Received: 6 June 2023 | Revised: 25 July 2023 | Accepted: 29 August 2023

DOI: 10.1002/sta4.617

ORIGINAL ARTICLE

WILEY

Homogeneity of marginal distributions for a large number of populations

M. V. Alba-Fernández [1] | M. D. Jiménez-Gamero [2]

[1]Department of Statistics and Operations Research, University of Jaén, Jaén, Spain

[2]Department of Statistics and Operations Research, University of Sevilla, Sevilla, Spain

Correspondence
M. V. Alba-Fernández, Department of Statistics and Operations Research, University of Jaén, B3-053, Campus las lagunillas, s/n, Jaén, Spain.
Email: mvalba@ujaen.es

Summary

Assume that a random vector (X,Y) is observed in k populations and independent samples of that random vector are available at each population. Assume that X and Y have the same dimension. Our purpose is to test the equality of the marginal distributions of X and Y in the k populations when k is large compared to the sample sizes. With this aim, we propose and study a test statistic that compares the empirical characteristic functions of the marginal distributions. Under the null, the test statistic is asymptotically free-distributed. An expression of the asymptotic power is also derived, which allows to study the consistency of the test. No assumption is made on the distribution of X and Y, which can be continuous, discrete or mixed; moreover, no assumption is made about moments. A simulation study investigates the finite sample performance of the new test. The proposal is applied to study air pollution levels that are directly related to environmental health, in all countries where observations are available.

KEYWORDS
asymptotic properties, many subpopulations, paired data, two-sample problem

1 | INTRODUCTION

The two-sample problem is a classical one in statistics. If X and Y are two random vectors of the same dimension with distribution functions F_X and F_Y, respectively, the two-sample problem consists in testing $H_{01}: F_X = F_Y$ versus $H_{11}: F_X \neq F_Y$, which is tantamount to $H_{01}: \varphi_X = \varphi_Y$ versus $H_{11}: \varphi_X \neq \varphi_Y$, where for any random vector $W \in \mathbb{R}^p$, φ_W denotes its characteristic function, that is, $\varphi_W(t) = \mathbb{E}(e^{\mathrm{i}t^\top W})$, $t \in \mathbb{R}^p$, with $\mathrm{i} = \sqrt{-1}$ and the superscript $\top$ means transposition of column vectors and matrices. Many approaches have been suggested in the statistical literature to deal with this problem. A number of them are based on comparing an estimator of a function that characterizes the population calculated at each sample. Examples are the chi-square test, based on comparing estimators of the probabilities for categorical data (see also Alba-Fernández & Jiménez-Gamero, 2009; Pardo et al., 1999, which use other divergence measures); tests based on comparing estimators of the cumulative distribution for univariate continuous data (Kiefer, 1959); tests based on comparing estimators of the quantile distribution for univariate continuous data (Kosorok, 1999); tests based on comparing estimators of the probability density function for continuous data (Anderson et al., 1994; Martínez-Camblor & de Uña Álvarez, 2009); tests based on comparing estimators of the characteristic function (Alba-Fernández et al., 2008; Baringhaus & Franz, 2004; Hušková & Meintanis, 2008; Jiménez-Gamero et al., 2017); and tests based on comparing estimators of the probability generating function for count data (Alba-Fernández et al., 2017). Most papers (including those previously cited) assume that independent samples are available from each population. The case where the data at hand are independent copies of a random vector (X,Y) has been less frequently considered. In this setting, the paper by Quessy and Éthier (2012) proposes procedures based on comparing the empirical distribution functions and the

characteristic functions of the marginal distributions of X and Y; the paper by Gaigall (2020) proposes procedures based on comparing the empirical distribution functions adapted to the case where the data have missing components. In both cases, independent and paired samples, the properties of the existing methods have been studied when the sample sizes increase.

This paper addresses the problem of testing the marginal homogeneity of a random vector (X,Y), but now, we assume that the population is divided into a large number of subpopulations, say k. The distribution of (X,Y) may vary across subpopulations, so we denote (X^j,Y^j) to the target vector when it is restricted to population j, $1 \le j \le k$. With this notation, our objective is to build a test of the null hypothesis:

$$\begin{aligned} H_0 &: F_{X^j} = F_{Y^j},\ 1 \le j \le k, \\ H_1 &: F_{X^j} \neq F_{Y^j},\ \text{for some } j \in \{1,\ldots,k\}, \end{aligned}$$

or equivalently,

$$\begin{aligned} H_0 &: \varphi_{X^j} = \varphi_{Y^j},\ 1 \le j \le k, \\ H_1 &: \varphi_{X^j} \neq \varphi_{Y^j},\ \text{for some } j \in \{1,\ldots,k\}. \end{aligned}$$

Notice that if the subpopulations are ignored, H_{01} may not be rejected, but H_0 may be rejected, as differences between the subpopulations could be compensated when they are considered as a whole.

Assume that $X,Y \in \mathbb{R}^p$, which entails that $\mathcal{X} = (X^{1\top},\ldots,X^{k\top})^\top$, $\mathcal{Y} = (Y^{1\top},\ldots,Y^{k\top})^\top \in \mathbb{R}^{kp}$. With this notation, we can write

$$\begin{aligned} H_0 &: F_{\mathcal{X}} = F_{\mathcal{Y}}, \\ H_1 &: F_{\mathcal{X}} \neq F_{\mathcal{Y}}. \end{aligned}$$

This way, H_0 can be seen as a two-sample problem for high-dimensional data. Some tests have been proposed in the statistical literature for such a problem. Examples are the tests in Hall and Tajvidi (2002), Liu and Modarres (2011), Chen and Friedman (2017), Cousido-Rocha et al. (2019) and Liu et al. (2022), just to cite a few. The tests in all these papers assume that the data consist of two independent samples, one from $\mathcal{X}$, say $\mathcal{X}_1,\ldots,\mathcal{X}_n$, and the other from $\mathcal{Y}$, say $\mathcal{Y}_1,\ldots,\mathcal{Y}_m$. To the best of our knowledge, the case of paired samples (that is our setting) has not been dealt with. Here, we study a special case, in which independent paired samples are available from each of the k components, $(X^1,Y^1),\ldots,(X^k,Y^k)$ of $(\mathcal{X},\mathcal{Y})$.

The rest of the paper is organized as follows. Section 2 is devoted to constructing the test statistic. For each subpopulation, an unbiased estimator of the squared of an L^2 type distance between the population characteristic functions of X and Y is built. Then, these estimators are combined to get the test statistic. As an approximation to the null distribution of the test statistic, its asymptotic distribution is derived, which turns out to be normal. Here, by asymptotic, it is meant when the number of groups k increases; the sample sizes of the data from each group can either stay bounded or grow with k. The asymptotic power of the proposed test is handled in Section 3. Section 4 presents some practical issues to facilitate users the application of the proposal. Section 5 reports the main findings of a large simulation study carried out to evaluate the finite sample performance of the test with critical region based on the asymptotic null distribution, with respect to both the level and the power. This section also contains a real data set application. Finally, Section 6 concludes. The proofs, some further tables related to the simulation study of Section 5 and the R code used to calculate the proposed test statistic are deferred to the Supporting Information.

Before ending this section, we introduce some notation: All 0s appearing in the paper represent vectors of the appropriate dimension with all its entries equal to 0; for $z \in \mathbb{C}$, the set of complex numbers, we write $z = \mathrm{Re}(z) + \mathrm{i}\,\mathrm{Im}(z)$ and $|z| = \left\{\mathrm{Re}(z)^2 + \mathrm{Im}(z)^2\right\}^{1/2}$ is the modulus of z; all limits in this paper are taken when $k \to \infty$; $\mathbb{E}$ and $\mathbb{V}$ denote expectation and variance, respectively, and $\mathbb{E}_0$ and $\mathbb{V}_0$ denote expectation and variance under the null hypothesis H_0, respectively.

2 | THE TEST STATISTIC

Assume that $(X,Y) \in \mathbb{R}^{2d}$, with $X,Y \in \mathbb{R}^d$. Let $\varphi = \varphi_{(X,Y)}$, that is, $\varphi(t,s)$, $t,s \in \mathbb{R}^d$, denotes the joint characteristic function of the random vector (X,Y). Then $\varphi_X(t) = \varphi(t,0)$, and $\varphi_Y(t) = \varphi(0,t)$, $t \in \mathbb{R}^d$. X and Y have the same marginal distribution if and only if $\varphi_X(t) = \varphi_Y(t)$, $\forall t \in \mathbb{R}^d$, which is tantamount to $\eta_w = 0$, with

$$\eta_w = \int |\varphi_X(t) - \varphi_Y(t)|^2 w(t)\mathrm{d}t, \tag{1}$$

where an unspecified integral denotes integration over $\mathbb{R}^d$ and w is a probability density function (pdf) defined on $\mathbb{R}^d$, which is positive everywhere. Let (X_1,Y_1) and (X_2,Y_2) be two independent copies of (X,Y). Proceeding as in Section 2 of Chen et al. (2022), it can be seen that

$$\begin{aligned}|\varphi_X(t)-\varphi_Y(t)|^2 &= \varphi_{X_1-X_2}(t)+\varphi_{Y_1-Y_2}(t)-2\varphi_{X_1-Y_2}(t)\\ &= \mathbb{E}\left[\cos\{t^\top(X_1-X_2)\}+\cos\{t^\top(Y_1-Y_2)\}-2\cos\{t^\top(X_1-Y_2)\}\right].\end{aligned} \tag{2}$$

From (2) and using Fubini's theorem, we have

$$\eta_w = \mathbb{E}\left(\int\left[\cos\{t^\top(X_1-X_2)\}+\cos\{t^\top(Y_1-Y_2)\}-2\cos\{t^\top(X_1-Y_2)\}\right]w(t)\mathrm{d}t\right). \tag{3}$$

Let $u(t)=\int\cos(t^\top x)w(x)\mathrm{d}x$, that is, u is the real part of the characteristic function of a random vector with pdf w. With this notation, (3) can be equivalently written as

$$\eta_w = \mathbb{E}\{u(X_1-X_2)+u(Y_1-Y_2))-2u(X_1-Y_2)\}. \tag{4}$$

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a random sample (independent copies) of (X,Y). From (4), η_w can be unbiasedly estimated by means of

$$\hat{\eta}_w = \frac{1}{n(n-1)}\sum_{1\le r\ne s\le n} h\{(X_r,Y_r),(X_s,Y_s)\}, \tag{5}$$

with

$$h\{(X_r,Y_r),(X_s,Y_s)\} = u(X_r-X_s)+u(Y_r-Y_s)-u(X_r-Y_s)-u(X_s-Y_r). \tag{6}$$

Remark 1. The statistic $\hat{\eta}_w$ is a bit different from that considered in Quessy and Éthier (2012). First, in that paper, it is assumed that $d=1$; this assumption causes no serious restriction since the extension to $d>1$ is easy. Second, the test statistic in that paper is built by replacing the characteristic functions in the expression (1) of η_w with their empirical versions obtaining, say, $\tilde{\eta}_w$. It can be easily checked that $\mathbb{E}(\tilde{\eta}_w)\to\eta_w$, as $n\to\infty$, but $\mathbb{E}(\tilde{\eta}_w)\ne\eta_w$, for each finite n.

Now, assume that there are k populations and that a random sample of size n_j, $\left(X^j_1,Y^j_1\right),\ldots,\left(X^j_{n_j},Y^j_{n_j}\right)$, is available from (X^j,Y^j), $1\le j\le k$. The k random samples are assumed to be independent. Let $\hat{\eta}_{j,w}$ denote the statistic in (5) when it is evaluated in the sample from population j, $1\le j\le k$, and define

$$T_k = \sum_{j=1}^{k}\hat{\eta}_{j,w}.$$

Let

$$\eta_{j,w} = \int|\varphi_{X^j}(t)-\varphi_{Y^j}(t)|^2 w(t)\mathrm{d}t,\ 1\le j\le k.$$

Notice that $\mathbb{E}(T_k)=\sum_{j=1}^{k}\eta_{j,w}\ge 0$, with $\mathbb{E}(T_k)=0$ if and only if H_0 is true. Thus, it is reasonable to reject the null hypothesis for large values of T_k. Now, to determine what are large values, we have to calculate its distribution under the null hypothesis, or at least an estimator.

As an approximation to the null distribution of T_k, the next theorem derives its asymptotic null distribution. With this aim, it will be assumed that the sample sizes of the data from each population are comparable in the following sense:

$$n_j = c_j m,\ m\ge 1,\ 0<c_0\le c_j\le C_0<\infty,\ 1\le j\le k, \tag{7}$$

where c_0 and C_0 are two fixed constants. In the above expression, m is allowed to vary with k, so, strictly speaking, it should be denoted as $m(k)$ but, in order to keep the notation as simple as possible, such dependence on k will be skipped.

Theorem 1. *Suppose that H_0 is true, that $n_j \geq 2$, $\forall j$, that (7) holds and that $\mathbb{E}\left[h^2\left\{\left(X_1^i, Y_1^i\right), \left(X_2^i, Y_2^i\right)\right\}\right] \geq \delta$, $\forall j$, for some $\delta > 0$. Then, $T_k / \sqrt{\mathbb{V}_0(T_k)} \xrightarrow{\mathcal{L}} Z$, where $Z \sim N(0,1)$.*

Therefore, if $\mathbb{V}_0(T_k)$ were a known quantity, the test that rejects H_0 when $T_k / \sqrt{\mathbb{V}_0(T_k)} \geq z_{1-\alpha}$, for some $\alpha \in (0,1)$, would have (asymptotic) level α, where $z_{1-\alpha}$ denotes the upper α-percentile of the standard normal distribution. From the independence of $\hat{\eta}_{1,w}, \ldots, \hat{\eta}_{k,w}$, it follows that

$$\mathbb{V}(T_k) = \sum_{j=1}^{k} \mathbb{V}(\hat{\eta}_{j,w}).$$

In particular, under H_0 (see the proof of Theorem 1),

$$\mathbb{V}_0(T_k) = 2 \sum_{j=1}^{k} \frac{1}{n_j(n_j-1)} \xi_j,$$

where $\xi_j = \mathbb{E}\left[h^2\left\{\left(X_1^i, Y_1^i\right), \left(X_2^i, Y_2^i\right)\right\}\right]$, $1 \leq j \leq k$, which are unknown quantities. Hence, for the result in Theorem 1 to be useful in order to get a critical region for testing H_0, a ratio consistent estimator of $\mathbb{V}_0(T_k)$ is needed. We will consider as estimator of $(1/k)\mathbb{V}_0(T_k)$ the sample variance of $\hat{\eta}_{1,w}, \ldots, \hat{\eta}_{k,w}$,

$$S_k^2 = \frac{1}{k} \sum_{j=1}^{k} \left(\hat{\eta}_{j,w} - \overline{\hat{\eta}_{\cdot,w}}\right)^2, \quad \overline{\hat{\eta}_{\cdot,w}} = \frac{1}{k} \sum_{j=1}^{k} \hat{\eta}_{j,w}.$$

Next proposition shows that, under H_0, S_k^2 is a ratio (strongly) consistent estimator of $(1/k)\mathbb{V}_0(T_k)$.

Proposition 1. *Suppose that assumptions in Theorem 1 hold. Then $kS_k^2 / \mathbb{V}_0(T_k) \xrightarrow{a.s.} 1$.*

As an immediate consequence of Theorem 1 and Proposition 1, we have the following result.

Corollary 1. *Suppose that assumptions in Theorem 1 hold. Then $T_k / \sqrt{k} S_k \xrightarrow{\mathcal{L}} Z$.*

Let $\alpha \in (0,1)$. For testing H_0 versus H_1, we consider the test that reject the null when

$$\frac{T_k}{\sqrt{k} S_k} \geq z_{1-\alpha}. \tag{8}$$

From Corollary 1, it has asymptotic level α.

Remark 2. Notice that to derive the previous results, it has only been assumed that the sample sizes are comparable, in the sense of (7), and that $n_j \geq 2$, $\forall j$. Hence, the stated results remain true if they increase arbitrarily or remain bounded.

Remark 3. Notice also that to derive the previous results, no assumption has been made on the distribution of X and Y, which could be continuous, discrete or mixed; moreover, no moment is assumed to exist.

Remark 4. The proof of Theorem 1 uses that $\mathbb{V}_0(mT_k) = \sum_{j=1}^{k} \mathbb{V}_0(m\hat{\eta}_{j,w}) \to \infty$, and the proof of Proposition 1 uses that $\mathbb{V}_0(mT_k)/k$ is a positive quantity. The assumption $\mathbb{E}_0\left[h^2((X_1, Y_1), (X_2, Y_2))\right] \geq \delta$, $\forall i$, ensures both requirements. Nevertheless, it can be replaced by any other assumption whenever those requirements are met, as, for example, that $\mathbb{E}\left[h^2\left\{\left(X_1^i, Y_1^i\right), \left(X_2^i, Y_2^i\right)\right\}\right] \geq \delta$, for a positive proportion of cases, for some $\delta > 0$.

Remark 5. All results in this section assume that the same pdf w is used in the definition of η_w in all subpopulations. Nevertheless, different pdfs could be taken at each population, say w_j, $1 \le j \le k$. In such a case, we still can use the critical region (8) with $\hat{\eta}_{j,w}$ replaced with $\hat{\eta}_{j,w_j}$, $1 \le j \le k$. The results in this section remain true whenever all kernels (6) associated to each w_j satisfy the required assumptions. Specifically, if h_j stands for the kernel (6) associated to w_j, then the assumption $\mathbb{E}\left[h^2\left\{\left(X_1^j, Y_1^j\right), \left(X_2^j, Y_2^j\right)\right\}\right] \ge \delta$, $\forall j$, for some $\delta > 0$, in Theorem 1 now becomes $\mathbb{E}\left[h_j^2\left\{\left(X_1^j, Y_1^j\right), \left(X_2^j, Y_2^j\right)\right\}\right] \ge \delta$, $\forall j$, for some $\delta > 0$, or at least for a positive proportion of cases, as observed in Remark 4.

Remark 6. To define T_k, we have chosen the sum of modified one-population test statistics (see Remark 1). The reason to opt for the sum is that, under the null hypothesis, we get an asymptotic free distributed test statistic (as shown in Corollary 1). The same is observed in other sum-type test statistics as those in Park and Park (2012), Zhan and Hart (2014), Jiménez-Gamero and Franco-Pereira (2021), Jiménez-Gamero et al. (2022) and Jiménez-Gamero (2023), just to cite a few. If one instead takes the maximum, the asymptotic null distribution of the resultant test statistic becomes more complicated, and one has to resort to resampling in order to approximate its null distribution. See, for example, Kim (2021).

Remark 7. Proposition 1 in Székely and Rizzo (2013) shows that the energy distance between two distributions is of the type considered in this paper: It is an L^2-type distance among the characteristic functions of the populations, as in (1). The main difference is that while the weight function w used here is a pdf, and thus it has finite integral, the weight function that uses the energy distance does not have a finite integral. To ensure the finiteness of the energy distance, the populations must have finite expectations. As observed in Remark 3, the proposal in this paper does not require the existence of any population moment. A similar test to the one studied in this paper could be designed by using the energy distance, but, as noticed before, its application would need stronger assumptions.

3 | POWER

This section deals with the asymptotic power of the test with the critical region in (8). With this aim, it will be also assumed w.l.o.g. that

$$F_{X^j} \ne F_{Y^j},\ 1 \le j \le v, \text{ and } F_{X^j} = F_{Y^j},\ v+1 \le j \le k, \tag{9}$$

or equivalently that

$$\varphi_{X^j} \ne \varphi_{Y^j},\ 1 \le j \le v, \text{ and } \varphi_{X^j} = \varphi_{Y^j},\ v+1 \le j \le k,$$

for some $1 \le v \le k$. If $v = k$, then $v+1 \le i \le k$ is understood to be the empty set. Here, v is allowed to vary with k, $v = v(k)$, but such dependence on k will be skipped. In order to derive the asymptotic power, we must study the asymptotic behaviour of T_k and S_k^2 under alternatives (9). The next result shows that, under alternatives and conveniently normalized, T_k also converges in law to a standard normal law.

Theorem 2. *Suppose that (7) and (9) hold, that* $\mathbb{V}\left(\mathbb{E}\left[h\left\{\left(X_1^j, Y_1^j\right), \left(X_2^j, Y_2^j\right)\right\} \middle| \left(X_2^j, Y_2^j\right)\right]\right) \ge \delta$, $1 \le j \le v$, *and* $\mathbb{E}\left[h^2\left\{\left(X_1^j, Y_1^j\right), \left(X_2^j, Y_2^j\right)\right\}\right] \ge \delta$, $v+1 \le j \le k$, *for some* $\delta > 0$, *and that either* $v \to \infty$ *or* v *is bounded and* $m/k \to 0$. *Then* $\left(T_k - \sum_{j=1}^{v} \eta_{j,w}\right) \Big/ \sqrt{\mathbb{V}(T_k)} \xrightarrow{\mathcal{L}} Z$.

The next proposition gives the limit in probability, under alternatives, of the variance estimator S_k^2. With this aim, we first write its expected value:

$$\mathbb{E}\left(S_k^2\right) = \frac{1}{k}\left(1 - \frac{1}{k}\right)\sum_{j=1}^{k} \mathbb{V}(\hat{\eta}_{j,w}) + \frac{1}{k}\sum_{j=1}^{k} \left(\eta_{j,w} - \overline{\eta_{\cdot,w}}\right)^2,\ \overline{\eta_{\cdot,w}} = \frac{1}{k}\sum_{j=1}^{k} \eta_{j,w}.$$

Notice that, under alternatives, S_k^2 is a biased estimator of $(1/k)\sum_{j=1}^{k} \mathbb{V}(\hat{\eta}_{j,w})$.

Proposition 2. *Suppose that assumptions in Theorem 2 hold. Then* $S_k^2 / \mathbb{E}\left(S_k^2\right) \xrightarrow{P} 1$.

Remark 8. Recall, as observed in Remark 2, that to derive results under the null hypothesis H_0, no condition was assumed on the sample sizes, except that they are comparable, in the sense of (5). By contrast, under alternatives, if v remains bounded, it is assumed that $m/k \to 0$. This is required because the order of the variance of $\hat{\eta}_{j,w}$ is different in populations not obeying the null (which is of order $O(m^{-1})$) and in homogeneous populations (which is of order $O(m^{-2})$). Such a condition is not severe at all, as it allows the sample sizes to remain bounded or increase with k, but at a lower rate. This is not a strong assumption, since in practice, when there is a large number of populations, it does not seem plausible to have many data coming from each population.

As a consequence of Theorem 2 and Proposition 2, the power of the test (8) can be approximated as follows:

$$\begin{aligned}\text{pwd} = P\left(\frac{T_k}{\sqrt{k}S_k} \geq z_{1-\alpha}\right) &= P\left(\frac{T_k - \sum_{j=1}^{v}\eta_{j,w}}{\sqrt{\mathbb{V}(T_k)}} \geq \frac{S_k}{\sqrt{\mathbb{V}(T_k)/k}}z_{1-\alpha} - \sqrt{k}\frac{\overline{\eta_{\cdot,w}}}{\sqrt{\mathbb{V}(T_k)/k}}\right)\\ &\approx \Phi\left(\sqrt{k}\text{coc}_1 + \text{coc}_2^{1/2}z_\alpha\right),\end{aligned}$$

where Φ denotes the cumulative distribution function of a standard normal law and

$$\text{coc}_1 = \frac{\overline{\eta_{\cdot,w}}}{\sqrt{\mathbb{V}(T_k)/k}},\ \text{coc}_2 = \frac{\mathbb{E}\left(S_k^2\right)}{\mathbb{V}(T_k)/k}.$$

Notice that $\text{coc}_2 \geq 1 - 1/k$. If the alternative is such that $\text{coc}_2 \leq M$, for some positive constant M, and $\text{coc}_1 > \nu$, for some positive constant ν, then $P\left(T_k/\sqrt{k}S_k \geq z_{1-\alpha}\right) \to 1$; that is, the test (8) is consistent against this sort of alternatives.

There are many possible configurations of alternative distributions. Next, we study the following particular case: Assume that $n_1 = \ldots = n_k = m > 2$, that $\eta_{1,w} = \ldots = \eta_{v,w} := \eta > 0$ and that

$$\begin{aligned}\mathbb{V}(\hat{\eta}_{j,w}) &= \frac{m-2}{m(m-1)}\xi_1 + \frac{1}{m(m-1)}\xi_2,\ 1 \leq j \leq v,\\ \mathbb{V}(\hat{\eta}_{j,w}) &= \frac{1}{m(m-1)}\xi_3,\ v+1 \leq j \leq k,\end{aligned}$$

for some $0 < \xi_1, \xi_2, \xi_3$. Since $|u| \leq 1$, and hence $|h| \leq 4$, it readily follows that $\xi_1, \xi_2, \xi_3 < M$, for some positive constant M. The expression of the variances of $\hat{\eta}_{j,w}$ has the above form (Serfling, 2009); what we are assuming (in order to make the analysis easier) is that the quantities ξ_1 and ξ_2 are the same in all alternative cases and that the quantity ξ_3 is the same in all null cases. Assume also that assumptions in Proposition 2 hold. In this setting,

$$\begin{aligned}\text{coc}_1 &= \sqrt{m}\frac{\eta\frac{v}{k}}{\sqrt{\frac{v}{k}\left(\frac{m-2}{m-1}\xi_1 + \frac{1}{m-1}\xi_2\right) + \left(1-\frac{v}{k}\right)\frac{1}{m-1}\xi_3}},\\ \text{coc}_2 &= 1 - \frac{1}{k} + \eta^2 m\frac{\frac{v}{k}\left(1-\frac{v}{k}\right)}{\frac{v}{k}\left(\frac{m-2}{m-1}\xi_1 + \frac{1}{m-1}\xi_2\right) + \left(1-\frac{v}{k}\right)\frac{1}{m-1}\xi_3}.\end{aligned}$$

Next, we consider three cases:

Case 1. Suppose that $v/k \to p \in (0,1]$. Then $\text{coc}_1/\sqrt{m}$ is a positive, bounded quantity, which implies that $\sqrt{k}\text{coc}_1/\sqrt{m} \to \infty$. We also have that coc_2/m is a positive, bounded quantity. Therefore,

$$\text{pwd} \approx \Phi\left(\sqrt{m}\left\{\sqrt{k}\frac{\text{coc}_1}{\sqrt{m}} + z_\alpha\left(\frac{\text{coc}_2}{m}\right)^{1/2}\right\}\right) \to 1;$$

that is, the test is consistent against this sort of alternatives.

Case 2. Suppose now that $v/k \to 0$ and that m remains bounded. Then,

$$\frac{k\mathrm{coc}_1^2}{\mathrm{coc}_2} \approx a^2 m^2 k \left(\frac{v}{k}\right)^2,\ a = \frac{\eta}{\sqrt{\xi_3}} \left(\frac{m-1}{m}\right)^{1/2},\ \mathrm{coc}_2 \to 1,$$

where the approximation is understood for large k. Therefore,

- if $v/\sqrt{k} \to \infty$, then $\mathrm{pwd} \to 1$,
- if $v/\sqrt{k} \to \mu \in [0,\infty)$, then $\mathrm{pwd} \to \Phi(\mu + z_\alpha)$.

Case 3. Suppose now that $v/k \to 0$ and $mv/k \to \mu \in (0,\infty)$, which implies that $m \to \infty$. Then,

$$\sqrt{k}\mathrm{coc}_1 \approx \theta\sqrt{v}\sqrt{m},\ \mathrm{coc}_2 \approx 1 + \theta^2 m,\ \theta = \eta/\sqrt{\xi_1 + \xi_3/\mu},$$

and thus, $\mathrm{pwd} \approx \Phi\left(\theta\sqrt{m}\left\{\sqrt{v} + z_\alpha\right\}\right)$. Therefore, if $\sqrt{v} + z_\alpha > 0$, then $\mathrm{pwd} \to 1$.
Summarizing, the test is consistent against alternatives whenever $v/k \to p \in (0,1]$, that is, when the proportion of alternative populations is positive; if such proportion goes to 0, then the test can be also consistent, have no power or have a power greater than α.

Remark 9. As observed in Remark 5, different pdfs could be taken at each population. The results in this section remain true whenever all kernels (6) associated to each w_j, say h_j, satisfy the required assumptions. Specifically, $\mathbb{V}\left(\mathbb{E}\left[h_j\left\{\left(X_1^j, Y_1^j\right), \left(X_2^j, Y_2^j\right)\right\} \middle| \left(X_2^j, Y_2^j\right)\right]\right) \geq \delta$, $1 \leq j \leq v$, and $\mathbb{E}\left[h_j^2\left\{\left(X_1^j, Y_1^j\right), \left(X_2^j, Y_2^j\right)\right\}\right] \geq \delta$, $v+1 \leq j \leq k$, for some $\delta > 0$.

Remark 10. That the sample sizes are comparable, in the sense of (7), is assumed for ease of exposition. If it is not fulfilled, the stated results remain true by adding appropriate conditions, which now become rather *ugly*. For example, to prove the asymptotic normality in Theorem 1, we show that Linbdeberg's condition is met. Now, if (7) is not supposed, it is difficult to find an *easy requisite* for that condition to hold (see Remark 4). If instead of choosing $T_k = \sum_{j=1}^k \hat{\eta}_{j,w}$, one takes $T_{k,new} = \sum_{j=1}^k n_i \hat{\eta}_{j,w}$ then it can be easily checked that Theorem 1 remains true without assuming (7). Nevertheless, when one faces the problem of determining the asymptotic distribution of $T_{k,new}$ under alternatives, the same problem arises again, due to the fact that the order of the variance of $\hat{\eta}_{j,w}$ differs under the null and under alternatives.

Remark 11. So far, we have assumed that the data at hand consist of n_j independent copies of (X^j, Y^j), $1 \leq j \leq k$. Now, assume that independent data are available from each vector, X^j and Y^j; that is, the data consist of $X_1^j, \ldots, X_{n_X^j}^j$, n_X^j independent copies of X^j, and $Y_1^j, \ldots, Y_{n_Y^j}^j$, n_Y^j independent copies of Y^j, with $X_1^j, \ldots, X_{n_X^j}^j$ and $Y_1, \ldots, Y_{n_Y^j}^j$ independent, $1 \leq j \leq k$. In this setting, an unbiased estimator of $\eta_{j,w}$ is given by

$$\hat{\eta}_{j,w} = \frac{1}{n_X^j\left(n_X^j - 1\right)} \sum_{1 \leq r \neq s \leq n_X^j} u\left(X_r^j - X_s^j\right) + \frac{1}{n_Y^j\left(n_Y^j - 1\right)} \sum_{1 \leq r \neq s \leq n_Y^j} u\left(Y_r^j - Y_s^j\right)$$

$$-2\frac{1}{n_X^j n_Y^j} \sum_{1 \leq r \leq n_X^j} \sum_{1 \leq s \leq n_Y^j} u\left(X_r^j - Y_s^j\right).$$

In the case of paired samples, $\hat{\eta}_{j,w}$ is a one-sample degree 2 U-statistic; while in the case of independent samples, $\hat{\eta}_{j,w}$ is a two-sample degree (2, 2) U-statistic. It can be easily checked that the results previously stated for the case of paired data keep on being true for the case of independent data.

4 | SOME PRACTICAL ISSUES

To calculate the test statistic $T_k/\sqrt{k}S_k$, the user must fix $w_1, \ldots, w_k$, where w_j is the pdf in (1) for population j, $1 \leq j \leq k$ (recall from Remarks 5 and 9 that different pdfs can be taken at each population). Two common choices for w in tests based on the empirical characteristic function are (Jiménez-Gamero et al., 2017, 2019; Meintanis, 2005):

- Normal weight:

$$w(t_1,\ldots,t_d)=\frac{1}{(2\pi)^{d/2}}\frac{1}{\prod_{r=1}^{d}a_r}\exp\left(-0.5\sum_{r=1}^{d}\frac{t_r^2}{a_r^2}\right),\ t_1,\ldots,t_d\in\mathbb{R},$$

for some $a_1,\ldots,a_d>0$, that is, w is the product of d univariate pdfs of normal laws with mean zero and variance a_r^2, $1\le r\le d$. This choice for w gives

$$u(t_1,\ldots,t_d)=\exp\left(-0.5\sum_{r=1}^{d}a_r^2t_r^2\right),\ t_1,\ldots,t_d\in\mathbb{R}.$$

- Laplace weight:

$$w(t_1,\ldots,t_d)=\frac{1}{\prod_{r=1}^{d}2c_r}\exp\left(\sum_{r=1}^{d}\frac{|t_r|}{c_r}\right),\ t_1,\ldots,t_d\in\mathbb{R},$$

for some $c_1,\ldots,c_d>0$, that is, w is the product of d univariate pdfs of Laplace laws with mean zero and variance $2c_r^2$, $1\le r\le d$. This choice for w gives

$$u(t_1,\ldots,t_d)=\prod_{r=1}^{d}\frac{1}{1+c_r^2t_r^2},\ t_1,\ldots,t_d\in\mathbb{R}.$$

Notice that the choice of the parameter a_r for the normal weight in population j, denote it by $a_{r,j}$, is coupled with the scaling of the rth component of (X^j,Y^j), say X_r^j,Y_r^j. So, it seems reasonable to take $a_{r,j}=\lambda_j/S_{r,j,pool}$, for some positive λ_j, where $S^2_{r,j,pool}$ is the sample variance of the rth component of (X^j,Y^j) in the pooled sample $X^j_{1,r},Y^j_{1,r},\ldots,X^j_{n_j,r},Y^j_{n_j,r}$. The same applies to the choice of the parameter c_r for the Laplace weight.

The above observation leads us to the choice of the proportionality constant λ_j. We have numerically analysed this issue through an extensive simulation study. We found that the level is not affected by the choice of the proportionality constant, but the power does; moreover, in some cases, the power increases with it, in other cases decreases, and in some further cases, it has a maximum at $\lambda_j=1$ for the normal weight and $\lambda_j=1/\sqrt{2}$ for the Laplace weight. As for these values of λ_j, we got reasonable powers in all tried cases, and we recommend taking $a_{r,j}=1/S_{r,j,pool}$, for the normal weight, and $c_{r,j}=1/\sqrt{2}S_{r,j,pool}$, for the Laplace weight. The next section shows up part of the simulation study we carried out, supporting our recommendation.

5 | EMPIRICAL RESULTS

5.1 | Simulated data: level

The test that rejects H_0 according to (8) has asymptotic level α. In order to check whether the proposal behaves well, in the sense of reaching the level for a small or moderate number of groups, an extensive simulation study was carried out for the bivariate case ($d=1$). Let (X,Y) be a bivariate random vector with joint distribution function $F_{(X,Y)}$ and let F_X, F_Y be the distribution functions of X and Y, respectively. If X and Y have continuous marginals, then Sklar's theorem (Nelsen, 2006) ensures that there exists a unique copula $C:[0,1]^2\to[0,1]$ that characterizes the dependence structure of $F_{(X,Y)}$, in the sense that $F_{(X,Y)}(x,y)=C(F_X(x),F_Y(y))$, $\forall x,y\in\mathbb{R}$. Artificial data have been generated from several patterns of dependence characterized by bivariate copulas. Particularly, we have taken three copulas: the normal copula (denoted in the tables as C^N), the Clayton copula (denoted in the tables as C^C) and the Gumbel copula (denoted in the tables as C^G). The parameters of the copulas were chosen so that Kendall's τ coefficient has the same value in all cases (see details in the Supporting Information). For each copula, the following marginals were considered: the univariate standard normal distribution, $N(0,1)$, and the exponential distribution with mean 1, $Exp(1)$. For the weight functions, we took those in Section 4 with a wide range of values for λ_j; in what follows, $a_{1,j}$, $c_{1,j}$ and $S_{1,j,pool}$ will be simply denoted as a_j, c_j and $S_{j,pool}$, respectively. Since in simulations, we observed that the level is not strongly affected by the choice of λ_j, the tables in this section (and those in the Supporting Information) display the results with $\lambda_j=\sqrt{2},1,1/\sqrt{2}$, for the normal weight, and $\lambda_j=1,1/\sqrt{2},1/2$, for the Laplace weight (these values correspond to variances 2, 1 and 0.5 of the weight functions).

TABLE 1 Estimated type I errors for nominal value $\alpha = 0.05$, $\tau = 1/3$ in all copulas and $Exp(1)$ marginals.

Normal weight										
		$a_j = \sqrt{2}/S_{j,pool}$			$a_j = 1/S_{j,pool}$			$a_j = 1/\sqrt{2}S_{j,pool}$		
k	n	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G
20	5	3.2	3.2	3.2	3.4	2.8	3.4	3.0	2.4	3.1
	10	2.9	3.2	3.0	2.9	2.8	2.9	2.9	3.2	2.9
50	5	3.5	3.4	3.6	3.3	3.3	3.5	3.2	3.1	3.0
	10	3.4	3.4	3.5	3.2	3.3	3.4	2.9	3.1	3.0
100	5	4.0	3.9	3.9	3.6	3.8	3.7	3.7	3.6	3.6
	10	3.7	3.9	3.8	3.6	3.8	3.6	3.7	3.6	3.6
200	5	4.2	3.9	4.1	4.3	3.8	3.7	4.2	3.7	3.6
	10	4.2	3.8	4.4	4.1	3.8	4.3	4.1	3.7	4.0
500	5	4.5	4.3	4.3	4.5	4.3	4.5	4.3	4.1	4.1
	10	4.4	4.3	4.4	4.3	4.3	4.3	4.3	4.1	4.2
Laplace weight										
		$c_j = 1/S_{j,pool}$			$c_j = 1/\sqrt{2}S_{j,pool}$			$c_j = 1/2S_{j,pool}$		
k	n	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G
20	5	3.3	3.2	3.4	3.2	2.8	3.4	3.1	2.5	3.0
	10	2.9	3.2	3.2	2.8	2.8	3.0	2.7	2.5	3.1
50	5	3.6	3.4	3.6	3.7	3.4	3.6	3.3	3.1	3.2
	10	3.4	3.4	3.5	3.3	3.4	3.6	3.0	3.1	3.3
100	5	3.9	3.9	3.8	3.9	3.8	3.7	3.7	3.7	3.6
	10	3.6	3.9	4.0	3.5	3.8	3.7	3.3	3.7	3.5
200	5	4.5	3.8	4.1	4.4	3.9	3.7	4.2	3.7	3.6
	10	4.0	3.8	4.3	4.0	3.9	4.2	4.0	3.7	4.1
500	5	4.5	4.3	4.2	4.5	4.2	4.3	4.3	4.2	4.2
	10	4.3	4.3	4.2	4.2	4.2	4.3	4.2	4.2	4.1

For each case, k random samples of size n_j, $(X_1^j, Y_1^j), \ldots, (X_{n_j}^j, Y_{n_j}^j)$ were generated by using the R package `copula` (Hofert et al., 2023) with sample sizes $n_j = 5{,}10$, $1 \le j \le k$, and $k = 20{,}50{,}100{,}200{,}500$. After calculating $T_k/\sqrt{k}S_k$, the p-value was computed using the normal approximation in Corollary 1. The experiment was repeated 10,000 times, and the percentage of p-values less than or equal to $\alpha = 0.05$ was collected, which estimates the probability of type I error. Table 1 reports the results obtained when the marginal distributions are $Exp(1)$ and $\tau = 1/3$. The results for $\tau = 2/3$ and those for standard normal marginals can be found in the Supporting Information. Looking at this table (and the tables in the Supporting Information), one can see that the empirical levels become closer to the nominal value $\alpha = 0.05$ as k increases. This finding does not seem to be significantly affected by the weight function.

5.2 | Simulated data: power

The power of the proposal has also been investigated by simulations. Artificial data were generated under the following conditions: (i) normal and Laplace weight functions with values a_j and c_j chosen proportional to $1/S_{j,pool}$, (ii) normal, Clayton and Gumbel copulas with parameters so that $\tau = 1/3$, (iii) $n_j = 5{,}10$ and $k = 50{,}100$ for the sample sizes and the number of groups, respectively, (iv) $100(1-p)\%$ of groups with $N(0,1)$ marginal distributions, and the rest of $100p\%$ of groups with different marginals chosen from logistic, beta, gamma, Laplace, normal and uniform distribution functions for a great variety of parameters, for $p = 0.2{,}0.4{,}0.6$. As before, each case was repeated 10,000 times, and the percentage of p-values less than or equal to $\alpha = 0.05$ was calculated.

From the battery of alternatives tried, we found three patterns for the power: For a first set of cases (setting 1), the powers increase with λ_j, for a second one (setting 2), they decrease with it, and in the third group of cases (setting 3), the higher estimated powers were obtained for $\lambda_j = 1$ and $\lambda_j = 1/\sqrt{2}$ for the normal and Laplace weight functions, respectively. Table 2 displays the results of an instance of the first set, with alternative cases $X \sim U(-1,1)$ and $Y \sim U(-0.75,1.25)$; Table 3 displays the results of an instance of the second set, with alternative cases $X \sim U(0,1)$ and $Y \sim \beta(2,2)$; and Table 4 displays the results of an instance of the third set, with alternative cases $X \sim N(0,1)$ and $Y \sim U(-1,1)$. These tables show

TABLE 2 Estimated power for alternative cases $X \sim U(-1,1)$ and $Y \sim U(-0.75,1.25)$ with normal weight.

			$a_j = \sqrt{2}/S_{j,pool}$			$a_j = 1/S_{j,pool}$			$a_j = 1/\sqrt{2}S_{j,pool}$		
p	k	n	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G
0.2	100	5	8.2	8.3	8.5	9.7	10.1	9.9	12.2	12.8	12.4
		10	17.5	17.0	17.5	22.3	22.4	22.5	31.3	30.8	31.6
	200	5	12.1	12.2	11.9	15.2	15.3	15.4	19.8	20.5	20.0
		10	28.3	29.2	29.0	37.9	38.2	38.6	53.5	52.7	54.1
0.3	100	5	11.0	11.6	11.6	14.1	14.9	14.5	19.5	20.2	19.9
		10	29.3	29.0	30.0	38.2	38.5	39.4	53.4	52.3	53.5
	200	5	17.8	18.6	18.2	23.7	24.0	24.3	34.2	33.5	34.2
		10	48.9	49.8	50.6	63.7	63.3	64.9	82.3	80.9	82.4
0.4	100	5	14.8	15.7	16.0	19.6	20.5	20.4	27.9	28.3	29.3
		10	43.5	43.4	44.2	56.5	55.8	56.4	73.1	72.7	73.2
	200	5	24.9	26.0	25.9	34.8	34.8	34.9	49.3	49.3	49.6
		10	68.9	70.4	70.9	84.0	83.4	84.2	95.0	95.0	95.0

TABLE 3 Estimated power for alternative cases $X \sim U(0,1)$ and $Y \sim \beta(2,2)$ with normal weight.

			$a_j = \sqrt{2}/S_{j,pool}$			$a_j = 1/S_{j,pool}$			$a_j = 1/\sqrt{2}S_{j,pool}$		
p	k	n	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G
0.2	100	5	8.6	9.2	9.2	8.2	8.9	8.7	7.1	7.3	7.1
		10	18.8	19.7	19.8	18.1	18.9	19.1	13.9	14.6	14.3
	200	5	13.1	13.4	13.1	12.4	12.7	12.6	9.8	10.5	10.1
		10	31.2	32.8	32.5	30.9	31.5	31.7	23.6	23.3	23.3
0.3	100	5	12.2	13.5	12.8	11.6	12.6	12.6	9.3	10.1	9.9
		10	31.6	33.1	33.6	31.1	32.4	32.9	23.2	23.7	24.6
	200	5	20.3	20.8	20.4	19.1	19.7	19.4	14.3	15.1	14.3
		10	53.5	55.7	56.3	56.0	54.3	55.6	40.2	40.6	41.7
0.4	100	5	16.8	18.5	17.8	15.9	17.4	17.0	11.9	13.1	12.9
		10	47.0	48.6	49.1	46.7	47.8	48.7	35.4	35.7	36.3
	200	5	28.9	29.5	29.9	27.6	27.8	28.3	20.1	20.1	20.0
		10	74.9	76.3	77.3	74.8	75.6	76.3	59.8	60.0	61.8

up the power results for the normal weight; the results for the Laplace weight exhibit the same pattern and can be found in the Supporting Information. From the whole simulation experiment, it can be highlighted that the estimated power increases with the sample size, the proportion of groups with different marginals and the number of groups. Besides, as in any practical situation users do not know the distribution of the data, we recommend using $a_j = 1/S_{j,pool}$ when the normal weight function is taken to apply the proposal, and $c_j = 1/\sqrt{2}S_{j,pool}$ for the Laplace weight function, as for these choices, we got reasonable powers in all cases.

The problem of testing H_0 can also be dealt with by testing each hypothesis that composes it, that is, testing $H_{0j}: F_{X^j} = F_{Y^j}$, against $H_{1j}: F_{X^j} \neq F_{Y^j}$, $1 \leq j \leq k$, obtaining $p_1,...,p_k$, the p-values for each test, and then applying some method to adjust them, as for example, the Bonferroni method, which controls the family-wise error rate, or the Benjamini-Hochberg method (Benjamini & Yekutieli, 2001), which controls the false discovery rate when the k tests are independent. Both procedures agree in rejecting H_0 if $\min_{1 \leq j \leq k} p_j \leq \alpha/k$. Another method is the higher criticism (HC), introduced by Tukey (Donoho & Jin, 2004), that rejects H_0, at the level $\alpha = 0.05$, for large values of $HC_{0.05,k} = \sqrt{k}\{(\text{fraction of} p_j \leq 0.05) - 0.05\}/\sqrt{0.05 \times 0.95}$. In our simulations, we rejected H_0 when $HC_{0.05,k} > z_{0.95}$. The point is that $p_1,...,p_k$ cannot be exactly calculated. Two methods to approximate them are the bootstrap and the permutation. In simulations, we observed that for small sample sizes, both methods give rather conservative p-value estimators, especially the bootstrap. Hence, when applied the above methods to adjust the obtained p-values, the procedures were very liberal. Table 5 displays the empirical levels obtained when $p_1,...,p_k$ were approximated with permutation, obtained by generating 1000 artificial samples in each case, with normal copula, $\tau = 1/3$, standard normal marginals, with the recommended values for a_j and c_j for the normal weight and the Laplace weight, respectively, and the whole experiment was repeated 1000 times. Looking at this table, one can see that, in most cases, the procedures are rather liberal, only for $k = 200$ and $n_j = 20,25$ (the sample sizes in

TABLE 4 Estimated power for alternative cases $X \sim N(0,1)$ and $Y \sim U(-1,1)$ with normal weight.

			$a_j = \sqrt{2}/S_{j,pool}$			$a_j = 1/S_{j,pool}$			$a_j = 1/\sqrt{2}S_{j,pool}$		
p	k	n	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G	C^N	C^{CL}	C^G
0.2	100	5	13.0	13.3	13.3	12.9	13.7	13.5	10.8	11.4	11.5
		10	32.3	32.4	33.7	37.0	36.5	38.4	34.4	33.8	35.2
	200	5	20.2	20.9	20.9	21.2	21.9	21.9	17.9	18.1	18.0
		10	53.8	54.7	56.2	61.2	61.0	63.0	57.8	57.1	59.0
0.3	100	5	20.1	21.0	21.2	21.2	22.1	22.7	17.4	17.8	18.2
		10	56.2	57.7	58.6	63.4	63.7	65.6	60.7	60.4	62.3
	200	5	34.1	35.5	35.7	36.3	37.1	37.3	29.8	30.1	30.4
		10	83.6	84.4	85.1	89.1	90.1	90.6	87.3	87.3	88.6
0.4	100	5	29.9	31.0	31.5	31.9	32.8	33.7	25.6	26.6	27.3
		10	77.4	78.5	79.2	83.6	84.6	85.6	82.3	82.2	83.4
	200	5	50.6	52.3	53.5	53.5	54.6	56.0	44.0	45.1	45.7
		10	96.6	96.8	97.3	98.5	98.5	98.7	98.0	98.0	98.2

TABLE 5 Estimated type I errors with Normal copula, $N(0,1)$ marginals, for the normal weight (N) and the Laplace weight (L), when $p_1,\ldots,p_k$ are estimated with 1000 permutation samples in each case, and the global decision is taken by using HC criterion (HC) and the Benjamini-Hochberg method (BH).

	$k = 100$				$k = 200$			
	HC		BH		HC		BH	
n_j	N	L	N	L	N	L	N	L
5	21.0	20.9	100	99.9	26.2	23.6	100	100
10	7.3	7.8	20.2	20.9	6.8	7.2	37.3	36.1
15	7.1	7.4	10.1	9.4	5.8	6.2	18.2	17.0
20	7.0	6.9	8.7	9.2	5.1	5.8	19.9	19.3
25	6.2	6.8	10.4	10.4	5.0	5.4	17.5	17.6

TABLE 6 Estimated powers with normal copula, $N(0,1)$ marginals, for the normal weight (N) and the Laplace weight (L), $p = 0.2$, $k = 200$, when using the new proposal (New) and when $p_1,\ldots,p_k$ are estimated with 1000 permutation samples in each case, and the global decision is taken by using HC criterion (HC).

	Setting 1				Setting 2				Setting 3			
	HC		New		HC		New		HC		New	
n_j	N	L	N	L	N	L	N	L	N	L	N	L
20	59.0	68.2	85.5	90.5	59.1	55.7	75.1	72.7	94.7	92.9	98.1	97.6
25	78.3	87.0	94.6	97.0	75.0	70.3	89.9	87.6	99.6	99.2	99.8	99.6

Note: In setting 1, the alternative cases are $X \sim U(-1,1)$ and $Y \sim U(-0.75,1.25)$; in setting 2, the alternative cases are $X \sim U(0,1)$ and $Y \sim \beta(2,2)$; in setting 3, the alternative cases are $X \sim N(0,1)$ and $Y \sim U(-1,1)$.

all populations were taken equal to n_j) with the HC method the estimated type I errors are close to the nominal value, $\alpha = 0.05$. In these cases, we compared the power with our proposal, in the same settings explored in Tables 2–4, but only for $p = 0.2$. Table 6 displays the empirical powers obtained. Looking at this table, one can see that, in all tried cases, the new procedure exhibits greater power.

5.3 | A real data set application

Air is one of the main elements for the continuation of human life, and its growing deterioration, mainly caused by pollution, has become a serious challenge in most countries. In this worry, the US Environmental Protection Agency defined and promoted a standard index of air quality called

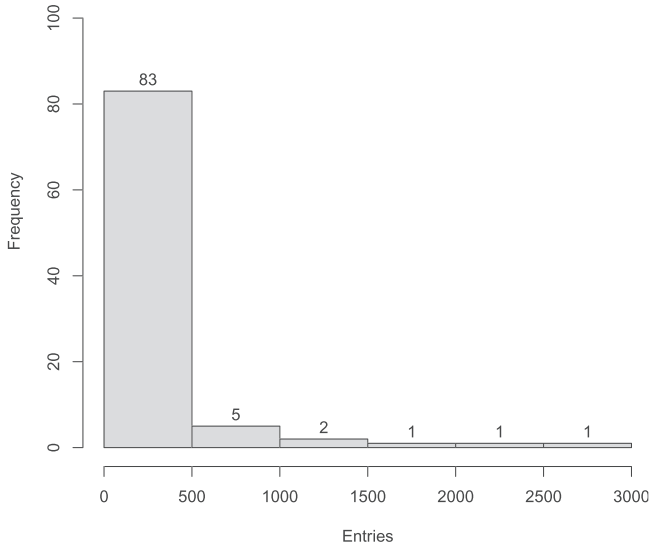

FIGURE 1 Histogram of sample sizes.

'Air Quality Index' that is mostly used by the rest of Environmental Protection Agencies to inform people about the quality of the air registered in big cities around the world. Details about how the AQI is calculated can be found in technical reports by the US Environmental Protection Agency (https://www.airnow.gov/).

The bulk of environmental agencies calculates AQI and makes it available to the general public through newspapers, web pages, apps or social networks. Readers are invited to visit the web page (https://aqicn.org/map/world/es/) to look for the available AQI values. Here, we consider AQI values of two pollutants: carbon monoxide (CO) and nitrogen dioxide (NO_2), two toxic substances produced as a result of the incomplete combustion of hydrocarbons and fossil fuels in industry, the combustion engines of vehicles and heating boilers, mainly. Both substances are largely responsible for the effect of traffic on pollution in large cities. The pair (CO, NO_2) was observed in a set of 23,036 big cities in 175 countries all over the world. The dataset is free (taken on 17 November 2022 from https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset).

Before applying the proposal, the dataset was cleaned up as follows: We started by removing those countries for which the sample variance of any pollutant is zero. So, from the initial 175 countries and 23,036 entries, we retained 121 countries and 22,070 entries.

In order to compare the results of the new proposal with the HC method, we did a further screening of the data. According to the findings in Section 5.2, we should have sample sizes greater than or equal to 20 to correctly apply the HC method to the p-values $p_1,\ldots,p_k$ (approximated by permutation as in the mentioned simulations). Attending to this requirement, the data set used ultimately consists of 93 countries with 21,768 entries. The median and mean of the sample sizes are 68 and 234.1, respectively. Figure 1 shows the histogram of the retained sample sizes. Then we tested H_0 versus H_1 for $(X,Y)=(CO,NO_2)$. To calculate the value of the test statistic in (8), the normal and Laplace weight functions were considered with the suggested tuning parameters ($a_j = 1/S_{j,pool}$ and $c_j = 1/\sqrt{2}S_{j,pool}$, $1 \le j \le k$, respectively), obtaining 9.81 and 11.17, respectively. Their associated p-values (calculated using the asymptotic null distribution) are 0 for both weight functions.

We also applied the HC method. With this aim, we first calculated the p-values $p_1,\ldots,p_k$ that were approximated with permutation (using 1000 artificial samples in each country). This procedure rejects H_0, at the level $\alpha = 0.05$, if $HC_{0.05,k} = \sqrt{k}\{(\text{fraction of} p_j \le 0.05) - 0.05\}/\sqrt{0.05 \times 0.95} > z_{0.95}$. In the case of (CO, NO_2), $HC_{0.05,93} = -2.2124$ for both weight functions, which is lower than 1.96. Therefore, according to this method, H_0 cannot be rejected. This is not surprising since, in line with the results in Table 6, the new procedure is more powerful than the HC method. In addition, after obtaining the permutation p-values for each weight function, we adjusted them using the Benjamini-Hochberg method. No hypothesis was rejected.

6 | CONCLUSIONS

This paper deals with the equality of marginals of a random vector (X,Y) when the target population is divided into k subpopulations or groups. Because the differences between marginals in some subpopulations could be compensated with others when they are considered as a whole population, we propose to test simultaneously the homogeneity of marginals in all groups. A procedure for carrying out that testing problem has been studied and analysed both theoretically and numerically.

It is worth mentioning that the requirement of independence between marginals is not needed. Furthermore, the proposal is very easy to implement, not requiring the use of complicated resampling methods to obtain the p-value. It applies whenever the sample sizes are comparable, allowing them to remain bounded or increase with k. These advantages of the proposal make it an attractive option to consider from a practical point of view.

AUTHOR CONTRIBUTIONS

Study concept and design: Maria Virtudes Alba-Fernández and Maria Dolores Jiménez-Gamero. *Drafting the manuscript:* Maria Virtudes Alba-Fernández and Maria Dolores Jiménez-Gamero. *Statistical analysis:* Maria Virtudes Alba-Fernández and Maria Dolores Jiménez-Gamero.

ACKNOWLEDGEMENTS

The authors thank two anonymous referees for their constructive comments and suggestions which helped to improve the presentation. M.V. Alba-Fernández acknowledges financial support from the grant PID2019-106195RB-100 (Ministerio de Ciencia, Innovación y Universidades, Spain) and M.D. Jiménez-Gamero acknowledges financial support from the grant PID2022-137818OB-I00 (Ministerio de Ciencia e Innovación, Spain).

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data set was taken from https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset, on 17 November 2022.

ORCID

M. V. Alba-Fernández https://orcid.org/0000-0002-4747-740X

M. D. Jiménez-Gamero https://orcid.org/0000-0002-8823-3292

REFERENCES

Alba-Fernández, M. V., Batsidis, A., Jiménez-Gamero, M. D., & Jodrá, P. (2017). A class of tests for the two-sample problem for count data. *Journal of Computational and Applied Mathematics*, *318*, 220–229. https://doi.org/10.1016/j.cam.2016.09.050

Alba-Fernández, M. V., & Jiménez-Gamero, M. D. (2009). Bootstrapping divergence statistics for testing homogeneity in multinomial populations. *Mathematics and Computers in Simulation*, *79*, 3375–3384. https://doi.org/10.1016/j.matcom.2009.04.002

Alba-Fernández, M. V., Jiménez-Gamero, M. D., & Muñoz García, J. (2008). A test for the two-sample problem based on empirical characteristic functions. *Computational Statistics & Data Analysis*, *52*(7), 3730–3748. https://doi.org/10.1016/j.csda.2007.12.013

Anderson, N. H., Hall, P., & Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *50*(1), 41–54. https://doi.org/10.1006/jmva.1994.1033

Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, *88*(1), 190–206. https://doi.org/10.1016/S0047-259X(03)00079-4

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Chen, F., Jiménez-Gamero, M. D., Meintanis, S. G., & Zhu, L. (2022). A general Monte Carlo method for multivariate goodness-of-fit testing applied to elliptical families. *Computational Statistics & Data Analysis*, *175*, Paper No. 107548. https://doi.org/10.1016/j.csda.2022.107548

Chen, H., & Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, *112*(517), 397–409. https://doi.org/10.1080/01621459.2016.1147356

Cousido-Rocha, M., de Uña Álvarez, J., & Hart, J. D. (2019). A two-sample test for the equality of univariate marginal distributions for high-dimensional data. *Journal of Multivariate Analysis*, *174*, 104537, 20. https://doi.org/10.1016/j.jmva.2019.104537

Donoho, D., & Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, *32*(3), 962–994. https://doi.org/10.1214/009053604000000265

Gaigall, D. (2020). Testing marginal homogeneity of a continuous bivariate distribution with possibly incomplete paired data. *Metrika*, *83*(4), 437–465. https://doi.org/10.1007/s00184-019-00742-5

Hall, P., & Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, *89*(2), 359–374. https://doi.org/10.1093/biomet/89.2.359

Hofert, M., Kojadinovic, I., Maechler, M., Yan, J., Nešlehová, J. G., & Morger, R. (2023). Multivariate dependence with copulas. R package version 1.1-2.

Hušková, M., & Meintanis, S. G. (2008). Tests for the multivariate k-sample problem based on the empirical characteristic function. *Journal of Nonparametric Statistics*, *20*(3), 263–277. https://doi.org/10.1080/10485250801948294

Jiménez-Gamero, M. D. (2023). Testing normality of a large number of populations. *Statistical Papers*. https://doi.org/10.1007/s00362-022-01384-y

Jiménez-Gamero, M. D., Alba-Fernández, M. V., & Ariza-López, F. J. (2019). Approximating the null distribution of a class of statistics for testing independence. *Journal of Computational and Applied Mathematics*, *354*, 131–143. https://doi.org/10.1016/j.cam.2018.03.011

Jiménez-Gamero, M. D., Alba-Fernández, M. V., Jodrá, P., & Barranco-Chamorro, I. (2017). Fast tests for the two-sample problem based on the empirical characteristic function. *Mathematics and Computers in Simulation*, *137*, 390–410. https://doi.org/10.1016/j.matcom.2016.09.007

Jiménez-Gamero, M. D., Cousido-Rocha, M., Alba-Fernández, M. V., & Jiménez-Jiménez, F. (2022). Testing the equality of a large number of populations. *Test*, *31*(1), 1–21. https://doi.org/10.1007/s11749-021-00769-9

Jiménez-Gamero, M. D., & Franco-Pereira, A. M. (2021). Testing the equality of a large number of means of functional data. *Journal of Multivariate Analysis*, *185*, 104778. https://doi.org/10.1016/j.jmva.2021.104778

Kiefer, J. (1959). *K*-sample analogues of the Kolmogorov-Smirnov and Cramér-V. Mises tests. *Annals of Mathematical Statistics*, *30*, 420–447. https://doi.org/10.1214/aoms/1177706261

Kim, I. (2021). Comparing a large number of multivariate distributions. *Bernoulli*, *27*(1), 419–441. https://doi.org/10.3150/20-BEJ1244

Kosorok, M. R. (1999). Two-sample quantile tests under general conditions. *Biometrika*, *86*(4), 909–921. https://doi.org/10.1093/biomet/86.4.909

Liu, L., Meng, Y., Wu, X., Ying, Z., & Zheng, T. (2022). Log-rank-type tests for equality of distributions in high-dimensional spaces. *Journal of Computational and Graphical Statistics*, *31*(4), 1384–1396. https://doi.org/10.1080/10618600.2022.2051530

Liu, Z., & Modarres, R. (2011). A triangle test for equality of distribution functions in high dimensions. *Journal of Nonparametric Statistics*, *23*(3), 605–615. https://doi.org/10.1080/10485252.2010.485644

Martínez-Camblor, P., & de Uña Álvarez, J. (2009). Non-parametric *k*-sample tests: Density functions vs distribution functions. *Computational Statistics & Data Analysis*, *53*(9), 3344–3357. https://doi.org/10.1016/j.csda.2009.02.009

Meintanis, S. G. (2005). Permutation tests for homogeneity based on the empirical characteristic function. *Journal of Nonparametric Statistics*, *17*(5), 583–592. https://doi.org/10.1080/10485250500039494

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

Pardo, L., Pardo, M. C., & Zografos, K. (1999). Homogeneity for multinomial populations based on ϕ-divergences. *Journal of the Japan Statistical Society*, *29*(2), 213–228. https://doi.org/10.14490/jjss1995.29.213

Park, J., & Park, D. (2012). Testing the equality of a large number of normal population means. *Computational Statistics & Data Analysis*, *56*(5), 1131–1149. https://doi.org/10.1016/j.csda.2011.08.017

Quessy, J. F., & Éthier, F. (2012). Cramér-von Mises and characteristic function tests for the two and *k*-sample problems with dependent data. *Computational Statistics & Data Analysis*, *56*(6), 2097–2111. https://doi.org/10.1016/j.csda.2011.12.021

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, Vol. 162. John Wiley & Sons.

Székely, G. J., & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, *143*(8), 1249–1272. https://doi.org/10.1016/j.jspi.2013.03.018

Zhan, D., & Hart, J. D. (2014). Testing equality of a large number of densities. *Biometrika*, *101*(2), 449–464. https://doi.org/10.1093/biomet/asu002

AUTHOR BIOGRAPHIES

M. V. Alba-Fernández holds a PhD in Mathematics from the University of Granada (Granada, Spain). She is a Professor at the Department of Statistics and Operations Research at the University of Jaén (Spain). Her research areas include goodness-of-fit tests and equality of distributions.

M. D. Jiménez-Gamero is a Professor at the Department of Statistics and Operations Research of the University of Sevilla (Sevilla, Spain). She holds master's and PhD degrees in Mathematics from the University of Sevilla. She has over 95 research articles published, and her main primary research interests include comparison of populations procedures, goodness-of-fit tests, specification tests and bootstrap.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Alba-Fernández, M. V., & Jiménez-Gamero, M. D. (2023). Homogeneity of marginal distributions for a large number of populations. *Stat*, *12*(1), e617. https://doi.org/10.1002/sta4.617