A Test of Homogeneity of Distributions when Observations are Subject to Measurement Errors

DongHyuk Lee[1], Soumendra N. Lahiri[2], and Samiran Sinha[3,*]

[1] Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute,
National Institutes of Health, Bethesda, MD, USA

[2]Department of Statistics, North Carolina State University, Raleigh, NC, USA

[3]Department of Statistics, Texas A&M University, College Station, TX, USA

*email: sinha@stat.tamu.edu

Summary

When the observed data are contaminated with errors, the standard two-sample testing approaches that ignore measurement errors may produce misleading results, including a higher Type-I error rate than the nominal level. To tackle this inconsistency, a nonparametric test is proposed for testing equality of two distributions when the observed contaminated data follow the classical additive measurement error model. The proposed test takes into account the presence of errors in the observed data, and the test statistic is defined in terms of the (deconvoluted) characteristic functions of the latent variables. Proposed method is applicable to a wide range of scenarios as no parametric restrictions are imposed either on the distribution of the underlying latent variables or on the distribution of the measurement errors. Asymptotic null distribution of the test statistic is derived which is given by an integral of a squared Gaussian process with a complicated covariance structure. For data based calibration of the test, a new nonparametric Bootstrap method is developed under the two-sample measurement error framework and its validity is established. Finite sample performance of the proposed test is investigated through simulation studies, and the results show superior performance of the proposed method than the standard tests that exhibit inconsistent behavior. Finally, the proposed method was applied to real datasets from the National Health and Nutrition Examination Survey. An `R` package `MEtest` is available on CRAN (`https://CRAN.R-project.org/package=MEtest`).

Key Words: Bootstrap; Characteristic function; Chi-square; Gaussian process; Power; Two sample test.

Running title: Homogeneity test under measurement error

1 Introduction

A common public health question is how the behavioral factors are associated with a biomarker, a health outcome, or a surrogate of a health outcome (Hogan et al., 2007; Puddy and Beilin, 2006; Primatesta et al., 2001). Suppose that we are interested in checking if alcohol consumption and the systolic blood pressure are associated. However, systolic blood pressure cannot be measured accurately, rather measured values are the underlying true blood pressure plus measurement errors. In an attempt to answer this question one may use the National Health and Nutrition Examination Survey (NHANES) data that contain multiple measurements on these variables that are subject to measurement uncertainty, and apply an existing two-sample test to the average of multiple measurements from the two behavioral groups, alcoholic and non-alcoholic.

There could be two cases, 1) where the interest is in comparing the means of the two groups and 2) where the interest is in comparing the two distributions corresponding to the two behavioral groups. Comparison of means makes sense when the two distributions have the similar shape. When observations are subject to classical measurement error (Carroll et al., 2006, Chap 1), existing two-sample tests of means are consistent. On the other hand, exiting two sample nonparametric tests, Kolmogorov-Smirnov test (Hodges, 1958) or Anderson-Darling test (Pettitt, 1976), for comparing two distributions are not consistent when observations are subject to classical measurement error. This lead us to develop a new two-sample testing method for checking equality of two distributions when the available data are measured with errors.

There are plenty of examples in epidemiology and medical sciences, where two distributions are compared, not just the means. In epidemiology, Jones (1997), Basu et al. (2015), just to name a few, compared groups via tests of equality of two distributions rather than the two mean parameters. In medical sciences, Hariharan et al. (2019) used Kolmogorov-Smirnov (K-S) test to compare images from low and high radiation dose, Liu et al. (2019) applied this test to compare two groups of electric signals for each genomic position, and Stephens et al. (2009) used the Mann-Whitney U-test and

K-sample Anderson-Darling (A-D) test in somatic rearrangement data from breast cancer genome, and these are a few among countless examples. Our formulation of the two sample testing problem will encompass the data structure like that of the NHANES data where repeated measurements are available on uncertain variables that include dietary intakes and other biomarkers.

Although errors in measured variables have received considerable attention from density estimation perspective (Delaigle and Hall, 2015; Carroll and Hall, 1988) and in the regression context (Carroll et al., 2006; Gustafson, 2003), to the best of our knowledge, no one has ever considered testing of homogeneity of two distributions when the observed data are subject to measurement errors.

Like errors-in-covariates in regression models, this problem can be tackled in several ways. First, one may model the distributions of true signals and measurement errors parametrically, and then test the homogeneity of distributions by checking equality of a set of parameters. However, any parametric approach may face misscpecification bias. Therefore, we do not wish to use any parametric model assumption. In the nonparametric context, one may estimate the two underlying densities from the two contaminated samples using any density deconvolution approach available in the literature (Delaigle et al., 2015; Delaigle et al., 2008). Then carry out a test based on the deconvoluted densities. Numerical instability is a well known phenomenon of deconvoluted density estimation, and that is due to the inverse transformation of the characteristic function (CF). To circumvent this problem, we design a test that is directly based on the CFs, and the test statistic itself does not depend on the deconvoluted density.

Briefly, our approach and organization of the rest of paper are given as follows. Section 2 contains the formulation of the test statistic based on the estimated CFs of the underlying true signals, and its asymptotic properties under the null hypothesis. The limit distribution has a complex form and it involves different unknown population parameters, making it less appealing to use for calibrating the test statistic. Due to this difficulty, in Section 3, we propose a novel Bootstrap approach under the measurement error framework that gives a theoretically valid data generation procedure under the null hypothesis, and that also constitutes an important contribution of this paper. Besides theoretical investigation of the large sample properties, finite sample properties of the test are judged via simulation studies (Section 4). The results of the simulation study show that the proposed testing method has competitive performance in terms of maintaining the size of the test, and superior power properties compared to its competitors, even when the two population distributions are not drastically different. An application of the methodology to an NHANES 2009-2010 survey data is given in Section 5, followed by some concluding remarks in Section 6. Proofs of the main results are given in the Appendix.

2 Testing methodology

2.1 Background and notation

Suppose that the cumulative distribution functions (CDFs) of X and Y are F_x and F_y, respectively. Our goal is to test the hypothesis $H_0 : F_x = F_y$ when the observed data are $D_w = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{n_x}\}$ and $D_v = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{n_y}\}$, where $\boldsymbol{W}_j^T = (W_{j1}, \ldots, W_{jm_x})$ and $\boldsymbol{V}_k^T = (V_{k1}, \ldots, V_{km_y})$ for $j = 1, \ldots, n_x$ and $k = 1, \ldots, n_y$. Assume that $m_x \geq 2$ and $m_y \geq 2$, and the observed W's and V's are related with the unobserved X's and Y's through the classical additive measurement error model (Carroll et al., 2006, Chap 1), that means,

$$W_{jl} = X_j + U_{x,jl} \text{ and } V_{kl'} = Y_k + U_{y,kl'},$$

for $j = 1, \ldots, n_x$, $l = 1, \ldots, m_x$, $k = 1, \ldots, n_y$, $l^{'} = 1, \ldots, m_y$. The measurement error $U_{x,jl}$'s are assumed to be iid, independent of X_j, and follows the distribution F_{u_x} that is symmetric around 0. Similarly, we assume that the measurement error $U_{y,kl}$ are iid, independent of Y_k, and follows the distribution F_{u_y} that is symmetric around 0. Further, X, Y, U_x and U_y are assumed to be independent. It is important to note that $\{X_1, \ldots, X_{n_x}\}$ and $\{Y_1, \ldots, Y_{n_y}\}$ are never observed. Also, the CDFs F_x, F_y, F_{u_x} and F_{u_y} are assumed to be absolutely continuous but otherwise left unknown.

Let ϕ_x, ϕ_y, ϕ_{u_x} and ϕ_{u_y} be the CFs of X, Y, U_x and U_y, respectively. Let $a_x(t)$ and $b_x(t)$ be the real and imaginary parts of $\phi_x(t)$, respectively. Similarly, define $a_y(t)$ and $b_y(t)$ from $\phi_y(t)$. For future reference, denote estimators of F_x, F_y, F_{u_x}, and F_{u_y}, by $\widehat{F}_x$, $\widehat{F}_y$, $\widehat{F}_{u_x}$, and $\widehat{F}_{u_y}$ respectively. Suppose that $F = F_x = F_y$ denotes the common distribution under H_0, and write $\widehat{F}$ to denote its estimator. Further, define $\overline{W}_j = \sum_{l=1}^{m_x} W_{jl}/m_x$, $\overline{V}_k = \sum_{l=1}^{m_y} V_{kl}/m_y$, $M_x = m_x(m_x-1)/2$, $M_y = m_y(m_y-1)/2$, $N_x = n_x M_x$, $N_y = n_y M_y$. Note that the CF of $\overline{W}_j$ is given by $\phi_1(t) = \phi_x(t)\{\phi_{u_x}(t/m_x)\}^{m_x}$, and that of $\overline{V}_k$ is $\phi_2(t) = \phi_y(t)\{\phi_{u_y}(t/m_y)\}^{m_y}$.

In the naive approach that ignores measurement errors in the observed data, one may first compute $\{\overline{W}_1, \dots, \overline{W}_{n_x}\}$ and $\{\overline{V}_1, \dots, \overline{V}_{n_y}\}$ and then apply any nonparametric testing procedure directly on these transformed data. Indeed, this naive method is usually inconsistent, that means, it fails to maintain the nominal type-I error level. If $m_x = m_y = m$, and $F_{u_x} = F_{u_y} = F_u$, then the CF of $\overline{W}_j$ and $\overline{V}_k$ are $\phi_1(t) = \phi_x(t)\{\phi_u(t/m)\}^m$ and $\phi_2(t) = \phi_y(t)\{\phi_u(t/m)\}^m$. Consequently the null hypothesis $H_0 : \phi_x(t) = \phi_y(t)$ implies $\phi_1(t) = \phi_2(t)$. That means, testing H_0 becomes equivalent to testing $H_0 : F_1 = F_2$, where F_1 and F_2 are the distribution functions of $\overline{W}_j$ and $\overline{V}_k$. Thus, when $m_x = m_y$ and $F_{u_x} = F_{u_y}$, the naive testing procedure is consistent for testing $H_0 : F_x = F_y$. However, if either $m_x \neq m_y$ or $F_{u_x} \neq F_{u_y}$, the naive test may not be consistent.

2.2 Development of the test statistic

We shall work under the standard condition (Delaigle et al., 2008) that $\phi_{u_x}(t)$ and $\phi_{u_y}(t)$ are real-valued function and do not vanish on $\mathbb{R}$, but do not impose any such conditions on the CFs ϕ_x and ϕ_y of the latent variables. The real valued CF condition results from the assumption that the error distribution is symmetric around zero. Further, as is well known (Stefanski and Carroll, 1990; Delaigle et al., 2008), the non vanishing assumption is also due to overcome the identifiability problem. Under these conditions, the CF for the measurement error can be recovered using the difference between two observations $W_1 - W_2$, where $W_1 = X + U_{x,1}$ and $W_2 = X + U_{x,2}$. Then, $\phi_{W_1-W_2}(t) = E[\exp\{it(W_1 - W_2)\}] = E[\exp\{it(U_{x,1} - U_{x,2})\}] = E[\exp\{itU_{x,1}\}]E[\exp\{-itU_{x,2}\}] = \{\phi_{u_x}(t)\}^2$, where $i^2 = -1$. Hence ϕ_{u_x} is estimable from the data by using all possible pairwise differences of the W_{jk} variables. On the other hand, $\phi_1(t)$ is directly estimable from the data, using the means of the replicated measurements. Consequently, $\phi_x(t)$ is estimable exploiting the relationship $\phi_1(t) = \phi_x(t)\{\phi_{u_x}(t/m_x)\}^{m_x}$. Specifically, estimators for $\phi_1(t)$ and $\phi_{u_x}(t)$ are given by $\widehat{\phi}_1(t) = n_x^{-1}\sum_{j=1}^{n_x}\exp(it\overline{W}_j)$,

$$\widehat{\phi}_{u_x}(t) = \sqrt{|\widehat{\phi}_{W_1-W_2}(t)|} = \sqrt{\left|\frac{1}{n_x}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{t(W_{jl_1} - W_{jl_2})\}\right|}, \tag{1}$$

respectively, where $\mathcal{S}_x = \{(l_1, l_2) : 1 \le l_1 < l_2 \le m_x\}$. Note that the non-vanishing and continuity assumption on $\phi_{u_x}(t)$, and $\phi_{u_x}(0) = 1$ imply that $\phi_{u_x}(t)$ is a positive real valued function. Thus, the above estimator of $\phi_{u_x}(t)$ is positive on compact subsets with high probability, for n_x large. Now, we propose to estimate $\phi_x(t)$ by

$$\widehat{\phi}_x(t) = \frac{\widehat{\phi}_1(t)}{\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}} = \frac{n_x^{-1}\sum_{j=1}^{n_x}\cos(t\overline{W}_j) + in_x^{-1}\sum_{j=1}^{n_x}\sin(t\overline{W}_j)}{|n_x^{-1}\sum_{j=1}^{n_x}M_x^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)(W_{jl_1} - W_{jl_2})\}|^{m_x/2}} = \widehat{a}_x(t) + i\widehat{b}_x(t),$$

where $\widehat{a}_x(t)$ and $\widehat{b}_x(t)$ are the real and imaginary part of $\widehat{\phi}_x(t)$, respectively, and we write

$$\widehat{a}_x(t) = \frac{n_x^{-1}\sum_{j=1}^{n_x}c_{jw}(t)}{\widehat{a}_{2x}(t)},\ \widehat{b}_x(t) = \frac{n_x^{-1}\sum_{j=1}^{n_x}d_{jw}(t)}{\widehat{a}_{2x}(t)},$$

with $c_{jw}(t) = \cos(t\overline{W}_j)$, $d_{jw}(t) = \sin(t\overline{W}_j)$, and

$$\widehat{a}_{2x}(t) = |n_x^{-1}\sum_{j=1}^{n_x}M_x^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)(W_{jl_1} - W_{jl_2})\}|^{m_x/2}.$$

Similarly, $\phi_y(t)$ can be estimated by $\widehat{\phi}_y(t) = \widehat{a}_y(t) + i\widehat{b}_y(t)$, where $\widehat{a}_y(t) = n_y^{-1}\sum_{j=1}^{n_y}c_{jv}(t)/\widehat{a}_{2y}(t)$ and $\widehat{b}_y(t) = n_y^{-1}\sum_{j=1}^{n_y}d_{jv}(t)/\widehat{a}_{2y}(t)$, with $c_{jv}(t) = \cos(t\overline{V}_j)$, $d_{jv}(t) = \sin(t\overline{V}_j)$, and

$$\widehat{a}_{2y}(t) = |n_y^{-1}\sum_{j=1}^{n_y}M_y^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_y}\cos\{(t/m_y)(V_{jl_1} - V_{jl_2})\}|^{m_y/2}.$$

Under the null hypothesis $F_x = F_y$, $(\widehat{a}_x(t), \widehat{b}_x(t))$ is expected to be close to $(\widehat{a}_y(t), \widehat{b}_y(t))$. When the null hypothesis does not hold, the difference between them is expected to be large, and this fact motivates us to form the following test statistic to test the hypothesis $H_0 : F_x = F_y$:

$$T_{n_x} = \int_{-\infty}^{\infty} n_x[\{\widehat{a}_x(t) - \widehat{a}_y(t)\}^2 + \{\widehat{b}_x(t) - \widehat{b}_y(t)\}^2]\omega(t)dt, \tag{2}$$

for a properly chosen non-negative weight function $\omega(t)$. The test function is

$$\Phi = \begin{cases} 1 & \text{if } T_{n_x} > t_{n_x,\alpha} \\ 0 & \text{otherwise,} \end{cases}$$

where the critical value $t_{n_x,\alpha}$ satisfies $\mathrm{pr}(T_{n_x} > t_{n_x,\alpha}) = \alpha$ under H_0, for a given $\alpha \in (0, 1)$.

In (2), the weight function $\omega(t)$ is used for ensuring the finiteness of the integral on the right side, and it is typically taken as a compactly supported function. As expected, the power of the test depends on the weight function $\omega(t)$. In a related work, Epps and Pulley (1983) proposed a test for normality based on the empirical CF of the observed data without measurement errors and described some desirable properties of $\omega(t)$. Here we follow Epps and Pulley (1983)'s guidance and take $\omega(t)$ to be a piece-wise continuous positive valued function with a compact support $[t_1, t_2]$ that includes 0, and $\omega(t) = 0$ for $t > t_2$ or $t < t_1$. For more details on some practical choices for t_1 and t_2, see the simulation and data analysis section.

2.3 Large Sample properties of the test statistic

The first result gives the null distribution of the test statistic.

Theorem 1. *Under the null hypothesis, as $n_x, n_y \to \infty$ and $\sqrt{n_x/n_y} \to \rho \in (0, \infty)$, the test statistic T_{n_x} converges to a random variable, given by*

$$\int [\xi_1(t)^2 + \xi_2(t)^2]\omega(t)dt$$

where $\xi_1(\cdot)$ and $\xi_2(\cdot)$ are independent zero mean Gaussian processes with continuous sample paths, with probability one. The covariance functions of $\xi_j(\cdot)$, $j = 1, 2$ are rational functions of the (real and imaginary parts of the) CFs of $\overline{W}_1$, $\overline{V}_1$, $U_{x,1}$ and $U_{y,1}$, and are given in the Appendix.

It follows from the statement of Theorem 1 that the limit distribution of the test statistic can also be expressed as an infinite sum of weighted, independent Chi-squared random variables with degrees of freedom 1. However, the weights in the infinite series representation or the covariance function of the Gaussian processes $\xi_j(\cdot)$, $j = 1, 2$ in the integral representation above are complicated functions of unknown population parameters that are difficult to estimate under the measurement error model. As a result, we shall develop a Bootstrap method to devise alternative approximations to the null distribution of the test statistic that can be used for calibrating the test.

The next result shows that under mild conditions, the power of the test statistic under alternative hypothesis tends to one. To state it, define $D_a(t) = a_x(t) - a_y(t)$ and $D_b(t) = b_x(t) - b_y(t)$.

Theorem 2. *Suppose that $\sqrt{n_x/n_y} \to \rho \in (0, \infty)$ and that the alternative hypothesis $\int\{D_a^2(t) + D_b^2(t)\}\omega(t)dt \neq 0$ holds. Then, for any $\alpha \in (0, 1)$, the power of the size α test, $pr(T_{n_x} > t_{n_x,\alpha})$ tends to 1 as $n_x, n_y \to \infty$.*

Remark 1. *In some cases the distribution of X and Y change with respect to other covariates. In these cases, the research question could be checking homogeneity of the two distributions after adjusting the effect of the covariates. This test after adjustments helps to identify any differences between the distributions of X and Y that are not generally accounted by the covariates. Suppose that $\boldsymbol{Z}$ denote a vector of covariates that is observed for every subject in the data. Also, we use a different notation $(A_{j1}, \ldots, A_{jm_x})^T$ to denote the replicated erroneous measurements, for the jth subject in group 1, for $j = 1, \ldots, n_x$. Similarly, define $(B_{k1}, \ldots, B_{km_y})^T$ to denote the replicated erroneous measurements, for the kth subject in group 2, $k = 1, \ldots, n_y$.*

We assume that observed data $A_{jl} = A_j$+measurement error, for $l = 1, \ldots, m_x$, $B_{kl} = B_k$+measurement error, for $l = 1, \ldots, m_y$. Next we assume that 1) $A_j = \beta_0 + \boldsymbol{Z}_j^T\boldsymbol{\beta}_1 + \varepsilon_{x,j}$, 2) $E(\varepsilon_{x,j}|\boldsymbol{Z}_j) = 0$, 3) $E(\varepsilon_{x,j}\varepsilon_{x,j'}|\boldsymbol{Z}_j, \boldsymbol{Z}_{j'}) = 0$, 4) $E(\varepsilon_{x,j}^2|\boldsymbol{Z}_j) = \tau_x^2$, 5) $B_k = \beta_0 + \boldsymbol{Z}_k^T\boldsymbol{\beta}_1 + \varepsilon_{y,k}$, 6) $E(\varepsilon_{y,k}|\boldsymbol{Z}_k) = 0$, 7) $E(\varepsilon_{y,k}\varepsilon_{y,k'}|\boldsymbol{Z}_k, \boldsymbol{Z}_{k'}) = 0$, 8) $E(\varepsilon_{y,k}^2|\boldsymbol{Z}_k) = \tau_y^2$, 9) $E(\varepsilon_{x,j}\varepsilon_{y,k}|\boldsymbol{Z}_j, \boldsymbol{Z}_k) = 0$. Note that

A_j's and B_k's are not observed, so we regress $(\overline{A}_1, \ldots, \overline{A}_{n_x}, \overline{B}_1, \ldots, \overline{B}_{n_y})^T$ on the covariates, and estimate $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ using weighted least square method. Let $\widehat{\boldsymbol{\beta}}$ be the weighted least square estimator of $\boldsymbol{\beta}$. Then, we define the residuals as $W_{jl} = A_{jl} - \widehat{\beta}_0 - \boldsymbol{Z}_j^T \widehat{\boldsymbol{\beta}}_1 \equiv X_j + U_{x,jl}$, $l = 1, \ldots, m_x$, $j = 1, \ldots, n_x$, and $V_{kl} = B_{kl} - \widehat{\beta}_0 - \boldsymbol{Z}_k^T \widehat{\boldsymbol{\beta}}_1 \equiv Y_j + U_{y,kl}$, $l = 1, \ldots, m_y$, $k = 1, \ldots, n_y$. Next, we apply the proposed test and the two naive tests on these residuals. This adjustment would work fine as long as the standard linear model assumptions given in (1)-(9) are valid.

3 The proposed Bootstrap method

3.1 Outline of the Bootstrap procedure

In this section, we describe a novel Bootstrap method for approximating the null distribution of the test statistic given in Theorem 1. Note that due to the presence of the measurement error, simple resampling from the original data will not capture the distributions of the latent variables and the error variables precisely. In addition, resampling the observations directly will also fail to ensure that the data are generated under the null hypothesis. Therefore, we propose to generate observations from a *suitable* estimated common distribution $\widehat{F}$ of the two populations for the latent variables, enforcing the null distribution. We also independently, generate observations from estimated distribution functions $\widehat{F}_{u_x}$ and $\widehat{F}_{u_y}$ of the two sets of error variables and combine them to define the Bootstrap analogues of W and V. Constructions of $\widehat{F}$ and $\widehat{F}_{u_x}$ (and $\widehat{F}_{u_y}$) require special care due to the complexities of the measurement error structure and are described in Sections 3.2 and 3.3 below; See also Remark 2 in Section 3.4 for some subtle issues and intricacies associated with formulation of the Bootstrap method.

Once the estimators $\widehat{F}$, $\widehat{F}_{u_x}$ and $\widehat{F}_{u_y}$ are specified, a Bootstrap sample will consist of $D_w^* = \{\boldsymbol{W}_1^*, \ldots, \boldsymbol{W}_{n_x}^*\}$ and $D_v^* = \{\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_{n_y}^*\}$, where $\boldsymbol{W}_j^* = (W_{j1}^*, \ldots, W_{jm_x}^*)^T$, $j = 1, \ldots, n_x$ and $\boldsymbol{V}_k^* = (V_{j1}^*, \ldots, V_{km_y}^*)^T$, $k = 1, \ldots, n_y$, with $W_{jl}^* = X_j^* + U_{x,jl}^*$ and $V_{kl}^* = Y_k^* + U_{y,kl}^*$. Here,

$X_1^*, \dots, X_{n_x}^*, Y_1^*, \dots, Y_{n_y}^*$ are iid draws from the estimated common distribution $\widehat{F}$, and $U_{x,jl}^*$ are iid draws from $\widehat{F}_{u_x}$ and $U_{y,kl}^*$ are iid draws from $\widehat{F}_{u_y}$. For each Bootstrap sample, we would compute the test statistic. Suppose that T_{b,n_x}^* denotes the test statistic corresponding to the bth Bootstrap sample. Then the estimated p-value is $\sum_{b=1}^B I(T_{b,n_x}^* > T_{n_x})/B$ based on B Bootstrap samples. We reject H_0 at the $100\alpha\%$ level of significance if the p-value is less than a given α. Now we describe how we estimate F, F_{u_x}, and F_{u_y} nonparametrically. Validity of the Bootstrap approximation is proved in Section 3.4.

3.2 Estimation of the common distribution F

Let g be a density function of $\overline{W}$, the mean of m_x repeated observations. Then for a symmetric kernel K and given bandwidth h_w, $\widehat{g}(w) = (n_x h_w)^{-1} \sum_{j=1}^{n_x} K\{(w - \overline{W}_j)/h_w\}$ is a kernel density estimator for g, and consequently the estimated characteristic function (CF) of $\overline{W}$ is

$$\widehat{\phi}_{\overline{W}}(t) = \int \exp(itw)\widehat{g}(w)dw = \frac{1}{n_x}\sum_{j=1}^{n_x} \exp(it\overline{W}_j) \int \exp(ith_w z)K(z)dz = \widehat{\phi}_1(t)\phi_K(h_w t),$$

where $\widehat{\phi}_1(t)$ is the empirical CF of $\overline{W}$ and $\phi_K(t)$ is the CF of the kernel K. Therefore, the estimated CF $\widehat{\phi}_x(t) = \widehat{\phi}_{\overline{W}}(t)/\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x} = \widehat{\phi}_1(t)\phi_K(h_w t)/\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}$. Note that $\widehat{\phi}_x(t)$ given in Section 2.2 does not satisfy the integrability condition needed for inverse Fourier transformation. Therefore, here we are using a different estimator for $\phi_x(t)$. Similarly, we estimate $\phi_y(t)$ by $\widehat{\phi}_y(t) = \widehat{\phi}_2(t)\phi_K(h_v t)/\{\widehat{\phi}_{u_y}(t/m_y)\}^{m_y}$. Although an estimator of the CF of the common distribution F can be defined in many ways, for simplicity we have decided to consider the estimator to be $\widehat{\phi}(t) = \{\widehat{\phi}_x(t) + \widehat{\phi}_y(t)\}/2$. Next using the inversion formula along with the conditions $\sup_t |\phi_K(t)/\phi_{u_x}(t/h_w)| < \infty$, $\int |\phi_K(t)/\phi_{u_x}(t/h_w)|dt < \infty$, $\sup_t |\phi_K(t)/\phi_{u_y}(t/h_v)| < \infty$ and $\int |\phi_K(t)/\phi_{u_y}(t/h_v)|dt < \infty$ for fixed $h_w, h_v > 0$ (Stefanski and Carroll, 1990), we have a deconvolution density estimator, given by:

$$\widehat{f}(r) \quad = \quad \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itr)\widehat{\phi}(t)dt$$

$$
\begin{aligned}
&= \frac{1}{4\pi}\int_{-\infty}^{\infty}\exp(-itr)\left[\frac{\sum_{j=1}^{n_x}\exp(it\overline{W}_j)\phi_K(h_w t)/n_x}{\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}} + \frac{\sum_{j=1}^{n_y}\exp(it\overline{V}_j)\phi_K(h_v t)/n_y}{\{\widehat{\phi}_{u_y}(t/m_y)\}^{m_y}}\right]dt \\
&= \sum_{j=1}^{n_x}\int_{-\infty}^{\infty}\frac{\exp\{-it(r-\overline{W}_j)\}\phi_K(h_w t)}{4\pi n_x\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}}dt + \sum_{j=1}^{n_y}\int_{-\infty}^{\infty}\frac{\exp\{-it(r-\overline{V}_j)\}\phi_K(h_v t)}{4\pi n_y\{\widehat{\phi}_{u_y}(t/m_y)\}^{m_y}}dt \\
&= \frac{1}{2n_x h_w}\sum_{j=1}^{n_x}L_x\left(\frac{r-\overline{W}_j}{h_w}\right) + \frac{1}{2n_y h_v}\sum_{j=1}^{n_y}L_y\left(\frac{r-\overline{V}_j}{h_v}\right),
\end{aligned}
$$

where $L_x(u) = (1/2\pi)\int_{-\infty}^{\infty}\exp(-itu)\phi_K(t)/\{\widehat{\phi}_{u_x}(t/h_w m_x)\}^{m_x}dt$ and $L_y(u) = (1/2\pi)\int_{-\infty}^{\infty}\exp(-itu)$ $\phi_K(t)/\{\widehat{\phi}_{u_y}(t/h_v m_y)\}^{m_y}dt$. Although the common population CDF F may not have a density, this density estimator is well defined. We are using this formula only to motivate the definition of the CDF estimator given next. Indeed, replacing ϕ_{u_x} by its estimator given in (1) and ϕ_{u_y} by the corresponding estimator, and replacing $\phi_K(t)$ by $(1-t^2)^3 1_{[-1,1]}(t)$, and using the integration formula (A.1) of Hall and Lahiri (2008), we obtain the estimator of the common distribution

$$
\begin{aligned}
\widehat{F}(r) &= \frac{1}{n_x}\sum_{j=1}^{n_x}\left[\frac{1}{2} + \int_{-\infty}^{\infty}\frac{\sin\{t(r-\overline{W}_j)\}(1-h_w^2t^2)^3 1_{[-1,1]}(h_w t)}{2\pi t|N_x^{-1}\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)(W_{jl_1}-W_{jl_2})\}|^{m_x/2}}dt\right] \\
&\quad + \frac{1}{n_y}\sum_{j=1}^{n_y}\left[\frac{1}{2} + \int_{-\infty}^{\infty}\frac{\sin\{t(r-\overline{V}_j)\}(1-h_v^2t^2)^3 1_{[-1,1]}(h_v t)}{2\pi t|N_y^{-1}\sum_{j=1}^{n_y}\sum_{(l_1,l_2)\in\mathcal{S}_y}\cos\{(t/m_y)(V_{jl_1}-V_{jl_2})\}|^{m_y/2}}dt\right] \\
&= \frac{1}{2} + \frac{1}{n_x\pi}\int_0^{1/h_w}\frac{(1-h_w^2t^2)^3\sum_{j=1}^{n_x}\sin\{t(r-\overline{W}_j)\}}{t|N_x^{-1}\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)(W_{jl_1}-W_{jl_2})\}|^{m_x/2}}dt \\
&\quad + \frac{1}{n_y\pi}\int_0^{1/h_v}\frac{(1-h_v^2t^2)^3\sum_{j=1}^{n_y}\sin\{t(r-\overline{V}_j)\}}{t|N_y^{-1}\sum_{j=1}^{n_y}\sum_{(l_1,l_2)\in\mathcal{S}_y}\cos\{(t/m_y)(V_{jl_1}-V_{jl_2})\}|^{m_y/2}}dt.
\end{aligned}
$$

It is important to point out that the density estimator $\widehat{f}$ can take negative values and hence, this distribution function estimator $\widehat{F}$ need not be monotone. As a result, $\widehat{F}$ can not be directly used for generating the Bootstrap observations. To overcome this limitation of the estimator $\widehat{F}$, we shall use a monotonized version of $\widehat{F}$, given by $\tilde{F}(r) = \sup\{\widehat{F}(r^*) : r^* \le r\}$, $r \in \mathbb{R}$, and use the inverse integral transform $\tilde{F}^{-1}(p) = \sup\{r : \tilde{F}(r) \le p\}$, $p \in (0,1)$, with iid Uniform (0,1) random variables to generate the Bootstrap versions of the X and Y variables under H_0.

Next we comment on the choice of h_w. We shall use Hall and Lahiri (2008)'s method that is relatively straight forward to apply. According to Theorem 4.1 of that paper, we choose the optimal h_w that minimizes $n_x^{-1}I(h) + B_x h^4$, where $2\pi I(h) = \int t^{-2}[1 - \phi_K(ht)/\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}]^2 dt$, $B_x = \kappa_2^2/(16\sqrt{\pi}\widehat{\sigma}_x^3)$ with $\kappa_2 = \int x^2 K(x)dx$. For our choice of kernel, $\kappa_2 = 6$. Also, $\mathrm{var}(\overline{W}) = \mathrm{var}(X) +$ $\mathrm{var}(U_x)/m_x$, so we estimate σ_x^2 by $\widehat{\sigma}_x^2 = \widehat{\sigma}_{\overline{W}}^2 - \widehat{\sigma}_{u_x}^2/m_x$, where $\widehat{\sigma}_{\overline{W}}^2 = (n_x - 1)^{-1}\sum_{j=1}^{n_x}(\overline{W}_j - \overline{W}_{..})^2$, $\widehat{\sigma}_{u_x}^2 = (n_x)^{-1}\sum_{j=1}^{n_x}(m_x - 1)^{-1}\sum_{l=1}^{m_x}(W_{jl} - \overline{W}_j)^2$, and $\overline{W}_{..} = (n_x m_x)^{-1}\sum_{j=1}^{n_x}\sum_{l=1}^{m_x} W_{jl}$. We use the numerical integration technique to evaluate $I(h_w)$. Similarly, we shall determine the optimal h_v.

Besides the cross-validation approach one may consider a plug-in or bootstrap based choice for (h_w, h_v) (Delaigle and Gijbels, 2004). However, based on our numerical experiences (the last paragraph of Section 4), the power or Type-I error probability of the proposed test are fairly insensitive towards different methods of bandwidth choices, specially when the sample size is large.

3.3 Estimation of F_{u_x} and F_{u_y}

We shall describe the estimation of F_{u_x} only. The estimation of F_{u_y} follows similar steps, and so it will be omitted. Observe that $W_{jl_1} - W_{jl_2} = U_{x,jl_1} - U_{x,jl_2}$, where U_{x,jl_1} and U_{x,jl_2} are iid copies of the random variable U_x and $(l_1, l_2) \in \mathcal{S}_x$. Hence the density of the difference of the iid copies can be estimated by the kernel method

$$\widehat{f}_{U_{x,1}-U_{x,2}}(u^*) = \frac{1}{hn_x}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x} K\{\frac{u^* - (W_{jl_1} - W_{jl_2})}{h}\},$$

where we take $h = 1.06\widehat{\sigma}_{d,u_x} n_x^{-1/5}$ (Sheather, 2004), where $\widehat{\sigma}_{d,u_x}^2 = (n_x - 1)^{-1}\sum_{j=1}^{n_x} 2\{m_x(m_x - 1)\}^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}[(W_{jl_1} - W_{jl_2}) - n_x^{-1}\sum_{j'=1}^{n_x} 2\{m_x(m_x - 1)\}^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}(W_{j'l_1} - W_{j'l_2})]^2$. Next we estimate the CF of $U_{x,1} - U_{x,2}$ by

$$\begin{aligned}\widehat{\phi}_{U_{x,1}-U_{x,2}}(t) &= \int \exp(itu^*)\widehat{f}_{U_{x,1}-U_{x,2}}(u^*)du^* \\ &= \frac{1}{n_x h}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x}\int \exp(itu^*)K\{\frac{u^* - (W_{jl_1} - W_{jl_2})}{h}\}du^*\end{aligned}$$

$$= \frac{1}{n_x}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x}\int \exp[it\{(W_{jl_1}-W_{jl_2})+hz\}]K(z)dz$$
$$= \frac{1}{n_x}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x}\exp\{it(W_{jl_1}-W_{jl_2})\}\phi_K(ht).$$

Since $E[\exp\{it(U_{x,1}-U_{x,2})\}] = \{\phi_{u_x}(t)\}^2$ due to the symmetry of U_x, and using $\phi_K(t) = (1-t^2)^3 1_{[-1,1]}(t)$, we estimate $\phi_{u_x}(t)$ by

$$\widehat{\phi}_{u_x}(t) = \sqrt{\widehat{\phi}_{U_{x,1}-U_{x,2}}(t)} = \sqrt{\left|\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x}\frac{2\cos\{t(W_{jl_1}-W_{jl_2})\}}{n_x m_x(m_x-1)}(1-h^2t^2)^3 1_{[-1,1]}(ht)\right|}. \quad (3)$$

Due to the presence of the indicator function, $\int|\widehat{\phi}_{u_x}(t)|dt < \infty$, and this integrability is a sufficient condition for the following inversion. Hence, using $\widehat{\phi}_{u_x}(t) = \widehat{\phi}_{u_x}(-t)$, we estimate $F_{u_x}(u)$ by

$$\widehat{F}_{u_x}(u) = \frac{1}{2}+\frac{1}{2\pi}\int_0^\infty \frac{\exp(itu)\widehat{\phi}_{u_x}(-t)-\exp(-itu)\widehat{\phi}_{u_x}(t)}{it}dt$$
$$= \frac{1}{2}+\frac{1}{2\pi}\int_0^\infty \frac{\widehat{\phi}_{u_x}(t)\{\exp(itu)-\exp(-itu)\}}{it}dt$$
$$= \frac{1}{2}+\frac{1}{\pi}\int_0^\infty \frac{\sin(tu)}{t}\sqrt{\left|\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x}\frac{2\cos\{t(W_{jl_1}-W_{jl_2})\}}{n_x m_x(m_x-1)}(1-h^2t^2)^3 1_{[-1,1]}(ht)\right|}dt$$
$$= \frac{1}{2}+\frac{1}{\pi}\int_0^{1/h} \frac{\sin(tu)}{t}\sqrt{\left|\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x}\frac{2\cos\{t(W_{jl_1}-W_{jl_2})\}}{n_x m_x(m_x-1)}\right|}(1-h^2t^2)^{3/2}dt.$$

We evaluate this integration by the Gauss-Legendre quadrature formula. As before, for simulating random numbers from this distribution we define the pth ($0 < p < 1$) quantile as $q = \sup\{r : \tilde{F}_{u_x}(r) \le p\}$, where $\tilde{F}_{u_x}(r) \equiv \sup\{\widehat{F}_{u_x}(r^*) : r^* \le r\}$.

We want to point out that although deconvoluted kernel density estimate is numerically unstable, the deconvoluted distribution function estimator after monotonization is quite stable numerically. The results of Section 4 of Hall and Lahiri (2008) demonstrate that the MSE of the quantile estimator based on the deconvoluted distribution function is quite stable for under (smaller values of bandwidth) or over-smoothing (larger values of bandwidth).

3.4 Validity of the Bootstrap

We now show that under some regularity conditions, the proposed Bootstrap method produces valid approximation to the distribution of the test statistic under the null. We shall denote the Bootstrap probability by P_*.

Theorem 3. *Suppose that $H_0 : F_x = F_y$ holds and as $n_x, n_y \to \infty$, $\sqrt{n_x/n_y} \to \rho \in (0, \infty)$. Also suppose that the bandwidths $h_w > 0$ and $h_v > 0$ are such that $[\{h_w + (n_x h_w)^{-1}\} + \{h_v + (n_x h_v)^{-1}\}] \to 0$. Then,*

$$\lim_{n_x \to \infty} \sup_{t \geq 0} \Big| P(T_{n_x} \leq t) - P_*(T^*_{n_x} \leq t) \Big| = 0, \text{ almost surely.}$$

Next, for $\alpha \in (0, 1)$, let $\widehat{t}_{n_x,\alpha}$ denote the $(1 - \alpha)$- quantile the Bootstrapped statistic $T^*_{n_x}$. Then, an immediate consequence of this result is that for any $\alpha \in (0, 1)$, $\widehat{t}_{n_x,\alpha} - t_{n_x,\alpha} \to 0$ almost surely. As a consequence, under the conditions of Theorem 3, $\mathrm{pr}(T_{n_x} > \hat{t}_\alpha) \to \alpha$. Thus, the Bootstrap method provides a valid method for calibrating the test statistic without having to estimate the covariance structure of the limit distribution of T_{n_x}. Finite sample properties of the Bootstrap approximation are presented in the next section.

Remark 2. *It may be noted that the formula for $\widehat{F}$ in Section 3.2 implicitly assumes that the median of $F(\cdot)$ is zero, i.e., the median of F_x and F_y are zero. However, this does not pose any problem for Bootstrapping the null distribution of the test statistic T_{n_x}. To appreciate why, note that $H_0 : F_x = F_y$ is equivalent to $H_0' : F_{x,a} = F_{y,a}$ for any $a \in \mathbb{R}$, where $F_{x,a}(t) = F_x(t + a)$ and $F_{y,a}(t) = F_y(t + a)$, $t \in \mathbb{R}$. Thus, if necessary, by subtracting a common constant $a \in \mathbb{R}$, we can, without loss of generality, assume that under the null hypothesis, the medians of F_x and F_y are zero. Indeed, noting that the test statistic T_{n_x} can be written as $T_{n_x} = \int |\widehat{\phi}_x(t) - \widehat{\phi}_y(t)|^2 \omega(t) dt$, it follows that T_{n_x} is invariant under a common location change. As a result, one gets a valid approximation to the null distribution of T_{n_x} by using the estimator $\widehat{F}$ in Section 3.2 even when the median of the common distribution F is different from zero. This observation also highlights the challenges and complexities associated with* *formulation of a valid Bootstrap method in the two sample testing problem in presence of measurement error.*

4 Simulation studies

Simulation designs: In this section, we present the numerical performance of the proposed test via Monte-Carlo simulations. We simulated datasets that consisted of two samples, $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{n_x}\}$ and $\{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{n_y}\}$, where $\boldsymbol{W}_j = (W_{j1}, \ldots, W_{jm_x})^T$ and $\boldsymbol{V}_k = (V_{k1}, \ldots, V_{km_y})^T$. We considered $n_x = n_y = 50, 200$ and 500 while we had two different scenarios corresponding to the number of repetitions: 1) $m_x = m_y = 2$ and 2) $m_x = 2, m_y = 3$. Type I error rate was examined in the following four designs (D1, D2, D3, D4), while power of the test was examined in designs D5, D6, D7, and D8. In addition, D9 and D10 were designed to explore robustness of the proposed method towards the symmetric measurement error assumption.

D1 $X, Y \sim \text{Normal}(0, 1)$ and $U_x, U_y \sim DE(0, 0.35)$

D2 $X, Y \sim \text{Normal}(0, 1)$ and $U_x, U_y \sim N(0, 0.5^2)$

D3 $X, Y \sim \text{Normal}(0, 1)$ and $U_x \sim DE(0, 0.35), U_y \sim N(0, 0.5^2)$

D4 $X, Y \sim (\chi^2_1 - 1)/\sqrt{2}$ and $U_x \sim DE(0, 0.35), U_y \sim DE(0, 0.2)$

D5 $X \sim \text{Normal}(0, 1), Y \sim \text{Normal}(0.2, 1)$ and $U_x, U_y \sim DE(0, 0.35)$

D6 $X \sim \text{Normal}(0, 1), Y \sim DE(0, 0.7)$ and $U_x, U_y \sim DE(0, 0.35)$

D7 $X \sim \text{Normal}(0, 1), Y \sim DE(0, 0.7)$ and $U_x \sim DE(0, 0.35), U_y \sim N(0, 0.5^2)$

D8 $X \sim 0.5\text{Normal}(-0.9, 0.45^2){+}0.5\text{Normal}(0.9, 0.45^2), Y \sim \text{Normal}(0, 1)$ and $U_x, U_y \sim DE(0, 0.35)$

D9 $X, Y \sim \text{Normal}(0, 1)$ and $U_x, U_y \sim EXP(0.5) - 0.5$

D10 $X \sim \text{Normal}(0, 1), Y \sim DE(0, 0.7)$ and $U_x, U_y \sim EXP(0.5) - 0.5$

Here $DE(a,b)$ stands for the double exponential distribution with mean a and variance $2b^2$ and $EXP(a)$ denotes the exponential distribution with mean a. In the first three designs, both measurement error variances associated with X and Y are 25% of the variability of X (or Y). In D4, both X and Y follow the modified chi-square distribution with degrees of freedom 1, mean 0 and variance 1. The choice of the true signals (the distribution of X or Y) and the measurement error variance were somewhat similar to that of Delaigle et al. (2008). In D4, measurement error variances corresponding to X and Y are different, and consequently the variances of the convoluted observations are different, i.e., $\text{var}(W_{jl}) \neq \text{var}(V_{kl^*})$. The designs are also different in terms of the smoothness of their measurement error distributions, we considered the ordinary smooth class (D1, D4, D6, D8), the supersmooth class (D2), the mixed case (D3, D5, D7). For the alternative hypotheses, we included cases where there are differences in the location (D5) and in the shape (D6, D7, D8). In D9 and D10, we considered centered exponential distribution for the measurement error with variability 25% of that of the true signal.

Method of analysis: For each dataset, we carried out hypothesis test at the 5% level of significance. For the proposed method we rejected the null hypothesis $H_0 : F_x(r) = F_y(r)$ against $H_a : F_x(r) \neq F_y(r)$ if the p-value calculated using $B = 1,000$ Bootstrap samples was less than $\alpha = 0.05$. We also analyzed each data set using the naive testing methods that included the two-sample K-S and A-D tests based on the averages $\{\overline{W}_j, j = 1, \dots, n_x\}$ and $\{\overline{V}_k, k = 1, \dots, n_y\}$. In the naive tests, $\{\overline{W}_j, j = 1, \dots, n_x\}$ and $\{\overline{V}_k, k = 1, \dots, n_y\}$ are considered as random samples from F_x and F_y, respectively.

Regarding the choice of $\omega(t)$, we highlight the desirable properties of $\omega(t)$ advocated by Epps and Pulley (1983). First, $\omega(t)$ should have more weight where the underlying difference between the two CFs is large, and that difference is usually large in an interval near zero. Second, the weight $\omega(t)$ should be large where the estimators $\widehat{a}_x(t) - \widehat{a}_y(t)$ and $\widehat{b}_x(t) - \widehat{b}_y(t)$ are highly precise. In fact, the precision decreases as t moves away from zero. Furthermore, for the ordinary smooth and supersmooth class of measurement error distributions (Fan, 1991), the characteristics functions are polynomially and exponentially decreasing, respectively. Consequently, for a small $\varepsilon > 0$, $|\widehat{\phi}_{u_x}(t)| \leq \varepsilon$ whenever $|t| \geq t*$ for some $t* > 0$, that in turn results in highly variable estimators $\widehat{a}_x(t)$, $\widehat{a}_y(t)$, $\widehat{b}_x(t)$, $\widehat{b}_y(t)$ when $|t| > t*$. Based on these considerations, for the proposed approach, we used different weights, the normal weight $\omega(t) = \exp(-t^2/2)I(t_1 < t < t_2)$ and the uniform weight $\omega(t) = I(t_1 < t < t_2)$. For each weight, we considered two sets of (t_1, t_2). In the first set we took $t_1 = \min\{F_x^{-1}(0.005), F_y^{-1}(0.005)\}$ and $t_2 = \max\{F_x^{-1}(0.995), F_y^{-1}(0.995)\}$, and the corresponding weights are referred to as $\text{norm}_{0.99}$ and $\text{unif}_{0.99}$ for the normal and uniform weight, respectively. In the second set we took $t_1 = \min\{F_x^{-1}(0.1), F_y^{-1}(0.1)\}$ and $t_2 = \max\{F_x^{-1}(0.9), F_y^{-1}(0.9)\}$, and the corresponding weights are referred to as $\text{norm}_{0.8}$ and $\text{unif}_{0.8}$. Results for these four different weights show how the performance of the test depends on the weight function.

Results: For each scenario we simulated 5,000 datasets, and for each scenario we computed the power, the proportion of times H_0 is rejected at the 5% level out of $5,000$ replications. Tables 1 and 2 contain the simulation results for 1) $m_x = m_y$ and 2) $m_x \neq m_y$ cases, respectively. The results indicate that the proposed test maintains the nominal level for all designs (D1 - D4) and for different weights. For D4, the naive tests fail to maintain the nominal level, and their power seems to be increasing with the sample size for both cases, 1) $m_x = m_y$ and 2) $m_x \neq m_y$. The intuitive reason is that although the means are the same $E(\overline{W}) = E(\overline{V})$, the variances are different, $\text{var}(\overline{W}) = 1 + 0.25/m_x$ and $\text{var}(\overline{V}) = 1 + 0.02/m_y$. Therefore, K-S or A-D test based on the empirical distributions of $(\overline{W}_1, \ldots, \overline{W}_{n_x})$ and $(\overline{V}_1, \ldots, \overline{V}_{n_y})$ are likely to reject H_0. For the scenarios D1-D3 when $m_x = 2$ and $m_y = 3$, although the type-I error rate of the K-S and A-D seems to be under the nominal level, a further simulation with $n_x = n_y = 2000$ revealed that the type-I error rate is exceeds the nominal level as powers for K-S (A-D) test are 0.0544 (0.0572), 0.0542 (0.061), and 0.054 (0.061) for designs D1, D2, and D3, respectively.

For the cases, where the alternative hypothesis holds, the power of the proposed test is increasing with the sample size. For D5, where the distribution of X and Y differ only by a location parameter, the power of the proposed test is somewhat lower than that of the naive approaches. Here is an intuitive explanation. Since the difference in the location parameters for the X and Y distributions is well reflected in the difference between the CDFs of $\overline{W}$ and $\overline{V}$ when the distribution of F_{u_x} and F_{u_y} are the same, the naive methods are capable of differentiating the two underlying distributions. Although the proposed method detects the difference between F_x and F_y in terms of the location parameter, the actual difference is somewhat masked out by the variability of the estimator of the CFs of the true signal and the measurement error. For scenarios D6, D7, and D8, the power of the proposed approach is significantly better than the other methods, even for sample size $n = 50$. In D6, D7, and D8, the mean and variance of the convoluted observations from the two samples are almost the same, $E(W_{jl}) = E(V_{kl^*}) = 0$ and $\text{var}(W_{jl}) \approx \text{var}(V_{kl^*})$, and also the first two moments of $\overline{W}_j$ are the same as that of $\overline{V}_k$, i.e., $E(\overline{W}_j) = E(\overline{V}_k) = 0$ and $\text{var}(\overline{W}_j) \approx \text{var}(\overline{V}_k)$ for $m_x = m_y$ case. Additionally, the shapes of the distribution of $\overline{W}_j$ and $\overline{V}_k$ are not dramatically different, especially for $m_x = m_y$ case. Therefore, the power of the K-S or A-D is lower than that of the proposed method. Naturally the power of the naive approaches improve from $m_x = m_y = 2$ to $m_x = 2, m_y = 3$ scenario as the variance of $\overline{W}$ and $\overline{V}$ become different due to different replications. For the asymmetric measurement error model (D9 and D10), the proposed test maintains the level and gives better power ($\text{unif}_{0.99}$ and $\text{norm}_{0.99}$) than the K-S or A-D tests. These results indicate that the proposed test is quite robust towards the violation of the symmetric error assumption.

In summary, the simulation results indicate that the proposed test is consistent, while the naive tests could be inconsistent. In the absence of any specific knowledge about the CF of the underlying distributions, in our opinion, the $\text{unif}_{0.99}$ weight is preferable as it covers a wide range of t-values and gives equal importance to the difference between the two CFs at any t.

Based on our numerical experiences, estimation of quantiles needed for the Bootstrap part is quite straight forward. Once the distribution functions F_x and F_y are estimated over a set of grid points ranging from $\min(\overline{W}_{..} - 4.5\widehat{\sigma}_x, \overline{V}_{..} - 4.5\widehat{\sigma}_y)$ to $\max(\overline{W}_{..} + 4.5\widehat{\sigma}_x, \overline{V}_{..} + 4.5\widehat{\sigma}_y)$, then the pth percentile point for X is calculated as $\max\{g : \widehat{F}_x(g) \le p\}$, where g denotes a set of grid points. Here $\overline{W}_{..}$ and $\overline{V}_{..}$ denote the overall mean of W and V, and $\widehat{\sigma}_x^2$ and $\widehat{\sigma}_y^2$ are given in the second last paragraph of Section 3.2. Similarly, we obtain the percentile points for Y. Finding these maxima over the set of grid points is easy in R (a single line R script). Also, the quantile estimators are numerically stable with respect to bandwidth (Hall and Lahiri, 2008).

Prompted by a reviewer's comment we conducted another simulation study to assess the sensitivity of our approach towards the tuning parameters (h_w, h_v) used in the deconvolution part. They also regulate t_1 and t_2 defined as the quantiles of the estimated distribution function. We generated data according to simulation scenarios D1 (H_0 holds) and D6 (H_0 does not hold). Besides the optimal choice of h_w as $h_w = \arg\min_h\{n_x^{-1}I(h) + B_x h^4, h = 0.01, 0.02, \dots\}$, we took a bad choice defined as $h_w = \arg$ 90th percentile of $\{n_x^{-1}I(h) + B_x h^4, h = 0.01, 0.02, \dots\}$. In these two choices of the bandwidth, h was varied over a grid of points. Under each choice of h_w, we calculated $\widehat{F}_x$. Similarly, under the bad and optimal choices of h_v we calculated $\widehat{F}_y$. Next, we calculated (t_1, t_2), test statistic, and p-value for every simulated dataset under both choices. Table 3 contains the power of the test for the optimal and bad choices of (h_w, h_v). The results for design D1 show that the empirical level (Type-I error rate) of the test is reasonably close to the nominal level for both choices of (h_w, h_v). Even bad choices of bandwidth can maintain type-I error rate. For $n_x = n_y = n = 50$, power changes between the two choices, and the change could be at most 20% on the relative scale. However, for larger sample sizes, the change of power is modest (at most 9% on the relative scale). Thus, we conclude that the power of the proposed method is fairly robust to the choices of the bandwidth (h_w, h_v), specially for a large sample size. Plots of (h_w, h_v), (t_1, t_2) and the test statistic under the two choices, bad and optimal, are given in the supplementary materials. The plots show that the bad choice of the bandwidth leads to a slightly wider difference between t_1 and t_2 compared to that of the optimal choice but test statistic's distributions remain unchanged under these two choices.

5 Numerical study using the NHANES data

We apply the proposed method to analyze the NHANES data that are publicly available at https://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm.

Blood pressure example: First we consider the NHANES 2009-2010 survey data, and focus only on non-Hispanic white males whose ages are between 35 and 55 years (middle-aged adults) so that we have a homogeneous demographic group. Our goal is to test equality of the distribution of systolic blood pressure between the two groups, non-alcoholic and alcoholic after adjusting the effect for two important covariates BMI and income that are readily available in the NHANES data. Alcohol consumption data are collected through two 24-hour recall interviews. We define a subject as non-alcoholic if both measurements are less than 14 grams, otherwise the subject is considered to be alcoholic. Since fourteen grams is considered to be the amount of alcohol in a standard drink, we use this value to define the two behavioral groups. This classification results in $n_x = 207$ (non-alcoholic) and $n_y = 126$ (alcoholic). Since an accurate measurement of blood pressure is difficult to obtain, at least three measurements were taken in the mobile examination center. For our analysis we consider the first three measurements for each subjects, i.e., $m_x = m_y = 3$. We define $A_{jk} = \log(\text{systolic blood pressure}_{jk})$ denote the logarithm of the kth blood pressure measurement of the jth individual, $k = 1, 2, 3$, and $j = 1, \ldots, n_x$, and similarly B_{jk} is defined, for $k = 1, 2, 3$ and $j = 1, \ldots, n_y$.

To adjust for the effect of body mass index (BMI), a continuous variable, and income, an ordinal categorical variable, we calculate the residuals as described in Remark 1. Then we apply the proposed test and the two naive tests on those residuals. As we discussed in the simulation study, we consider the $\text{unif}_{0.99}$ weight function. To calculate t_1 and t_2 we use the deconvoluted distribution functions $\widehat{F}_x$ and $\widehat{F}_y$ instead of F_x and F_y as the later two are unknown in the real data.

The resulting p-values are given in the first row of Table 4. At the 5% level, the proposed method strongly rejects H_0 while the naive approaches contradict each other so that it is difficult to make a decision. For this and the next application, we use $10,000$ Bootstrap samples and the $\text{unif}_{0.99}$ weight function to calculate the p-value for our proposed method. The conclusion based on the proposed test affirms the medical science that usually alcohol consumption and high blood pressure are associated. Moreover, repeated binge drinking for a long time may cause elevated blood pressure (http://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/blood-pressure/faq-20058254).

Albumin-to-creatinine ratio (ACR) example: Here we check if the distribution of albumin-to-creatinine ratio (ACR) differs by smoking status. Albumin is a protein and creatinine is a chemical waste, and their ratio ACR is used to assess renal functionality. Usually higher level of ACR is associated with a higher risk of renal events. Our interest is in testing equality of the distribution of ACR among non-smoking and smoking group after adjusting the effect of the confounding variables BMI and income.

In the NHANES study (2009-2010 survey data), urinary albumin and creatinine were measured twice for each participants, the first sample was collected in the mobile examination center (MEC) and the second sample was collected during the interview at home. We consider these two measurements (samples) as the two noisy measurements of the same underlying truth, and hence $m_x = m_y = m = 2$.

For this test we consider only non-Hispanic white males who are older than 60 years as the renal issue is more prevalent in the older group. We define a person as a non-smoker if he smoked less than 100 cigarettes in his lifetime, otherwise the person is called a smoker, and based on this classification we obtain $n_x = 161$ (non-smoking) and $n_y = 290$ (smoking). We define $A_{jk} = \log(\text{albumin}_{jk}/\text{creatinine}_{jk})$ for $k = 1, 2$, and $j = 1, \ldots, n_x$ for the non-smoking group and $B_{jk} = \log(\text{albumin}_{jk}/\text{creatinine}_{jk})$ for $k = 1, 2$, and $j = 1, \ldots, n_y$ for the smoking group. First, we obtain residuals as described in Remark 1 to remove the effect of BMI and income. Then we apply the proposed test and the two naive tests on the residuals. As in the previous application, we consider the $\text{unif}_{0.99}$ weight function, where t_1 and t_2 are calculated from the deconvoluted distribution functions $\widehat{F}_x$ and $\widehat{F}_y$.

The resulting p-values are given in the second row of Table 4. For the proposed test, we get p-value 0.009 so that we conclude that smoking status and ACR are related. At the 5% level, the A-D fails to reject H_0 while the p-value for the K-S test is barely below the nominal level. Therefore, as a whole the naive test could be misleading. The test result based on the proposed method is consistent with the finding of Hogan et al. (2007) who considered a similar issue with different smoking groups and have used the data from the NHANES III survey (1988-1994).

Simulation study that mimics the NHANES data: To show the effectiveness of the adjustment method, when it is necessary, mimicking the real dataset on the systolic blood pressure example, we conducted a simulation study. We generated two covariates T_1 and T_2 by mimicking the distributions of BMI and income. Specifically, T_1 was generated from the Gamma distribution with shape 26.7 and rate 0.9, T_2 from the multinomial distribution with the cell probability same as the observed relative frequency from the data. Next, we defined $B_j^X = \widehat{\beta}_0 + \widehat{\beta}_1 T_{1j} + \widehat{\beta}_2 T_{2j} + X_j$, $B_k^Y = \widehat{\beta}_0 + \widehat{\beta}_1 T_{1k} + \widehat{\beta}_2 T_{2k} + Y_k$, $A_{jl} = B_j^X + U_{x,jl}$ and $B_{kl} = B_k^Y + U_{y,kl}$, for $l = 1, 2, 3$, $m_x = m_y = 3$, $(n_x, n_y) = (200, 120), (400, 240)$, where $X_j, Y_k, U_{x,jl}$ and $U_{y,kl}$ were specified by some designs given in Section 4. Here $\widehat{\boldsymbol{\beta}}$ denotes the estimated $\boldsymbol{\beta}$ in the first data example.

For checking the type-I error rate, we considered designs D3 and D4, and a new design, D11: $X_j, Y_k \sim \widehat{F}$, $U_{x,jl} \sim \widehat{F}_{u_x}$, $U_{y,kl} \sim \widehat{F}_{u_y}$, where $\widehat{F}$ is the estimator of the common distribution of X and Y in the first data example, and $\widehat{F}_{u_x}$ and $\widehat{F}_{u_y}$ are the corresponding estimator of the measurement error distributions. For checking power, we considered designs D6, D8, and a new design, D12: $X_j \sim \widehat{F}_x$, $Y_k \sim \widehat{F}_y$, $U_{x,jl} \sim \widehat{F}_{u_x}$, $U_{y,kl} \sim \widehat{F}_{u_y}$, where $\widehat{F}_x$ and $\widehat{F}_y$ are the deconvolution estimator of X and Y, respectively, for the first data example. Each dataset was analyzed using the adjustment approach (Remark 1). Table 5 contains this simulation results. We find the patterns are similar to those in Tables 1 and 2. One remarkable result in this simulation is that naive approaches cannot control the nominal level even when $X, Y \sim \widehat{F}$ as in the case D4. Overall the proposed method shows consistent behavior, and much superior performance than the other approaches.

6 Concluding remarks

This article considers the test of homogeneity of two distributions when observed data are contaminated with the classical measurement error. To extract the true signals from the error contaminated data, we have applied a non-parametric method that does not make any assumption regarding the true signal. Also, other than symmetry and non-vanishing characteristic function (CF) over the entire real line, no other assumption was used for the measurement errors. We have proposed a valid Bootstrap approach to calibrate the test statistic and calculate the (estimated) p-value of the test.

The benefit of the proposed approach is shown through simulation studies. The simulation studies also show that the test maintains the nominal size and has good power properties in moderately large samples over a wide range of measurement error distributions. Further, weight functions with a shorter effective support (e.g., $\text{unif}_{0.8}$ vs. $\text{unif}_{0.99}$) may lead to less accurate results. We have applied the proposed method to analyze two real datasets obtained from the NHANES 2009-2010 study. Since this data was collected from the nationally representative sample, the results of the data analysis is applicable to a broader section of the population.

The proposed method can be extended to the scenarios, 1) where the number of replications (m_x or m_y) is varying by subjects, and 2) where instead of the replicated measurements, a validation dataset is available. Alternatively the test statistic could be defined as the supremum of the squared distances between the two estimated characteristic functions over a compact set which does not involve with the weight function ω. A similar bootstrap method can be used to assess its null distribution.

The method is based on two critical assumptions, 1) the measurement error is independent of the true signals, and 2) the measurement error distribution is symmetric and has a non-vanishing CF over the entire real line. In observational studies, often replicated measurements are collected via self-reporting. Variability in self-reporting likely to change by subjects resulting in a dependence between the true signal and the measurement error. This, known as heteroscedastic measurement error, is a difficult problem (see Rudolph and Stuart (2018) for regression and Staudenmayer et al. (2008) for density estimation) that lacks a complete nonparametric solution to date. The second critical assumption excludes any skewed distribution or an uniform distribution over a compact set to be the measurement error distribution. In future we will develop methods where these assumptions will be relaxed to some extent.

Acknowledgment

This research was partially supported by NSF Grant No: DMS 1613192.

References

Basu, S., Hong, A. and Siddiqi, A. (2015). Using decomposition analysis to identify modifiable racial disparities in the distribution of blood pressure in the United States. *American Journal of Epidemiology*, **182**, 345–353.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association* **83**, 1184–1186.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Delaigle, A. and Gijbels, I. (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics & Data Analysis*, **45**, 249–267.

Delaigle, A. and Hall, P. (2015). Methodology for nonparametric deconvolution when the error distribution is unknown. *Journal of the Royal Statistical Society, Series B* **78**, 231–252.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics* **36**, 665–685.

Epps, T. W. and Pulley, L. B. (1983). A test for normality based on the empirical characteristic function. *Biometrika* **70**, 723–726.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics* **19**, 1257–1272.

Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman and Hall/CRC Press.

Hall, P. and Lahiri, S. N. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *The Annals of Statistics* **36**, 2110–2134.

Hariharan, S. G., Strobel, N., Kaethner, C., Kowarschik, M., Fahrig, R., and Navab, N. (2019). An analytical approach for the simulation of realistic low-dose fluoroscopic images. *International Journal of Computer Assisted Radiology and Surgery*, **14**, 601–610.

Hodges, J. L. (1958). The significance probability of the Smirnov two-sample test. Arkiv för Matematik **3**, 469–486.

Hogan, S. L., Vupputuri, S., Guo, X., Cai, J., Colindres, R. E., Heiss, G., and Coresh, J. (2007). Association of cigarette smoking with albuminuria in the United States: The third National Health and Nutrition Examination Survey. *Renal Failure* **29**, 133–142.

Jones, C. P. (1997). Living beyond our "means": new methods for comparing distributions. *American Journal of Epidemiology*, **146**, 1056–1066.

Liu, Q., Georgieva, D. C., Egli, D., and Wang, K. (2019). NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC genomics*, **20 (Suppl 1)**, 78.

Pettitt, A. N. (1976). A two-sample Anderson-Darling rank statistic. *Biometrika*, **63**, 161–168.

Primatesta, P., Falaschetti, E., Gupta, S., Marmot, M. G., and Poulter, N. R. (2001). Association between smoking and blood pressure: evidence from the health survey for England. *Hypertension* **37**, 187–193.

Puddy, I. B. and Beilin, L. J. (2006). Alcohol is bad for blood pressure. *Clinical and Experimental Pharmacology and Physiology* **33**, 847–852.

Rudolph, K. E. and Stuart, E. A. (2018). Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods. *American Journal of Epidemiology*, **187**, 604–613.

Sheather, S. J. (2004). Density estimation. *Statistical Science* **19**, 588–597.

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. P. (2008). Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association*, **103**, 726–736.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21**, 165–184.

Stephens, P. J. et al. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.

Supporting Information

Appendix referred in this manuscript is given in a separate file, Supplementary Materials for "A Test of Homogeneity of Distributions when Observations are Subject to Measurement Errors". The real datasets, the code, and a README.txt file are provided in a .zip file. All computations were done using `R`, and an `R` package `MEtest` is available on CRAN (`https://CRAN.R-project.org/package=MEtest`), and some illustrative examples can be found at `https://www.stat.tamu.edu/~sinha/research.html`.

Table 1: The entries of the table show the proportion of the rejection of H_0 at the 5% level for the simulation study with sample sizes $n_x = n_y = n$ and $m_x = m_y = 2$ based on $5,000$ replications. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively. The entries corresponding to designs D1-D4 show the Type-I error rate, and the other entries are power.

				C-F			
	n	K-S	A-D	unif$_{0.99}$	unif$_{0.8}$	norm$_{0.99}$	norm$_{0.8}$
D1	50	0.038	0.044	0.033	0.036	0.038	0.037
	200	0.039	0.050	0.038	0.043	0.044	0.044
	500	0.052	0.052	0.047	0.050	0.048	0.048
D2	50	0.039	0.050	0.029	0.036	0.036	0.037
	200	0.038	0.052	0.039	0.043	0.041	0.044
	500	0.049	0.051	0.037	0.043	0.041	0.044
D3	50	0.038	0.049	0.033	0.042	0.037	0.042
	200	0.039	0.053	0.039	0.047	0.043	0.047
	500	0.055	0.051	0.045	0.046	0.045	0.045
D4	50	0.041	0.056	0.010	0.038	0.035	0.037
	200	0.052	0.083	0.033	0.036	0.037	0.036
	500	0.120	0.198	0.040	0.036	0.039	0.035
D5	50	0.099	0.137	0.053	0.108	0.092	0.115
	200	0.327	0.443	0.156	0.369	0.322	0.393
	500	0.728	0.820	0.401	0.749	0.695	0.775
D6	50	0.063	0.069	0.095	0.053	0.085	0.051
	200	0.147	0.154	0.439	0.110	0.315	0.095
	500	0.423	0.470	0.863	0.275	0.745	0.222
D7	50	0.060	0.069	0.082	0.054	0.085	0.051
	200	0.133	0.149	0.403	0.107	0.300	0.089
	500	0.391	0.425	0.845	0.262	0.738	0.207
D8	50	0.097	0.084	0.218	0.054	0.101	0.053
	200	0.326	0.251	0.812	0.066	0.392	0.061
	500	0.828	0.832	0.997	0.132	0.896	0.104
D9	50	0.042	0.047	0.038	0.044	0.039	0.042
	200	0.037	0.050	0.042	0.045	0.047	0.045
	500	0.046	0.046	0.047	0.045	0.047	0.045
D10	50	0.062	0.069	0.109	0.052	0.092	0.046
	200	0.150	0.159	0.450	0.121	0.319	0.099
	500	0.443	0.472	0.875	0.280	0.752	0.217

Table 2: The entries of the table show the proportion of the rejection of H_0 at the 5% level for the simulation study with sample sizes $n_x = n_y = n$ and $m_x = 2, m_y = 3$ based on $5,000$ replications. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively. The entries corresponding to designs D1-D4 show the Type-I error rate, and the other entries are power.

				C-F			
	n	K-S	A-D	$\text{unif}_{0.99}$	$\text{unif}_{0.8}$	$\text{norm}_{0.99}$	$\text{norm}_{0.8}$
D1	50	0.038	0.043	0.037	0.038	0.038	0.038
	200	0.039	0.048	0.041	0.047	0.047	0.045
	500	0.055	0.056	0.046	0.051	0.051	0.052
D2	50	0.039	0.050	0.032	0.038	0.037	0.038
	200	0.046	0.052	0.039	0.043	0.042	0.044
	500	0.050	0.053	0.038	0.046	0.041	0.047
D3	50	0.036	0.049	0.036	0.038	0.039	0.039
	200	0.038	0.054	0.042	0.046	0.046	0.048
	500	0.049	0.053	0.044	0.048	0.047	0.048
D4	50	0.049	0.066	0.009	0.039	0.038	0.040
	200	0.082	0.162	0.026	0.034	0.035	0.035
	500	0.287	0.551	0.036	0.034	0.034	0.034
D5	50	0.105	0.141	0.059	0.112	0.100	0.120
	200	0.342	0.457	0.179	0.384	0.339	0.409
	500	0.746	0.828	0.434	0.764	0.718	0.786
D6	50	0.071	0.075	0.116	0.055	0.095	0.051
	200	0.194	0.210	0.483	0.113	0.328	0.099
	500	0.583	0.647	0.900	0.285	0.778	0.230
D7	50	0.070	0.075	0.112	0.059	0.089	0.056
	200	0.190	0.210	0.475	0.113	0.318	0.094
	500	0.558	0.621	0.890	0.282	0.769	0.219
D8	50	0.104	0.088	0.243	0.057	0.104	0.056
	200	0.364	0.312	0.837	0.073	0.420	0.064
	500	0.869	0.873	0.998	0.131	0.904	0.099
D9	50	0.042	0.046	0.041	0.041	0.039	0.042
	200	0.038	0.050	0.046	0.047	0.047	0.046
	500	0.046	0.050	0.051	0.046	0.045	0.045
D10	50	0.067	0.076	0.130	0.052	0.094	0.048
	200	0.195	0.215	0.497	0.123	0.343	0.101
	500	0.575	0.636	0.906	0.288	0.774	0.226

Table 3: The entries of the table show the proportion of the rejection of H_0 at the 5% level for the simulation study with sample sizes $n_x = n_y = n$ and $m_x = m_y = 2$ based on $5,000$ replications. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively. The entries corresponding to design D1 show the Type-I error rate, and the other entries are power.

Design	n	$\text{unif}_{0.99}$		$\text{norm}_{0.99}$	
		Bad	Optimal	Bad	Optimal
	50	0.029	0.033	0.033	0.034
D1	200	0.035	0.037	0.039	0.045
	500	0.048	0.048	0.040	0.049
	50	0.078	0.097	0.070	0.085
D6	200	0.399	0.448	0.281	0.315
	500	0.835	0.879	0.706	0.741

Table 4: The table shows the p-values for testing of hypothesis using the real data. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively. Also, ACR≡ Albumin-to-creatinine ratio, and BP≡ Systolic blood pressure. For the proposed method, the number of Bootstrap samples was $10,000$.

Variable	K-S	A-D	C-F
BP	0.058	0.001	0.001
ACR	0.043	0.139	0.009

Table 5: The entries of the table show the proportion of the rejection of H_0 at the 5% level for the simulation study where simulated datasets mimicked the blood pressure dataset ($m_x = m_y = 3$) given in Section 5. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively.

		$n_x = 200, n_y = 120$			$n_x = 400, n_y = 240$		
		K-S	A-D	C-F	K-S	A-D	C-F
	D3	0.039	0.049	0.046	0.043	0.051	0.048
Type-I error rate	D4	0.050	0.060	0.039	0.064	0.083	0.043
	D11	0.053	0.060	0.038	0.065	0.073	0.046
	D6	0.125	0.121	0.398	0.261	0.273	0.716
Power	D8	0.282	0.220	0.716	0.609	0.587	0.969
	D12	0.575	0.814	0.821	0.882	0.986	0.984