0.103	0.098	0.085	0.097	0.098	0.094
		0.5	0.518	0.523	**0.626**	0.542	0.521	0.580	**0.653**	0.578	0.533	**0.882**	0.736	0.708
		1.0	0.976	0.965	**0.994**	0.976	0.981	0.983	**0.995**	0.988	0.986	**1.000**	0.999	0.998
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(50, 200)	0.0	0.089	0.097	0.091	0.100	0.086	0.097	0.091	0.112	0.086	0.097	0.101	0.100
		0.5	0.724	0.688	**0.812**	0.722	0.714	0.770	**0.827**	0.763	0.718	**0.978**	0.898	0.873
		1.0	0.999	0.999	**1.000**	0.999	0.998	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(T-R)	(30, 30)	0.0	0.088	0.114	0.113	0.105	0.085	0.115	0.114	0.107	0.085	0.115	0.114	0.111
		0.5	0.365	**0.366**	0.093	0.124	0.318	**0.413**	0.100	0.131	0.305	**0.740**	0.153	0.152
		1.0	**0.879**	0.860	0.242	0.333	0.849	**0.899**	0.290	0.377	0.854	**0.995**	0.491	0.480
		1.5	**0.993**	0.989	0.676	0.701	**0.993**	0.992	0.734	0.757	0.997	**1.000**	0.920	0.899
	(30, 120)	0.0	0.089	0.099	0.103	0.101	0.091	0.099	0.097	0.103	0.086	0.099	0.100	0.102
		0.5	**0.529**	0.511	0.136	0.190	0.524	**0.565**	0.141	0.228	0.527	**0.869**	0.235	0.304
		1.0	**0.975**	0.958	0.442	0.600	**0.978**	0.974	0.507	0.664	0.981	**1.000**	0.751	0.815

Table 13 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
		1.5	**1.000**	0.999	0.932	0.911	**1.000**	**1.000**	0.959	0.946	**1.000**	**1.000**	0.990	0.993
	(50, 50)	0.0	0.090	0.097	0.107	0.108	0.085	0.097	0.103	0.098	0.085	0.097	0.098	0.094
		0.5	0.541	**0.543**	0.131	0.176	0.547	**0.611**	0.150	0.184	0.551	**0.899**	0.224	0.230
		1.0	**0.984**	0.967	0.429	0.510	**0.983**	0.982	0.500	0.558	0.985	**1.000**	0.760	0.706
		1.5	**1.000**	0.999	0.922	0.899	**1.000**	**1.000**	0.955	0.925	**1.000**	**1.000**	0.995	0.990
	(50, 200)	0.0	0.089	0.097	0.091	0.100	0.086	0.097	0.091	0.112	0.086	0.097	0.101	0.100
		0.5	**0.715**	0.683	0.183	0.265	0.726	**0.762**	0.196	0.302	0.736	**0.986**	0.338	0.388
		1.0	**0.999**	0.997	0.706	0.775	**0.999**	**0.999**	0.770	0.828	0.999	**1.000**	0.937	0.939
		1.5	**1.000**	**1.000**	0.999	0.987	**1.000**	**1.000**	0.999	0.993	**1.000**	**1.000**	**1.000**	**1.000**

Table 14 Simulation results for the significance level 1%. Power of each test in case of the bivariate skew-normal distribution with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(SN-C)	(30, 30)	0.0	0.008	0.007	0.007	0.007	0.007	0.008	0.008	0.009	0.007	0.008	0.012	0.012
		0.5	0.123	**0.168**	0.163	0.158	0.122	**0.210**	0.183	0.163	0.123	**0.590**	0.252	0.243
		1.0	0.701	**0.788**	0.734	0.739	0.701	**0.862**	0.781	0.755	0.704	**0.999**	0.894	0.875
		1.5	0.988	**0.997**	0.992	0.993	0.988	**1.000**	0.996	0.994	0.988	**1.000**	**1.000**	**1.000**
	(30, 120)	0.0	0.010	0.009	0.009	0.010	0.010	0.011	0.008	0.010	0.008	0.010	0.010	0.011
		0.5	0.253	**0.330**	0.287	0.286	0.256	**0.407**	0.309	0.278	0.255	**0.876**	0.444	0.406
		1.0	0.958	**0.975**	0.959	0.962	0.958	**0.990**	0.956	0.952	0.958	**1.000**	0.993	0.990
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(SN-C.S)	(30, 30)	0.0	0.008	0.010	0.008	0.010	0.007	0.010	0.009	0.010	0.007	0.008	0.009	0.009
		0.5	0.128	0.165	**0.195**	**0.195**	0.129	0.203	**0.217**	0.212	0.128	**0.563**	0.306	0.301
		1.0	0.715	0.779	**0.812**	0.810	0.714	**0.862**	0.848	0.834	0.717	**1.000**	0.934	0.928
		1.5	0.988	0.990	**0.994**	**0.994**	0.988	**0.998**	**0.998**	**0.998**	0.988	**1.000**	**1.000**	**1.000**
	(30, 120)	0.0	0.009	0.011	0.007	0.010	0.010	0.011	0.008	0.009	0.007	0.010	0.009	0.009
		0.5	0.298	0.363	0.395	**0.405**	0.296	**0.436**	0.411	0.396	0.299	**0.881**	0.538	0.496
		1.0	0.946	**0.984**	0.974	0.975	0.945	**0.992**	0.982	0.980	0.947	**1.000**	0.998	0.997
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 15 Simulation results for the significance level 10%. Power of each test in case of the bivariate skew-normal distribution with equal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(SN-C)	(30, 30)	0.0	0.087	0.099	0.102	0.091	0.085	0.107	0.103	0.097	0.085	0.107	0.109	0.111
		0.5	0.395	**0.509**	0.483	0.473	0.391	**0.560**	0.506	0.493	0.383	**0.886**	0.607	0.586
		1.0	0.923	**0.957**	0.955	0.958	0.923	**0.983**	0.961	0.955	0.931	**1.000**	0.985	0.983
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(30, 120)	0.0	0.102	0.101	0.108	0.099	0.093	0.096	0.107	0.103	0.090	0.108	0.100	0.108
		0.5	0.613	**0.686**	0.656	0.655	0.609	**0.750**	0.673	0.655	0.620	**0.978**	0.773	0.736
		1.0	0.998	**0.999**	0.998	0.998	0.998	**1.000**	0.997	0.997	0.999	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
(SN-C.S)	(30, 30)	0.0	0.086	0.101	0.091	0.092	0.086	0.102	0.094	0.095	0.087	0.103	0.099	0.100
		0.5	0.388	0.505	**0.559**	0.547	0.387	0.553	**0.567**	0.546	0.382	**0.887**	0.666	0.657
		1.0	0.918	0.956	**0.970**	0.965	0.921	**0.985**	0.974	0.971	0.926	**1.000**	0.995	0.994
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(30, 120)	0.0	0.093	0.109	0.107	0.113	0.092	0.107	0.104	0.110	0.086	0.113	0.101	0.101
		0.5	0.636	0.705	**0.745**	0.734	0.630	**0.768**	0.760	0.739	0.635	**0.979**	0.844	0.827
		1.0	0.998	**0.999**	0.998	**0.999**	0.998	**1.000**	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 16 Simulation results for the significance level 1%. Power of each test in case of the bivariate normal distribution with unequal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-COV)	(30, 30)	0.0	0.010	0.013	0.009	0.008	0.009	0.012	0.010	0.009	0.007	0.013	0.008	0.008
		0.5	0.012	0.012	0.022	**0.030**	0.013	0.012	0.021	**0.031**	0.011	0.012	0.025	**0.033**
		1.0	0.017	0.012	0.039	**0.070**	0.016	0.012	0.048	**0.094**	0.015	0.012	0.067	**0.111**
		1.5	0.031	0.009	0.065	**0.170**	0.022	0.009	0.077	**0.205**	0.021	0.009	0.143	**0.249**
	(30, 120)	0.0	0.009	0.010	0.007	0.008	0.008	0.010	0.009	0.007	0.012	0.010	0.011	0.011
		0.5	0.010	0.013	0.013	**0.024**	0.010	0.013	0.014	**0.032**	0.010	0.013	0.021	**0.048**
		1.0	0.018	0.013	0.026	**0.091**	0.013	0.013	0.039	**0.144**	0.017	0.013	0.091	**0.221**
		1.5	0.030	0.010	0.075	**0.264**	0.022	0.010	0.103	**0.372**	0.019	0.010	0.253	**0.524**
(N-C.V1)	(30, 30)	0.0	0.010	0.015	0.009	0.008	0.009	0.013	0.010	0.009	0.007	0.013	0.008	0.008
		0.5	0.105	0.114	**0.211**	0.190	0.096	0.133	**0.187**	0.164	0.095	**0.361**	0.191	0.181
		1.0	0.479	0.399	**0.700**	0.685	0.463	0.494	**0.666**	0.630	0.462	**0.889**	0.701	0.677
		1.5	0.824	0.726	**0.947**	0.931	0.808	0.812	**0.928**	0.911	0.807	**0.995**	0.944	0.936
	(30, 120)	0.0	0.009	0.010	0.007	0.008	0.008	0.010	0.009	0.007	0.012	0.010	0.011	0.011
		0.5	0.198	0.261	**0.333**	**0.333**	0.193	**0.315**	0.294	0.271	0.178	**0.718**	0.339	0.305
		1.0	0.818	0.860	**0.926**	0.924	0.817	**0.918**	0.902	0.890	0.821	**1.000**	0.927	0.913
		1.5	0.992	0.995	**0.999**	**0.999**	0.992	**0.999**	0.997	0.998	0.993	**1.000**	0.999	0.998
(N-C.V2)	(30, 30)	0.0	0.008	0.013	0.009	0.009	0.011	0.013	0.010	0.010	0.008	0.015	0.008	0.009
		0.5	0.035	0.046	**0.051**	0.047	0.033	**0.056**	0.054	0.053	0.029	**0.135**	0.064	0.066
		1.0	0.218	0.252	**0.296**	0.273	0.215	0.305	**0.335**	0.318	0.207	**0.729**	0.429	0.450
		1.5	0.802	0.745	**0.849**	0.835	0.804	0.819	**0.900**	0.886	0.802	**0.994**	0.962	0.969
	(30, 120)	0.0	0.008	0.010	0.011	0.009	0.009	0.010	0.008	0.009	0.009	0.010	0.009	0.008
		0.5	0.055	0.064	**0.097**	0.088	0.059	0.075	**0.110**	0.097	0.054	**0.207**	0.148	0.140
		1.0	0.432	0.375	**0.534**	0.516	0.437	0.446	**0.590**	0.571	0.433	**0.865**	0.707	0.694

Table 16 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
		1.5	0.957	0.826	**0.965**	0.960	0.957	0.889	**0.985**	0.978	0.958	**0.999**	0.997	0.998
(N-R.V)	(30, 30)	0.0	0.010	0.013	0.009	0.008	0.009	0.013	0.010	0.009	0.007	0.012	0.008	0.008
		0.5	0.112	0.115	0.012	**0.129**	0.118	**0.135**	0.012	0.126	0.103	**0.363**	0.034	0.085
		1.0	0.464	0.394	0.032	**0.713**	0.460	0.478	0.058	**0.722**	0.470	**0.890**	0.194	0.587
		1.5	0.822	0.744	0.142	**0.970**	0.817	0.822	0.236	**0.969**	0.820	**0.994**	0.576	0.952
	(30, 120)	0.0	0.009	0.010	0.007	0.008	0.008	0.010	0.009	0.007	0.012	0.010	0.011	0.011
		0.5	0.236	0.255	0.012	**0.280**	0.220	**0.314**	0.019	0.307	0.233	**0.727**	0.088	0.294
		1.0	0.855	0.862	0.060	**0.967**	0.846	0.915	0.128	**0.966**	0.866	**1.000**	0.502	0.950
		1.5	0.994	0.997	0.435	**1.000**	0.995	0.999	0.608	**1.000**	0.997	**1.000**	0.955	**1.000**

Table 17 Simulation results for the significance level 10%. Power of each test in case of the bivariate normal distribution with unequal covariances. Numbers in bold denote the highest power among CB, HT, UK and GK for each simulation setting except $\delta = 0$ where numbers denote the size of test

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-COV)	(30, 30)	0.0	0.094	0.102	0.101	0.101	0.094	0.102	0.103	0.098	0.087	0.102	0.095	0.095
		0.5	0.127	0.113	0.156	**0.190**	0.116	0.113	0.168	**0.198**	0.093	0.113	0.183	**0.206**
		1.0	0.173	0.101	0.269	**0.373**	0.163	0.101	0.283	**0.385**	0.132	0.101	0.335	**0.412**
		1.5	0.242	0.109	0.417	**0.571**	0.231	0.109	0.442	**0.592**	0.186	0.109	0.532	**0.631**
	(30, 120)	0.0	0.088	0.105	0.100	0.098	0.090	0.105	0.098	0.092	0.086	0.105	0.100	0.097
		0.5	0.093	0.093	0.156	**0.215**	0.091	0.093	0.156	**0.247**	0.077	0.093	0.199	**0.287**
		1.0	0.159	0.101	0.303	**0.484**	0.167	0.101	0.331	**0.560**	0.133	0.101	0.444	**0.652**
		1.5	0.253	0.092	0.496	**0.734**	0.265	0.092	0.549	**0.815**	0.215	0.092	0.704	**0.892**
(N-C.V1)	(30, 30)	0.0	0.094	0.102	0.101	0.101	0.094	0.102	0.103	0.098	0.087	0.102	0.095	0.095
		0.5	0.392	0.392	**0.545**	0.515	0.389	0.439	**0.510**	0.483	0.354	**0.724**	0.519	0.496
		1.0	0.813	0.784	**0.929**	0.917	0.822	0.839	**0.907**	0.890	0.823	**0.987**	0.921	0.907
		1.5	0.976	0.953	**0.995**	0.993	0.978	0.973	**0.994**	0.991	0.974	**1.000**	0.996	0.995
	(30, 120)	0.0	0.086	0.105	0.100	0.098	0.090	0.105	0.098	0.092	0.087	0.105	0.100	0.097
		0.5	0.546	0.617	**0.727**	0.712	0.552	0.673	**0.680**	0.659	0.560	**0.950**	0.685	0.668
		1.0	0.971	0.980	**0.997**	**0.997**	0.975	**0.994**	0.991	0.991	0.974	**1.000**	0.993	0.991
		1.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		0.5	0.176	0.225	**0.258**	0.251	0.176	0.255	**0.277**	0.265	0.163	**0.437**	0.307	0.304
		1.0	0.560	0.619	**0.688**	0.664	0.564	0.675	**0.738**	0.733	0.552	**0.952**	0.811	0.814
		1.5	0.960	0.943	**0.979**	0.976	0.968	0.969	**0.990**	**0.990**	0.970	**1.000**	0.998	0.999
	(30, 120)	0.0	0.087	0.105	0.107	0.097	0.091	0.105	0.096	0.097	0.091	0.105	0.096	0.098
		0.5	0.266	0.297	**0.353**	0.338	0.270	0.326	**0.370**	0.355	0.254	**0.553**	0.425	0.403
		1.0	0.771	0.725	**0.842**	0.827	0.772	0.789	**0.875**	0.863	0.775	**0.983**	0.925	0.920
		1.5	0.995	0.969	**1.000**	0.999	0.998	0.988	**1.000**	**1.000**	0.998	**1.000**	**1.000**	**1.000**

Table 17 (continued)

Case	(m, n)	δ	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-R.V)	(30, 30)	0.0	0.094	0.102	0.101	0.101	0.094	0.102	0.103	0.098	0.087	0.102	0.095	0.095
		0.5	0.392	0.383	0.147	**0.501**	0.383	0.435	0.176	**0.493**	0.391	**0.715**	0.283	0.419
		1.0	0.828	0.783	0.386	**0.935**	0.837	0.841	0.470	**0.947**	0.835	**0.989**	0.729	0.934
		1.5	0.978	0.953	0.775	**0.998**	0.979	0.974	0.844	**0.997**	0.977	**1.000**	0.973	0.999
	(30, 120)	0.0	0.088	0.105	0.100	0.098	0.090	0.105	0.098	0.092	0.088	0.105	0.100	0.097
		0.5	0.577	0.612	0.199	**0.750**	0.612	0.674	0.228	**0.758**	0.585	**0.944**	0.420	0.753
		1.0	0.983	0.981	0.704	**0.998**	0.984	0.992	0.791	**0.997**	0.984	**1.000**	0.948	1.000
		1.5	**1.000**	**1.000**	0.990	**1.000**	**1.000**	**1.000**	0.996	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Appendix 2: Power comparison of UK, GK$_h$, and GK$_{h_{\max}}$

In this section, we compare the powers of the UK, the GK$_h$ and the GK$_{h_{\max}}$ tests, where GK$_{h_{\max}}$ and GK$_h$ denote the tests based on the Gaussian kernel estimates with the proposed $h_{\max}$ and the optimal bandwidths separately chosen by Jeon et al. (2015), respectively.

To be more specific, as described in Jeon et al. (2015), the Gaussian kernel density estimator is a smoothing version of the uniform kernel estimator and selects h that minimizes the following Kullback–Leibler loss:

$$-\int f_n^{UK}(x)\log f_n(x;h)dx,$$

Table 18 Simulation results of UK, GK$_h$ and GK$_{h_{\max}}$ for the significance level 5%. Power of each test is reported when (m,n) is (30, 30)

Case	δ	$\rho = 0$			$\rho = 0.8$		
		UK	GK$_h$	GK$_{h_{\max}}$	UK	GK$_h$	GK$_{h_{\max}}$
(N-C)	0.0	0.048	0.046	0.046	0.046	0.041	0.047
	0.5	0.402	0.393	0.387	0.520	0.507	0.498
	1.0	0.928	0.923	0.916	0.982	0.976	0.979
	1.5	0.998	0.999	0.998	1.000	1.000	1.000
(N-R)	0.0	0.048	0.046	0.046	0.046	0.041	0.047
	0.5	0.043	0.041	0.071	0.089	0.077	0.083
	1.0	0.109	0.115	0.315	0.409	0.314	0.433
	1.5	0.537	0.487	0.822	0.952	0.883	0.978
(T-C)	0.0	0.059	0.050	0.049	0.057	0.056	0.055
	0.5	0.356	0.299	0.256	0.428	0.395	0.398
	1.0	0.870	0.831	0.775	0.940	0.924	0.925
	1.5	0.994	0.990	0.979	0.998	0.999	0.998
(T-R)	0.0	0.059	0.050	0.049	0.057	0.056	0.055
	0.5	0.044	0.042	0.067	0.083	0.077	0.085
	1.0	0.119	0.111	0.214	0.329	0.246	0.316
	1.5	0.436	0.340	0.531	0.809	0.667	0.798
(N-COV)	0.0	0.048	0.046	0.046	0.046	0.041	0.047
	0.5	0.080	0.084	0.104	0.098	0.100	0.115
	1.0	0.155	0.161	0.227	0.217	0.236	0.280
	1.5	0.239	0.278	0.408	0.384	0.415	0.496
(N-C.V1)	0.0	0.048	0.046	0.046	0.046	0.041	0.047
	0.5	0.420	0.398	0.401	0.402	0.395	0.374
	1.0	0.865	0.862	0.837	0.865	0.859	0.841
	1.5	0.988	0.977	0.981	0.990	0.981	0.987
(N-R.V)	0.0	0.048	0.046	0.046	0.046	0.041	0.047
	0.5	0.067	0.074	0.351	0.166	0.167	0.256
	1.0	0.195	0.243	0.886	0.552	0.517	0.859
	1.5	0.517	0.540	0.993	0.905	0.863	0.995

where $f_n^{UK}(x)$ is the uniform kernel estimator (UK) and $f_n(x;h)$ is the Gaussian Kernel estimator in Jeon et al. (2015). In the numerical study, we select the optimal bandwidth h for each sample and denote the chosen bandwidths for $\mathbf{X}$ and $\mathbf{Y}$ as h_X and h_Y, respectively. We reported the results of powers for the case of $(m, n) = (30, 30)$, $\rho = (0.0, 0.8)$, and (N-C), (N-R), (T-C), (T-R), (N-COV), (N-C, V1), (N-R.V) in Table 18. Note that we have reduced the number of cases of the numerical study since the cost of GK_h test is expensive due to the fact that it needs to apply the bandwidth selection for every permutation to estimate null distribution.

From the result in Table 18, GK_h performs very similar to the UK for all cases we consider. This is caused by the bandwidth selection procedure choosing h that minimizes the Kulback–Leibler loss between the UK and the GK. Thus, as described in the numerical study of the main manuscript, $\mathrm{GK}_{h_{\max}}$ has similar powers to the GK_h for the cases of center change and much larger powers than GK_h for the cases of range change.

References

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society Series B (Methodological)*, *61*(3), 579–602.

Bertrand, P., & Goupil, F. (2000). Descriptive statistics for symbolic data. In H.-H. Bock & E. Diday (Eds.), *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data* (pp. 106–124). Berlin: Springer.

Blanco-Fernández, A., & Winker, P. (2016). Data generation processes and statistical management of interval data. *AStA Advances in Statistical Analysis*, *100*(4), 475–494.

Choi, H., Lim, J., Kwak, M., & Park, S. (2019). Testing for stochastic order in interval-valued data. *The Korean Journal of Applied Statistics*, *32*, 879–887.

Feller, W. (1948). On the Kolmogorov–Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics*, *19*(2), 177–189.

Grzegorzewski, P., & Śpiewak, M. (2017). The Mann–Whitney test for interval-valued data. In *EUSFLAT 2017, IWIFSGN 2017: Advances in fuzzy logic and technology* (pp. 188–199).

Grzegorzewski, P. (2018). The Kolmogorov–Smirnov goodness-of-fit test for interval-valued data. In E. Gil, et al. (Eds.), *The mathematics of the uncertain* (pp. 615–627). Berlin: Springer.

Grzegorzewski, P., & Śpiewak, M. (2019). The sign test and the signed-rank test for interval-valued data. *International Journal of Intelligent Systems*, *34*, 2122–2150.

Jeon, Y., Ahn, J., & Park, C. (2015). A nonparametric kernel approach to interval-valued data analysis. *Technometrics*, *57*(4), 566–575.

Perolat, J., Couso, I., Loquin, K., & Strauss, O. (2015). Generalizing the Wilcoxon rank-sum test for interval data. *International Journal of Approximate Reasoning*, *56*, 108–121.

Præstgaard, J. T. (1995). Permutation and bootstrap Kolmogorov–Smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, *22*(3), 305–322.