DOI: 10.1111/rssc.12529

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY | Series C Applied Statistics

Multiple comparison procedures for discrete uniform and homogeneous tests

Marta Cousido-Rocha[1] | **Jacobo de Uña-Álvarez**[2] | **Sebastian Döhler**[3]

[1]CINBIO, Universidade de Vigo, Grupo SiDOR (Inferencia Estadística, Decisión e Investigación Operativa), Vigo, Spain

[2]CINBIO, Universidade de Vigo, Departamento de Estadística e Investigación Operativa, Grupo SiDOR (Inferencia Estadística, Decisión e Investigación Operativa), Vigo, Spain

[3]Darmstadt University of Applied Sciences, CCSOR and Faculty of Mathematics and Science, Darmstadt, Germany

Correspondence
Marta Cousido-Rocha, CINBIO, Universidade de Vigo, Grupo SiDOR (Inferencia Estadística, Decisión e Investigación Operativa), 36310 Vigo, Spain
Email: martacousido@uvigo.es

Abstract

Discrete uniform and homogeneous *p*-values often arise in applications with multiple testing. For example, this occurs in genome wide association studies whenever a non-parametric one-sample (or two-sample) test is applied throughout the gene loci. In this paper, we consider multiple comparison procedures for such scenarios based on several existing estimators for the proportion of true null hypotheses, π_0, which take the discreteness of the *p*-values into account. The theoretical guarantees of the several approaches with respect to the estimation of π_0 and the false discovery rate control are reviewed. The performance of the discrete procedures is investigated through intensive Monte Carlo simulations considering both independent and dependent *p*-values. The methods are applied to three real data sets for illustration purposes too. Since the particular estimator of π_0 used to compute the *q*-values may influence its performance, relative advantages and disadvantages of the reviewed procedures are discussed. Practical recommendations are given.

KEYWORDS

discrete *p*-values, discrete *q*-values, false discovery rate, homogeneous *p*-values

1 | INTRODUCTION

In many modern applications, several hypotheses are simultaneously tested leading to a sequence of p-values. Classical approaches to deal with the multiplicity problem focus on the control of the number of false positives. Two well-known error rates that multiple comparison procedures (MCP) aim to control are the family-wise error rate (FWER), which is the probability of having at least one false positive, and the false discovery rate (FDR), which is the expected proportion of true null hypotheses rejected out of all rejected hypotheses (see Benjamini & Hochberg, 1995). Research on FDR-controlling procedures has been booming; see Benjamini (2010) for existing proposals up to that date. The majority of these procedures have been developed in the setting of continuously distributed test statistics; such procedures can be overly conservative when the p-values follow a discrete distribution. For example, for continuous p-values the FDR of Benjamini and Hochberg (1995) procedure, henceforth referred to the BH method, is $(m_0/m)\alpha$ when applied at nominal level α. Here m and m_0 denote the number of hypotheses and the number of true null hypotheses, respectively. For discrete p-values, the FDR of the BH method may be much smaller than $(m_0/m)\alpha$ (see Heller & Gur, 2012, Section 1), thus yielding a conservative decision rule and, consequently, a loss in power. This can be prevented, however, by developing procedures that appropriately incorporate the discreteness of the p-values. Indeed, by exploiting the discrete nature of the p-values dramatic improvements in power can be achieved, especially when the p-values are highly discrete.

Even though discrete p-values arise in many applications, few papers explicitly deal with this aspect of multiple testing. Heyse (2011) introduced a discrete BH procedure, known as Heyse's method, which takes advantage of the discrete distribution of the p-values. However, Heyse's method may be anti-conservative, that is, the actual FDR level may be larger than nominal. Döhler et al. (2018) constructed similar BH-type procedures that incorporate the discrete and heterogeneous structure of the data and guarantee FDR-control, filling the gap of Heyse (2011). On the other hand, Heller and Gur (2012) proposed a step-down procedure that exploits the discreteness of the p-values and obtains FDR levels closer in magnitude to the nominal level. Their method can be considered as a discrete version of the classical method of Benjamini and Liu (1999) which controls the FDR for continuous p-values under independence or positive dependence. Recently, Chen and Sarkar (2020) investigated the BH procedure when applied to mid p-values, providing in this way a correction of the BH method for discrete p-values. More precisely, they proved the FDR control of the BH procedure applied to two-sided mid p-values of Binomial tests and Fisher's exact tests. In the same line of research, Chen (2020) proposed a new BH procedure which controls the FDR when applied to mid-p-values and to p-values with general distributions.

In this article, we investigate a particular type of discrete p-values, which are homogeneous (i.e. identically distributed) and which we term discrete uniform in the sense of Definition 1 below. To formalise things, suppose that one tests a number of null hypotheses, m, and that the resulting p-values $\{pv_1, \ldots, pv_m\}$ are observations of the random variables $PV_i, i = 1, \ldots, m$. Assume that all the p-values are identically distributed under the null hypothesis sharing a common support $A = \{t_1, \ldots, t_s, t_{s+1}\}$ with $t_0 \equiv 0 < t_1 < \ldots < t_s < t_{s+1} \equiv 1$. Furthermore, throughout the paper, it is assumed that the p-values follow the cumulative distribution function (cdf) introduced in the following definition.

Definition 1 (*Discrete uniform cdf*). Given $A = \{t_1, \ldots, t_s, t_{s+1}\}$ with $t_0 \equiv 0 < t_1 < \ldots < t_s < t_{s+1} \equiv 1$ (the support set of the distribution of the p-values), the discrete uniform cdf with support A, $H_A \equiv H_{\{t_1,\ldots,t_s,t_{s+1}\}}$, is defined as

$$H_{\{t_1,\dots,t_s,t_{s+1}\}}(x) = \begin{cases} 0 & \text{for } x < t_1 \\ t_j & \text{for } x \in [t_j, t_{j+1}) \\ 1 & \text{for } x \geq 1 \end{cases}$$

Note that H_A is a step function that jumps up by $t_j - t_{j-1}$ at t_j for $j = 1, \dots, s+1$. The classical discrete uniform cdf is H_A where A contains equally spaced points, that is, $A = \{1/N, 2/N, \dots, (N-1)/N, 1\}$, $N \in \mathbb{N}$. Therefore, Definition 1 generalises this concept to possibly non-equidistant support points. Summarising, we refer to any member of the class $\mathcal{H} = \{H_A | A \subset (0, 1], A \text{ countable}\}$ as discrete uniform distribution.

In practice, p-values whose cdf belongs to the class $\mathcal{H}$ are often found. One-sided p-values derived from homogeneous and discrete tests are necessarily homogeneous discrete uniform (hdu); see Section 4.3 for a real data example. The same applies to two-sided homogeneous and discrete p-values whenever the null distribution of the test statistic is symmetric. Nevertheless, discrete uniform p-values are found when analysing continuous data through non-parametric tests too. For example, in the two-sample problem with low sample size and large dimension addressed in Sections 3.1 and 4.1, several non-parametric two-sample tests have been applied: for location differences, the Wilcoxon rank-sum test (Gibbons & Chakraborti, 1992) and the absolute group mean difference test in Liang (2016) (abbr. abs); for scale differences, the Siegel–Tukey (Siegel & Tukey, 1960) and the Ansari–Bradley test (Ansari & Bradley, 1960); and, for detecting any type of differences, the well-known Kolmogorov–Smirnov test (abbr. KS) (see Section 7.3 of Gibbons & Chakraborti, 1992) and the non-parametric test based on the L_2-distance between the two empirical characteristic functions proposed by Cousido-Rocha et al. (2019b) (called J_i test). The distribution of the statistics of all the tests mentioned above is determined using a permutation test when the sample sizes are small, hence the corresponding p-values are discrete, specifically, hdu. Table 1 shows the corresponding support points of the distributions of the p-values for some sample sizes as an illustration.

Discrete corrections of MCP like those in Döhler et al. (2018) and Heller and Gur (2012) are irrelevant for hdu p-values, which are special to this regard. Indeed, the adjusted discrete p-values of Heller and Gur (2012) and Heyse (2011) reduce to the ones for continuous p-values in Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001), respectively, when applied to hdu p-values, leaving the results unchanged. The same holds true for the method of Chen (2020). Therefore, we decide to focus our research on the q-value approach proposed by Storey (2003) but based on estimators of the proportion of true null hypotheses, π_0, which take the discreteness of the p-values into account. The estimators of π_0 considered are well-suited for hdu p-values and generally lead to a power increase when compared to standard estimators for continuous p-values; see Section 3 for more on this. The implementation of the q-value method by Storey and Tibshirani (2003) that uses the FDR estimator of Storey et al. (2004) is equivalent to applying the Benjamini and Hochberg (1995) method at level $\alpha/\hat{\pi}_0$ (being $\hat{\pi}_0$ any acceptable π_0 estimator), and this method is known as adaptive Benjamini and Hochberg (adaptive BH). Furthermore, keep in mind that the adaptive BH is equivalent to the 'adaptive Storey's procedure' of Storey et al. (2004) on basis of their Lemma 2. Therefore, the q-value method, the adaptive BH method and the adaptive Storey's procedure are equivalent. This is further discussed in Section 2 of the Supplementary Material.

The paper is organised as follows. In Section 2, we review the q-value method and several corrections of such approach for hdu p-values. The theoretical guarantees of the proposed methods with respect to the estimation of the proportion of true null hypotheses, the estimation

TABLE 1 Support points of the p-values derived from the J_i permutation test, abs test, KS test, Wilcoxon test, Ansari–Bradley test and Siegel–Tukey test for different sample sizes

	J_i permutation test	**abs test**
n_1, n_2		
4,4	$\left\{\frac{1}{35}, \frac{2}{35}, \dots, \frac{35}{35}\right\}$	$\left\{\frac{1}{35}, \frac{2}{35}, \dots, \frac{35}{35}\right\}$
5,5	$\left\{\frac{1}{126}, \frac{2}{126}, \dots, \frac{126}{126}\right\}$	$\left\{\frac{1}{126}, \frac{2}{126}, \dots, \frac{126}{126}\right\}$
	KS test	**Wilcoxon test**
n_1, n_2		
4,4	$\left\{\frac{1}{35}, \frac{8}{35}, \frac{27}{35}, \frac{35}{35}\right\}$	$\left\{\frac{1}{35}, \frac{2}{35}, \frac{4}{35}, \frac{7}{35}, \frac{12}{35}, \frac{17}{35}, \frac{24}{35}, \frac{31}{35}, \frac{35}{35}\right\}$
5,5	$\left\{\frac{1}{126}, \frac{10}{126}, \frac{45}{126}, \frac{110}{126}, \frac{126}{126}\right\}$	$\left\{\frac{1}{126}, \frac{2}{126}, \frac{4}{126}, \frac{7}{126}, \frac{12}{126}, \frac{19}{126}, \frac{28}{126}, \frac{39}{126}, \frac{53}{126}, \frac{69}{126}, \frac{87}{126}, \frac{106}{126}, \frac{126}{126}\right\}$
	Siegel–Tukey test	**Ansari–Bradley test**
n_1, n_2		
4,4	$\left\{\frac{1}{35}, \frac{2}{35}, \frac{4}{35}, \frac{7}{35}, \frac{12}{35}, \frac{17}{35}, \frac{24}{35}, \frac{31}{35}, \frac{35}{35}\right\}$	$\left\{\frac{1}{35}, \frac{5}{35}, \frac{14}{35}, \frac{26}{35}, \frac{35}{35}\right\}$
5,5	$\left\{\frac{1}{126}, \frac{2}{126}, \frac{4}{126}, \frac{7}{126}, \frac{12}{126}, \frac{19}{126}, \frac{28}{126}, \frac{39}{126}, \frac{53}{126}, \frac{69}{126}, \frac{87}{126}, \frac{106}{126}, \frac{126}{126}\right\}$	$\left\{\frac{2}{126}, \frac{6}{126}, \frac{18}{126}, \frac{38}{126}, \frac{68}{126}, \frac{104}{126}, \frac{126}{126}\right\}$

of the FDR and the FDR control are summarised too. The performance of the discrete q-values is investigated through intensive Monte Carlo simulations in Section 3. In Section 4, we illustrate the behaviour of the proposed methods through three real data examples. Finally, in Section 5, we give the main conclusions of our comparative study and we provide some practical recommendations. The methods investigated in this paper have been implemented in the user-friendly `DiscreteQvalue` package (Cousido-Rocha et al., 2019a) of the free software `R`.

2 | MULTIPLE COMPARISON PROCEDURES: q-VALUE METHOD

2.1 | q-value method revisited

In this section, we review the q-value method and several ways of estimating q-values when the p-values are hdu. Consider a family of m null hypotheses H_{0i}, $i = 1, \dots, m$, with associated p-values pv_i, $i = 1, \dots, m$, which are observations of the random variables PV_i, $i = 1, \dots, m$. The number of true null hypotheses is denoted by m_0; R_m is the number of rejected null hypotheses, while V_m the number of true null hypotheses which are rejected (Type I errors). The most popular error rates to control the Type I errors in a simultaneous way are the FWER and the FDR. The q-value method aims at controlling the latter, which is defined as the expected value of the proportion of Type I errors among the rejected hypotheses, that is, $\text{FDR} = E\left[V_m/R_m\right]$. The q-value method decides whether each one of the H_{0i}, $i = 1, \dots, m$, should be rejected or not based on a measure of each feature's significance (referred to as its q-value) that automatically takes multiplicity into

account. The q-value of a feature i is defined as the infimum FDR that can be attained when declaring that feature significant:

$$q(pv_i) = \inf_{t \geq pv_i} \text{FDR}(t), \tag{1}$$

where FDR(t) denotes the FDR when one rejects the hypotheses with p-values smaller than or equal to t.

Note that the FDR is undefined if $R_m = 0$; actually, the formal definition of the FDR is given by FDR $= E[(V_m/R_m)|R_m > 0]\text{P}(R_m > 0)$. However, since the q-value is interpreted under the assumption that the feature is called significant, the inclusion of the term $\text{P}(R_m > 0)$ in the definition of the FDR is strange. Hence, the q-value is most technically defined as the minimum positive false discovery rate, pFDR $= E[(V_m/R_m)|R_m > 0]$, at which the feature can be called significant. In practice, FDR(t) is unknown and must be estimated. Hence, one can estimate the q-value of a feature i by plugging a FDR estimator in (1). We consider the FDR estimator employed in Storey et al. (2004) which is

$$\widehat{\text{FDR}}(t) = \frac{m\hat{\pi}_0 t}{\#\{i|pv_i \leq t\}}, \tag{2}$$

where $\hat{\pi}_0$ is an estimator of the proportion of true null hypotheses $\pi_0 = m_0/m$.

As we mentioned above, once a FDR estimator is available the q-values are estimated by plugging a FDR estimator in (1), that is we define

$$\hat{q}(pv_i) = \min_{t \geq pv_i} \widehat{\text{FDR}}(t). \tag{3}$$

In Equation (3), the left-hand side is an estimate of (1) that is constructed initially by replacing FDR by its estimate in (2). Since $\widehat{\text{FDR}}(t)$ is the ratio between a linear function and a step function that jumps at the p-values, we can replace infimum by minimum. In practice, we compute expression (3) of the estimated q-values using the easily implemented and fully automated algorithm of Storey and Tibshirani (2003) described in Remark B of their Appendix. The steps of the algorithm for estimating q-values from a list of p-values are:

1. Let $pv_{(1)} \leq pv_{(2)} \leq \ldots \leq pv_{(m)}$ be the ordered p-values.
2. Compute an estimate of π_0 ($\hat{\pi}_0$).
3. Calculate

$$\hat{q}(pv_{(m)}) = \min_{t \geq pv_{(m)}} \widehat{\text{FDR}}(t) = \min_{t \geq pv_{(m)}} \frac{m\hat{\pi}_0 t}{\#\{j|pv_j \leq t\}} = \hat{\pi}_0 pv_{(m)}. \tag{4}$$

4. For $i = m-1, m-2, \ldots, 1$, calculate

$$\hat{q}(pv_{(i)}) = \min_{t \geq pv_{(i)}} \widehat{\text{FDR}}(t) = \min_{t \geq pv_{(i)}} \frac{m\hat{\pi}_0 t}{\#\{j|pv_j \leq t\}} = \min\left(\frac{m\hat{\pi}_0 pv_{(i)}}{i}, \hat{q}(pv_{(i+1)})\right). \tag{5}$$

5. The estimated q-value for the i-th most significant feature is $\hat{q}(pv_{(i)})$.

Once the estimated q-values are computed following the previous algorithm, the q-value method rejects the null hypotheses whose estimated q-values are less than or equal to the nominal

level α. This is equivalent to applying the Benjamini and Hochberg (1995) method at level $\alpha/\hat{\pi}_0$, which is the so-called adaptive BH; see Section 2 of the Supplementary Material. Hence, for a given nominal level α, the q-value method is more powerful than the Benjamini and Hochberg (1995) method except when $\hat{\pi}_0 = 1$ (they are equivalent in this case), or when the estimator of π_0 is unacceptable because it reports values greater than 1.

Different versions of the q-value method can be defined depending on which π_0 estimator is plugged in (2). In Section 2.2, two different versions of the q-value method for continuous p-values are reviewed. Furthermore, we consider in Section 2.3 three different versions of the q-value method for hdu p-values. The q-value versions in Sections 2.2 and 2.3 are based on the q-value algorithm explained above and the unique difference among them is the specific π_0 estimator plugged in (2).

2.2 | π_0 estimators for continuous p-values

The classical π_0 estimator proposed in Storey (2002) is

$$\hat{\pi}_0(\lambda) = \frac{\#\{pv_i > \lambda; i = 1, \ldots, m\} + 1}{m(1-\lambda)}, \tag{6}$$

where $\lambda \in [0, 1]$ is well-chosen according to some procedure. A standard choice for λ, for continuous p-values, is 1/2 (Storey, 2002). Henceforth, we refer to the π_0 estimator given by (6) and $\lambda = 1/2$ as standard Storey estimator (abbr. $\hat{\pi}_0^{SS}$), and to the corresponding q-value method as standard Storey (SS) q-value method. Blanchard and Roquain (2009) recommend λ equal to the nominal level α instead of $\lambda = 1/2$ since it leads to a more robust procedure under positive dependence, but at the price of being more conservative.

Additionally Storey and Tibshirani (2003) proposed an automatic method to estimate π_0 which avoids the selection of the λ parameter in (6). Specifically they suggested $\hat{\pi}_0^{ST} = \widehat{f}(1)$, where $\widehat{f}$ is the natural cubic spline with 3 degrees of freedom of $\hat{\pi}_0(\lambda)$ on λ, with $\lambda = 0{,}0.01, 0.02, \ldots, 0.95$ (or another sequence of λ values between 0 and 1) and $\hat{\pi}_0(\lambda)$ is the estimator in (6). Henceforth, we refer to this estimator and the corresponding q-value method as the ST (Storey and Tibshirani) estimator and the ST q-value method, respectively.

The two π_0 estimators presented in this subsection are suitable for continuous p-values but can be overly conservative for discrete p-values (see Gilbert, 2005). For this reason, in the next subsection we consider three π_0 estimators which take into account the discrete distribution of the p-values.

2.3 | π_0 estimator for discrete p-values

In Section 2.3.1, the q-value method based on the π_0 estimator of Liang (2016) is considered. To the best of our knowledge, the performance of the q-value method based on such estimator is studied for the first time in this paper (Section 3), to wit, it is the first time that both approaches, q-value method and π_0 estimator of Liang (2016), are combined to address the problem of testing m null hypothesis through a sequence of hdu p-values. In Section 2.3.3, the q-value method based on a π_0 estimator based on randomised p-values is considered. On the other hand, the q-values which arise from the π_0 estimator in Section 2.3.2 can be regarded as a simplification of the adaptive

FDR-procedure in Chen et al. (2017) for hdu p-values. The procedures in Sections 2.3.2 and 2.3.3 have been investigated previously; however, this has been never done in the important setting with hdu p-values considered in this paper.

2.3.1 | q-values based on the Liang π_0 estimator

Liang (2016) proposed a π_0 estimator for hdu p-values. Let $B = \{b_1, \ldots, b_{s+1}\}$ be the sample frequencies of every element in A, that is, $b_i = \#\{pv_j : pv_j = t_i\}$ for $i = 1, \ldots, s{+}1$. His procedure is based on finding the smallest support point such that the b_i's to its right are roughly equal, that is, it is a right-boundary procedure. The method finds the smallest λ for which $\hat{\pi}_0(\lambda)$ stops decreasing, where λ is chosen from a subset of $\{t_0, \ldots, t_s\} = A \setminus t_{s+1}$ (see Definition 1).

Formally, the Liang π_0 estimator is $\hat{\pi}_0(\lambda_L)$, where $\hat{\pi}_0(\lambda)$ is the estimator in (6) and λ_L is defined in Definition 2

Definition 2 Let $\Lambda = \{\lambda_1, \ldots, \lambda_\nu\} \subseteq \{t_1, \ldots, t_s\} = A \setminus t_{s+1}$, see Definition 1, be a candidate set for λ such that $0 \equiv \lambda_0 < \lambda_1 < \ldots < \lambda_\nu < \lambda_{\nu+1} \equiv 1$. Then, the λ chosen is λ_L where $L = \min\{1 \le i \le \nu - 1 : \hat{\pi}_0(\lambda_i) \ge \hat{\pi}_0(\lambda_{i-1})\}$ if $\hat{\pi}_0(\lambda_i) \ge \hat{\pi}_0(\lambda_{i-1})$ for some $i = 1, \ldots, \nu{-}1$ and $\lambda_L = \lambda_\nu$ otherwise.

In order to illustrate Liang's method, we report in Figure 1 the histogram of the p-values in the application in Liang (2016), Section 6. In this example $A = \{0.1, \ldots, 0.9, 1\}$, $\Lambda = \{0, 0.1, \ldots, 0.5\}$, $\lambda_L = 0.5$ and $\hat{m}_0 = 9474$; the dotted horizontal line is the expected number of true null p-values at every support point, 947.

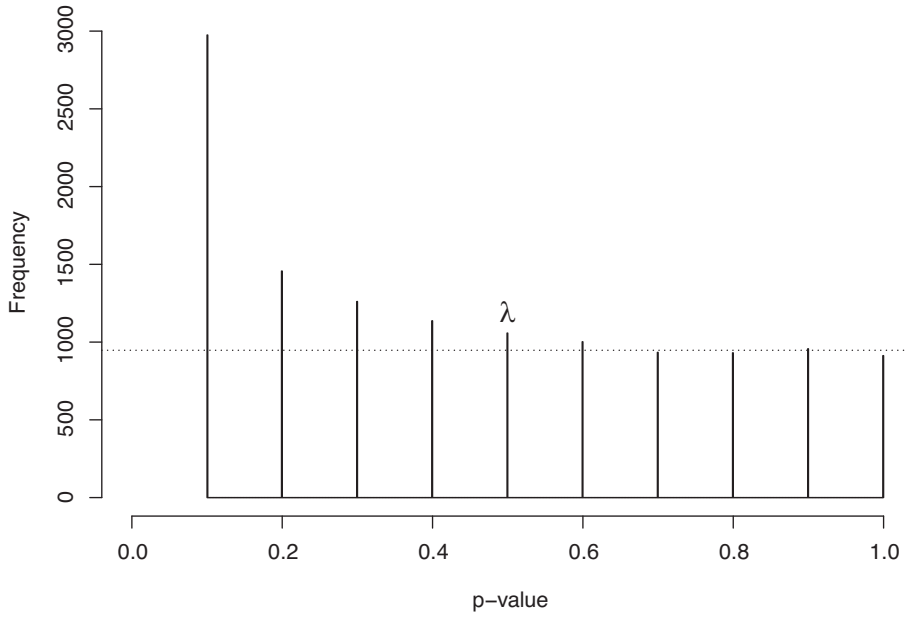

FIGURE 1 The histogram of the p-values in the application in Liang (2016), Section 6. His method takes $\lambda_L = 0.5$, and the dotted horizontal line is the expected number of true null p-values at every support point

2.3.2 | q-values based on the Chen π_0 estimator

Chen et al. (2017) proposed a π_0 estimator for p-values which follow discrete and possibly heterogeneous null distributions. We present a simplified version of Chen's algorithm for the case of hdu p-values.

Chen et al. (2017) studied the bias of the π_0 estimator (6) in the discrete paradigm. In order to reduce this bias they followed an idea similar to that in Liang (2016) but, instead of choosing a single λ parameter, they suggested to consider several λ's and then to average the resulting estimates for π_0. The steps of Chen's algorithm are (with A as in Definition 1):

1. Set $q = \inf\{c : c \in A\}$. Pick a sequence of B increasing, equally spaced 'guiding values' $\{\tau_j\}_{j=1}^{B}$ such that $q = \tau_0 \leq \tau_1 \leq \ldots \leq \tau_B < 1$.
2. For each $j \in \{1, \ldots, B\}$, set $T_j = \{\lambda \in A : \lambda \leq \tau_j\}$ and $\lambda_j = \sup\{\lambda : \lambda \in T_j\}$. For each $j \in \{1, \ldots, B\}$, define the 'trial estimator' $\beta(\tau_j) = 1/((1-\tau_j)m) + (1/m)\sum_{i=1}^{m} I\{pv_i > \lambda_j\}/(1-\lambda_j)$. Truncate $\beta(\tau_j)$ at 1 when it is greater than 1.
3. Set $\hat{\pi}_0^G = (1/B)\sum_{j=1}^{B}\beta(\tau_j)$ as the estimate of π_0.

The first term in $\beta(\tau_j)$ is technical and only useful to prove theoretical properties of adaptive MCP's. The sequence $\{\tau_j\}_{j=1}^{B}$ used in Chen et al. (2017) is $\tau_1 = \tau_0 + 0.5 \times (0.5 - \tau_0)$, $B = 100$ if $\tau_0 < 0.5$, otherwise set $\tau_1 = \tau_B = 0.5$ and $B = 1$. An in depth study of the sensitivity of the Chen method to the choice of $\{\tau_j\}_{j=1}^{B}$ may be of practical interest, but it is beyond the scope of the present work. However, it is worth to mention that we checked via simulation the behaviour of Chen's $\hat{\pi}_0$ based on different sequences of 'guiding values' (results not shown). First, we tried Chen's $\hat{\pi}_0$ with $\{\tau_j\}_{j=1}^{B} = A$, and the mean squared error (MSE) was always larger than that obtained using the $\{\tau_j\}_{j=1}^{B}$ recommended by Chen et al. (2017). This is probably related to the fact that, for large values in A, the π_0 estimator is based on few p-values, leading to a poor performance. Secondly, we fixed $\{\tau_j\}_{j=1}^{B}$ to be the support points smaller than 1/2, and the MSE was approximately equal to that attached to the sequence proposed by Chen et al. (2017). Further investigation is required before reaching solid conclusions to this regard.

2.3.3 | q-values based on the randomised π_0 estimator

Other approaches to take the discreteness into account have been suggested in the literature. Kulinskaya and Lewin (2009) and Habiger (2015), among others, suggested procedures based on randomised p-values. Habiger (2015) extends to the multiple testing setting the randomised P-value, (non-randomised) mid P-value and abstract randomised P-value which are recommended when the test statistic has a discrete distribution. Kulinskaya and Lewin (2009) introduce fuzzy MCP's as a solution to the problem of multiple comparisons for discrete test statistics. The randomised p-values follow a continuous uniform distribution under the global null hypothesis, and therefore classical methods to estimate π_0 as (6) can be applied. The randomised procedure used here is a simple one described in the next steps. It uses the definition of randomised p-values in Dickhaus et al. (2012). Suppose that we want to define the randomised version of pv_i with $i \in \{1, \ldots, m\}$. Remember that the support of the p-values is denoted by $A = \{t_1, \ldots, t_s, t_{s+1}\}$ with $t_0 = 0 < t_1 < \ldots < t_s < t_{s+1} = 1$ (see Definition 1).

1. Generate an observation u from a $U(0, 1)$.
2. Suppose $pv_i = t_k$, $k \in \{1, \ldots, s{+}1\}$; then, the randomised *p-value* is defined by

$$pv_i^{Rand} = pv_i - u(t_k - t_{k-1}).$$

Applying this algorithm to each *p-value*, we obtain a set of randomised p-values $\{pv_i^{Rand}, i = 1, \ldots, m\}$. The next step is to compute (6) using the randomised p-values and $\lambda = 0.5$. This procedure can be repeated a large number of times L reporting L values of (6) which can be summarised using the average and reported it as our final estimator, that is, $\hat{\pi}_0^{Rand}(\lambda) = (1/L)\sum_{j=1}^{L} \hat{\pi}_{0,j}^{Rand}(\lambda)$ where $\hat{\pi}_{0,j}^{Rand}(\lambda)$ is the estimator in (6) computed using the randomised p-values obtained in the j-th simulation run. We refer to the q-value method which plug in this π_0 estimator as *randomised q-value method* (abbr. Rand).

In the setting of multiple testing, it is important to distinguish three different issues: (a) conservativeness of the π_0 estimator; (b) conservativeness of the FDR estimator (2); and (c) FDR control of the q-value method based on (1) and (2). Below we discuss these issues for each of the q-value methods. In the standard setting with continuous and independent p-values, the estimators for π_0 and FDR(t) are known to be conservative when λ is fixed; FDR control has been established too (Storey et al., 2004). Under special forms of weak dependence, conservativeness of the estimator for FDR(t) and FDR control have been proved asymptotically (Storey et al., 2004). Under weak dependence FDR control with the data-driven λ employed by the ST method holds asymptotically too (Storey & Tibshirani, 2003), but the conservativeness of the estimators for π_0 and FDR(t) is not guaranteed in this case. Other data-driven choices for λ have been validated theoretically however (Liang & Nettleton, 2012).

The known theoretical properties of the methods for discrete p-values include the conservativeness of the estimators for π_0 (Liang, Chen) and FDR(t) (Liang), as well as the FDR control (Chen) (see Liang, 2016, for Liang' method, and Blanchard & Roquain, 2009; Chen et al., 2017; and Chen & Doerge, 2020, for Chen's method). These properties are restricted to independent p-values. Under weak dependence, Liang's method is asymptotically conservative for the estimation of FDR(t) (Liang, 2016). Properties of the Chen method under dependence have been investigated only through simulations; in particular, violation of FDR control with blockwise dependence has been reported for this method (Chen et al., 2017). In Table 2 we indicate the state

TABLE 2 Conservativeness of the several estimators for π_0 and for the FDR introduced in Section 2, and FDR-control of the corresponding q-values

	Independence			**Dependence**		
	$\hat{\pi}_0$	**FDR**	***q*-value**	$\hat{\pi}_0$	**FDR**	**FDR control**
SS	T	T	T		$T^{A,W}$	$T^{A,W}$
ST			$T^{A,W}$			$T^{A,W}$
Liang	T	T			$T^{A,W}$	
Chen	T		T	S^{B}		$S^{\times,B}$
Rand						

For each case, the table reports 'T' if a theoretical proof is available in the literature, and 'S' if so far the result is only supported by simulation studies. Empty cells correspond to missing theoretical or by-simulation validation. Superscript A means that the results hold in the asymptotic setting (m tends to ∞). Superscript W means that the type of dependence is the weak dependence defined in Storey and Tibshirani (2003), Liang (2016) and Storey et al. (2004), whereas B denotes block-wise dependence.
$^{\times}$ means that the results provide evidences against the property

of the art regarding the properties of the methods considered in this paper. Our simulation studies in Section 3 bring new knowledge on the guarantees behind the reviewed procedures. This is summarised in the Discussion section.

3 | SIMULATION STUDY

As mentioned before, hdu p-values appear when one-sided tests or two-sided tests (if the distribution of the test statistic is symmetric) are applied to discrete data and also analysing continuous data through non-parametric tests (small sample size). Hence, the current simulation addresses both types of frameworks, continuous data and discrete data settings, in Sections 3.1 and 3.2, respectively.

3.1 | Hdu p-values derived from non-parametric tests applied to continuous data

In this section we consider the two-sample problem with low sample size along a large number of variables, which mimics the real data setting of Section 4.1. In the Supplementary Material, additional simulations for the one-sample problem in the same low sample size and high dimensional setting are provided too; these simulations mimic the real data setting of Section 4.2. The aim of the simulation study is to compare the performance of the different q-value methods and π_0 estimators in Section 2.

We consider a vector autoregressive model of order 1 (or multivariate autoregressive model), VAR(1), defined as $W_i = AW_{i-1} + \varepsilon_i$, where $W_i = (W_{i1}, \cdots, W_{i\eta})^T$, $A = (a_{ij})$ is an $\eta \times \eta$ design matrix such that the process $(W_i)_{i\in\mathbb{N}}$ is stationary, η is the sample size, and $\varepsilon_i \in \mathbb{R}^\eta$ are i.i.d. random vectors (the innovations). We generate a time series of length m from the vector autoregressive model with innovations $\varepsilon_i \sim N_\eta(0, I_\eta)$ and initial point $W_0 \sim N_\eta(0, \Sigma)$ where Σ is the stationary covariance matrix, i.e., $\Sigma = A^T\Sigma A + I_\eta$ (Lyapunov equation; see Hamilton, 1994). The vectors X_i (resp. Y_i), $i = 1, \ldots, m$, consist on i.i.d. observations W_i, $i = 1, \ldots, m$. Specifically, $X = [X_1, \ldots, X_m]^T$ and $Y = [Y_1, \ldots, Y_m]^T$ are based on a standardisation of $W = [W_1, \ldots, W_m]^T$. The simulated dependence tries to mimic the type of weak dependence considered by Storey and Tibshirani (2003) and Liang (2016).

Depending on the choice of the design matrix A, a particular degree of dependence is obtained. In this study, we consider two possibilities for A, each of which is an $\eta \times \eta$ lower triangular matrix with elements a_{ij} satisfying $a_{ij} = 0$ for $i - j > 1$ ($\eta = n_1$ or $\eta = n_2$ depending on whether one is simulating X or Y):

- Independence is simulated by setting

$$a_{ii} = 0,\ i = 1, \ldots, \eta, \text{ and } a_{i,i-1} = 0,\ i = 2, \ldots, \eta. \tag{7}$$

- Medium dependence of X_{ij} and X_{kj} for $i \neq k$ and strong dependence of X_{ij} and X_{lk} for $i \neq l$ and $j \neq k$ is simulated by setting

$$a_{ii} = 0.5,\ i = 1, \ldots, \eta, \text{ and } a_{i,i-1} = 0.4,\ i = 2, \ldots, \eta. \tag{8}$$

In order to simulate X_i, we first define $X_i^{(0)} = \Sigma^{-1/2} W_i$, where W_i are the vectors generated from the VAR(1) model with stationary covariance matrix Σ. Let $\{f_1, f_2, f_3, f_4\}$ be a collection of four densities, and let $I = \{I_j : j \in \{1, \ldots, m\}\}$ be a sequence of i.i.d. random variables such that $P(I_1 = j) = \omega_j$, with $\omega_j = 1/4$ for $j = 1, \ldots, 4$. Then, we take $X_i = F_{I_i}^{-1}(\Phi(X_i^{(0)}))$, where F_i is the cdf corresponding to the density $f_i, i = 1, \ldots, 4$, and Φ stands for the cdf of the standard normal. On the other hand, the data set Y is generated as $Y_i = F_{L_i}^{-1}(\Phi(Y_i^{(0)}))$, where $Y_i^{(0)} = \Sigma^{-1/2} W_i$ and where $L = \{L_j : j \in \{1, \ldots, m\}\}$ is a sequence of i.i.d random variables defined in the following way: given $I = i$, L takes the same value with probability $P(L_1 = i | I_1 = i) = r_{ii} = 1 - \delta$, and a different value with probabilities $P(L_1 = j | I_1 = i) = r_{ij}$, where $r_{31} = r_{42} = r_{13} = r_{24} = \delta$ and $r_{ij} = 0$ otherwise; here we take $\delta = 0, 0.3, 0.5$. Note that the proportion of null hypotheses in these settings is $\pi_0 = 1 - \delta$.

The family of densities $\{f_1, f_2, f_3, f_4\}$ is chosen in order to simulate differences in location, scale or shape. Specifically,

- $\{f_1, f_2, f_3, f_4\} = \{N(0, 1), N(0, 1/4), N(\mu, 1), N(\mu, 1/4)\}$ with $\mu = 2$ or $\mu = 3$ for location;
- $\{f_1, f_2, f_3, f_4\} = \{N(0, 1/4), N(3, 1/4), N(0, 4), N(3, 9)\}$ for scale;
- $\{f_1, f_2, f_3, f_4\} = \{N(2.5, 1/4), N(3.5, 1/4), Exp(1/2), Exp(1/3)\}$ for shape.

The third scenario involves differences in scale too, location differences being minor otherwise. The dimension is $m = 100$ or $m = 1000$. The proportion of true null hypothesis $\pi_0 = 1 - \delta$ is 1, 0.7 or 0.5. The sample sizes are $n_1 = n_2 = 4$ and $n_1 = n_2 = 5$ for location differences, and are increased to $n_1 = n_2 = 8$ for scale and shape differences in order to get some statistical power. The number of Monte Carlo replicates is 1000. The non-parametric test applied are the ones mentioned in Section 1. Additionally, the parametric Student's t-test (see Section 9.1 of Gibbons & Chakraborti, 1992), F-test (see e.g. Section 10.1 of Gibbons & Chakraborti, 1992) and Levene test (Levene, 1960) are also applied for comparative purposes.

Under the global null hypothesis ($\pi_0 = 1$), all the tests control the FDR at the nominal level (results not shown). The FDR is approximately zero for the non-parametric tests, whereas for the parametric ones the FDR is about 0.03. These results suggest that the tests are overly conservative. The full set of simulation results for $\pi_0 < 1$ (i.e. $\delta > 0$) is provided along seventeen Tables in the Supplementary Material. In general, it is seen that the statistical power increases with the proportion of non-true nulls. The same holds true for the effect μ in the case of location differences. However, the power remains roughly the same when moving from the scenario with $m = 100$ hypotheses to that with $m = 1000$. In Figures 2 and 3 (location differences), Figures 4 (scale differences) and Figure 5 (shape differences) we graphically display results on the FDR and power for selected scenarios. The Monte Carlo bias and standard deviation of the several estimators of π_0 in one of the location scenarios are given in Table 3.

Among the several q-value procedures, the best results for hdu p-values are achieved by the Chen method. Indeed, the power of the Chen method is comparable to (and sometimes larger than) that corresponding to the benchmark method which uses the true π_0 (labelled as *Real* in Figures and Tables). The Liang and Rand methods perform correctly too. However, the q-value methods for continuous p-values, SS and ST, report the smallest power values when applied to discrete uniform p-values (equal to 0 in some cases); an exception is found in settings where the discreteness of the p-values is weak; in such settings the power can be similar to the attained by the q-value methods for hdu p-values. Generally speaking, it is seen that the discrete methods improve their continuous counterparts regardless the particular non-parametric test which is employed.

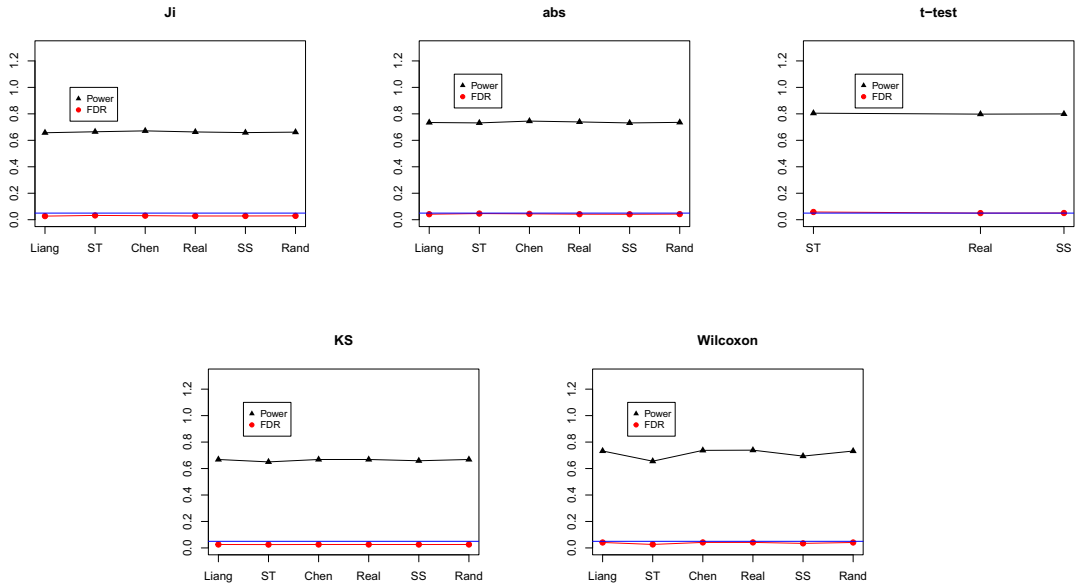

FIGURE 2 Location differences with $n_1 = n_2 = 5$, $m = 100$, $\delta = 0.3$, $\mu = 2$ and A given by (7). The Monte Carlo estimator of the false discovery rate and power are reported for each test and q-value method. The blue line corresponds to $\alpha = 0.05$ [Colour figure can be viewed at wileyonlinelibrary.com]

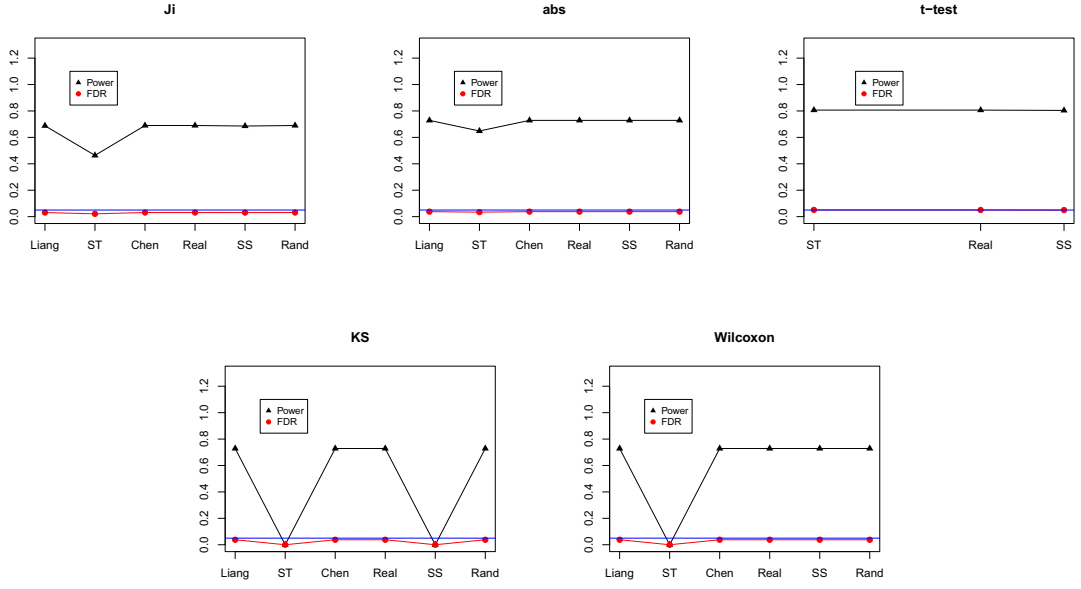

FIGURE 3 Location differences with $n_1 = n_2 = 4$, $m = 1000$, $\delta = 0.5$, $\mu = 2$ and A given by (7). The Monte Carlo estimator of the false discovery rate and power are reported for each test and q-value method. The blue line corresponds to $\alpha = 0.05$ [Colour figure can be viewed at wileyonlinelibrary.com]

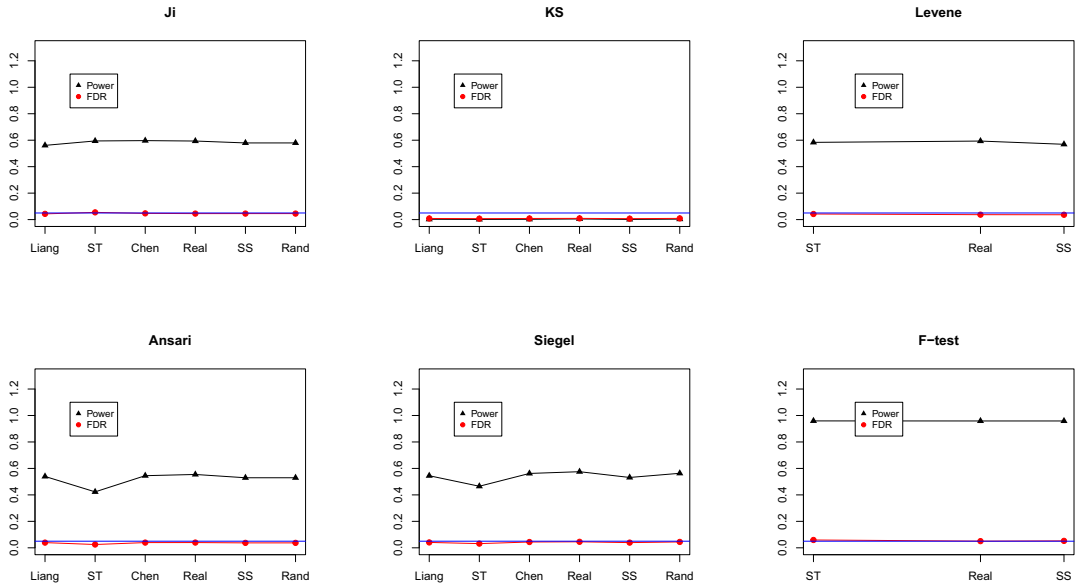

FIGURE 4 Scale differences with $n_1 = n_2 = 8$, $m = 100$, $\delta = 0.5$ and A given by (7). The Monte Carlo estimator of the false discovery rate and power are reported for each test and q-value method. The blue line corresponds to $\alpha = 0.05$ [Colour figure can be viewed at wileyonlinelibrary.com]

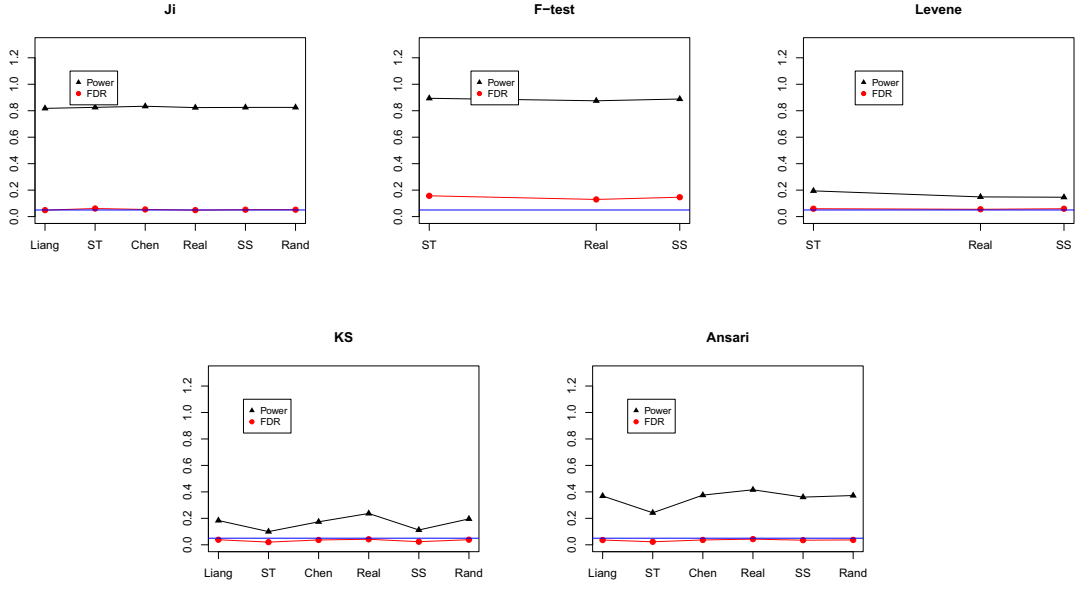

FIGURE 5 Shape differences with $n_1 = n_2 = 8$, $m = 100$, $\delta = 0.5$ and A given by (8). The Monte Carlo estimator of the false discovery rate and power are reported for each test and q-value method. The blue line corresponds to $\alpha = 0.05$ [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Location differences with $n_1 = n_2 = 5$, $m = 100$ and $m = 1000$, $\delta = 0.5$, $\mu = 2$ and A given by (7). The Monte Carlo bias and standard deviation of each π_0 estimator are provided

	J_i		abs		t-test		KS		Wilcoxon	
$\hat{\pi}_0$	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD
Two-sample tests ($m = 100$)										
Liang	0.0367	0.0459	0.0167	0.0418	-	-	0.0137	0.0523	0.0215	0.0467
ST	0.0430	0.1954	0.0315	0.1972	0.0121	0.1871	0.4969	0.0192	0.3855	0.1670
Chen	−0.0032	0.0536	−0.0193	0.0532	-	-	0.0229	0.0453	0.0026	0.0481
SS	0.0266	0.0714	0.0104	0.0707	0.0025	0.0715	0.1604	0.0672	0.0880	0.0692
Rand	0.0185	0.0718	0.0028	0.0711	-	-	0.0131	0.0536	0.0067	0.0653
Two-sample tests ($m = 1000$)										
Liang	0.0266	0.0199	0.0099	0.0172	-	-	0.0134	0.0176	0.0095	0.0193
ST	0.0371	0.0631	0.0280	0.0680	0.0009	0.0655	0.5000	0.0000	0.4683	0.0548
Chen	0.0097	0.0190	−0.0049	0.0179	-	-	0.0149	0.0175	0.0044	0.0163
SS	0.0254	0.0228	0.0121	0.0221	0.0042	0.0222	0.1601	0.0226	0.0873	0.0225
Rand	0.0170	0.0229	0.0041	0.0219	-	-	0.0126	0.0179	0.0058	0.0211

With respect to the estimation of π_0 it is seen that, for continuous p-values (i.e. for the parametric tests), both the ST and SS procedures report estimates with a small positive bias which decreases as m increases, the standard deviation being decreasing too. The bias of ST is somehow smaller than that of SS (this is particularly clear for $m = 1000$), while the SS approach entails a smaller variance (see e.g. Table 3). For the discrete tests, the behaviour of the ST and SS q-value procedures is not so promising. Even when their standard deviations decrease for an increasing m, they exhibit a large positive bias which remains roughly constant when moving from $m = 100$ to $m = 1000$. This suggests the inconsistency of such $\hat{\pi}_0$'s. On the other hand, among the three estimators proposed for discrete p-values, the method with the smallest bias is Chen, Rand being competitive in most of the scenarios. It should be noted however that the Chen method shows a systematic bias in the simulated settings, although of small magnitude (Table 3).

The additional simulation results obtained for the one sample problem (Supplementary Material) were in agreement to those of the two-sample setting. The only exception was a relatively smaller bias of the Liang estimator for π_0 compared to the Chen approach.

3.2 | Hdu p-values derived from discrete data

We simulated the following situation with count data. First, 225 independent Poisson variables with rate parameters μ_j, $1 \leq j \leq 225$, were generated. Then, $m = 45$ groups of size $k = 5$ or $m = 15$ groups of size $k = 15$ were made up, and the within sums of the Poisson variables X_i, $1 \leq i \leq m$, were considered. Note that these X_i are Poisson too. The Poisson p-values of the one-sided test for $H_{0i} : \lambda_i = \mu k$ against $H_{1i} : \lambda_i < \mu k$ for $1 \leq i \leq m$ were computed from the X_i's, were $\mu = 1.15$ was taken. This mimics the application to count data from a DNA-sequencing experiment described in Section 4.3, in which grouping was performed in order to get some statistical power.

TABLE 4 Configurations for the scenario S1 in the simulation study: frequency table for ν_i, $1 \leq i \leq m$

		ν_i							
m	model	0.2	0.4	0.5	0.6	0.7	0.8	0.9	1
45	A1				12	6	9	9	9
	A2	3	4						38
	A3		18	18					9
	A4	18	3						24
15	B1			1	1				13
	B2						4	2	9

TABLE 5 False discovery rate computed along 1000 Monte Carlo replicates for the several multiple testing methods. The nominal FDR is $\alpha = 0.05$

	$m = 45$					$m = 15$		
	S0	A1	A2	A3	A4	S0	B1	B2
Chen	0.001	0.0059	0.0215	0.0296	0.0361	0.022	0.0368	0.0200
Liang	0.001	0.0069	0.0225	0.0358	0.0361	0.021	0.0387	0.0200
Rand	0.001	0.0069	0.0230	0.0369	0.0371	0.028	0.0422	0.0245
SS	0.001	0.0054	0.0215	0.0360	0.0364	0.025	0.0409	0.0232
ST	0.004	0.0208	0.0250	0.0743	0.0442	0.118	0.1374	0.1437

Several configurations for the m null and alternative hypotheses were considered. These were:

S0. $\mu_j = \mu$, $1 \leq j \leq 225$; this is the complete null, since the within rate parameter λ_i is equal to μk for $1 \leq i \leq m$.

S1. $\mu_j = \rho_j \mu$ for some constants $\rho_j \in [0, 1]$, $1 \leq j \leq 225$, with $\sum_{j=1}^{225} \rho_j < 225$; this is a situation in which some of the H_{0i}'s are false, since the true within rate is $\lambda_i = \nu_i \mu k$, where ν_i denotes the average of the ρ_j's in group i, and by force $\nu_i < 1$ for some i's.

Scenario S1 may occur in sequencing experiments due to the presence of deletions in the DNA region or the existence of sampling biases. See Section 5.3 for further motivation. The specific configurations of scenario S1 are given in Table 4.

The several multiple testing methods were applied to the m one-sided Poisson p-values $P(\xi \leq X_i)$, $1 \leq i \leq m$, where ξ is Poisson with rate μk. The FDR and the power were computed along 1000 Monte Carlo replicates, as in Section 3.1. Note that in the simulated settings the p-values are independent and hdu as we have argued in Section 1 and also at the beginning of Section 3. In Tables 5 and 6, we report the FDR and power, respectively. From these Tables it is seen that:

- The ST method does not respect the nominal FDR. This can be seen in one of the scenarios with $m = 45$ (A3) and in all the cases with $m = 15$. So this method should not be applied in the current setting. In the case $m = 15$ the ST method was extremely anticonservative under scenario S0, rejecting all the nulls in 55 out of the 1000 replicates. This violation of ST of the FDR

TABLE 6 Power computed along 1000 Monte Carlo replicates for the several multiple testing methods. The nominal false discovery rate is $\alpha = 0.05$

	$m = 45$				$m = 15$	
	A1	**A2**	**A3**	**A4**	**B1**	**B2**
Chen	0.0070	0.0753	0.3817	0.6001	0.2375	0.0233
Liang	0.0070	0.0750	0.4350	0.5920	0.2340	0.0217
Rand	0.0071	0.0783	0.4509	0.5941	0.2510	0.0270
SS	0.0066	0.0744	0.4379	0.5829	0.2400	0.0245
ST	0.0571	0.0836	0.5702	0.4986	0.3510	0.2055

control also occurred in the scenarios of Section 3.1 when taking $m = 45$ or $m = 15$ (results not shown). On the other hand, additional simulations in the current setting of count data were performed for ST with $m = 100$, and the method respected the FDR in this case. All these findings suggest that the ST method should not be applied in the non-asymptotic setting in which m is small or moderate. The other four methods respect the nominal FDR in the performed simulations.

- Among the valid procedures, Rand was the most powerful, with the only exception of scenario A4 in which it was beaten by Chen. Although, differences in power were not very large.
- In general the power of all the methods was small. This was particularly clear in scenarios A1, A2 and B2, where the amount of non-true nulls was small or the true alternatives were relatively close to the corresponding nulls (see Table 4).

We performed further simulations to investigate the potential effect of dependence in the setting of count data too. Specifically, Poisson counts X_i were simulated from the model $X_i = X_{i-1} + E_i$, where E_i is another Poisson outcome independent of X_{i-1}. This situation is encountered in sequencing experiments when the coverages (cumsums of the read counts) rather than the read counts themselves are used as test statistics. In this scenario with such a strong dependence structure all the methods lost their FDR control.

4 | REAL DATA ANALYSIS

In this section, we consider three real data examples, the first and second one corresponds to setting where hdu p-values appear analysing continuous data through non-parametric tests, whereas the third one addresses count data.

The first is a genetic data set which consists of a large number of gene expression levels measured on two groups of patients with breast cancer, classified according to BRCA mutation type. Then, the framework in this first real data set is the two-sample problem setting considered in Section 3.1. The second real data example is a economic data set which have the daily log return of the five Spanish banks with highest capitalisation for approximately one thousand days. In this case we have a one-sample setting since the aim is to test whether or not the expectation of the log returns is zero (more details in Section 4.2). As we mentioned previously, simulations based on the one-sample setting, where the aim is to test a null hypothesis related with the mean of each of the m variables, are available in the Supplementary Material. Finally, the third data set

has been derived from a DNA sequencing experiment, and the aim of the analysis is the detection of deletions and/or sampling biases in the DNA region using the read counts provided by the sequencing experiment. Then, the framework of this last real data example is the one simulated in Section 3.2.

4.1 | Genetic data

We consider the microarray study of hereditary breast cancer in Hedenfalk et al. (2001). The data set consists of $m = 3226$ logged gene expression levels measured on $n_1 = 7$ patients with breast tumours having BRCA1 mutations, on $n_2 = 8$ patients with breast tumours having BRCA2 mutations and on patients with sporadic breast cancer, which we did not use. Following Storey and Tibshirani (2003), we eliminate all the genes whose measurement exceed 20; the final number of genes is $m = 3170$. We are interested in testing the null hypothesis that the distribution of each of the $m = 3170$ genes is the same for the two types of tumour, BRCA1 tumour and BRCA2 tumour.

Before conducting the analysis, we investigated the dependence structure in the data. Then, we treat the data of each patient as a time series, and we compute the sample autocorrelation function for each patient. For the first lags, the autocorrelation between genes was significant different than 0, whereas such autocorrelation lessened as the number of lags increased. On basis of this results, it seems that the type of dependence present in the data follows the weak dependence considered by Storey and Tibshirani (2003) and Liang (2016), which is the dependence considered in our simulation study through the vector autoregressive model of order 1. In order to avoid unreliable conclusions, our recommendation is to study the dependence structure in the data before conducting the analysis, as we have carried out in our real data set.

Previous analysis of this data set rejected the complete null hypothesis, so one or more genes out of the 3170 are differently distributed; see Cousido-Rocha et al. (2019b) and references therein. Table 7 reports the π_0 estimates for the several methods investigated in this paper. Note that the p-values derived from the application of the t-test and F-test are continuous and hence only the ST and SS estimators can be applied. Table 7 shows that the tests designed to detect scale differences report very conservative results, with $\hat{\pi}_0 = 1$ or $\hat{\pi}_0 > 0.9$, thus suggesting that the main differences between the distributions are not in scale. The number of rejections for such tests at FDR level $\alpha = 0.05$ is zero for any of the q-value approaches. On the other hand, the values $\hat{\pi}_0$ for the remaining tests indicate that the proportion of true null hypotheses is rather large. The number of rejections of each of the remaining methods are 9 for J_i, 96 for abs, 75 for t-test and 18 for KS (all the q-value methods report the same value), whereas Wilcoxon test reports 61 rejections for all the q-value methods except ST, for which the result is zero rejections.

Based on Table 7 and on the aforementioned number of rejections for each test, one may conclude that the differences between the distribution of the genes are basically due to location. Then, we may also conclude that the final result depends mainly on which individual test is applied instead of the selected method for estimating π_0 (except if we apply the ST method to discrete uniform distributed p-values.).

Regarding the q-value method, in this application the number of rejections is the same for all tests regardless of the q-value method, except for the Wilcoxon test. This is explained by the fact that, when $n_1 = 7$ and $n_2 = 8$, the total number of permutations N is 6435 and then the discreteness of the p-values of the tests is not very strong. However, the Wilcoxon test has a 'more pronounced discreteness' than the J_i permutation test or the absolute value test, so it is not surprising that the ST method performs badly reporting zero rejections. Figure 6 depicts the number

TABLE 7 The π_0 estimates obtained by each method for the Hedenfalk data

	Two-sample tests								
	J_i	abs	t-test	F-test	KS	Wilcoxon	Ansari	Siegel	Levene
$\hat{\pi}_0$									
Liang	0.7513	0.6907	-	-	0.8648	0.7568	1	1	-
ST	0.6705	0.6888	0.6885	0.9297	0.7558	1	1	1	1
Chen	0.7508	0.6891	-	-	0.7635	0.7254	1	1	-
SS	0.7514	0.6909	0.6871	0.9495	0.8259	0.7470	1	1	1
Rand	0.7511	0.6908	-	-	0.8259	0.7467	1	1	-

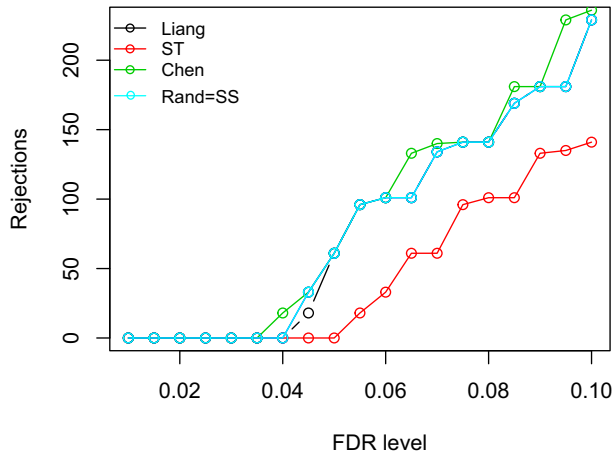

FIGURE 6 Number of rejections of the Wilcoxon test depending on the nominal false discovery rate level. The number of rejections of the Liang, Chen and ST methods are not shown when they overlap the corresponding to the Rand and SS methods [Colour figure can be viewed at wileyonlinelibrary.com]

of rejections reported by the Wilcoxon test for each of the q-value methods along a sequence of nominal levels ($\alpha = 0.010, 0.015, \ldots, 0.095, 0.100$). From Figure 6, it is seen that the ST method is too conservative, whereas the SS method behaves surprisingly well in this case; this does not happen in the second real data application considered in Section 4.2, were the application of the SS method is misleading too.

4.2 | Financial data

In this section, we provide a real data illustration, corresponding to the one sample setting. We consider daily log returns of the five Spanish banks with highest market capitalisation (Santander, BBVA, Bankinter, Caixabank and Sabadell) from 1 January 2015, (first date registered) to 4 June 2018, and from 4 June 2018 to 1 December 2018. The first period corresponds to the term of a right-wing party in the Spanish government, while the second period relates the term of a left-wing party. The data are available at `https://finance.yahoo.com/q?s=ibm`.

The variable log return of an asset at time i is defined as $r_i = \log(P_i) - \log(P_{i-1})$, where P_i is the price of an asset at time i. The first goal of our illustrative application is to explore if for any of these two terms (right-wing, left-wing) the efficiency of the financial market is violated, and to which extent. The second goal is to identify the particular period of time where the financial market lived the worst situation in terms of efficiency; this could allow for association studies with respect to economic or political events.

A classical assumption in finance is that the markets are efficient. This means that the price of assets contains all the information available (Fama, 1970). However, this theoretical assumption is not always true in practice. For example, inefficiency can be a consequence of transactions costs or due to arrival information about the assets (see French & Roll, 1986; Grossman & Stiglitz, 1980). The expectation of the returns must be close to zero if the market is efficient. For this reason, the aforementioned goals are addressed by testing if the expectation of the log returns is zero or not for each time instant. More specifically, we conclude that the market is efficient on day i if $\mu_i \equiv E(r_i) = 0$ where r_i is the log return of the asset at time i (see Tomasz & Tomasz, 2012).

We fix some notation. The data set with the information of the right-wing term is denoted by $X = [X_1, \ldots, X_m]^T$, where $X_i = (X_{i1}, \ldots, X_{i5})$ contains the log returns of the 5 banks at time i which are considered as observations (sample) of the same variable r_i, $i = 1, \ldots, m$, for $m = 873$ (the length of the right-wing party period, after a data cleaning process). On the other hand, the data set with the information of the left-wing term is denoted by $Y = \left[Y_1, \ldots, Y_q\right]^T$, where $Y_i = (Y_{i1}, \ldots, Y_{i5})$ contains the log returns of the 5 banks at time i which are considered as observations (sample) of the same variable r_i, $i = 1, \ldots, q$, $q = 128$ (the length of the left-wing party term, after a data cleaning process). Note that we assume that the observations in X_i are independent for $i = 1, \ldots, m$. This assumption has sense in this economic example since the log return of a bank at time i depends, among others, on the behaviour of the banks at previous time instants but not on the situation at time i. In other words, the financial contagion, that is, the spread of market disturbances, does not occur immediately. Furthermore, prior to carrying out our analysis, we have checked the dependence structure treating the data of each bank as a time series and computing their autocorrelation and cross-correlation functions. Such analysis leads to conclude that our economic data sets do not present a clear dependence structure, the data can be assumed almost independent.

In order to test for $E(r_i) = 0$, we consider two different test statistics: the parametric one sample t-test and the non-parametric one sample Wilcoxon test. The results attained by the several q-values at FDR level $\alpha = 0.05$ for the X and Y samples are reported in Table 8. We can see that the parametric test reports the largest number of rejections for both samples. However, the t-test assumes that the sample is normally distributed, and it seems that this assumption is violated in this setting. Applying the Shapiro–Wilk normality test to the pooled sample of standardised daily log returns yields a *p-value* smaller than 2.2×10^{-16}. This is why a non-parametric test such as Wilcoxon is of interest.

The number of rejections reported by the non-parametric test may be as low as zero when the q-values for continuous tests are naively applied; however, the discrete q-values give almost as many rejections as with the parametric t-test. In this illustrative application, the Liang, Chen and Rand corrections report the same amount of rejections. These results are in agreement with what we have observed in our simulated scenarios (Supplementary Material). Summarising, one may say that the application of the improved q-values may be critical whenever the p-values are discrete, which is the situation with non-parametric tests and small sample sizes; the SS and ST methods for continuous tests cannot be recommended in such a setting.

TABLE 8 The estimates for π_0 (left) and the number of rejections (right) given by each method. Financial data

	Right-wing party (X)		Left-wing party (Y)			Right-wing party (X)		Left-wing party (Y)	
	t-test	Wilcoxon	*t*-test	Wilcoxon		*t*-test	Wilcoxon	*t*-test	Wilcoxon
$\hat{\pi}_0$					$\hat{\pi}_0$				
Liang	-	0.2704	-	0.3611	Liang	-	578	-	62
ST	0.2084	0.5121	0.2303	1	ST	612	0	78	0
Chen	-	0.2725	-	0.3750	Chen	-	578	-	62
SS	0.2467	0.3042	0.4219	0.4062	SS	582	499	56	0
Rand	-	0.2708	-	0.3802	Rand	-	578	-	62

We have compared the proportion of true null hypothesis for the right-wing party and left-wing party. The estimates of π_0 corresponding to the Wilcoxon test with improved q-values are 0.27 (right-wing party) and $0.36 - 0.38$ (left-wing). Hence, the proportion of inefficient days in each period, $1 - \hat{\pi}_0$, is 0.73 (right-wing party) and 0.62–0.64 (left-wing). This result could suggest an association between efficiency of the Spanish financial market and the particular party in the Government. Regarding the particular time period in which the market efficiency is violated, the inspection of the q-values reveals that the period between 4 December 2015 and 28 August 2016, reports the largest number of inefficient days. Interestingly, during this period, two successive elections took place (due to failed negotiations), with a new government agreed precisely by 28 August 2016. Therefore, the political instability would have influenced the performance of the market along these 9 months.

4.3 | DNA sequencing experiment

In order to illustrate the application of the multiple testing procedures with count data, we consider the real targeted resequencing single-end experiment in Li et al. (2012). The data correspond to a DNA region with $f = 315$ bases located in chromosome 1, obtained from the individual NA11893. Jiménez-Otero and de Uña Álvarez (2019) used a Poisson distribution with rate $\lambda = 1.15$ to model the read counts of this DNA region.

Since the length of the reads was concentrated at 90, the number of read counts provided by the sequencing experiment along the DNA region reduces to $m = 225$ (the ones from base 1 to base 225). These read counts are displayed in Figure 7. In this setting, it is interesting to test for the null hypotheses $H_{0i} : \lambda_i = 1.15$ against the one-sided alternatives $H_{1i} : \lambda_i < 1.15$, where λ_i denotes the true expectation for the read count at base i, $1 \leq i \leq m$. Rejection of H_{0i} could suggest the presence of deletion at base i (homozygotic or heterozygotic), or indicate an undersampling of the reads due to a changeable guanine/cytosine content, as suggested by Jiménez-Otero and de Uña Álvarez (2019). However, testing for the individual H_{0i}'s from a single Poisson observation with a small sequencing effort (λ) is complicated; note that the Poisson distribution with rate $\lambda = 1.15$ locates a large mass at zero (0.3166), which eliminates the possibility of rejection for the one-sided test at the usual significance levels. In order to solve this issue, we considered 45 groups of 5 read counts each (bases {1, 2, 3, 4, 5}, {6, 7, 8, 9, 10} and so on), and we took the within sums of the counts as test statistics; note that the updated null rate parameter is $5\lambda = 5.75$. We also tried

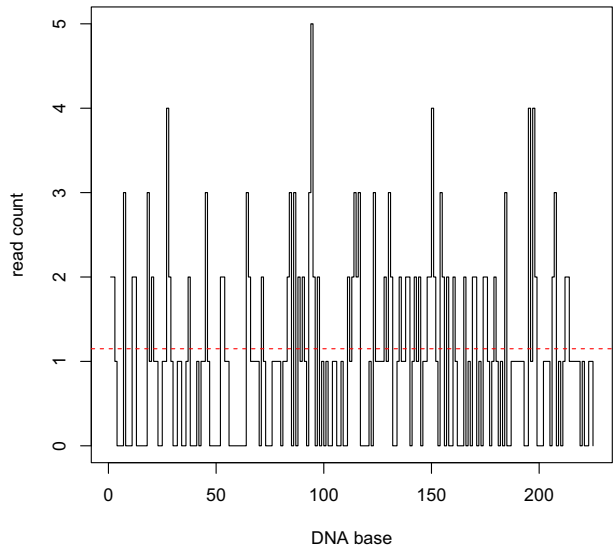

FIGURE 7 Read counts along the DNA region of individual NA11893, bases from 1 to 225. Horizontal dashed line indicates the expected value [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 9 Two smallest q-values, number of rejected nulls $R(\alpha)$ depending on the nominal false discovery rate level α, significant DNA bases at 15% of FDR, and estimated proportion of true nulls for the several multiple testing methods. DNA region of individual NA11893

Groups	Method	$qv_{(1)}$	$qv_{(2)}$	$R(0.05)$	$R(0.10)$	$R(0.15)$	Bases	$\hat{\pi}_0$
$m = 45$	Chen	0.1170	0.2633	0	0	1	56:60	0.8169
	Liang	0.1116	0.2511	0	0	1	56:60	0.7792
	Rand	0.1102	0.2480	0	0	1	56:60	0.7694
	SS	0.1146	0.2578	0	0	1	56:60	0.8000
	ST	0.1399	0.3148	0	0	1	56:60	0.9768
$m = 15$	Chen	0.0500	0.3305	0	1	1	46:60	0.7171
	Liang	0.0418	0.2762	1	1	1	46:60	0.5994
	Rand	0.0357	0.2360	1	1	1	46:60	0.5120
	SS	0.0465	0.3072	1	1	1	46:60	0.6667
	ST	0.0075	0.0493	4	15	15	1:225	0.1069

a grosser grouping by taking 15 groups of size 15; the updated rate parameter in this case was $15\lambda = 17.25$.

In Table 9, we report the two smallest q-values for each method in each of these two grouping scenarios, together with the number of rejected nulls at FDR levels 0.05, 0.10 and 0.15, the groups of DNA bases declared significant at FDR 0.15, and the corresponding estimation of π_0. With 45 groups of bases, all the methods found a unique group of significant bases, and this only happened with an FDR of 15%. Among the several methods, according to the smallest q-value and to the value of $\hat{\pi}_0$, the procedure with greatest power was Rand. With 15 groups, the Rand method was

again the most powerful procedure, with the exception of ST, which should not be regarded as valid in this setting since it does not respect the nominal FDR. Indeed, ST rejected all the nulls already at FDR level of 10%, which is weird; this is in agreement with the simulation results in Section 3.2. All the valid methods declared one group of significant bases regardless the nominal FDR (exception: the Chen method, which was unable to find any significance with an FDR of 5%). The DNA bases found significant were the ones from 56 to 60 when $m = 45$, and the ones between 46 and 60 when $m = 15$. A reasonable conclusion is that there exists some undersampling of the experimental reads close to (and to the left of) base 60; since individual NA11893 provides an unaltered DNA region, this could be explained from a guanine/cytosine sampling bias rather than by a deletion issue; see Jiménez-Otero and de Uña Álvarez (2019) for further study of the potential sampling bias in the current setting.

5 | DISCUSSION

Standard MCP for continuous tests may be inaccurate when applied to discrete p-values. In this paper, we have investigated MCP for hdu p-values. The three methods (Liang, Chen and Rand) performed correctly in the three simulated scenarios: the one-sample and the two-sample non-parametric tests applied to continuous data, and the count data setting. Among the three discrete methods, the optimal choice depends on the particular setting. However, it can be generally said that the differences among the three procedures are minor whatever the case and that, in practice, any choice will do the job well.

On the other hand, the results of our simulation study lead to conclude that the SS and ST methods must not be applied to hdu p-values. More precisely, the results in the continuous data simulation settings (Section 3.1), for which the discreteness of the resulting hdu p-values is strong, show that the SS and ST methods can be overly conservative reporting even 0 rejections under the alternative. When the discreteness of the p-values is soft as in the count data simulation settings, SS performs similar to the three q-value methods for hdu p-values, whereas the ST method does not respect the nominal FDR, the method was anticonservative. This behaviour of the ST method is related to the value of m since also in the continuous data simulation settings the ST method does not respect the FDR nominal level if m is small to moderate. Hence, in practice not only the degree of discreteness of the p-values is relevant, but also the value of m can be crucial. The final conclusion is that the SS and ST methods must not be applied when the discreteness of the p-values is high since they have no power, furthermore, the ST method should also not be applied for moderate m since it loses the FDR control.

In our simulation, we have verified that the behaviour of the three q-value methods for hdu p-values is correct even when the p-values are dependent. However, it is important to stand out that the simulation study only considers the weak dependence in Storey and Tibshirani (2003) and Liang (2016), hence, in practice, we must analyse the type of dependence in our data, since there are no guarantees about the correct performance of the methods under strong dependence structures. In fact, as mentioned in Section 3.2, additional simulations based on highly correlated count data have been carried out (results not shown) and in such setting all methods lose the control of FDR.

As a by-product, the two-sample simulation setting has revealed that the J_i test is competitive, and may perform even better than other well-known two-sample tests. For example, our simulation results suggest that the KS test should not be used when the sample sizes are small and the

differences are other than location (see also Song-Hee & Ward, 2015). The accuracy of the result, in settings as the one simulated in Section 3.1, will depend not only on a suitable choice of the q-value method but also on the selection of an appropriate test, so particular attention should be paid to this regard.

The application of the permutation approach to calibrate the null distribution of a two-sample test leads to hdu P-values. The conclusions of our research hold in general, going beyond the classical Kolmogorov–Smirnov and Wilcoxon tests, although the relative performance of the investigated methods may be sensitive to the particular test at hand. This general validity of the results for permutation tests is not affected by the covariance structures behind the simulated and real data considered in this paper, which correspond to m-dimensional sequences locally independent (i.e. the observations corresponding to each time point are statistically independent). Care is needed, however, when such local independence among the sampled individuals is violated; the permutation approach may be inconsistent in such a case.

ACKNOWLEDGEMENTS

The authors acknowledge comments and suggestions from a reviewer and an Associate Editor that have greatly improved the manuscript. This work has received financial support of the Call 2015 Grants for PhD contracts for training of doctors of the Ministry of Economy and Competitiveness, cofinanced by the European Social Fund (Ref. BES-2015-074958). We acknowledge support from MTM2014-55966-P project, Ministry of Economy and Competitiveness, and MTM2017-89422-P project, Ministry of Economy, Industry and Competitiveness, State Research Agency, and Regional Development Fund, UE. We also acknowledge the financial support provided by the SiDOR research group through the grant Competitive Reference Group, 2016-2019 (ED431C 2016/040), funded by the 'Consellería de Cultura, Educación e Ordenación Universitaria. Xunta de Galicia'. The first author thanks the University of Vigo, and its Escola Internacional de Doutoramento (EIDO) for the financial support provided through mobility doctorate grants. Funding for open access charge: Universidade de Vigo/CISUG.

The authors also thank José Carlos Soage, research support technician in SiDOR group, for helping them in the analysis of the financial data.

ORCID


Marta Cousido-Rocha https://orcid.org/0000-0002-4587-8808

REFERENCES

Ansari, H. & Bradley, R. (1960) Rank-sum tests for dispersions. *Annals of Mathematical Statistics*, 31, 1174–1189.

Benjamini, Y. (2010) Discovering the false discovery rate. *Journal of the Royal Statistical Society*, 72, 405–416.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.

Benjamini, Y. & Liu, W. (1999) A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82, 163–170.

Benjamini, Y. & Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependence. *Annals of Statistics*, 29, 1165–1188.

Blanchard, G. & Roquain, E. (2009) Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10, 2837–2871.

Chen, X. (2020) False discovery rate control for multiple testing based on discrete p-values. *Biometrical Journal*, 62, 1060–1079.

Chen, X., Doerge, R. & Heyse, J. (2017) Multiple testing with discrete data: proportion of true null hypotheses and two adaptive FDR procedures. *Biometrical Journal*, 60, 761–779.

Chen, X. & Doerge, R.W. (2020) Comments on Dr. Aniket Biswas' letter to the editor. *Biometrical Journal*, 62, 2034–2035.

Chen, X. & Sarkar, S.K. (2020) On Benjamini-Hochberg procedure applied to mid p-values. *Journal of Statistical Planning and Inference*, 205, 34–45.

Cousido-Rocha, M., de Uña-Álvarez, J. & Döhler, S. (2019a) DiscreteQvalue: Improved q-values for discrete uniform and homogeneous tests. R package version 1.0.

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

Dickhaus, T., Strassburger, K., Schunk, D., Morcillo-Suarez, C., Illig, T. & Navarro, A. (2012) How to analyze many contingency tables simultaneously in genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, 11, Article 12.

Döhler, S., Durand, G. & Roquain, E. (2018) New FDR bounds for discrete and heterogeneous tests. *Electronic Journal of Statistics*, 12, 1867–1900.

Fama, E. (1970) Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25, 383–417.

French, K.R. & Roll, R. (1986) Stock return variance: the arrival of information and the reaction of traders. *Journal of Financial Economics*, 17, 5–26.

Gibbons, J.D. & Chakraborti, S. (1992) *Nonparametric statistical inference*, 3rd end. New York: Marcel Dekker, Inc.

Gilbert, P.B. (2005) A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1), 143–158.

Grossman, S.J. & Stiglitz, J.E. (1980) On the impossibility of informationally efficient markets. *The American Economic Review*, 70, 393–408.

Habiger, J.D. (2015) Multiple test functions and adjusted p-values for test statistics with discrete distributions. *Journal of Statistical Planning and Inference*, 167, 1–13.

Hamilton, J. (1994) *Time series analysis*. Princeton: Princeton University Press.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R. et al. (2001) Gene- expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344, 539–548.

Heller, R. & Gur, H. (2012) False discovery rate controlling procedures for discrete tests. *arxiv:1112.4627v2*.

Heyse, J.F. (2011) A false discovery rate procedure for categorical data. In: Zhang, H. (Ed.) *Recent advancements in biostatistics*. New Jersey: World Scientific Publishing Company, pp. 43–58.

Jiménez-Otero, N., de Uña Álvarez, J., Pardo-Fernández, J.C. (2019) Goodness-of-fit tests for disorder detection in NGS experiments. *Biometrical Journal*, 61, 424–441.

Kulinskaya, E. & Lewin, A. (2009) On fuzzy familywise error rate and false discovery rate procedures for discrete distributions. *Biometrika*, 96, 201–211.

Levene, H. (1960) Robust tests for equality of variances. In: Olkin, I. (Ed.) *Contributions to probability and statistics*, Palo Alto, Calif: Stanford University Press, pp. 278–92.

Li, J., Lupat, R., Amarasinghe, K.C., Thompson, E.R., Doyle, M.A., Ryland, G.L. et al. (2012) Contra: copy number analysis for targeted resequencing. *Bioinformatics*, 28, 1307–1313.

Liang, K. (2016) False discovery rate estimation for large scale homogeneous discrete p-values. *Biometrics*, 72, 639–648.

Liang, K. & Nettleton, D. (2012) Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 163–182.

Siegel, S. & Tukey, J. W. (1960) A non-parametric sum of ranks procedure for relative spread in unpaired samples. *Journal of the American Statistical Association*, 55, 429–445.

Song-Hee, K. & Ward, W. (2015) The power of alternative Kolmogorov-Smirnov tests based on transformations of the data. *ACM Transactions on Modeling and Computer Simulation*, 25, 1–22.

Storey, J. (2002) A non-parametric sum of ranks procedure for relative spread in unpaired samples. *Statistical Methodology Series B*, 64, 479–498.

Storey, J. (2003) The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31, 2013–2035.

Storey, J., Taylor, J. & Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rate: a unified approach. *Journal of the Royal Statistical Society*, 66, 187–205.

Storey, J. & Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of National Academy of Science*, 100, 9440–9445.

Tomasz, P. & Tomasz, S. (2012) Empirical test of the strong form efficiency of the warsaw stock exchange the analysis of WIG 20 index shares. *South-Eastern Europe Journal of Economics, Association of Economic Universities of South and Eastern Europe and the Black Sea Region*, 10, 155–172.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Cousido-Rocha, M., de Uña-Álvarez, J. & Döhler, S. (2022) Multiple comparison procedures for discrete uniform and homogeneous tests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(1), 219–243. Available from: https://doi.org/10.1111/rssc.12529