TwoSampleTest.HD: An R Package for the Two-Sample Problem with High-Dimensional Data

by Marta Cousido-Rocha and Jacobo de Uña-Álvarez

Abstract The two-sample problem refers to the comparison of two probability distributions via two independent samples. With high-dimensional data, such comparison is performed along a large number p of possibly correlated variables or outcomes. In genomics, for instance, the variables may represent gene expression levels for p locations, recorded for two (usually small) groups of individuals. In this paper we introduce TwoSampleTest.HD, a new R package to test for the equal distribution of the p outcomes. Specifically, TwoSampleTest.HD implements the tests recently proposed by (Cousido-Rocha, Uña-Álvarez, and Hart 2019) for the low sample size, large dimensional setting. These tests take the possible dependence among the p variables into account, and work for sample sizes as small as two. The tests are based on the distance between the empirical characteristic functions of the two samples, when averaged along the p locations. Different options to estimate the variance of the test statistic under dependence are allowed. The package TwoSampleTest.HD provides the user with individual permutation p-values too, so feature discovery is possible when the null hypothesis of equal distribution is rejected. We illustrate the usage of the package through the analysis of simulated and real data, where results provided by alternative approaches are considered for comparison purposes. In particular, benefits of the implemented tests relative to ordinary multiple comparison procedures are highlighted. Practical recommendations are given.

1 Introduction

One of the most important questions in modern statistics is how to efficiently deal with the low sample size, high dimensional setting, in which a large number p of variables are measured for a relatively small number of individuals. This type of high-dimensional data arises in many different areas of science, such as genetics, medicine, pharmacy and social sciences. In microarray data, for example, the variables typically represent the expression levels of a large set of genes. In such a context, a usual goal is to compare the distributions of the gene expression levels for individuals with two different tumour types. Hence, a formal two-sample test in the low sample size and large dimension setting is required. More precisely, the aim is to test for the equality of the p marginal distributions for the two groups. In other words, one can regard the null hypothesis as the intersection of the p null hypotheses corresponding to each of the p locations (genes).

In the majority of examples with high-dimensional data the large number of variables or outcomes are not independent. In genetics, for example, dependency among expression levels of different genes on the same individual is often observed. Several two-sample tests have been developed for the high-dimensional setting under dependence; see for instance Biswas and Gosh (2014), Mondal et al. (2015), Biswas et al. (2014), Liu et al. (2015) and Wei et al. (2016). Nevertheless, all of these proposals have at least one of the following disadvantages: (1) the null hypothesis asserts that the p-variate distribution is the same for the two groups being compared instead of testing the equality of the univariate marginals; (2) the dependence structure is not considered or is too restrictive; (3) the theoretical results are only suitable for normally distributed data. Besides, to the best of our knowledge none of these methods are available in R. While gaps (2) and (3) are limiting in applications, issue (1) is more fundamental; note that usually the focus is more on the marginal outcome distribution than on the within-group correlation structure. Therefore, new ideas are needed.

Recently, Cousido-Rocha et al. (2019) overcame the aforementioned flaws by introducing a non-parametric omnibus test that, with focus on the marginal distributions, included the dependent case through mixing conditions (Doukhan, 1995). This type of dependence, being fairly general, has been frequently used in the goodness-of-fit testing literature; see for example Neumann and Paparoditis (2000) and Dehling et al. (2015). Mixing conditions imply that the dependence between the variables softens at distant locations. In genetics, for instance, this means that the correlation among expression levels of different genes lessens as the distance between the biological function of the genes increases, which is a flexible, realistic assumption for such applications.

In this paper we introduce the **TwoSampleTest.HD** R package which implements the tests proposed in Cousido-Rocha et al. (2019) for testing the (global, or intersection) null hypothesis of equality of the p univariate marginals in the two populations. The basic test statistic is the L_2-distance between the empirical characteristic functions pertaining to the two groups, when averaged along the p locations.

Several approaches to estimate the variance of the test statistic under dependence lead then to slightly different procedures. At the same time, **TwoSampleTest.HD** provides the user with permutation p-values for each location. When the null hypothesis is rejected, these p-values can be used to rank the locations according to their contribution to the global significance, or for feature selection by performing multiple testing (Dudoit and van der Laan, 2007). Finally, a different test statistic for the global null hypothesis based on the average of the permutation p-values is implemented within **TwoSampleTest.HD**. All of these procedures are fully illustrated in this piece of work.

Alternative nonparametric approaches for the two-sample problem include Kolmogorov-Smirnov and Cramér-von Mises tests. These methods compare the empirical distribution functions, rather than the empirical characteristic functions, of the two groups. Empirical characteristic functions are related to smooth tests, which have been found preferable to distribution-based tests in many settings due to their greater power (Martínez-Camblor and de Uña-Álvarez, 2009). More importantly, whenever a test is performed locally and repeatedly, multiple comparison procedures (MCP) are needed in order to keep the type I error under control. Unfortunately, such approach may not be optimal when testing for different distributions in a global way. This relays the fact that feature discovery is more difficult than testing for the intersection null, and explains why the test based on permutation p-values implemented in **TwoSampleTest.HD** can exhibit a power lower than that of the averaged L_2-type tests within the package. Efforts to efficiently summarize local p-values for testing an intersection null hypothesis include the so-called Higher Criticism (HC) approach, see Zhang et al. (2020) and references therein; however, the performance of HC may be dramatically affected by dependence (Hall and Jin, 2008). Further discussion of these issues is provided within this paper on the basis of empirical results.

The rest of the paper is organized as follows. In Section 2.2 the methodological background is introduced, and the tests proposed by Cousido-Rocha et al. (2019) are presented in detail. In Section 2.3 the **TwoSampleTest.HD** package is described, and its usage is illustrated through the analysis of simulated data and microarray data derived from a hereditary breast cancer study. Finally, Section 2.4 reports the main conclusions of this work.

2 Methodology

In this section we describe the four two-sample tests proposed by Cousido-Rocha et al. (2019), which are implemented in the **TwoSampleTest.HD** package. Three of the methods are based on the average of p individual L_2-distances between the empirical characteristic functions computed from the two samples. The three versions of this test differ in the way in which the variance is estimated; in all the cases, the variance estimate takes the possible dependence among the p outcomes into account. The fourth method is based on the average of the permutation p-values derived for the individual L_2-distances.

We consider two random matrices $X = \left[X_1, \ldots, X_p\right]^T$ and $Y = \left[Y_1, \ldots, Y_p\right]^T$ of respective dimensions $p \times n$ and $p \times m$, where $X_k = (X_{k1}, \ldots, X_{kn})$ and $Y_k = (Y_{k1}, \ldots, Y_{km})$, $k = 1, \ldots, p$, are the sample values for the p target variables; the sample sizes are n and m. The variables X_k and Y_k may be discrete or continuous; normality is not assumed in the continuous case. Given sequences of characteristic functions $\{C_{X_1}, C_{X_2}, \ldots, C_{X_p}\}$ and $\{C_{Y_1}, C_{Y_2}, \ldots, C_{Y_p}\}$, it is assumed that $X_{k1}, \ldots, X_{kn}$ and $Y_{k1}, \ldots, Y_{km}$ are independent random samples from C_{X_k} and C_{Y_k}, respectively. Observations X_{ij} and X_{sl} for $s \neq i$ can be dependent in our framework, and similar comments apply to the components of Y.

The focus is in testing for the intersection null hypothesis

$$H_0 = \bigcap_{k=1}^{p} H_{0k},$$

where, for $1 \leq k \leq p$, H_{0k} states that C_{X_k} and C_{Y_k} coincide. As indicated by Cousido-Rocha et al. (2019), the L_2-distance between the empirical characteristic functions of X_k and Y_k is given by

$$\begin{aligned} J_k &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l=1, l\neq j}^{n} \exp\left(-\frac{1}{2}\left(\frac{X_{kj} - X_{kl}}{\sqrt{2}b}\right)^2\right) \\ &+ \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{l=1, l\neq j}^{m} \exp\left(-\frac{1}{2}\left(\frac{Y_{kj} - Y_{kl}}{\sqrt{2}b}\right)^2\right) \\ &- \frac{2}{nm} \sum_{j=1}^{n} \sum_{l=1}^{m} \exp\left(-\frac{1}{2}\left(\frac{X_{kj} - Y_{kl}}{\sqrt{2}b}\right)^2\right), \end{aligned} \tag{1}$$

where $b \in \mathbb{R}^+$. The first term in J_k is an intra-sample parameter estimate for the sample X_k and the second term is an intra-sample parameter estimate for the sample Y_k, whereas the third term is an *inter*-samples parameter estimate, since X_{kj} and $Y_{k\ell}$ come from different samples, X_k and Y_k, respectively.

We consider the test statistic $T_p = \sum_{k=1}^{p} J_k/\sqrt{p}$ in order to test for H_0. Cousido-Rocha et al. (2019) introduced two variance estimators for T_p when $(X_k, Y_k)_{k\in\{1,\ldots,p\}}$ comes from a strictly stationary sequence $(X_k, Y_k)_{k\in\mathbb{N}}$. The first variance estimator $\widehat{\sigma}_S$ is based on the spectral density estimate. Indeed, the problem of estimating the variance of a sample mean based on dependent data is the same as that of estimating the spectrum of the process at frequency zero. Hence, $\mathrm{Var}(T_p)$ can be approximated through the classical estimator of the spectrum of the process at frequency zero. It holds that $T_p/\widehat{\sigma}_S \xrightarrow{D} \mathcal{N}(0,1)$ as p tends to ∞. Since the expectation of T_p is strictly positive under the alternative hypothesis, the test is one-sided, and rejects H_0 at nominal level α when $T_p/\widehat{\sigma}_S$ is larger than the $1-\alpha$ quantile of the standard normal distribution. We refer to this test as spectral test.

The second variance estimator is derived from the block bootstrap procedure proposed by Carlstein (1996) to estimate the variance of a general statistic computed from a strictly stationary α-mixing sequence (see Bosq, 1998). More precisely, this resampling method defines no overlapping blocks of length l on the sequence $(J_k)_{k\in\{1,\ldots,p\}}$ and computes the statistic T_p in each of the blocks, with l being the maximum lag of significant autocorrelation on the $(J_k)_{k\in\{1,\ldots,p\}}$ sequence according to the procedure in Politis and White (2004). Finally the block bootstrap variance estimator $\widehat{\sigma}_B$ is simply defined as the sample variance of the values of T_p in each of the block. As before, the standardized version of T_p based on $\widehat{\sigma}_B$, is asymptotically distributed as a $\mathcal{N}(0,1)$ random variate. Hence, the block bootstrap test rejects the null hypothesis when $T_p/\widehat{\sigma}_B$ is larger than the $1-\alpha$ quantile of the standard normal distribution.

Cousido-Rocha et al. (2019) also introduced an alternative variance estimator suitable for possibly non-stationary sequences based on U-statistics theory. More precisely, the variance of T_p can be written as the sum of two terms; the first one is the average of the variance of the statistics $(J_k)_{k\in\{1,\ldots,p\}}$, whereas the second one comprises the covariance terms arising from such sequence. The first term can be estimated by replacing the unknown theoretical expectations by their corresponding sample means; on the other hand, an unbiased estimator for the covariance term $Cov(J_j, J_{j+k})$ is given by $J_j J_{j+k}$, $j = 1,\ldots,p-l$, $k = 1,\ldots,l$. The standardized version of T_p based on such variance estimator, $T_p/\widehat{\sigma}_U$, is asymptotically distributed as a $\mathcal{N}(0,1)$ as $p \to \infty$. Hence, the U-statistic test rejects the null hypothesis when $T_p/\widehat{\sigma}_U$ is larger than the $1-\alpha$ quantile of the standard normal distribution.

Interestingly, the test statistics J_k can be used locally to test for the null hypotheses H_{0k} that X_{k1} and Y_{k1} have the same distribution, $1 \le k \le p$. Since the common density under H_{0k} is unknown, a permutation test can be used to calibrate the null distribution of J_k. The application of the permutation test to each of the H_{0k}, $k \in \{1,\ldots,p\}$ yields a set of p-values $\{P_1,\ldots,P_p\}$. Cousido-Rocha et al. (2019) proposed a test statistic based on the average of the permutation p-values, $\bar{P} = \sum_{k=1}^{p} P_k/p$. The idea of combining the individual p-values to obtain a global test statistic is old, dating to Fisher (1934); Stouffer et al. (1949) and others. The statistic P is standardized taking into account that its expectation is $(N+1)/2N$, with N being the number of permutations that lead to a different value of the statistic J_k, and using a variance estimator based on the spectral analysis. The standardized version of $\bar{P}$, say $T_p^{pv} = \sqrt{p}\,(\bar{P} - (N+1)/(2N))/\widehat{\sigma}_P$, is asymptotically distributed as a standard normal as $p \to \infty$. It is assumed that, under H_0, $(X_k, Y_k)_{k\in\{1,\ldots,p\}}$ comes from a strictly stationary and strongly mixing process $(X_k, Y_k)_{k\in\mathbb{N}}$. The null hypothesis is rejected when T_p^{pv} is smaller than the α quantile of the standard normal distribution. One advantage of the permutation p-values is that, when the intersection null is rejected, they can be used to rank the null hypotheses H_{0k} according to their contribution to the significance. In genomics, for instance, this ranking may reveal the genes which express differently between two tumors. Finally, a formal MCP can be applied to the set of permutation p-values to get rigorous conclusions on the individual nulls H_{0k}, $k \in \{1,\ldots,p\}$. Note that the ranking provided by the p-values is related, but not equal, to the ranking based on the J_k's; this is because the target outcome may be differently distributed along the p locations, and J_k is not distribution free. The situation for the Hedenfalk data example in Section 2.3.1 is depicted in Figure 1; the shift of the p-values distribution compared to uniform suggests that some genes are differently expressed in the two groups considered.

For the implementation of the aforementioned test statistics the parameter b in (1), which plays the role of a smoothing parameter or bandwidth, is set to $\widehat{b} = 1.144 s_{pool}\,((n+m)/2)^{-1/5}$, where s^2_{pool} is the average of $\left((n-1)s^2_{X_k} + (m-1)s^2_{Y_k})\right)/(n+m-2)$, $k = 1,\ldots,p$, and $s^2_{X_k}$ and $s^2_{Y_k}$ are the sample variances of X_k and Y_k, respectively, $k = 1,\ldots,p$. When the permutation p-values of the statistics J_k are to be computed, a local bandwidth can be used instead; specifically, the local bandwidth for J_k is given by $\widehat{b}_k = 1.144 s_{pool}\,((n+m)/2)^{-1/5}$, with $s^2_{pool} = \left((n-1)s^2_{X_k} + (m-1)s^2_{Y_k})\right)/(n+m-2)$, and $s^2_{X_k}$ and $s^2_{Y_k}$ are the sample variances of X_k and Y_k.

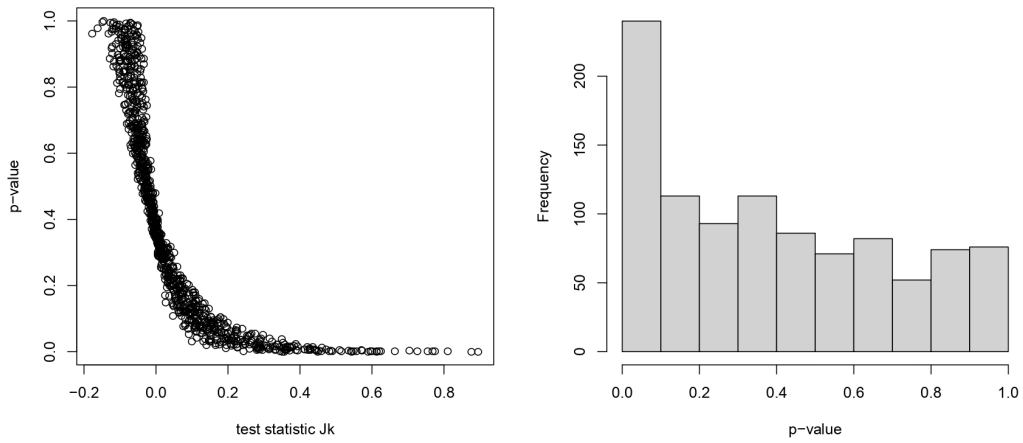

Figure 1: Left: p-values vs test statistics J_k. Right: histogram of the p-values. Hedenfalk data.

	Spectral			Block bootstrap		
$p/n, m$	$2, 2$	$5, 5$	$10, 10$	$2, 2$	$5, 5$	$10, 10$
100	0.01	0.02	0.01	0.01	0.02	0.03
500	0.03	0.03	0.06	0.03	0.05	0.10
1000	0.07	0.08	0.11	0.07	0.07	0.12
	U-statistic			Permutation		
$p/n, m$	$2, 2$	$5, 5$	$10, 10$	$2, 2$	$5, 5$	$10, 10$
100	0.01	0.11	0.58	0.03	0.89	1182.40
500	0.11	0.57	2.51	0.015	3.55	5122.53
1000	0.24	1.04	4.87	0.27	8.13	8006.70

Table 1: Execution time (in seconds) for the several versions of the proposed two-sample test. Simulated data with dimension p and sample sizes n and m.

3 Package TwoSampleTest.HD in practice

In this section the main features of **TwoSampleTest.HD** package are described. We also consider two examples with high-dimensional data in order to explain how to use **TwoSampleTest.HD** in practice. The first example refers a large number of gene expression levels measured on two groups of patients with breast cancer, classified according to BRCA mutation type. The second example is a simulation scenario in which the target outcome is differently distributed in the two groups for 10% of the p locations. This second example serves in particular to illustrate the smaller power of ordinary MCP when compared to the tests based on the averaged L_2-distances between the empirical characteristic functions of the two groups.

3.1 Hedenfalk data

In this subsection we consider the microarray data set of hereditary breast cancer in Hedenfalk et al. (2001). The data set consists of $p = 3226$ logged gene expression levels measured on $n = 7$ patients with breast tumors having BRCA1 mutations and on $m = 8$ patients with breast tumors having BRCA2 mutations. The goal is to test the null hypothesis that the distribution of the p genes is the same for the two types of tumor, BRCA1 tumor and BRCA2 tumor. Since the example is merely illustrative, we only consider the first 1000 genes in order to save computational time. With 1000 locations, the execution time is reduced to < 1 second for the block bootstrap and spectral tests, to < 5 seconds for the U-statistic test and to 9 minutes for the permutation p-values test, in a laptop provided with a i5-1135G7 CPU. The waiting time of the permutation test is relatively long since $n = 7$ and $m = 8$ lead to 6435 permutations which must be carried out for each of the $p = 1000$ genes. For additional inspection, in Table 1 execution times for simulated data with several dimensions p and sample sizes n and m are reported.

The main function of the package is **TwoSampleTest.HD**. This function computes, among other things, the value of the selected test statistic and the corresponding p-value. The list of arguments of **TwoSampleTest.HD** is given in Table 2. The required arguments are `X` and `Y`, matrices where each row is one of the p-samples in the first group and second group, respectively; the other arguments have a default value. When the user forgets to include the argument `X` or `Y` in the function, the following message is returned:

```
Call:
TwoSampleTest.HD(X = X)
'us' method used by default
'global' bandwidth used by default
Error in ncol(Y) : argument "Y" is missing, with no default

Call:
TwoSampleTest.HD(Y= Y)
'us' method used by default
'global' bandwidth used by default
Error in ncol(X) : argument "X" is missing, with no default
```

With `method="spect"` the two-sample spectral test described in Section 2.2 is applied; the option `"method=spect_ind"` corresponds to a simplified version that pre-assumes the independence among the outcomes. On the other hand, the options `method="us"` and `method="us_ind"` apply the two-sample U-statistic test explained in Section 2.2 for dependent data and its simplification for independent variables, respectively. The last option based on the average of the p individual statistics, J_k, corresponding to each of the p variables is `method="boot"` which implements the two-sample block bootstrap test (Section 2.2). Finally, the function also performs the alternative test based on the average of the permutation p-values corresponding to the individual statistics J_k through the argument `method="perm"`. In our experience, the most powerful test for independent outcomes is the U-statistic test, whereas under dependence the more powerful tests are the spectral and block bootstrap tests. We also observed that the block bootstrap and spectral tests, which were developed assuming that $(X_k, Y_k)_{k\in\mathbb{N}}$ is a strictly stationary process, performed well when stationarity is violated. The `"us"` method has been defined as the default one.

When choosing `method="perm"`, the sequence of permutation p-values is computed and reported. On the other hand, the computation of the permutation p-values must be explicitly requested using `I.permutation.p.values=TRUE` argument for the U-statistic, spectral or block bootstrap tests. As mentioned, these individual p-values may be used to rank the outcomes according to their significance. Argument `b_I.permutation.p.values` allows the user to select the bandwidth b. The option `b_I.permutation. p.values="global"` computes a global bandwidth $\widehat{b}$ and uses it to evaluate the J_k's, whereas option `b_I.permutation.p.values="individual"` estimates the bandwidth for each variable separately; see details in Section 2.2. The default option is `b_I.permutation.p.values="global"`. In Table 3 a summary of the results provided by the function **TwoSampleTest.HD** is given. The `I.statistics` object contains the individual statistics J_k , $k = 1, \ldots, p$ described in the previous section, while the `I.permutation.p.values` object reports the permutation p-values $\{P_1, \ldots, P_p\}$.

Hedenfalk data are available within **Equalden.HD** package (Cousido-Rocha and de Uña-Álvarez, 2022). In order to analyze this dataset, we load this package together with **TwoSampleTest.HD** package. For the investigation of the possible dependence among the gene expression levels, we treat the data of each patient as a time series, and we compute the sample autocorrelation function. For the first lags the autocorrelation between genes was significantly different from zero, whereas it lessened as the number of lags increased. The estimates of the autocorrelation were computed using the `acf` function of the R package **stats**. On the basis of these results, the weak dependence assumption behind the tests implemented in the **TwoSampleTest.HD** seems realistic. The four tests designed for weak dependence (spectral test, U-statistic test, block bootstrap test and permutation test) can be performed by using the following code lines:

```
> library(Equalden.HD)
> data("Hedenfalk")
> X=log(Hedenfalk[,1:7])
> Y=log(Hedenfalk[,8:15])
>
> X=X[1:1000,]
> Y=Y[1:1000,]
> library(TwoSampleTest.HD)
> res1 <- TwoSampleTest.HD(X, Y, method = "spect")
Call:
```

Usage of the function:

```
TwoSampleTest.HD(X, Y, method = c("spect",
"spect_ind", "boot", "us","us_ind", "perm"),
I.permutation.p.values = FALSE,
b_I.permutation.p.values = c("global", "individual"))
```

`X`	A matrix where each row is one of the p-samples in the first group.
`Y`	A matrix where each row is one of the p-samples in the second group.
`method`	The two-sample test. By default the "us" method is computed.
`I.permutation.p.values`	Logical. Default is FALSE. A variable indicating whether to compute the permutation p-values or not when the selected method is not "perm".
`b_I.permutation.p.values`	The bandwidth method used to compute the individual statistics on which are based the permutation p-values.

Table 2: Usage and list of the arguments of the **TwoSampleTest.HD** function.

`standardized statistic`	the value of the standardized statistic.
`p.value`	the p-value for the test.
`statistic`	the value of the statistic.
`variance`	the value of the variance estimator.
`p`	number of samples or populations.
`n`	sample size in the first group.
`m`	sample size in the second group.
`method`	a character string indicating which two sample test is performed.
`I.statistics`	the p individual statistics.
`I.permutation.p.values`	the p individual permutation p-values.
`data.name`	a character string giving the name of the data.

Table 3: Summary of the results reported by **TwoSampleTest.HD** function.

```
TwoSampleTest.HD(X = X, Y = Y, method = "spect")
'global' bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = 11.536, p-value < 2.2e-16

> res2 <- TwoSampleTest.HD(X, Y, method = "boot")
Call:
TwoSampleTest.HD(X = X, Y = Y, method = "boot")
'global' bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = 11.515, p-value < 2.2e-16

> res3 <- TwoSampleTest.HD(X, Y, method = "us")
Call:
```

```
TwoSampleTest.HD(X = X, Y = Y, method = "us")
'global' bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = 12.104, p-value < 2.2e-16

> res4 <- TwoSampleTest.HD(X, Y, method = "perm")
Call:
TwoSampleTest.HD(X = X, Y = Y, method = "perm")
'global'bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = -10.955, p-value < 2.2e-16
```

The output of the function **TwoSampleTest.HD** shows that $T_p/\widehat{\sigma}_S = 11.536$, $T_p/\widehat{\sigma}_B = 11.515$, $T_p/\widehat{\sigma}_U = 12.104$ and $T_p^{pv} = -10.955$, whereas the corresponding p-values are almost zero. The negative value of the T_p^{pv} statistic means that the average of the permutation p-values is lower than the expected mean of their (uniform) null distribution. Hence, the null hypothesis is rejected; the conclusion is that one or more genes are differently expressed depending on the tumor type. The object derived from **TwoSampleTest.HD** function is a list which saves, as usual with `R` functions, relevant information. Besides the standardized statistic and the p-value printed in the console when running the function (as shown above), the list of saved objects comprises the value of the statistic T_p (or $\sqrt{p}\bar{P}$ if one runs the permutation test), the variance ($\widehat{\sigma}_S$, $\widehat{\sigma}_B$, $\widehat{\sigma}_U$ or $\widehat{\sigma}_P$), the number of variables (p), the sample sizes (n and m), the method used for the data analysis, and the values of the statistics $J_1, \ldots, J_p$. Below, we display such information for the spectral test as an illustrative example. The values of the statistics $J_1, \ldots, J_p$ are plotted in Figure 2 instead of reporting them through the console.

```
> res1\$statistic
[1] 1.827471
> res1\$variance
[1] 0.02509699
> res1\$p
[1] 1000
> res1\$n
[1] 7
> res1\$m
[1] 8
> res1\$method
[1] "spect"
> library(ggplot2)
>
> data=data.frame(Jk=res1\$I.statistics,Genes=1:res1\$p)
> ggplot(data, aes(x=Genes, y=Jk)) +
+   geom_point(shape=21, col=8) + geom_rug()+ggtitle("Individual test statistics")
```

Since the null hypothesis is rejected for Hedenfalk data, the next natural aim is to identify which genes are not equally distributed in both types of tumors. For this, we first rank the null hypotheses H_{0i} according to their contribution to the significance by using the sequence of permutation p-values. This sequence has only been computed for the permutation test, since for the remaining tests it is only computed when the argument `I.permutation.p.values` is equal to `TRUE` and, in our previous applications of the tests such argument has not been specified hence the default option `I.permutation.p.values=FALSE` has been used. Therefore, `res4` is the unique object which has the sequence of permutation p-values. Note that, since the argument `b_I.permutation.p.values` has not been used, the default option `b_I.permutation.p.values="global"` has been considered, and then the global $\widehat{b}$ has been employed to compute each one of the J_k, $k = 1, \ldots, p$, for which the permutation p-values are calculated. Below, the code used to determine which are the 10 genes of lowest p-values is reported.

```
> pv=res4\$I.permutation.p.values
> order(pv)[1:10]
```

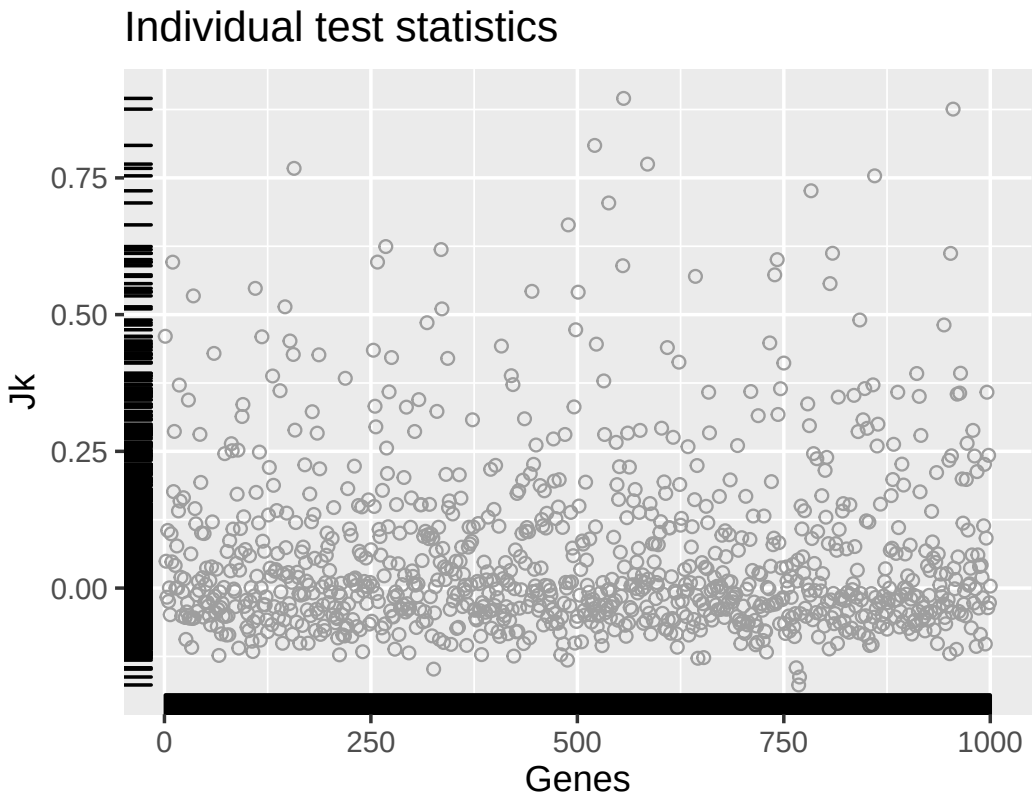

Figure 2: The values of the statistics $J_1, \ldots, J_p$ for the two-sample spectral test. Hedenfalk data.

```
[1] 556 733 952 955 445 555 914 963 118 157
```

Although the above list can be informative, any rigorous procedure should keep the type I error under control. The `p.adjust` function, available within **stats** package, implements the well-known Benjamini and Hochberg (1995) false discovery rate (FDR) controlling procedure. The application of this method to the sequence of 1000 permutation p-values at 5% FDR level reports 13 discoveries (see code lines below). Note that, although Benjamini and Hochberg (1995) has been initially studied for independent p-values, subsequent research has shown that it remains valid under weaker assumptions.

```
> alpha=0.05
> sum(p.adjust(pv,method = "BH")<=alpha)
[1] 13
```

One interesting question is whether nonparametric two-sample test statistics alternative to J_k could perform better in the multiple testing setting. As a by-product of their research, Cousido-Rocha et al. (2021) proved through simulations that the J_k test statistic performs similarly or even better than other well-known two-sample tests. For example, simulation results in the referred paper suggest that the Kolmogorov–Smirnov test should not be used when the sample sizes are small and the differences are other than location. For illustrative purposes, we have tested each one of the null hypothesis H_{0k}, $k \in \{1, \ldots, p\}$, through Student's t test, Wilcoxon test, Levene test and Kolmogorov- Smirnov test (see Gibbons and Chakraborti, 1992 and Levene, 1960). Then, Benjamini and Hochberg (1995) has been applied to the corresponding p-values sequence (code below).

```
> p=res1\$p;n=res1\$n; m=res1\$m
> pv_t.test=1:p
> pv_KS=1:p
> pv_Wilcoxon=1:p
> pv_Levene=1:p
>
> library(car)
> library(exactRankTests)
> library(coin)
```

```
>
> for (i in 1:p){
        +   pv_Wilcoxon[i]=wilcox.exact(X[i,],Y[i,])\$p.value
        +   pv_t.test[i]=t.test(X[i,],Y[i,],var.equal = F)\$p.value
        +   pv_KS[i]=ks.test(X[i,],Y[i,])\$p.value
        +   pv_Levene[i]=leveneTest(c(X[i,],Y[i,]),
        +   as.factor(c(rep(1,n),rep(2,m))))\$`Pr(>F)`[1]
        + }
> sum(p.adjust(pv_Wilcoxon,method = "BH")<=alpha)
[1] 13
> sum(p.adjust(pv_t.test,method = "BH")<=alpha)
[1] 1
> sum(p.adjust(pv_KS,method = "BH")<=alpha)
[1] 0
> sum(p.adjust(pv_Levene,method = "BH")<=alpha)
[1] 0
```

From results above it is seen that the Kolmogorov–Smirnov (KS) test is unable to provide any discovery at 5% of FDR. However, these results should be taken with some caution since exact KS p-values could not be computed by `ks.test` function due to the presence of ties in Hedenfalk data; note that the asymptotic distribution of the KS test may be inaccurate for small sample sizes. The lack of power of KS in the multiple testing setting has been pointed out in Cousido-Rocha et al. (2021) too. Similarly as for KS, Levene test does not declare any gene as differently expressed in the two tumor groups; this is not surprising, since differences between the two groups are mainly due to a location shift (Hedenfalk et al., 2001). The number of discoveries of the t-test is very low (only one rejection); on the contrary, Wilcoxon test provides as many discoveries as the J_k test statistic. In order to better summarize the relative power of the several testing procedures, Figure 3 depicts the number of rejections along a sequence of nominal levels for the FDR ($\alpha = 0.001, 0.002, \ldots, 0.10$). Interestingly, Figure 3 supports previous comments on the poor performance of KS test.

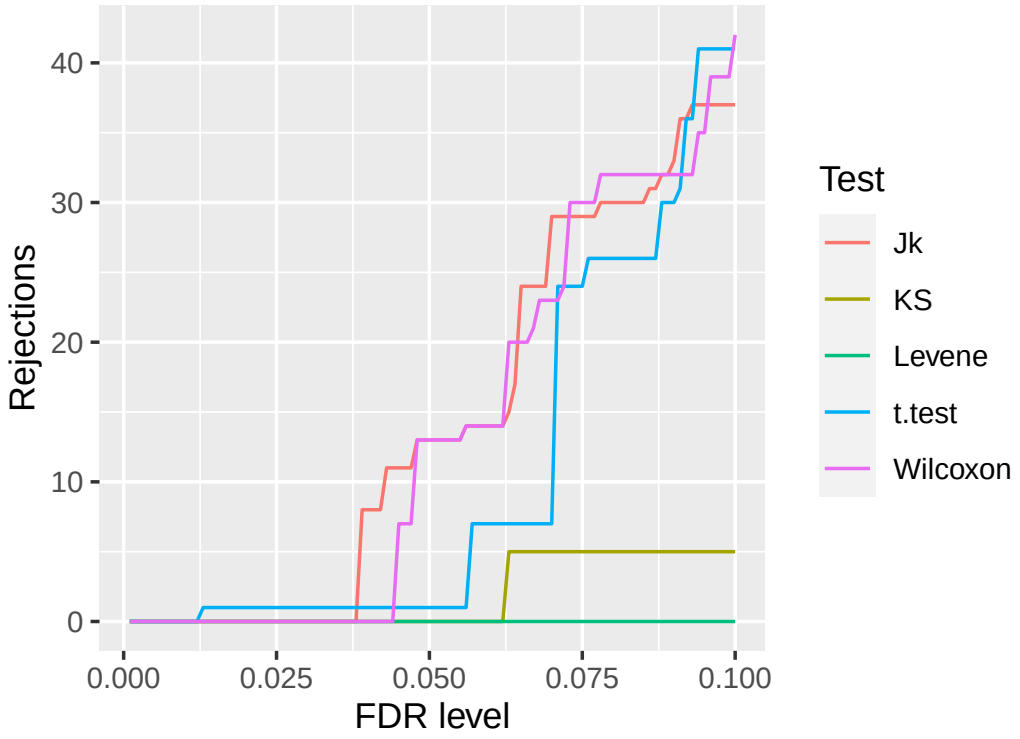

Figure 3: Number of rejections of Wilcoxon test, Kolmogorov-Smirnov test, t-test, Levene test and J_k permutation test depending on the nominal FDR level. Hedenfalk data.

3.2 Simulated data

We simulated $p = 1000$ independent variables with sample sizes $n = m = 4$ under the alternative hypothesis. More precisely, the p samples in the first group (X) were generated in 4 blocks from the following distributions, respectively: $N(0,1)$, $N(0,2)$, $N(1,1)$ and $N(1,2)$. In the second group (Y), 90% of the p samples were generated exactly as for X (true individual nulls), whereas for simulating the remaining 10% of the samples the distributions were interchanged, with a location shift as result (non-true individual nulls). To be specific, in the case $X \sim N(0,1)$, the Y was generated from a $N(1,1)$, and vice versa; when $X \sim N(0,2)$, the Y was generated from a $N(1,2)$, and vice versa. The code for the simulation is provided in Appendix .1.

Below, the two-sample tests implemented in **TwoSampleTest.HD** are applied to test the null hypothesis that the distribution of each of the samples is the same in the groups. All of the tests reject the null hypothesis. The results suggest that the simpler versions which make use of the independence assumption, `"spect_ind"` and `"us_ind"`, are slightly more powerful than their counterparts for dependent data.

```
> TwoSampleTest.HD(X, Y, method = "spect")\$p.value
Call:
TwoSampleTest.HD(X = X, Y = Y, method = "spect")
'global' bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = 2.2275, p-value = 0.01296

[1] 0.01295652
> TwoSampleTest.HD(X, Y, method = "spect_ind")\$p.value
Call:
TwoSampleTest.HD(X = X, Y = Y, method = "spect_ind")
'global' bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = 2.2821, p-value = 0.01124

[1] 0.01124119
> TwoSampleTest.HD(X, Y, method = "boot")\$p.value
Call:
TwoSampleTest.HD(X = X, Y = Y, method = "boot")
'global' bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = 2.2643, p-value = 0.01178

[1] 0.01177765
> TwoSampleTest.HD(X, Y, method = "us")\$p.value
Call:
TwoSampleTest.HD(X = X, Y = Y, method = "us")
'global' bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = 2.3058, p-value = 0.01056

[1] 0.01056058
> TwoSampleTest.HD(X, Y, method = "us_ind")\$p.value
Call:
TwoSampleTest.HD(X = X, Y = Y, method = "us_ind")
'global' bandwidth used by default
```

```
A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = 2.423, p-value = 0.007696

[1] 0.007695904
> res=TwoSampleTest.HD(X, Y, method = "perm")
Call:
TwoSampleTest.HD(X = X, Y = Y, method = "perm")
'global' bandwidth used by default

A two-sample test for the equality of distributions for high-dimensional data

data:  c(X, Y)
standardized statistic = -2.2287, p-value = 0.01292
```

As done for Hedenfalk data, one can individually test for H_{0k}, $k \in \{1, \ldots, p\}$, at 5% level of FDR by using the J_k statistic. In this case, the number of rejections is zero. The same occurs when using the Wilcoxon test, the Kolmogorov-Smirnov test, the t-test, or the Levene test (see code lines below). This highlights once again the need for global two-sample tests as those implemented in **TwoSampleTest.HD**.

```
> pvalues=res$I.permutation.p.values
>
> alpha=0.05
>
> sum(p.adjust(pvalues,method = "BH")<=alpha)
[1] 0
>
>
>
> pv_t.test=1:p
> pv_KS=1:p
> pv_Wilcoxon=1:p
> pv_Levene=1:p
>
> library(car)
> library(exactRankTests)
> library(coin)
>
> for (i in 1:p){
        +   pv_Wilcoxon[i]=wilcox.exact(X[i,],Y[i,])\$p.value
        +   pv_t.test[i]=t.test(X[i,],Y[i,],var.equal = F)\$p.value
        +   pv_KS[i]=ks.test(X[i,],Y[i,])\$p.value
        +   pv_Levene[i]=leveneTest(c(X[i,],Y[i,]),
        +   as.factor(c(rep(1,n),rep(2,m))))\$`Pr(>F)`[1]
        + }
>
> sum(p.adjust(pv_Wilcoxon,method = "BH")<=alpha)
[1] 0
> sum(p.adjust(pv_t.test,method = "BH")<=alpha)
[1] 0
> sum(p.adjust(pv_KS,method = "BH")<=alpha)
[1] 0
> sum(p.adjust(pv_Levene,method = "BH")<=alpha)
[1] 0
```

An interesting question is the necessary computation time for **TwoSampleTest.HD**. For the simulated example, "spect" and "spect_ind" methods run in 0.16 and 0.15 seconds, respectively; "boot" method in 0.15 seconds, "us" and "us_ind" methods in 2.91 and 2.86 seconds, respectively; and, finally, the "perm" needed 4.55 seconds for running the analysis. The results shown in the current example match the general performance of the main function within the package; the spectral and block bootstrap tests are the most efficient from a computational point of view, followed by the U-statistic test and finally by the permutation test. As we increased the number of variables or (more

critically) the sample sizes, these differences in computational efficiency became more evident. On the other hand, the simplified versions for independent data did not result in a visible reduction of the run time.

4 Conclusions

Package **TwoSampleTest.HD** implements a two-sample test for the null hypothesis that all the marginal distributions of the p-variate outcome of interest coincide on the two groups. The two-sample test takes advantage of the large p, in the sense that it uses a null Gaussian distribution that holds as p goes to infinity; interestingly however, in our experience the asymptotic approximation is also correct for p as small as 20. On the other hand, the implemented test statistic is just an average of the L_2-type deviations between the empirical characteristic functions pertaining to the two samples along the p margins. Each of these p deviations can be used to perform a local two-sample test through the preliminary computation of permutation p-values. These permutation p-values can be used to introduce an alternative testing procedure (also implemented in **TwoSampleTest.HD**), by using the asymptotic null Gaussian distribution of their average as p grows. An interesting question here is if this sequence of p-values can be used in another fashion to introduce a more powerful testing method. In principle, standard multiple comparison procedures are not competitive, since they focus (not only on the intersection null but also) on identifying the margins in which the two groups differ. However, some multiple comparison procedures have been specifically designed to test for the intersection null, and these methods could be competitive in our setting. This is an interesting open question at the time of writing.

Summarizing, **TwoSampleTest.HD** package implements for the first time omnibus two-sample tests for the high-dimensional setting under dependence. The package is user-friendly, and it is hoped that it will serve the scientific community by providing a simple and powerful tool for the analysis of high-dimensional data. Clear advice for a correct use of the package and fully illustrative examples have been given.

Acknowledgements

The authors acknowledge financial support from the Grant PID2020-118101GB-I00, Ministerio de Ciencia e Innovación.

1 Appendix: Simulated data set code

The code employed for generating the described simulated data set in Section 2.3.2 can be found below.

```
> n=m=4
> p=1000
>
> set.seed(123)
>
> p <- 1000
> n = m = 4
> inds <- sample(1:4, p, replace = TRUE)
> X <- matrix(rep(0, n * p), ncol = n)
> for (j in 1:p){
        +   if (inds[j] == 1){
                +     X[j, ] <- rnorm(n)
                +   }
        +   if (inds[j] == 2){
                +     X[j, ] <- rnorm(n, sd = 2)
                +   }
        +   if (inds[j] == 3){
                +     X[j, ] <- rnorm(n, mean = 1)
                +   }
        +   if (inds[j] == 4){
                +     X[j, ] <- rnorm(n, mean = 1, sd = 2)
                +   }
        + }
```

```
> rho <-  0.1
> ind <- sample(1:p, rho * p)
> li <- length(ind)
> indsy <- inds
> for (l in 1:li){
        +   if (indsy[ind[l]]==1){
                +     indsy[ind[l]]=3
                +   } else {
                +     if (indsy[ind[l]]==2){
                        +       indsy[ind[l]]=4
                        +     } else {
                        +       if (indsy[ind[l]]==3){
                                +         indsy[ind[l]]=1
                                +       } else {
                                +         indsy[ind[l]] = 2
                                +       }
                        +     }
                +   }
        + }
> Y <- matrix(rep(0, m * p), ncol = m)
> for (j in 1:p){
        +   if (indsy[j] == 1){
                +     Y[j,] <- rnorm(m)}
        +   if (indsy[j] == 2){
                +     Y[j, ] <- rnorm(m, sd = 2)
                +   }
        +   if (indsy[j]==3){
                +     Y[j, ] <- rnorm(m, mean = 1)
                +   }
        +   if (indsy[j] == 4){
                +     Y[j,] <- rnorm(m, mean = 1, sd = 2)
                +   }
        + }
```

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995. [p86]

M. Biswas and A. K. Gosh. A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, 123:160–171, 2014. [p79]

M. Biswas, M. Mukhopadhyay, and A. K. Ghosh. A distribution-free two-sample run tests applicable to high-dimensional data. *Biometrika*, 101:913–926, 2014. [p79]

D. Bosq. *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction. Second Edition*. Springer-Verlag, New York, 1998. [p81]

E. Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics*, 4:1171–1179, 1996. [p81]

M. Cousido-Rocha and J. de Uña-Álvarez. `Equalden.HD`: An `R` package for testing the equality of a high dimensional set of densities. *Computer Methods and Programs in Biomedicine*, 217:106694, 2022. doi: https://doi.org/10.1016/j.cmpb.2022.106694. [p83]

M. Cousido-Rocha, J. de Uña-Álvarez, and J. Hart. A two-sample test for the equality of distributions for high-dimensional data. *Journal of Multivariate Analysis*, 174:104537, 2019. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2019.104537. URL https://www.sciencedirect.com/science/article/pii/S0047259X19300521. [p79, 80, 81]

M. Cousido-Rocha, J. de Uña-Álvarez, and S. Döhler. Multiple comparison procedures for discrete uniform and homogeneous tests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71: 219–243, 2021. doi: https://doi.org/10.1111/rssc.12529. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12529. [p86, 87]

H. Dehling, R. Fried, I. Garcia, and M. Wendler. *Change-point Detection Under Dependence Based on Two-Sample U-Statistics*. In: Dawson D., R. Kulik, M. Ould Haye, B. Szyszkowicz, Y. Zhao (eds) Asymptotic Laws and Methods in Stochastics. Fields Institute Communications, vol 76. Springer, New York, NY. 2015. [p79]

P. Doukhan. *Mixing: Properties and Examples*. Springer-Verlag, New York, 1995. [p79]

S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures and Applications to Genomics*. Spinger, New York, 2007. [p80]

R. A. Fisher. *Statistical Methods for Research Workers. Fourth Edition*. Oliver and Boyd, Edinburgh, 1934. [p81]

J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference. Third Edition*. Marcel Dekker, Inc, New York, 1992. [p86]

P. Hall and J. Jin. Properties of higher criticism under strong dependence. *The Annals of Statistics*, 36: 381–402, 2008. [p80]

I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. Kallioniemi, B. Wilfond, A. Borg, J. Trent, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, G. Gruvberger, N. Loman, O. Johannsson, H. Olsson, and G. Sauter. Gene-Expression Profiles in Hereditary Breast Cancer. *New England Journal of Medicine*, 344:539–548, 2001. [p82, 87]

H. Levene. Robust tests for equality of variances. In I. Olkin, editor, *Contributions to Probability and Statistics*, pages 278–92. Stanford University Press, Palo Alto, Calif., 1960. [p86]

Z. Liu, X. Xia, and W. Zhou. A test for equality of two distributions via jackknife empirical likelihood and characteristic functions. *Computational Statistics and Data Analysis*, 92:97–114, 2015. [p79]

P. Martínez-Camblor and J. de Uña-Álvarez. Nonparametric k-sample tests: Density functions vs distribution functions. *Computational Statistics and Data Analysis*, 53:3344–3357, 2009. [p80]

P. K. Mondal, M. Biswas, and A. K. Ghosh. On high dimensional two-sample tests based on nearest neighbors. *Journal of Multivariate Analysis*, 141:168–178, 2015. [p79]

M. H. Neumann and E. Paparoditis. On bootstrapping L_2-type statistics in density testing. *Statistics & Probability Letters*, 50:137–147, 2000. [p79]

D. N. Politis and H. White. Automatic block-length selection for the dependent bootstrap. *American Economic Review*, 23:53–70, 2004. [p81]

S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams. *The American Soldier. Adjustment during Army Life*. Princeton University Press, England, 1949. [p81]

S. Wei, C. Lee, L. Wichers, and J. S. Marron. Direction-projection-permutation for high-dimensional hypothesis tests. *Journal of Computational and Graphical Statistics*, 25:549–569, 2016. [p79]

H. Zhang, J. Jin, and Z. Wu. Distributions and power of optimal signal-detection statistics in finite case. *IEEE Transactions on Signal Processing*, 68:1021–1033, 2020. doi: https://doi.org/10.1109/TSP.2020.2967179. [p80]

Marta Cousido-Rocha
Instituto Español de Oceanografía (IEO, CSIC), Centro Oceanográfico de Vigo
Subida a Radio Faro 50–52, Vigo 36390
Spain
ORCiD: 0000-0002-4587-8808
marta.cousido@ieo.csic.es

Jacobo de Uña-Álvarez
CINBIO, Universidade de Vigo, SiDOR Research Group
Campus Lagoas-Marcosende, Vigo 36310
Spain
ORCiD: 0000-0002-4686-8417
jacobo@uvigo.es