Annals of the Institute of Statistical Mathematics (2022) 74:289–319
https://doi.org/10.1007/s10463-021-00797-0

On the usage of randomized p-values in the Schweder–Spjøtvoll estimator

Anh-Tuan Hoang[1] · **Thorsten Dickhaus**[1]

Received: 17 April 2020 / Revised: 21 January 2021 / Accepted: 26 March 2021 / Published online: 28 April 2021

Abstract
We consider multiple test problems with composite null hypotheses and the estimation of the proportion π_0 of true null hypotheses. The Schweder–Spjøtvoll estimator $\hat{\pi}_0$ utilizes marginal p-values and relies on the assumption that p-values corresponding to true nulls are uniformly distributed on [0, 1]. In the case of composite null hypotheses, marginal p-values are usually computed under least favorable parameter configurations (LFCs). Thus, they are stochastically larger than uniform under non-LFCs in the null hypotheses. When using these LFC-based p-values, $\hat{\pi}_0$ tends to overestimate π_0. We introduce a new way of randomizing p-values that depends on a tuning parameter $c \in [0, 1]$. For a certain value $c = c^\star$, the resulting bias of $\hat{\pi}_0$ is minimized. This often also entails a smaller mean squared error of the estimator as compared to the usage of LFC-based p-values. We analyze these points theoretically, and we demonstrate them numerically in simulations.

Keywords Bias · Composite null hypotheses · Mean squared error · Multiple testing · Proportion of true null hypotheses

1 Introduction

In multiple test problems with composite null hypotheses, to account for type I errors, marginal tests are usually calibrated with respect to least favorable parameter configurations (LFCs). These are parameter values in (or on the boundary of) the corresponding null hypotheses under which the marginal tests are most likely to reject. Under certain assumptions, the resulting marginal LFC-based p-values are then uniformly distributed on [0, 1] (Uni[0, 1]-distributed) under LFCs, but

✉ Thorsten Dickhaus
dickhaus@uni-bremen.de

Anh-Tuan Hoang
anhtuan.hoang@uni-bremen.de

[1] Institute for Statistics, University of Bremen, D-28344 Bremen, Germany

stochastically larger than Uni[0, 1] under non-LFCs in the null hypothesis. Under the alternative, LFC p-values usually tend to be stochastically smaller than Uni[0, 1].

While the latter property is desirable in terms of protecting against type II errors, the deviation from uniformity under null hypotheses is problematic for some estimators of the proportion π_0 of true null hypotheses that use the empirical cumulative distribution function (ecdf) of all marginal p-values. We will denote the latter ecdf by $\hat{F}_m$ throughout the remainder, where m is the number of all null hypotheses. One ecdf-based estimator for π_0 was introduced by Schweder and Spjøtvoll (1982), and it is given by

$$\hat{\pi}_0 \equiv \hat{\pi}_0(\lambda) = \frac{1 - \hat{F}_m(\lambda)}{1 - \lambda}, \tag{1}$$

where $\lambda \in (0, 1)$ is a tuning parameter.

In this work, we will investigate the bias, given by $\mathrm{bias}_\vartheta(\hat{\pi}_0) = \mathbb{E}_\vartheta[\hat{\pi}_0] - \pi_0$, and the mean squared error (MSE), given by $\mathrm{MSE}_\vartheta[\hat{\pi}_0] = \mathbb{E}_\vartheta[(\hat{\pi}_0 - \pi_0)^2]$, of $\hat{\pi}_0$ under various statistical models, where ϑ denotes the model parameter. Notice that $\mathrm{MSE}_\vartheta[\hat{\pi}_0] = \mathrm{Var}_\vartheta(\hat{\pi}_0) + \mathrm{bias}^2_\vartheta(\hat{\pi}_0)$. In the case that $\mathrm{bias}_\vartheta(\hat{\pi}_0) = 0$, $\hat{\pi}_0$ is called unbiased. Under the restriction of valid p-values (i.e., p-values that are stochastically not smaller than Uni[0, 1] under null hypotheses), $\hat{\pi}_0(\lambda)$ can only be unbiased, if the marginal p-values that correspond to the true null hypotheses are Uni[0, 1]-distributed. The Schweder–Spjøtvoll estimator is an unbiased estimator if, in addition, all p-values that correspond to the false null hypotheses are smaller than λ with probability one. In general, $\hat{\pi}_0(\lambda)$ is non-negatively biased if used with valid p-values. The aforementioned properties of $\hat{\pi}_0$ follow, for example, from the calculations in Appendix I of Dickhaus et al (2012). It is also known for a longer time (cf., e.g., the discussion by Storey et al (2004) after their Eq. (4)) that the variance of $\hat{\pi}_0(\lambda)$ increases with increasing λ in most cases.

Non-uniformity of p-values under null hypotheses happens for instance in case of discrete models, which has been, among others, investigated by Finner and Strassburger (2007), Habiger and Pena (2011), Dickhaus et al (2012), and Habiger (2015). The randomization approach proposed by Dickhaus et al (2012) results in uniformly distributed p-values under simple (i.e., one-elementary) nulls. In case of composite null hypotheses, the deviation of p-values from uniformity occurs, when marginal test statistics do not have a unique distribution under the null hypotheses and the marginal tests hence cannot be calibrated precisely with respect to their type I error probabilities. To provide more uniform p-values under composite null hypotheses, Dickhaus (2013) proposed randomized p-values that result from a data-dependent mixing of the LFC-based p-values and additional Uni[0, 1]-distributed random variables that are (stochastically) independent of the data. In certain models, these randomized p-values can be simplified to have a linear structure (cf. Hoang and Dickhaus 2021).

While accurate estimations of π_0 are valuable in themselves, they can also improve the power of existing multiple test procedures. Namely, many of such procedures are (implicitly) calibrated to control the familywise error rate (FWER) or the false discovery rate (FDR), respectively, for the case that every null hypothesis

is true, that is, in case of $\pi_0 = 1$, which is often the worst case. If some null hypotheses are false, these procedures become overconservative. Adjusting them according to a pre-estimate of π_0 can improve the overall power of these tests. Benjamini and Hochberg (2000) discuss these so-called adaptive procedures where the original procedure is the linear step-up test from Benjamini and Hochberg (1995). Storey (2003) proved that applying the linear step-up test by Benjamini and Hochberg (1995) at an adjusted level controls the FDR if the p-values are independent. Finner and Gontscharuk (2009) investigated the use of estimators of π_0 as plug-in estimators in single-step or step-down procedures and proved that the Bonferroni procedure at an adjusted level controls the FWER if the marginal p-values are independent. Further results and references on adaptive multiple tests (for FDR control) can be found in Heesen and Janssen (2015, 2016), and MacDonald et al (2019).

We focus on the case of composite null hypotheses and present a new way of randomizing LFC-based p-values. To this end, we utilize a set of stochastically independent and identically Uni[0, 1]-distributed random variables $U_1, \dots, U_m$, which are (stochastically) independent of the data X, as well as a set of constants $c_1, \dots, c_m$, where $c_j \in [0, 1]$ for all $1 \le j \le m$. For a (continuously distributed) LFC-based p-value $p_j^{LFC}(X)$, we propose randomized p-values defined as

$$p_j^{rand}(X, U_j, c_j) = U_j \mathbf{1}\{p_j^{LFC}(X) \ge c_j\} + p_j^{LFC}(X) c_j^{-1} \mathbf{1}\{p_j^{LFC}(X) < c_j\}, \tag{2}$$

$j = 1 \dots, m$.

In many models, this definition comprises the one of Dickhaus (2013) for certain values of $c_j \in [0, 1]$ (cf. Hoang and Dickhaus 2021). It is clear that c_j determines how close p_j^{rand} is to either U_j or p_j^{LFC}. The choices $c_j = 0$ and $c_j = 1$ lead to $p_j^{rand} = U_j$ (by convention) or $p_j^{rand} = p_j^{LFC}$ (with probability one), respectively. Under certain conditions, it holds $U_j \le_{\text{st}} p_j^{rand} \le_{\text{st}} p_j^{LFC}$ under the jth null hypothesis and $p_j^{LFC} \le_{\text{st}} p_j^{rand} \le_{\text{st}} U_j$ under the jth alternative for all $c_j \in [0, 1]$, where $\le_{\text{st}}$ denotes the stochastic order (see, e.g., Theorem 2). For definitions and notations of the stochastic order $\le_{\text{st}}$ and further ones, we refer to the appendix of Hoang and Dickhaus (2021). While Uni[0, 1]-distributed p-values are desirable under null hypotheses, we want to keep them small under alternatives. When using $p_1^{rand}(X, U_1, c_1), \dots, p_m^{rand}(X, U_m, c_m)$ in $\hat{\pi}_0$, we discuss how the choice of the constants $c_1, \dots, c_m$ affects the bias of $\hat{\pi}_0$. Under the restriction of identical c_j's, we find that there exists a $c^\star \in [0, 1]$ for which $\hat{\pi}_0$ has minimal bias when using $p_1^{rand}(X, U_1, c^\star), \dots, p_m^{rand}(X, U_m, c^\star)$. We mainly focus on the bias instead of the MSE, since it turns out that $c^\star$ is close to the MSE-minimizing value of c, especially if m is large. Furthermore, if the LFC-based p-values are positively dependent, the variance of the Schweder–Spjøtvoll estimator is much higher when using the LFC-based p-values instead of the randomized p-values. The problem of minimizing the MSE may in this case lead to the trivial choice of $c = 0$.

The rest of the work is organized as follows. In Sect. 2, we provide the model framework. In Sect. 3, we analyze properties of our proposed randomized p-values and compare them to the LFC-based ones. Section 4 presents theoretical and numerical results regarding the bias and the MSE of $\hat{\pi}_0$ when used with the proposed

randomized p-values. Section 5 illustrates the performance of resulting data-adaptive multiple tests for control of the FDR. In Sect. 6, we compare our proposed methodology with other approaches from the literature. We conclude with a discussion in Sect. 7.

2 Model setup

We consider a statistical model $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, where ϑ denotes the parameter of the model and Θ the corresponding parameter space. In the context of multiple testing, we define a derived parameter $\theta = \theta(\vartheta) = (\theta_1(\vartheta), \ldots, \theta_m(\vartheta))^\top$ with values in $\mathbb{R}^m$, $m \geq 2$. The jth component $\theta_j(\vartheta)$ of this derived parameter is assumed to be the object of interest in the jth null hypothesis H_j, $j = 1, \ldots, m$, where the family of m null hypotheses $H_1, \ldots, H_m$ and the family of their corresponding alternatives $K_1, \ldots, K_m$ consist of non-empty Borel sets of $\mathbb{R}$. For each $j = 1, \ldots, m$, we test $\theta_j(\vartheta) \in H_j$ against $\theta_j(\vartheta) \in K_j = \mathbb{R} \setminus H_j$.

We assume that for each $j = 1, \ldots, m$ a test statistic $T_j : \Omega \to \mathbb{R}$ and a rejection region $\Gamma_j(\alpha) \subset \mathbb{R}$ are given, where $\alpha \in (0, 1)$ denotes a fixed, local significance level. We denote by $x \in \Omega$ the realization of X. The test statistics $\{T_j(X)\}_{1\leq j\leq m}$ are assumed to have absolutely continuous distributions with respect to the Lebesgue measure under any $\vartheta \in \Theta$. The marginal tests φ_j for testing H_j versus K_j are given by $\varphi_j(X) = \mathbf{1}\{T_j(X) \in \Gamma_j(\alpha)\}$, where $\varphi_j(x) = 1$ means rejection of H_j in favor of K_j and $\varphi_j(x) = 0$ means that H_j is retained, for observed data x and $1 \leq j \leq m$. Note that we do not make any (general) assumptions about the dependence structure among the different test statistics at this point.

Furthermore, we make the following additional general assumptions:

(*A*1) Nested rejection regions: For every $j = 1, \ldots, m$ and $\alpha' < \alpha$, it holds that $\Gamma_j(\alpha') \subseteq \Gamma_j(\alpha)$.

(*A*2) For every $j = 1, \ldots, m$, it holds $\sup_{\vartheta:\theta_j(\vartheta)\in H_j} \mathbb{P}_\vartheta(T_j(X) \in \Gamma_j(\alpha)) = \alpha$.

(*A*3) The set of LFCs for φ_j, i.e., the set of parameter values that yield the supremum in (*A*2), does not depend on α.

Under assumption (*A*1), rejections at significance levels α' always imply rejections at larger significance levels $\alpha > \alpha'$. Assumption (*A*2) means that under any LFC for φ_j the rejection probability is exactly α.

LFC-based p-values for the marginal tests $\{\varphi_j\}_{1\leq j\leq m}$ are formally defined as

$$p_j^{LFC}(x) = \inf_{\{\tilde{\alpha}\in(0,1):\, T_j(x)\in\Gamma_j(\tilde{\alpha})\}} \sup_{\{\vartheta:\,\theta_j(\vartheta)\in H_j\}} \mathbb{P}_\vartheta(T_j(X) \in \Gamma_j(\tilde{\alpha})).$$

Under assumptions (*A*1)–(*A*3), we obtain that

$$p_j^{LFC}(X) = \inf\{\tilde{\alpha} \in (0, 1) : T_j(X) \in \Gamma_j(\tilde{\alpha})\},\ j = 1, \ldots, m. \tag{3}$$

With assumption $(A2)$, any such LFC-based p-value $p_j^{LFC}(X)$ is uniformly distributed on [0, 1] under any LFC for φ_j; cf. Lemma 3.3.1 of Lehmann and Romano (2005). Let F_ϑ be the cumulative distribution function (cdf) of $T_j(X)$ under $\vartheta \in \Theta$. If the rejection region $\Gamma_j(\alpha)$ is given by $(F_{\vartheta_0}^{-1}(1-\alpha), \infty)$, where ϑ_0 is an LFC for φ_j, then the definition in (3) simplifies to $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$. Rejection regions of that type are typical if test statistics tend to larger values under alternatives, which is often the case.

As examples, we give two models that fulfill the general assumptions $(A1)$–$(A3)$.

Example 1 (*Multiple Z-tests model*) We consider $X = (X_{i,j} : i = 1, \dots, n_j, j = 1, \dots, m)$, where $(n_j)_{j=1,\dots,m}$ are fixed sample sizes. For all j the random variables $X_{1,j}, \dots, X_{n_j,j}$ are assumed to be stochastically independent and identically normally distributed as $N(\theta_j(\vartheta), 1)$, where $\vartheta = (\vartheta_1, \dots, \vartheta_m)^\top \in \Theta = \mathbb{R}^m$ is the (main) parameter of the model and $\theta(\vartheta)$, given by $\theta_j(\vartheta) = \vartheta_j$ for $1 \le j \le m$, is the derived parameter. For each $1 \le j \le m$, we are interested in the null hypothesis $H_j : \vartheta_j \le 0$ against its alternative $K_j : \vartheta_j > 0$ and consider the test statistic $T_j(X) = n_j^{-1} \sum_{i=1}^{n_j} X_{i,j} \sim N(\vartheta_j, n_j^{-1})$. Furthermore, we let $\Gamma_j(\alpha) = (\Phi_{(0,n_j^{-1})}^{-1}(1-\alpha), \infty)$, leading to the LFC-based p-value $p_j^{LFC}(X) = 1 - \Phi_{(0,n_j^{-1})}(T_j(X))$, where $\Phi_{(\mu,\sigma^2)}$ denotes the cdf of the normal distribution on $\mathbb{R}$ with parameters μ and σ^2. For each $j = 1, \dots, m$, the set of LFCs for φ_j is $\{\vartheta \in \Theta : \vartheta_j = 0\}$, independently of α. As mentioned before, we do not specify the dependence structure of $T_{j_1}(X)$ and $T_{j_2}(X)$ for $1 \le j_1 \ne j_2 \le m$. The latter dependence structure may be regarded as a further (nuisance) parameter of the model.

Example 2 (*Two-sample means comparison model*) Let $j = 1, \dots, m$ be fixed. For given sample sizes $n_{1,j} \ge 2$ and $n_{2,j} \ge 2$, let $X_{1,j}, \dots, X_{n_{1,j},j}$ and $Y_{1,j}, \dots, Y_{n_{2,j},j}$ be jointly stochastically independent, observable random variables. Assume that $X_{1,j}, \dots, X_{n_{1,j},j}$ are identically distributed with $X_{1,j} \sim N(\theta_{1,j}(\vartheta), \sigma_j^2)$ and that $Y_{1,j}, \dots, Y_{n_{2,j},j}$ are identically distributed with $Y_{1,j} \sim N(\theta_{2,j}(\vartheta), \sigma_j^2)$, where $\sigma_j^2 > 0$ is unknown. Similarly as in Example 1, the parameter vector ϑ consists of all unknown means and all unknown variances of the model. For each $1 \le j \le m$, we compare the means of the two samples. To this end, we let $\theta_j(\vartheta) = \theta_{1,j}(\vartheta) - \theta_{2,j}(\vartheta)$ and assume that $H_j : \theta_j(\vartheta) \le 0$ versus $K_j : \theta_j(\vartheta) > 0$ is the marginal test problem of interest. Let $\bar{X}_j = n_{1,j}^{-1} \sum_{i=1}^{n_{1,j}} X_{i,j}$, $\bar{Y}_j = n_{2,j}^{-1} \sum_{i=1}^{n_{2,j}} Y_{i,j}$ and

$$S_j(X) = \frac{1}{n_{1,j} + n_{2,j} - 2} \Big[\sum_{i=1}^{n_{1,j}} (X_{i,j} - \bar{X}_j)^2 + \sum_{i=1}^{n_{2,j}} (Y_{i,j} - \bar{Y}_j)^2 \Big].$$

Under an LFC for φ_j, that is, any $\vartheta \in \Theta$ with $\theta_j(\vartheta) = 0$, the test statistic

$$T_j(X) = \sqrt{\frac{n_{1,j} n_{2,j}}{n_{1,j} + n_{2,j}}} (\bar{X}_j - \bar{Y}_j)/S_j$$

follows Student's t-distribution with $n_{1,j} + n_{2,j} - 2$ degrees of freedom, denoted by $t_{n_{1,j}+n_{2,j}-2}$. The corresponding rejection region is $\Gamma_j(\alpha) = (F^{-1}_{t_{n_{1,j}+n_{2,j}-2}}(1-\alpha), \infty)$ and the LFC-based p-value is given by $p_j^{LFC}(X) = 1 - F_{t_{n_{1,j}+n_{2,j}-2}}(T_j(X))$, where $F_{t_{n_{1,j}+n_{2,j}-2}}$ denotes the cdf of $t_{n_{1,j}+n_{2,j}-2}$. Again, the aforementioned set of LFCs for φ_j does not depend on α, for each $1 \leq j \leq m$. For the dependence structure among different coordinates $j_1 \neq j_2$, we argue as in Example 1.

3 The randomized *p*-values

3.1 General properties

Definition 1 Let a model as in Sect. 2 and a set of random variables $U_1, \ldots, U_m$ that are defined on the same probability space as X, jointly stochastically independent, identically Uni[0, 1]-distributed (under any $\vartheta \in \Theta$), and stochastically independent of the data X, be given. For each $j = 1, \ldots, m$ and given constants $c_1, \ldots, c_m$ with $cj \in [0, 1]$ for all $1 \leq j \leq m$, we define our randomized p-values as in Eq. (2), where $p_j^{rand}(X, U_j, 0) = U_j$ by convention.

For a more general definition of these p-values, we refer to the appendix. Before we discuss the properties of these randomized p-values and compare them to LFC-based ones, we give a few remarks.

Remark 1 (*a.*) If $p_j^{LFC}(X)$ is stochastically large, then it is likely that $p_j^{rand}(X, U_j, c_j) = U_j$ holds. This means that under the null hypothesis H_j, the distribution of p_j^{rand} will typically be close to a Uni[0, 1]-distribution. On the other hand, if K_j is true and $p_j^{LFC}(X)$ is stochastically small, the randomized p-value $p_j^{rand}(X, U_j, c_j)$ is more likely to be equal to $p_j^{LFC}(X)/c_j \geq p_j^{LFC}(X)$ than it is to be equal to U_j.

(*b.*) Under an LFC ϑ_0 for φ_j, the randomized p-value $p_j^{rand}(X, U_j, c_j)$ is uniformly distributed on [0, 1] for any $1 \leq j \leq m$. Namely, it holds that

$$\begin{aligned}
&\mathbb{P}_{\vartheta_0}(p_j^{rand}(X, U_j, c_j) \leq t) \\
&\quad = \mathbb{P}_{\vartheta_0}(U_j \leq t)\, \mathbb{P}_{\vartheta_0}(p_j^{LFC}(X) \geq c_j) + \mathbb{P}_{\vartheta_0}(p_j^{LFC}(X) < t c_j) \\
&\quad = t(1 - c_j) + t\, c_j = t,
\end{aligned}$$

where we have used that $p_j^{LFC}(X)$ is Uni[0, 1]-distributed under any LFC ϑ_0 for φ_j, due to assumptions (*A*1)–(*A*2), and that U_j is always Uni[0, 1]-distributed, no matter the value of ϑ.

As mentioned in Sect. 1, the use of valid p-values in the Schweder–Spjøtvoll estimator ensures that the latter has a nonnegative bias; cf. Lemma 1 of Dickhaus et al (2012). Therefore, it is of interest to give some conditions for the validity of our randomized p-values.

Theorem 1 *Let a model as in Sect. 2 be given and $j \in \{1, \ldots, m\}$ be fixed. Then, $p_j^{rand}(X, U_j, c_j)$ is a valid p-value for a given $c_j \in [0, 1]$ if and only if the following condition* (1.) *is fulfilled. Furthermore, either of the following conditions* (2.) *and* (3.) *is a sufficient condition for the validity of $p_j^{rand}(X, U_j, c_j)$ for any $c_j \in [0, 1]$.*

(1.) *For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds*

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t\, c_j) \leq t\, \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq c_j)$$

for all $t \in [0, 1]$.

(2.) *For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, $\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t)/t$ is non-decreasing in t.*
(3.) *The cdf of $p_j^{LFC}(X)$ is convex under any parameter $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$.*

If the LFC-based p-value is given by $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$, where $\vartheta_0 \in \Theta$ is an LFC for φ_j, then the following condition (4.) *is equivalent to condition* (2.)*, while condition* (5.) *is equivalent to condition* (3.)*.*

(4.) *For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds $T_j(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_j(X)^{(\vartheta_0)}$.*
(5.) *For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$.*

With $\leq_{\mathrm{hr}}$ and $\leq_{\mathrm{lr}}$, we mean the hazard rate order and the likelihood ratio order, respectively. The notation $T_j(X)^{(\vartheta)}$ refers to the distribution of $T_j(X)$ under $\vartheta \in \Theta$. The relationship $T_j(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_j(X)^{(\vartheta_0)}$ is equivalent to $(1 - F_{\vartheta_0}(t))/(1 - F_\vartheta(t))$ being non-decreasing in t, and $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$ is equivalent to $f_{\vartheta_0}(t)/f_\vartheta(t)$ being non-decreasing in t, where f_ϑ denotes the Lebesgue density of $T_j(X)$ under $\vartheta \in \Theta$.

The Proof of Theorem 1 is given in the appendix.

Corollary 1 *Under the models from Examples 1 and 2, the randomized p-values $(p_j^{rand}(X, U_j, c_j))_{1 \leq j \leq m}$ are valid for any $(c_1, \ldots, c_m)^\top \in [0, 1]^m$.*

Proof The multiple Z-tests model from Example 1 fulfills the general assumptions (A1)–(A3) from Sect. 2. Let $j \in \{1, \ldots, m\}$ be arbitrarily chosen. For a parameter value $\vartheta \in \Theta$ with $\theta_j(\vartheta) = \vartheta_j \in H_j$, i.e., $\vartheta_j \leq 0$, it is easy to show that $f_0(t)/f_{\vartheta_j}(t)$ is non-decreasing in t, where f_z denotes the Lebesgue density of the $N(z, n_j^{-1})$-distribution. Following Theorem 1, $p_j^{rand}(X, U_j, c_j)$ is valid for any constant $c_j \in [0, 1]$. The choice of $c_j = 1/2$ for all $1 \leq j \leq m$ results in the randomized p-values from Dickhaus (2013) for this model.

The two-sample means comparison model from Example 2 fulfills the general assumptions (*A*1) – (*A*3), too. Again, let $j \in \{1, \ldots, m\}$ be arbitrarily chosen. Under any parameter value $\vartheta \in \Theta$, it holds that $T_j(X) \sim t_{\tau_j, n_{1,j}+n_{2,j}-2}$, where $\tau_j = \sqrt{\frac{n_{1,j}n_{2,j}}{n_{1,j}+n_{2,j}}}\theta_j(\vartheta)/\sigma_j$, and $t_{\tau,\nu}$ denotes the non-central t-distribution with non-centrality parameter τ and ν degrees of freedom. The family $(t_{\tau,n_{1,j}+n_{2,j}-2})_{\tau\in\mathbb{R}}$ of distributions possesses the monotone likelihood ratio (MLR) property, i.e., it holds $t_{\tau_1,n_{1,j}+n_{2,j}-2} \leq_{\mathrm{lr}} t_{\tau_2,n_{1,j}+n_{2,j}-2}$ if and only if $\tau_1 \leq \tau_2$; cf. Karlin (1956) and Karlin and Rubin (1956). For a parameter value $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, i.e., $\theta_{1,j}(\vartheta) \leq \theta_{2,j}(\vartheta)$, it holds that $\tau_j = \sqrt{\frac{n_{1,j}n_{2,j}}{n_{1,j}+n_{2,j}}}\theta_j(\vartheta)/\sigma_j \leq 0$ and therefore $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$, where ϑ_0 is an LFC for φ_j, i.e., $\theta_{1,j}(\vartheta_0) = \theta_{2,j}(\vartheta_0)$. According to Theorem 1, $p_j^{rand}(X, U_j, c_j)$ is valid for any choice of the constant $c_j \in [0, 1]$ in this model. □

3.2 A comparison between the LFC-based and the randomized *p*-values

For any $1 \leq j \leq m$, we want to compare the cdfs of $p_j^{LFC}(X)$ and $p_j^{rand}(X, U_j, c_j)$. Due to the discussion below (2), this comparison is trivial for $c_j = 0$ and for $c_j = 1$, respectively. Therefore, let us assume here that c_j is bounded away from zero and from one. For example, one may for the moment assume that $c_j = 0.5$ is chosen, for concreteness.

We first note that

$$\begin{aligned}
&\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t) \\
&\quad = \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t \mid p_j^{LFC}(X) > c_j)\, \mathbb{P}_\vartheta(p_j^{LFC}(X) > c_j) \\
&\qquad + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t,\, p_j^{LFC}(X) \leq c_j),
\end{aligned} \tag{4}$$

$$\begin{aligned}
&\mathbb{P}_\vartheta(p_j^{rand}(X, U_j, c_j) \leq t) \\
&\quad = \mathbb{P}_\vartheta(U_j \leq t)\, \mathbb{P}_\vartheta(p_j^{LFC}(X) > c_j) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t c_j).
\end{aligned} \tag{5}$$

Now, if the value of the derived parameter $\theta_j(\vartheta)$ is so "deep inside" H_j that $\mathbb{P}_\vartheta(p_j^{LFC}(X) > c_j)$ is large, then the first summands in (4) and (5) dominate the second ones, and we see that

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t \mid p_j^{LFC}(X) > c_j) \leq \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t) \leq t = \mathbb{P}_\vartheta(U_j \leq t).$$

Thus, provided that $p_j^{rand}(X, U_j, c_j)$ is a valid p-value, its distribution under H_j will typically be closer to Uni[0, 1] than that of $p_j^{LFC}(X)$.

However, if ϑ is such that K_j is true instead and that $\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq c_j)$ is large, it holds that

$$\begin{aligned}\mathbb{P}_\vartheta(p_j^{LFC}(X) \le t,\, p_j^{LFC}(X) \le c_j) &= \mathbb{P}_\vartheta(p_j^{LFC}(X) \le \min(t, c_j)) \\ &\ge \mathbb{P}_\vartheta(p_j^{LFC}(X) \le t c_j).\end{aligned}$$

Thus, under K_j the cdf of $p_j^{LFC}(X)$ will typically be pointwise larger than the cdf of $p_j^{rand}(X, U_j, c_j)$.

The former heuristic argumentation cannot be made mathematically rigorous in general. However, if condition (3.) in Theorem 1 is fulfilled, p_j^{rand} does indeed always lie between U_j and p_j^{LFC} under the null hypothesis H_j, in the sense of the stochastic order. The same holds under the alternative K_j, if a condition similar to (3.) is fulfilled in the case of $\theta_j(\vartheta) \in K_j$.

Theorem 2 *Let a model as in Sect. 2 be given and $j \in \{1, \dots, m\}$ be fixed.*

If the cdf of $p_j^{LFC}(X)$ is convex under a fixed $\vartheta \in \Theta$, then

$$p_j^{rand}(X, U_j, c_j)^{(\vartheta)} \le_{\mathrm{st}} p_j^{rand}(X, U_j, \tilde{c}_j)^{(\vartheta)}$$

for any $0 \le c_j \le \tilde{c}_j \le 1$.

If the cdf of $p_j^{LFC}(X)$ is concave under a fixed $\vartheta \in \Theta$, then it holds that

$$p_j^{rand}(X, U_j, \tilde{c}_j)^{(\vartheta)} \le_{\mathrm{st}} p_j^{rand}(X, U_j, c_j)^{(\vartheta)}$$

for any $0 \le c_j \le \tilde{c}_j \le 1$.

We give the Proof of Theorem 2 in the appendix.

Remark 2 Let $j \in \{1, \dots, m\}$ be fixed.

1. If the jth LFC-based p-value is given by $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$, where ϑ_0 is an LFC for φ_j, then $p_j^{LFC}(X)$ has a convex cdf under $\vartheta \in \Theta$ if and only if $T_j(X)^{(\vartheta)} \le_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$ and a concave cdf under $\vartheta \in \Theta$ if and only if $T_j(X)^{(\vartheta_0)} \le_{\mathrm{lr}} T_j(X)^{(\vartheta)}$ (cf. the Proof of Theorem 1 in appendix).
2. If condition (3.) from Theorem 1 is fulfilled, then Theorem 2 implies

$$U_j \le_{\mathrm{st}} p_j^{rand}(X, U_j, c_j)^{(\vartheta)} \le_{\mathrm{st}} p_j^{LFC}(X)^{(\vartheta)}$$

for all $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$ and any $c_j \in [0, 1]$. This also implies the validity of $p_j^{rand}(X, U_j, c_j)$, as it was claimed in Theorem 1.

3. The cdf of $p_j^{LFC}(X)$ can never be concave under H_j.

Corollary 2 *For the multiple Z-tests model and the two-sample means comparison model from Examples 1 and 2, respectively, it holds for any* $1 \leq j \leq m$ *and any* $c_j \in [0,1]$, *that*

$$U_j \leq_{\mathrm{st}} p_j^{rand}(X, U_j, c_j)^{(\vartheta)} \leq_{\mathrm{st}} p_j^{LFC}(X)^{(\vartheta)}$$

under any $\vartheta \in \Theta$ *with* $\theta_j(\vartheta) \in H_j$, *as well as*

$$p_j^{LFC}(X)^{(\vartheta)} \leq_{\mathrm{st}} p_j^{rand}(X, U_j, c_j)^{(\vartheta)} \leq_{\mathrm{st}} U_j$$

under any ϑ *with* $\theta_j(\vartheta) \in K_j$.

We conclude this section by illustrating the assertions of Theorem 2 and Corollary 2 under the multiple Z-tests model. In Fig. 1, we compare the cdfs of $p_j^{rand}(X, U_j, c)$ for an arbitrary $j \in \{1, \ldots, m\}$ with $c = 0, 0.25, 0.5, 0.75$ and 1 under $\vartheta \in \Theta$, where we set $\theta_j(\vartheta) = -1/\sqrt{n_j}$ or $\theta_j(\vartheta) = 1/\sqrt{n_j}$ for $n_j = 50$, respectively. It is apparent that the cdfs move from that of the Uni$[0,1]$-distribution to the one of $p_j^{LFC}(X)$ with increasing c.

4 Estimation of the proportion of true null hypotheses

4.1 The expected value of the Schweder–Spjøtvoll estimator

We consider the usage of $\{p_j^{rand}(X, U_j, c_j)\}_{1 \leq j \leq m}$ in the Schweder–Spjøtvoll estimator $\hat{\pi}_0 \equiv \hat{\pi}_0(\lambda)$ defined in (1). It can easily be seen from the representation on the right-hand side of (1) that the bias of $\hat{\pi}_0(\lambda)$ decreases if $\mathbb{E}_\vartheta[\hat{F}_m(\lambda)]$ increases, under any $\vartheta \in \Theta$. Thus, in terms of bias reduction of $\hat{\pi}_0(\lambda)$ (for a fixed, given value of λ)

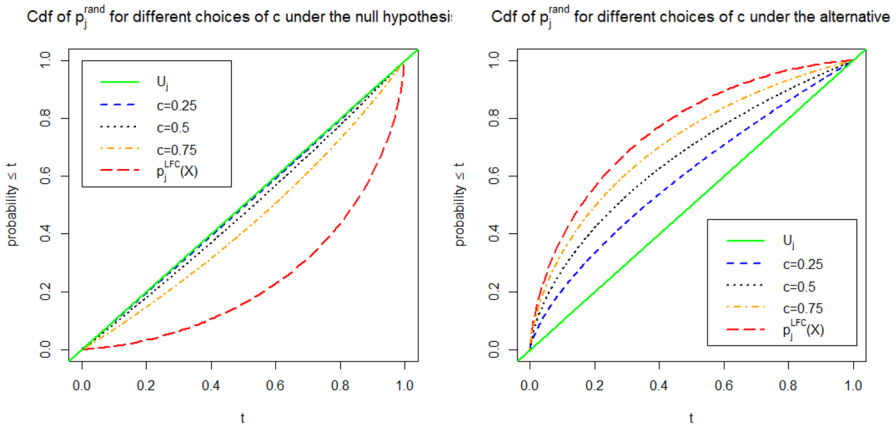

Fig. 1 A comparison of the cdfs of $p_j^{rand}(X, U_j, c)$ for $c \in \{0, 0.25, 0.5, 0.75, 1\}$ under the multiple Z-tests model. In the left graph, $\theta_j(\vartheta) = -1/\sqrt{n_j}$ and in the right graph, $\theta_j(\vartheta) = 1/\sqrt{n_j}$, where $n_j = 50$. The value of $j \in \{1, \ldots, m\}$ is arbitrary

stochastically small (randomized) p-values (with pointwise large cdfs) are most suitable. In order to avoid a negative bias of $\hat{\pi}_0(\lambda)$, we furthermore have to ensure validity of the p-values utilized in $\hat{\pi}_0(\lambda)$. Hence, if the cdfs of the LFC-based p-values are convex under null hypotheses and concave under alternatives, the optimal ("oracle") value of c_j is zero whenever H_j is true and one whenever K_j is true; cf. Theorem 2. This is also in line with Remark 6 of Dickhaus et al (2012), who showed that $\hat{\pi}_0(\lambda)$ is unbiased if the p-values utilized in $\hat{\pi}_0(\lambda)$ are Uni[0, 1]-distributed under true null hypotheses and almost surely smaller than λ under false null hypotheses. Under the restriction of identical c_j's, i.e., $c_1 = c_2 = \ldots = c_m \equiv c$, one may expect that an optimal ("oracle") value of c (leading to a small, but nonnegative bias of $\hat{\pi}_0(\lambda)$) should be close to $1 - \pi_0$. The restriction $c_1 = c_2 = \ldots = c_m \equiv c$ will be made throughout the remainder for computational convenience and feasibility.

Definition 2 The Schweder–Spjøtvoll estimator $\hat{\pi}_0(\lambda)$, if used with $p_1^{rand}(X, U_1, c)$, $\ldots, p_m^{rand}(X, U_m, c)$, will be denoted by $\hat{\pi}_0(\lambda, c)$ throughout the remainder. Notice that in the estimators $\hat{\pi}_0(\lambda, 0)$ and $\hat{\pi}_0(\lambda, 1)$, respectively, we use $U_1, \ldots, U_m$ (as the marginal p-values) and $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$, respectively. Furthermore, we consider the function $h_\vartheta : [0, 1) \times [0, 1] \to \mathbb{R}$, given by $h_\vartheta(\lambda, c) = \mathbb{E}_\vartheta[\hat{\pi}_0(\lambda, c)]$, where $\vartheta \in \Theta$ is the underlying parameter value.

Lemma 1 *For every $\lambda \in [0, 1)$ and under any $\vartheta \in \Theta$, $h_\vartheta(\lambda, 0) = 1$. If the cdfs of the $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$ are continuous under ϑ, then there exists a minimizing argument $c^\star \in [0, 1]$ of $h_\vartheta(\lambda, \cdot)$.*

Proof In the case of $c = 0$, $p_j^{rand}(X, U_j, 0) = U_j$ for each $j \in \{1, \ldots, m\}$ and $\mathbb{E}_\vartheta[\hat{\pi}_0(\lambda, 0)] = (1 - \lambda)/(1 - \lambda) = 1$, proving the first assertion.

In order to show the second assertion, we note that under any $\vartheta \in \Theta$

$$\mathbb{E}_\vartheta[\hat{F}_m(\lambda)] = \sum_{j=1}^{m} \left[\lambda \, \mathbb{P}_\vartheta(p_j^{LFC}(X) \geq c) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq c\lambda) \right]. \tag{6}$$

The right-hand side of (6) is continuous in c if the cdfs of the p-values $p_1^{LFC}(X)$, $\ldots, p_m^{LFC}(X)$ are continuous under ϑ. Since [0, 1] is a compact set, the function $h_\vartheta(\lambda, \cdot)$ attains a minimum on [0, 1], by the extreme value theorem. □

For an illustration, let us consider the multiple Z-tests model from Example 1, where we set the total number of null hypotheses to $m = 1{,}000$ and the sample sizes to $n_j = 50$ for all $j = 1, \ldots, m$. As mentioned before, the choice of $c = 1/2$ leads to the randomized p-values as defined in Dickhaus (2013) for this model. Figure 2 displays the graphs of the function $c \mapsto h_\vartheta(1/2, c)$ for two different parameter values $\vartheta \in \Theta$ under this model. In both cases, $\pi_0 = 0.7$ (meaning that 700 null hypotheses are true and 300 are false) and $\theta_j(\vartheta) = 2.5/\sqrt{50}$ whenever H_j is false.

In the left graph of Fig. 2, $\theta_j(\vartheta) = -1/\sqrt{50}$ whenever H_j is true. The minimum of $c \mapsto h_\vartheta(1/2, c)$ is attained at $c^\star = 0.3276$ and yields $\mathbb{E}_\vartheta[\hat{\pi}_0(1/2, c^\star)] = 0.7508$. It is apparent that $h_\vartheta(1/2, c)$ is largest for $c = 1$, that is, when utilizing the LFC-based

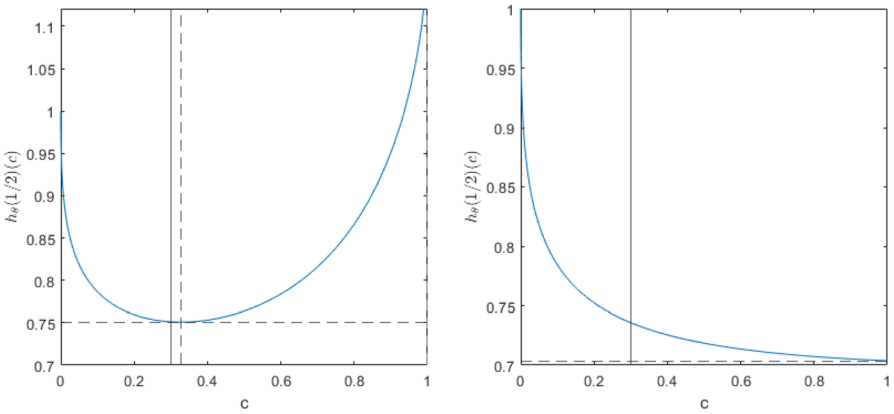

Fig. 2 Two plots of $c \mapsto h_\vartheta(1/2, c)$ for $c \in [0, 1]$ under the multiple Z-tests model. We set $\pi_0 = 0.7$ and $\vartheta \in \Theta$ such that $\theta_j(\vartheta) = 2.5/\sqrt{50}$ if K_j is true, $j = 1, \ldots, m = 1{,}000$. The parameter value under each null is $\theta_j(\vartheta) = -1/\sqrt{50}$ in the left graph and $\theta_j(\vartheta) = 0$ (leading to uniform null p-values) in the right graph. The solid vertical line indicates $c = 1 - \pi_0$, while the dashed one indicates the minimizing argument $c^\star$ of $c \mapsto h_\vartheta(1/2, c)$. The dashed horizontal line indicates $h_\vartheta(1/2, c^\star)$

p-values $\{p_j^{LFC}(X)\}_{1 \le j \le m}$. Finally, we see that the optimal bias of $\hat{\pi}_0(1/2)$ when using the same $c_j \equiv c$ for all $1 \le j \le m$ is larger than zero.

In the right graph of Fig. 2, $\theta_j(\theta) = 0$ whenever H_j is true. In this case, the estimator $\hat{\pi}_0(1/2, 1)$ has the lowest bias among all estimators $\{\hat{\pi}_0(1/2, c) : c \in [0, 1]\}$, meaning that $c^\star = 1$. This is because for every j with $\theta_j(\vartheta) \in H_j$, ϑ is an LFC for φ_j and thus $p_j^{LFC}(X)$ is Uni$[0, 1]$-distributed under ϑ. In such cases, $p_j^{rand}(X, U_j, c)$ is Uni$[0, 1]$-distributed for any c under H_j, while $p_j^{LFC}(X)^{(\vartheta)} \le_{\text{st}} p_j^{rand}(X, U_j, c)^{(\vartheta)}$ if K_j is true, due to Theorem 2.

4.2 Minimizing the MSE

From a decision-theoretic perspective, the bias alone is not enough to judge the estimation quality of $\hat{\pi}_0$. A more commonly used criterion for the quality of an estimator is its MSE. Therefore, we investigate the MSE of $\hat{\pi}_0$, when using the randomized p-values, in this section. The bias of $\hat{\pi}_0(\lambda)$ does not depend on the dependence structure of the utilized marginal p-values $p_1, \ldots, p_m$. To see this, notice that $\mathbb{E}_\vartheta[\hat{F}_m(\lambda)] = m^{-1} \sum_{j=1}^m \mathbb{P}_\vartheta(p_j \le \lambda)$. Calculating the variance of $\hat{\pi}_0(\lambda)$, however, requires the knowledge of the dependence structure of the p-values, because

$$\mathrm{Var}_\vartheta\big(\hat{F}_m(\lambda)\big) = \frac{1}{m^2}\left[\sum_{j=1}^m \mathrm{Var}_\vartheta\big(\mathbf{1}\{p_j \le \lambda\}\big) + \sum_{i \ne j} \mathrm{Cov}_\vartheta\big(\mathbf{1}\{p_i \le \lambda\}, \mathbf{1}\{p_j \le \lambda\}\big)\right].$$

4.2.1 The independent case

Here, we present some results regarding the variance of $\hat{\pi}_0$ when the marginal LFC-based p-values are stochastically independent. In this, we assume that the cdf of p_j^{LFC} is convex under H_j and concave under K_j. This assumption is often fulfilled, especially in the models studied before.

Lemma 2 *Assume that $p_1^{LFC}(X), \dots, p_m^{LFC}(X)$ are stochastically independent. Then, the variance of $\hat{\pi}_0(1/2)$ is monotonically decreasing in $c \in [0,1]$ if used with $p_1^{rand}(X, U_1, c), \dots, p_m^{LFC}(X, U_m, c)$. Furthermore, the maximum variance of $\hat{\pi}_0(1/2)$, among all $c \in [0,1]$, equals $1/m$.*

Proof Since $\mathrm{Var}_\vartheta\big(\hat{\pi}_0(\lambda)\big) = (1-\lambda)^{-2}\mathrm{Var}_\vartheta\big(\hat{F}_m(\lambda)\big)$, we have to show that the latter is decreasing in c, where $\hat{F}_m(\lambda)$ is the ecdf of the randomized p-values $p_1^{rand}(X, U_1, c), \dots, p_m^{LFC}(X, U_m, c)$ at point λ.

If the p-values are independent, then so are the randomized p-values, thus

$$\mathrm{Var}_\vartheta\big(\hat{F}_m(\lambda)\big) = \frac{1}{m^2}\sum_{j=1}^{m}\mathrm{Var}_\vartheta\Big(\mathbf{1}\{p_j^{rand}(X, U_j, c) \le \lambda\}\Big).$$

We can show that each summand $\mathrm{Var}_\vartheta\Big(\mathbf{1}\{p_j^{rand}(X, U_j, c) \le \lambda\}\Big)$ is decreasing in c, $j = 1, \dots, m$.

For a fixed j, it holds $\mathrm{Var}_\vartheta\Big(\mathbf{1}\{p_j^{rand}(X, U_j, c) \le \lambda\}\Big) = f(c) - f(c)^2$, where f is the cdf of $p_j^{rand}(X, U_j, c)$ at point λ. Furthermore, it holds

$$\frac{d}{dc}\big(f(c) - f(c)^2\big) = f'(c)(1 - 2f(c)).$$

Due to Theorem 2 $f'(c)$ is non-positive under H_j and nonnegative under K_j for all $c \in [0,1]$. More particularly, since $f(0) = \mathbb{P}\big(U_j \le \lambda = 1/2\big) = 1/2$ the term $1 - 2f(c)$ is non-positive under H_j and nonnegative under K_j. In total, it holds

$$\frac{d}{dc}\big(f(c) - f(c)^2\big) \le 0.$$

For the maximum variance, we have to plug $c = 0$ into the variance formula, for which the randomized p-values are the uniformly distributed $U_1, \dots, U_m$. For these, it holds $\mathrm{Var}_\vartheta\big(\hat{F}_m(\lambda)\big) = \lambda(1-\lambda)/m$ and $\mathrm{Var}_\vartheta\big(\hat{\pi}_0(\lambda)\big) = \lambda/((1-\lambda)m) = 1/m$ if $\lambda = 1/2$. □

Lemma 2 implies that in case of stochastically independent LFC-based p-values, randomization increases the variance of $\hat{\pi}_0(1/2)$. However, if m is large, the variance of $\hat{\pi}_0(1/2)$ has a small impact on the MSE, because the bias of $\hat{\pi}_0(1/2)$ does not explicitly depend on m. We demonstrate this with an example. As in Sect. 3.1, we consider the multiple Z-tests model from Example 1 for the

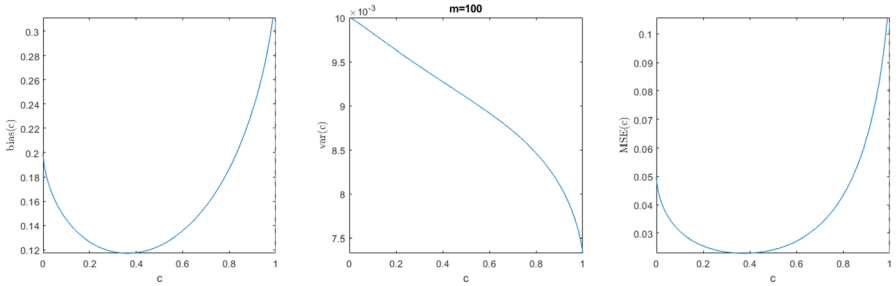

Fig. 3 Plots of the bias, the variance and the MSE of $\hat{\pi}_0(1/2)$, respectively, as functions of c, when used with the p-values $p_1^{rand}(X, U_1, c), \ldots, p_m^{LFC}(X, U_m, c)$ under the multiple Z-tests model as described in Sect. 4.2 with $m = 100$. Here, we assume stochastically independent test statistics

Table 1 MSE-minimizing value c^{MSE} of c for different numbers of hypotheses m under the model and the parameter setting described in Sect. 4.2.1. Independently of m, the bias minimizing parameter $c^\star$ equals 0.3626 here

m	c^{MSE}
1	1
10	0.4759
50	0.3858
100	0.3742
500	0.3649
1, 000	0.3638
10, 000	0.3627
∞	$c^\star = 0.3626$

choices of $\theta_j(\vartheta) = -0.5/\sqrt{50}$ if H_j is true and $\theta_j(\vartheta) = 1.5/\sqrt{50}$ otherwise. We set $\pi_0 = 0.8$, $\lambda = 1/2$ and $m = 100$.

Figure 3 displays the bias, the variance and the MSE of $\hat{\pi}_0(\lambda)$ when used with the randomized p-values $p_1^{rand}(X, U_1, c), \ldots, p_m^{LFC}(X, U_m, c)$ as functions of $c \in [0, 1]$. The bias curve starts at $1 - \pi_0$ for $c = 0$ and has its minimum 0.1171 at $c = c^\star = 0.3626$. For $c > c^\star$, it monotonically increases to 0.3331 at $c = 1$. The variance curve starts at $1/m = 0.01$ for $c = 0$ and decreases with increasing c, as expected. The MSE curve is mostly affected by the (squared) bias curve. The MSE-minimizing value c^{MSE} of c equals 0.3742, which is slightly larger than $c^\star$. For our choice of the parameters, the MSE has its minimum at $c = 1$ only for $m = 1$; the MSE curve has a u-shape for all $m \geq 2$.

Since the variance of $\hat{\pi}_0(1/2)$ decreases with c and is upper-bounded by $1/m$, the MSE-minimizing value c^{MSE} of c converges to $c^\star$ from above. We calculated c^{MSE} for some values of m, where we used the same model and settings as before, cf. Table 1.

4.2.2 The positively dependent case

In this section, we consider positively dependent p-values $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$. We calculate the variance and MSE curves for varying degrees of positive

dependence. Again, we consider the multiple Z-tests model from Example 1, where the test statistics $T_1(X), \ldots, T_m(X)$ now have a pairwise correlation coefficient $\mathrm{Corr}(T_i(X), T_j(X)) = \rho \in [0, 1]$, for all $i \neq j$. One further model class, referring to Gumbel–Hougaard copula dependency structures, is considered in the supplementary material.

Apart from ρ, we choose the same model settings as in the previous section. As mentioned before, the bias curve as well as $c^\star$ do not depend on the dependence structure of the p-values. However, we included it in the following figure to facilitate the comparison of MSE and (squared) bias. We considered $\rho = 0.5$ in Fig. 4. The independent case, which we have considered in the previous section, corresponds to $\rho = 0$.

As displayed in Fig. 4, the variance of $\hat{\pi}_0(\lambda)$ increases monotonically in c here, which means that the variance of $\hat{\pi}_0(\lambda)$ decreases with the amount of randomization. The overall magnitude of the variances is also much higher now when compared with the case of $\rho = 0$. The MSE-minimizing value of c is 0.2642 here, meaning that the optimal amount of randomization is higher for $\rho = 0.5$ than for $\rho = 0$.

This example illustrates that the presence of positive dependence among the marginal p-values can deteriorate the performance of the Schweder–Spjøtvoll estimator. Similar results in this direction have recently been obtained by Neumann et al (2021). Since the randomized p-values are a mix of the LFC-based p-values and the stochastically independent random variables $U_1, \ldots, U_m$, the degree of dependence is lower among the randomized p-values and thus also the variance of the Schweder–Spjøtvoll estimator when used with the latter.

4.3 Estimating π_0 in practice

The expected value in $h_\vartheta(\lambda, c) = \mathbb{E}_\vartheta[\hat{\pi}_0(\lambda, c)]$ discussed in Sect. 4.1 refers to the joint distribution of $\{U_j\}_{1 \leq j \leq m}$ and the data X under ϑ. In practice, the distribution of X under ϑ is unknown, but we have a realized data sample $X = x \in \Omega$ at hand, from which $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$ can be computed. Throughout this section, let us assume a statistical model such that any of the conditions (2.)–(5.) from Theorem 1 is fulfilled, so that $p_1^{rand}(X, U_1, c), \ldots, p_m^{rand}(X, U_m, c)$ are valid p-values for any $c \in [0, 1]$.

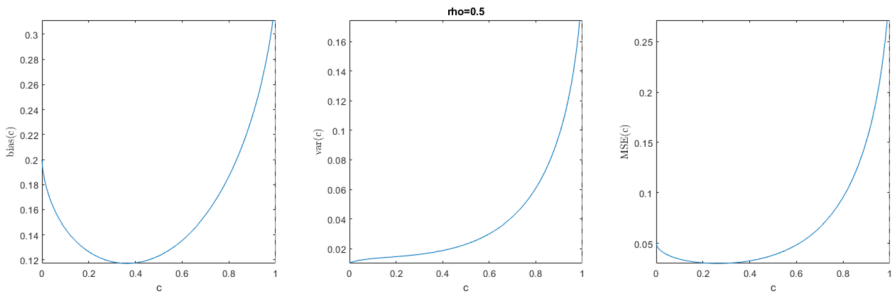

Fig. 4 Plots of the bias, the variance and the MSE of $\hat{\pi}_0(1/2)$, respectively, as functions of c, when used with the p-values $p_1^{rand}(X, U_1, c), \ldots, p_m^{LFC}(X, U_m, c)$ under the multiple Z-tests model as described in Sect. 4.2 with $m = 100$. Here, the test statistics have a pairwise correlation of $\rho = 0.5$

In analogy to (6), we obtain that the conditional expected value (with respect to the U_j's) of $\hat{\pi}_0(\lambda, c)$ under the condition $X = x$ is given by

$$\mathbb{E}[\hat{\pi}_0(\lambda, c) \mid X = x] = \frac{1}{1-\lambda}\left[1 - \frac{1}{m}\sum_{j=1}^{m}\left[\lambda\,\mathbf{1}\{p_j^{LFC}(x) \geq c\} + \mathbf{1}\{p_j^{LFC}(x) \leq \lambda c\}\right]\right]. \tag{7}$$

Our proposal for practical purposes is to minimize (7) with respect to $c \in [0, 1]$, for fixed $\lambda \in [0, 1)$. Thus, this approach focuses on minimizing the conditional bias of $\hat{\pi}_0$, given the data. Denoting the solution of this minimization problem by c_0, we then propose to utilize $p_1^{rand}(x, U_1, c_0), \ldots, p_m^{rand}(x, U_m, c_0)$ in $\hat{\pi}_0(\lambda)$.

Minimizing (7) with respect to $c \in [0, 1]$ is equivalent to maximizing the function $c \mapsto g_x(\lambda, c)$, given by

$$g_x(\lambda, c) = \sum_{j=1}^{m}\left(\lambda\,\mathbf{1}\{p_j^{LFC}(x) \geq c\} + \mathbf{1}\{p_j^{LFC}(x) \leq \lambda c\}\right), \tag{8}$$

with respect to $c \in [0, 1]$. Hence, the solution c_0 is such that most of the (realized) LFC-based p-values are outside of the interval $(\lambda c_0, c_0)$. An optimal choice c_0 can be determined numerically by either evaluating $g_x(\lambda, \cdot)$ on a given grid $0 = c_0 < \cdots < c_N = 1$ or on the set $\{p_1^{LFC}(x), \ldots, p_m^{LFC}(x), p_1^{LFC}(x)/\lambda, \ldots, p_m^{LFC}(x)/\lambda\}$ (excluding values larger than 1). Notice that $g_x(\lambda, \cdot)$ can only change its values at points from the second set.

We demonstrate this procedure with an example. Again, consider the multiple Z-tests model and the same parameter setting as for deriving the left graph in Fig. 2. Under these settings, we randomly drew one sample $x \in \Omega$ and applied the proposed procedure with $\lambda = 1/2$. After the removal of elements exceeding one from the set $\{p_1^{LFC}(x), \ldots, p_m^{LFC}(x), 2p_1^{LFC}(x), \ldots, 2p_m^{LFC}(x)\}$, 1, 406 relevant points remained for the evaluation of $g_x(1/2, \cdot)$. As displayed in Fig. 5, the maximum of $g_x(1/2, \cdot)$ is for the observed x attained at $c_0 = 0.3286$. This is an optimal c given the realized values $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$. For comparison, recall that we have

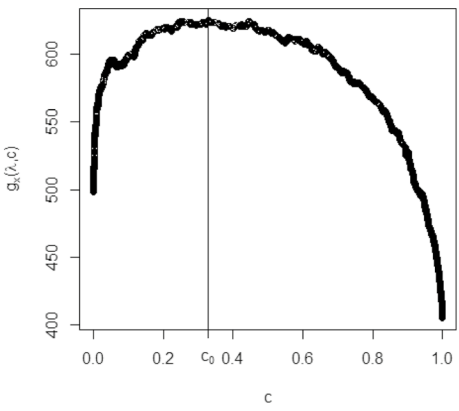

Fig. 5 A plot of the function $c \mapsto g_x(\lambda, c)$, for $\lambda = 1/2$, evaluated on those 1,406 elements of the set $\{p_1^{LFC}(x), \ldots, p_m^{LFC}(x), p_1^{LFC}(x)/\lambda, \ldots, p_m^{LFC}(x)/\lambda\}$ which are not larger than one. Here, $g_x(\lambda, \cdot)$ attains its maximum at $c_0 = 0.3286$. The underlying data x have randomly been drawn under the multiple Z-tests model

seen in Sect. 4.1 that $c^\star = 0.3276$ minimizes the bias of $\hat{\pi}_0(1/2, c)$ on average over $X \sim \mathbb{P}_\vartheta$.

Figure 6 displays the ecdfs pertaining to $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$ and $p_1^{rand}(x, u_1, c_0), \ldots, p_m^{rand}(x, u_m, c_0)$, respectively, where $\{u_1, \ldots, u_m\}$ is one particular set of realizations of the random variables $U_1, \ldots, U_m$. Furthermore, the two dotted vertical lines in Fig. 6 indicate the interval $[c_0/2, c_0]$. Recall that c_0 is chosen such that most of the (realized) LFC-based p-values are outside of the latter interval. This can visually be confirmed, since the ecdf pertaining to $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$ is rather flat on $[c_0/2, c_0]$.

For any ecdf $t \mapsto \hat{F}_m(t)$ utilized in $\hat{\pi}_0(\lambda)$, the offset at $t = 0$ of the straight line connecting the points (1, 1) and $(\lambda, \hat{F}_m(\lambda))$ equals $1 - \hat{\pi}_0(\lambda)$; cf., e.g., Figure 3.2.(b) in Dickhaus (2014). We therefore obtain an accurate estimate of π_0 if the ecdf $t \mapsto \hat{F}_m(t)$ utilized in $\hat{\pi}_0(\lambda)$ is at $t = \lambda$ close to the straight line connecting the points (1, 1) and $(0, 1 - \pi_0)$. The latter "optimal" line is the expected ecdf of marginal p-values that are Uni[0, 1]-distributed under the null and almost surely equal to zero under the alternative. In Fig. 6, the ecdf pertaining to $p_1^{rand}(x, u_1, c_0), \ldots, p_m^{rand}(x, u_m, c_0)$ is much closer to that optimal line than the ecdf pertaining to $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$. Consequently, for this particular dataset the estimation approach based on $p_1^{rand}(x, u_1, c_0), \ldots, p_m^{rand}(x, u_m, c_0)$ leads to a much more precise estimate of $\pi_0 = 0.7$ then the one based on $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$. The estimate based on $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$ even exceeds one in this example. We have

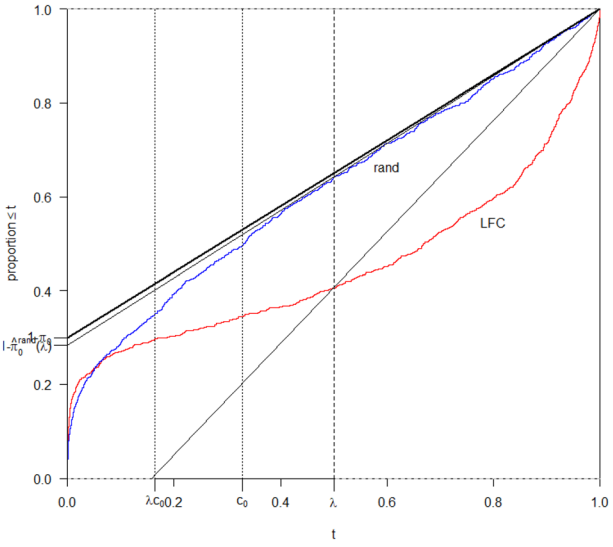

Fig. 6 Ecdfs $\hat{F}_m$ of $(p_j^{LFC}(x))_{j=1,\ldots,m}$ and $(p_j^{rand}(x, u_j, c_0))_{j=1,\ldots,m}$, respectively, under the multiple Z-tests model for $\pi_0 = 0.7$. The underlying data x are the same as in Fig. 5. The thicker straight line connects the points $(0, 1 - \pi_0)$ and (1, 1), while the two thinner straight lines connect $(\lambda, \hat{F}_m(\lambda))$ with (1, 1) for the two aforementioned ecdfs. The offset of each of the two thinner lines at $t = 0$ equals $1 - \hat{\pi}_0(\lambda)$ for the respective ecdf, where $\lambda = 1/2$. The two dotted vertical lines indicate the interval $[\lambda c_0, c_0]$, where c_0 is as in Fig. 5

repeated this simulation several times (results not included here) and the conclusions have always been rather similar.

5 Impact on data-adaptive multiple tests

A multiple test procedure $\varphi : \Omega \to \{0,1\}^m$ is a measurable mapping, such that, for each j, $\varphi_j(x) = 1$ means that we reject H_j on the basis of the observed data x and $\varphi_j(x) = 0$ means that H_j is retained. The false discovery rate (FDR) of a given multiple test procedure φ under ϑ is given by

$$\mathrm{FDR}_\vartheta(\varphi) = \mathbb{E}_\vartheta\left[\frac{V}{\max(R,1)}\right],$$

where V is the (random) number of false rejections and R is the total (random) number of rejections. We say that φ controls the FDR at level $\alpha \in (0,1)$, if $\sup_{\vartheta\in\Theta}\mathrm{FDR}_\vartheta(\varphi) \leq \alpha$ holds true.

Throughout this section, we consider the so-called linear step-up (LSU) procedure by Benjamini and Hochberg (1995), which works as follows. Given the ordered marginal p-values $p_{(1)} \leq \cdots \leq p_{(m)}$, the correspondingly ordered null hypotheses $H_{(1)}, \ldots, H_{(m)}$, and thresholds $\delta_j = \alpha j/m$, $j = 1, \ldots, m$, let $k = \max\{1 \leq j \leq m : p_{(j)} \leq \delta_j\}$. If there is no such k, we set $k = 0$. Then, we reject all null hypotheses $H_{(j)}$ for which $j \leq k$.

If the marginal p-values are jointly stochastically independent, the LSU test controls the FDR at level $\alpha\pi_0$. Thus, it is possible to replace the thresholds δ_j by δ_j/π_0, $j = 1, \ldots, m$, while still controlling the FDR at level α. Note that this leads to larger thresholds for each j. Indexwise larger thresholds increase the power of the resulting multiple test. However, as π_0 is unknown, we instead use $\tilde{\delta}_j = \delta_j G$, where G is an estimator for $1/\pi_0$ based on the marginal p-values. A step-up procedure with these kinds of thresholds is called a data-adaptive multiple test procedure.

An important result is Theorem 11 in Blanchard and Roquain (2009). It provides a condition under which a given measurable, coordinatewise non-increasing function $G : [0,1]^m \to (0,\infty)$ and jointly independent p-values leads to FDR control at level α of the resulting data-adaptive LSU test. This condition is given by

$$\mathbb{E}_\vartheta[G_{j\to 0}] \leq \pi_0^{-1} \tag{9}$$

for all $j \in \{1, \ldots, m\}$ and all parameter values $\vartheta \in H_j$. The notation $G_{j\to 0}$ means that the jth p-value is replaced by zero when G is applied.

The estimator $G = 1/\hat{\pi}_0(\lambda)$ based on the Schweder–Spjøtvoll estimator does not fulfill the condition in (9), but $G = 1/\hat{\pi}_0^+(\lambda)$ based on the modified, more conservative Storey estimator

$$\hat{\pi}_0^+(\lambda) = \frac{1 - \hat{F}_m(\lambda) + 1/m}{1-\lambda} = \hat{\pi}_0(\lambda) + \frac{1}{(1-\lambda)m}$$

does, cf. Storey (2002). Since this result applies to any set of valid p-values, we can use our randomized p-values in the data-adaptive LSU test with $G = 1/\hat{\pi}_0^+(\lambda)$.

Remark 3 In the previous sections, we focused on the Schweder–Spjøtvoll estimator. However, it is clear that $\mathrm{Var}(\hat{\pi}_0^+(\lambda)) = \mathrm{Var}(\hat{\pi}_0(\lambda))$ and $\mathrm{bias}(\hat{\pi}_0^+(\lambda)) = \mathrm{bias}(\hat{\pi}_0(\lambda)) + [(1-\lambda)m]^{-1}$. Consequently, the same bias minimizing value $c^\star$ of c applies to both estimators.

In the remainder of this section, we assess how well the adaptive LSU test with $G = 1/\hat{\pi}_0^+(\lambda)$ performs when used with our randomized p-values with $c = c^\star$ and when used with the LFC-based p-values. We consider both the independent and the positively dependent case, although FDR control is not guaranteed in the latter case. To this end, we employ the same model as in Sect. 4.2 and in Section 3.4 of Blanchard and Roquain (2009). We set $\pi_0 = 0.75$ and consider pairwise correlations $\rho \in \{0, 0.25, 0.5, 0.75\}$ of the test statistics. The underlying parameter values have been set to $\theta_j(\vartheta) = -0.2r$ if H_j is true, and $\theta_j(\vartheta) = 1 + 0.25r$ if K_j is true, $j = 1, \ldots, m$, where $r \in \{1, \ldots, 10\}$ denotes a signal strength which is expressed in units of the standard deviation of the test statistics.

Figures 7, 8 and 9 display the averaged π_0 estimations (i.e., the values of $1/G = \hat{\pi}_0^+$), the FDR and the "performances," respectively. Under the performance of a multiple test, we mean a function that increases in the power of the test and decreases in its FDR. We chose

$$\mathrm{Performance}_\vartheta(\varphi) = \mathrm{Power}_\vartheta(\varphi) - \frac{\mathrm{FDR}_\vartheta(\varphi)}{\alpha}, \tag{10}$$

where power denotes the proportion of correctly rejected null hypotheses. All values displayed in Figs. 7, 8, and 9 have been calculated via Monte Carlo simulations with 100, 000 repetitions. The subplots in each figure correspond to different values for ρ, and we varied the signal strength parameter r on the horizontal axis. In each simulation, the randomized p-values with $c = c^\star$ have only been employed in G, while the LFC-based p-values have been used in the comparison with $(\tilde{\delta}_j)_{1 \le j \le m}$. The reason for

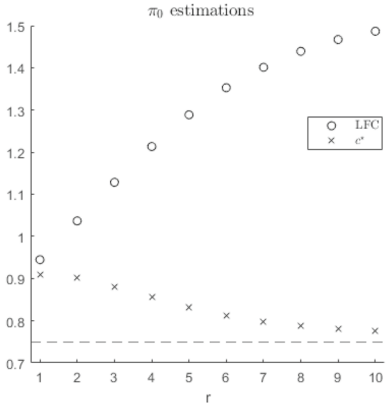

Fig. 7 Monte Carlo averages (referring to 100, 000 simulation runs) of the π_0 estimations that are utilized in the adaptive LSU tests described in Sect. 5. Considered is the modified Storey estimator with $\lambda = 1/2$, used with the LFC-based p-values (circles) or the randomized p-values with $c = c^\star$ (crosses), respectively. The signal strength is denoted by r. The horizontal dashed line indicates π_0

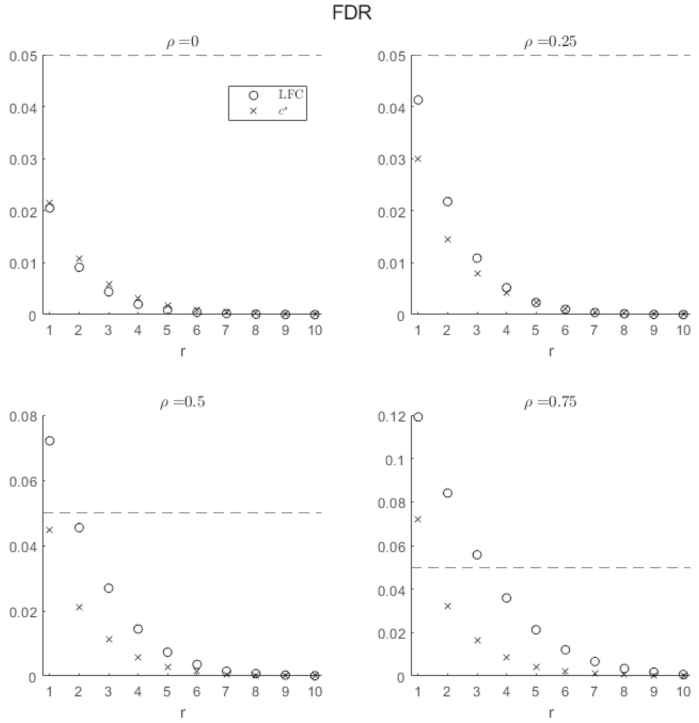

Fig. 8 FDR of the adaptive LSU tests. The subplots differ in the value of the correlation parameter ρ. Considered is the modified Storey estimator with $\lambda = 1/2$, used with the LFC-based p-values (circles), or the randomized p-values with $c = c^\star$ (crosses), respectively. The signal strength is denoted by r. The horizontal dashed lines are at $\alpha = 0.05$

this procedure is that randomized p-values are irreproducible and should therefore not be used for making test decisions.

As for the Schweder–Spjøtvoll estimator, the expected value of the modified Storey estimator does not depend on the particular dependence structure of the utilized marginal p-values. Hence, Fig. 7 is representative for any ρ. For all considered values of r, the LFC-based p-values lead to uninformative π_0 estimations. For $r = 1$ the estimations corresponding to the two different sets of p-values are closest, since for small r the null p-values are close to being uniform. Using $c = c^\star$ leads to the lowest π_0 estimations, on average, among all $c \in [0, 1]$, because a nonnegative bias is guaranteed here. With increasing r, the π_0 estimations get worse when using the LFC-based p-values and better when using the randomized p-values.

In Fig. 8, we display the realized values of the FDR of the procedures described above. As expected, the FDR values in cases with independent LFC-based p-values ($\rho = 0$) are all smaller than $\alpha = 0.05$. The two approaches perform similarly in this case. Under positive dependence ($\rho > 0$), the FDR of the multiple test using the LFC-based p-values is higher than that of the multiple test using randomized p-values. Especially for low r and high ρ, the FDR is not always controlled when

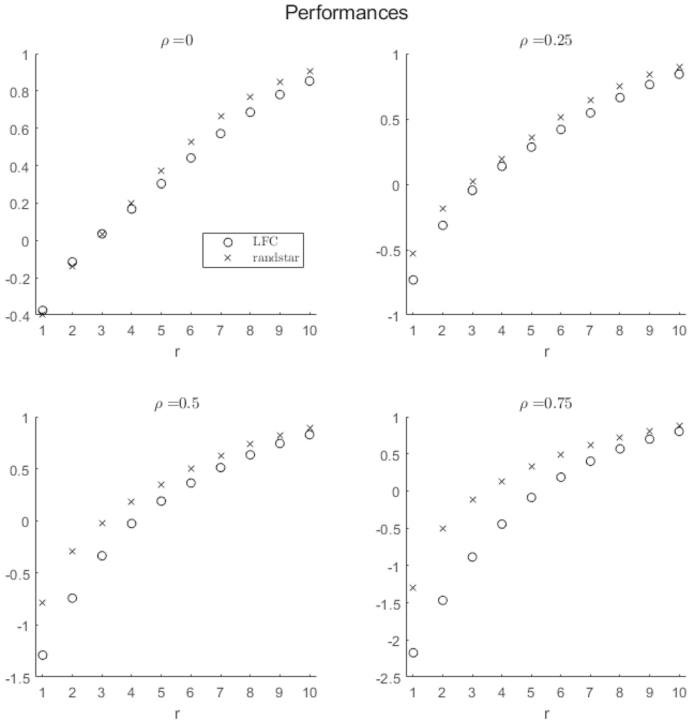

Fig. 9 Performances of the adaptive LSU tests. The subplots differ in the value of the correlation parameter ρ. Considered is the modified Storey estimator with $\lambda = 1/2$, used with the LFC-based p-values (circles), or the randomized p-values with $c = c^\star$ (crosses), respectively. The signal strength is denoted by r

using the LFC-based p-values. Furthermore, we notice that the FDR decreases with increasing signal strength r in all approaches. Even though the expected π_0 estimations are the same regardless of the dependence structure of the LFC-based p-values, we notice that the FDR increases with increasing correlation ρ. This is due to the increase in the variance of $1/G$, cf. also Sect. 4.2.

Figure 9 displays the performances in the sense of (10) of the adaptive LSU tests utilizing the different sets of p-values. Under independence, the performances of the different approaches are close to each other. Under positive dependence, however, the performance when using the LFC-based p-values is inferior to that of the randomized p-values, especially for low values of r.

6 Relationships to other approaches

There are several ways to draw connections between our proposed methodology and existing literature. One way is to consider statistical methods presented in previous literature that assume uniformly distributed p-values under null hypotheses and to discuss the advantages of our randomized p-values compared to conservative

p-values. Another way is to compare our approach with further strategies that help make conservative p-values more uniformly distributed under null hypotheses.

6.1 Implementation of our proposed approach into existing procedures

Our proposed methodology of randomizing p-values can be used in connection with any estimator of π_0 which relies on the ecdf of marginal p-values. Such estimators constitute a major class of estimators of π_0; cf., e.g., Table 1 in Chen (2019).

For example, Meinshausen and Rice (2006) considered the problem of estimating a lower bound $\hat{\pi}_1$ such that

$$\mathbb{P}_{\vartheta}\left(\hat{\pi}_1 \leq \pi_1\right) \geq 1-\alpha, \tag{11}$$

where $\pi_1 = 1 - \pi_0$ is the proportion of false null hypotheses. They propose the estimator

$$\hat{\pi}_1 = \sup_{\lambda\in(0,1)}\left[\frac{\hat{F}_m(\lambda) - \lambda - c_m(\lambda)}{1-\lambda}\right] = \sup_{\lambda\in(0,1)}\left[1 - \hat{\pi}_0(\lambda) - \frac{c_m(\lambda)}{1-\lambda}\right],$$

where $c_m(\lambda)$ is a constant which is independent of the model. For valid p-values, $1 - \hat{\pi}_0(\lambda)$ is a non-positively biased estimator for π_1, and the constant $c_m(\lambda)/(1-\lambda)$ is to make sure that (11) is satisfied. However, if the null p-values are conservative rather than uniformly distributed, we face an analogous problem as with the Schweder–Spjøtvoll (point) estimator. For stochastically increasing p-values, the expected value of the estimator $\hat{\pi}_1$ gets smaller and therefore provides a less informative lower bound. Using randomized p-values that come closer to uniformity under nulls instead of conservative LFC-based p-values can therefore help to minimize the size of the confidence region.

Uniformly distributed p-values under null hypotheses are also beneficial in a different, but related context. Ghosal and Roy (2011) considered a mixture model for the p-values. For a parameter θ_0 and a parameter space Θ_1 where $\theta_0 \notin \Theta_1$, let $\left\{h_\theta : \theta \in \Theta = \{\theta_0\} \cup \Theta_1\right\}$ be a parametrized family of Lebesgue densities. Assume that each (transformed) p-value in the multiple testing problem has an overall density

$$h(t) = \pi_0 h_{\theta_0}(t) + (1-\pi_0)\int_{\theta\in\Theta_1} h_\theta(t)\mathrm{d}G(\theta),$$

where G is a distribution over the parameter space Θ_1. Thus, each p-value has a probability of π_0 of being null and the density h_{θ_0} is such that the p-values are uniformly distributed under the (simple) nulls. Among other conditions, if it holds $h_\theta(t)/h_{\theta_0}(t) \to 0$ for $|t| \to \infty$ and each parameter $\theta \in \Theta_1$, then h uniquely identifies π_0. An analogous result holds for the characteristic functions of the densities h_θ, $\theta \in \Theta$. In practice, assuming conservative p-values under the null hypotheses instead of uniformly distributed ones, leads to an overestimation of $h(t)$ for large t

and therefore to overestimations of π_0. Switching to our more uniformly distributed p-values can therefore result in more accurate estimations for π_0 in this context.

6.2 Comparison with other approaches that address conservativity of *p*-values

In previous literature, there have been other approaches that deal with the problem of conservative p-values. For example, Zhao et al (2019) propose discarding p-values that are larger than a constant $\tau \in (0, 1)$ and multiplying the remaining p-values with $1/\tau$. This allows them to use a smaller number of p-values in the test. They show that these transformed p-values $\{p_j^{LFC}/\tau \mid p_j^{LFC} \leq \tau\}$ are valid if the p-value p_j^{LFC} is uniformly conservative, which is equivalent to condition (1.) in Theorem 1, where $\tau = c_j$, $j = 1, \ldots, m$. More particularly, if any of the conditions in Theorem 1 hold, both the transformed p-value p_j^{LFC}/τ, given $p_j^{LFC} \leq \tau$, and our randomized p-value $p_j^{rand}(X, U_j, \tau)$ are valid, $j = 1, \ldots, m$. The difference is that instead of discarding the p-values that are larger than τ, we replace them with the uniformly distributed random variables $U_1, \ldots, U_m$.

Among other things, Zhao et al (2019) consider the global null hypothesis $H_0 = \bigcap_{j=1}^m H_j$ that all individual null hypotheses are true, and present several examples of global p-values for testing H_0. One method is the so-called Bonferroni correction, in which we reject H_0 if any of the marginal p-values is smaller than α/m, where $\alpha \in (0, 1)$ is a significance level. In other words, we reject H_0 if

$$\min\{p_j^{LFC} \mid j = 1, \ldots, m\} \cdot m \leq \alpha.$$

Let $S_\tau = \{j \mid p_j^{LFC} \leq \tau\}$ be the set of non-discarded indices, then the conditional, global Bonferroni p-value, based on the adjusted p-values by Zhao et al (2019), is $\min\{p_j^{LFC}/\tau \mid j = 1, \ldots, m\}|S_\tau|$. Using our randomized p-values with $c_j = \tau$ for each j yields

$$\min\left\{p_j^{LFC}\mathbf{1}\{p_j^{LFC} \leq \tau\}/\tau + U_j\mathbf{1}\{p_j^{LFC} > \tau\} \mid j = 1, \ldots, m\right\} \cdot m$$

instead. Thus, using our randomized p-values leads to the smaller (not larger) min -term, however, $|S_\tau| \leq m$, so none of the two methods is strictly better than the other in every setting. Similarly, if applying Fisher's combination, the adjusted p-values by Zhao et al (2019) replace the discarded p-values by 1, where our randomized p-values replace them by $U_j \leq 1$, thus making the statistic that tests the global null H_0 larger when using our randomized p-values. However, the discard method gains an advantage by replacing m with $|S_\tau|$.

In Tian and Ramdas (2019), a similar method is employed in the context of online testing based on the concept of α consumption. Consider a sequence of null hypotheses $H_1, H_2, \ldots$, and denote I_0 the set of indices j for which H_j is true. Furthermore, denote the false discovery proportion (FDP) at point $t > 0$ by $\mathrm{FDP}(t) = |R(t) \cap I_0|/\max\{|R(t)|, 1\}$, where $R(t)$ is the set of indices $j \leq t$ for which H_j has been rejected. For some sequence $(\alpha_j)_{j=1}^\infty$ in the range [0, 1], they first consider the oracle estimate

$$\mathrm{FDP}^{\star}(t) = \frac{\sum_{j=1}^{t} \alpha_j \mathbf{1}\{j \in I_0\}}{\max\{|R(t)|, 1\}} \tag{12}$$

for the FDP.

Notice, however, that the set I_0 is unknown. A conservative approach would be to replace $\mathbf{1}\{j \in I_0\}$ by one in (12). In a certain sense, this may be interpreted as estimating π_0 as one. This conservative approach has then been improved by utilizing the estimate

$$\widehat{\mathrm{FDP}}_{\mathrm{SAFFRON}}(t) = \frac{\sum_{j \le t} \alpha_j \frac{\mathbf{1}\{p_j > \lambda\}}{1-\lambda}}{\max\{|R(t)|, 1\}}. \tag{13}$$

Comparing the numerators in (12) and (13), we conclude that the SAFFRON approach is similar to estimating π_0 by the Schweder–Spjøtvoll estimator $\hat{\pi}_0(\lambda)$, when $t = m$. However, as pointed out in the previous sections, $\hat{\pi}_0(\lambda)$ overestimates π_0 if the null p-values are conservative. Therefore, Tian and Ramdas (2019) propose to use $\mathbf{1}\{\lambda\tau < p_j \le \tau\}$ instead of $\mathbf{1}\{p_j > \lambda\}$, where $\tau \in (0, 1)$, resulting in

$$\widehat{\mathrm{FDP}}_{\mathrm{ADDIS}}(t) = \frac{\sum_{j \le t} \alpha_j \frac{\mathbf{1}\{\lambda\tau < p_j \le \tau\}}{\tau - \lambda\tau}}{\max\{|R(t)|, 1\}}.$$

Again, translating this into a π_0 estimator, when $t = m$, we get

$$\hat{\pi}_0^{discard}(\lambda, \tau) = \frac{1}{m} \sum_{j=1}^{m} \frac{\mathbf{1}\{\lambda\tau < p_j \le \tau\}}{\tau(1-\lambda)} = \frac{1}{\tau m} \sum_{j=1}^{m} \mathbf{1}\{p_j \le \tau\} \frac{\mathbf{1}\{\lambda < p_j/\tau\}}{1-\lambda}.$$

The estimator $\hat{\pi}_0^{discard}(\lambda, \tau)$ is an unbiased estimator for π_0 if the non-null p-values are almost surely smaller than $\lambda\tau$, and the null p-values are uniformly distributed. However, $\hat{\pi}_0^{discard}(\lambda, \tau)$ is generally not non-negatively biased if the p-values are merely valid. This means that this estimator does not satisfy (9).

Furthermore, the estimator $\hat{\pi}_0^{discard}(\lambda, \tau)$ is similar to using the adjusted p-values from Zhao et al (2019) in the Schweder–Spjøtvoll estimator $\hat{\pi}_0(\lambda)$, where instead of $|S_\tau|$, the number of non-discarded p-values, it employs τm. Albeit, Tian and Ramdas (2019) showed that

$$\mathbb{E}_\vartheta\left[\frac{\mathbf{1}\{\lambda\tau < p_j \le \tau\}}{\tau(1-\lambda)}\right] \le \mathbb{E}_\vartheta\left[\frac{\mathbf{1}\{\lambda < p_j\}}{1-\lambda}\right],$$

holds, if the cdf of p_j is convex under ϑ, and p_j is thus uniformly conservative, cf. condition (3.) in Theorem 1. Therefore, if π_0 is large and the null p-values are uniformly conservative, $\hat{\pi}_0^{discard}(\lambda, \tau)$ has a lower bias than $\hat{\pi}_0(\lambda)$ (but possibly negative), and $\widehat{\mathrm{FDP}}_{\mathrm{ADDIS}}(t) \le \widehat{\mathrm{FDP}}_{\mathrm{SAFFRON}}(t)$.

7 Discussion

We have demonstrated how randomized p-values can be utilized in the Schweder–Spjøtvoll estimator $\hat{\pi}_0$. Whenever composite null hypotheses are under consideration, our proposed approach leads to a reduction of the bias and of the MSE of $\hat{\pi}_0$, when compared to the usage of LFC-based p-values, at least in our simulations. Furthermore, our approach also robustifies $\hat{\pi}_0$ against dependencies among $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$. The latter property is important in modern high-dimensional applications, where the biological and/or technological mechanisms involved in the data-generating process virtually always lead to dependencies (cf. Stange et al 2016), especially in studies with multiple endpoints which are all measured for the same observational units. Furthermore, we have explained in detail how the proposed methodology can be applied in practice. Worksheets, with which all results of the present work can be reproduced, are available as supplementary material.

Statistical models that fulfill any of the conditions (2.)–(5.) from Theorem 1 admit valid randomized p-values $\{p_j^{rand}(X, U_j, c_j)\}_{1\leq j\leq m}$ for any choice of the constants $(c_j)_{1\leq j\leq m} \in [0,1]^m$. We gave two such models in Examples 1 and 2. These models have a variety of applications, for instance in the life sciences; cf., e.g., Part II of Dickhaus (2014). Closely related examples are the replicability models considered in Hoang and Dickhaus (2021). Identifying additional model classes that have that property is a topic for future research. Furthermore, in models for which the jth LFC-based p-value is of the form $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$ for $1 \leq j \leq m$ and in which $(T_j(X)^{(\vartheta)})_{\theta_j(\vartheta)}$ is an MLR family, the cdf of $p_j^{rand}(X, U_j, c_j)$ is always between those of Uni$[0,1]$ and $p_j^{LFC}(X)$. The latter follows from point 1 in Remark 2 and Theorem 2. Distributions with the MLR property include exponential families, for example the family of univariate normal distributions with fixed variance and the family of Gamma distributions (cf. Karlin and Rubin 1956). Also, the family of non-central t-distributions and the family of non-central F-distributions have the MLR property with respect to their non-centrality parameters (cf. Karlin 1956). It is of interest to deeper investigate properties of our randomized p-values in such models.

There are several further possible extensions of the present work. First, in Sect. 4 we only considered the usage of $p_1^{rand}(X, U_1, c_1), \ldots, p_m^{rand}(X, U_m, c_m)$ in $\hat{\pi}_0$ for identical constants $c_1 = \cdots = c_m \equiv c$. In future work, it may be of interest to develop a method for choosing each c_j individually, for instance depending on the size of the jth LFC-based p-value. Second, we have chosen c_0 in Sect. 4.3 such that the conditional (to the observed data $X = x$) bias of $\hat{\pi}_0(\lambda)$ is minimized. Another approach, which can be pursued in future research, is to choose a c_0 that minimizes the MSE of $\hat{\pi}_0(\lambda)$ instead. Third, we restricted our attention to the Schweder–Spjøtvoll estimator $\hat{\pi}_0(\lambda)$. However, there exists a wide variety of other ecdf-based estimators in the literature (see, for instance, Table 1 in Chen (2019) for a recent overview), which are prone to suffer from the same issues as $\hat{\pi}_0(\lambda)$ when used with LFC-based p-values in the context of composite null hypotheses. One other ecdf-based estimator for π_0 is the more conservative estimator $\hat{\pi}_0^+(\lambda) = \hat{\pi}_0(\lambda) + 1/(m(1-\lambda))$ proposed by Storey (2002). The bias of $\hat{\pi}_0^+$ when used with the randomized p-values $p_1^{rand}(X, U_1, c), \ldots, p_m^{rand}(X, U_m, c)$ is minimized for the same $c = c^\star$ from Sect. 4.

Thus, the same algorithm as outlined in Sect. 4.3 can be applied to $\hat{\pi}_0^+$ in practice. In future research, randomization approaches for other ecdf-based estimators can be investigated.

We have not elaborated on the choice of λ in the present work. The standard choice of $\lambda = 1/2$ seemed to work reasonably well in connection with our proposed randomized p-values. We have also performed some preliminary sensitivity analyses (not included here) with respect to λ, which indicated that the sensitivity of $\hat{\pi}_0$ with respect to λ is less pronounced for the case of randomized p-values than for the case of LFC-based p-values. Investigating this phenomenon deeper, both from the theoretical and from the numerical perspective, is also a worthwhile topic for future research.

Finally, in case of composite null hypotheses under discrete models, one may apply two stages of randomization: In the first stage, the discreteness of the model is addressed by applying a randomization procedure as proposed by Dickhaus et al (2012). Under LFCs, this leads to exactly uniformly distributed randomized p-values. Then, in a second stage of randomization, the approach proposed in this work helps to alleviate the problem of conservative p-values (under non-LFCs) resulting from the composite nature of the null hypotheses.

Appendix

The more general randomized *p*-values

Definition

Let $U_1, \ldots, U_m$ and X be as before. For a set of stochastically independent (not necessarily identically distributed) random variables $R_1, \ldots, R_m$ with values in [0, 1] that are defined on the same probability space as X, stochastically independent of the U_j's and the data X, and whose distributions do not depend on ϑ, we define

$$p_j^{rand}(X, U_j, R_j) = U_j \mathbf{1}\{p_j^{LFC}(X) \geq R_j\} + \frac{p_j^{LFC}(X)}{R_j} \mathbf{1}\{p_j^{LFC}(X) < R_j\}, \qquad (14)$$

$j = 1, \ldots, m$. This definition includes the case $R_j \equiv c_j$ from Definition 1 for any constant $c_j \in [0, 1]$, $j = 1, \ldots, m$. We generalize and prove Theorems 1 and 2 for the randomized p-values $\{p_j^{rand}(X, U_j, R_j)\}_{1 \leq j \leq m}$.

Theorem 1 ′

Let a model as in Sect. 2 be given and $j \in \{1, \ldots, m\}$ be fixed. Then, the jth randomized p-value $p_j^{rand}(X, U_j, R_j)$ as in (14) is a valid p-value for a given random variable R_j with values in [0, 1] if and only if condition (0.) is fulfilled. Furthermore, either of the following conditions (1.′), (2.) and (3.) is a sufficient condition for the validity of $p_j^{rand}(X, U_j, R_j)$ for any random variable R_j with values in [0, 1].

(0.) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \le tR_j) \le t\mathbb{P}_\vartheta(p_j^{LFC}(X) \le R_j)$$

for all $t \in [0, 1]$.

(1.′) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \le tu) \le t\mathbb{P}_\vartheta(p_j^{LFC}(X) \le u)$$

for all $u, t \in [0, 1]$.

(2.) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, $\mathbb{P}_\vartheta(p_j^{LFC}(X) \le t)/t$ is non-decreasing in t.

(3.) The cdf of $p_j^{LFC}(X)$ is convex under any $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$.

Let F_ϑ be the cdf of $T_j(X)$ under $\vartheta \in \Theta$. If the LFC-based p-value is given by $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$, where $\vartheta_0 \in \Theta$ is an LFC for φ_j, then the following condition (4.) is equivalent to condition (2.), while condition (5.) is equivalent to condition (3.).

(4.) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds $T_j(X)^{(\vartheta)} \le_{\mathrm{hr}} T_j(X)^{(\vartheta_0)}$.

(5.) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds $T_j(X)^{(\vartheta)} \le_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$.

Proof First, we show that condition (0.) is equivalent to $p_j^{rand}(X, U_j, R_j)$ being valid. For $p_j^{rand}(X, U_j, R_j)$ to be valid, it has to hold that

$$\mathbb{P}_\vartheta(p_j^{rand}(X, U_j, R_j) \le t) \le t,$$

for all $t \in [0, 1]$ and all $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$. It holds that

$$\begin{aligned} &\mathbb{P}_\vartheta(p_j^{rand}(X, U_j, R_j) \le t) \\ &\quad = \mathbb{P}_\vartheta(U_j \le t)\mathbb{P}_\vartheta(p_j^{LFC}(X) > R_j) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \le tR_j) \\ &\quad = t\mathbb{P}_\vartheta(p_j^{LFC}(X) > R_j) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \le tR_j). \end{aligned} \tag{15}$$

Now, the term in (15) is not larger than t if and only if it holds

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \le tR_j) \le t\big[1 - \mathbb{P}_\vartheta(p_j^{LFC}(X) > R_j)\big] = t\mathbb{P}_\vartheta(p_j^{LFC}(X) \le R_j),$$

which is condition (0.).

Let G be the cdf of R_j. From condition (1.′) it follows

$$\int_0^1 \mathbb{P}_\vartheta(p_j^{LFC}(X) \le tu)\mathrm{d}G(u) \le t\int_0^1 \mathbb{P}_\vartheta(p_j^{LFC}(X) \le u)\mathrm{d}G(u),$$

for every $t \in [0, 1]$, thus condition (1.′) implies (0.).

Substituting $z = tu$ in condition (1.′) leads to

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \le z) \le z\frac{\mathbb{P}_\vartheta(p_j^{LFC}(X) \le u)}{u}$$

for all $0 \le z < u \le 1$ and all $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, which is equivalent to condition (2.).

Now, we show that condition (3.) implies condition (1.′). Let $u \in [0,1]$ be fixed. The inequality in (1.′) is always satisfied for $t = 0$ and $t = 1$. Since $t \mapsto t\mathbb{P}_\vartheta(p_j^{LFC}(X) \le u)$ is a linear function and $t \mapsto \mathbb{P}_\vartheta(p_j^{LFC}(X) \le tu)$ is a convex function, if (3.) is fulfilled, it holds

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \le tu) \le t\mathbb{P}_\vartheta(p_j^{LFC}(X) \le u)$$

for all $t \in [0,1]$.

Now we assume that $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$. At first, we show that conditions (2.) and (4.) are equivalent. To this end, notice that the term

$$\frac{\mathbb{P}_\vartheta(p_j^{LFC}(X) \le t)}{t} = \frac{\mathbb{P}_\vartheta(p_j^{LFC}(X) \le t)}{\mathbb{P}_{\vartheta_0}(p_j^{LFC}(X) \le t)} = \frac{\mathbb{P}_\vartheta(T_j(X) \ge F_{\vartheta_0}^{-1}(1-t))}{\mathbb{P}_{\vartheta_0}(T_j(X) \ge F_{\vartheta_0}^{-1}(1-t))}$$

is non-decreasing in t if and only if $\mathbb{P}_\vartheta(T_j(X) \ge z)/\mathbb{P}_{\vartheta_0}(T_j(X) \ge z) = (1 - F_\vartheta(z))/(1 - F_{\vartheta_0}(z))$ is non-increasing in z.

Lastly, we show that conditions (3.) and (5.) are equivalent. Let f_ϑ be the Lebesgue density of $T_j(X)$ under $\vartheta \in \Theta$. Let $\vartheta \in \Theta$ be such that $\theta_j(\vartheta) \in H_j$ holds. The convexity of $t \mapsto \mathbb{P}_\vartheta(p_j^{LFC}(X) \le t)$ is equivalent to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}_\vartheta(T_j(X) \ge F_{\vartheta_0}^{-1}(1-t)) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[1 - F_\vartheta(F_{\vartheta_0}^{-1}(1-t))\right] \\ &= \frac{f_\vartheta(F_{\vartheta_0}^{-1}(1-t))}{f_{\vartheta_0}(F_{\vartheta_0}^{-1}(1-t))}\end{aligned}$$

being non-decreasing in t, or $f_\vartheta(z)/f_{\vartheta_0}(z)$ being non-increasing in z, which is equivalent to condition (5.); cf. the remarks after Theorem 1. □

In Theorem 1', the conditions (2.)–(5.) are the same as in Theorem 1. Condition (1.′) is equivalent to condition (1.) in Theorem 1 holding for all $c_j \in [0,1]$. Thus, $p_j^{rand}(X, U_j, c_j)$ being valid for all $c_j \in [0,1]$ implies the validity of $p_j^{rand}(X, U_j, R_j)$ for any random variable R_j on $[0, 1]$, $j = 1, \ldots, m$. The reverse is also true, thus, the randomized p-value $p_j^{rand}(X, U_j, R_j)$ is valid for any random variable R_j on $[0, 1]$ if and only if it is valid for $R_j \equiv c_j$, for all $c_j \in [0,1]$, $j = 1, \ldots, m$.

In the following, we show that Theorem 2 still holds if we replace the constants $c_j \le \tilde{c}_j$ by the random variables $R_j \le_{\mathrm{st}} \tilde{R}_j$.

Theorem 2 ′

Let a model as in Sect. 2 be given and $j \in \{1, \ldots, m\}$ be fixed. If the cdf of $p_j^{LFC}(X)$ is convex under a fixed $\vartheta \in \Theta$, then it is

$$p_j^{rand}(X, U_j, R_j)^{(\vartheta)} \leq_{\text{st}} p_j^{rand}(X, U_j, \tilde{R}_j)^{(\vartheta)}$$

for any random variables $R_j, \tilde{R}_j$ on [0, 1], with $R_j \leq_{\text{st}} \tilde{R}_j$.

If the cdf of $p_j^{LFC}(X)$ is concave under a fixed $\vartheta \in \Theta$, then it holds that

$$p_j^{rand}(X, U_j, \tilde{R}_j)^{(\vartheta)} \leq_{\text{st}} p_j^{rand}(X, U_j, R_j)^{(\vartheta)}$$

for any random variables R_j and $\tilde{R}_j$ with values in [0, 1] and with $R_j \leq_{\text{st}} \tilde{R}_j$.

Proof We first show both statements in Theorem 2' for constants $0 \leq c_j \leq \tilde{c}_j \leq 1$ instead of random variables R_j and $\tilde{R}_j$, which amounts to the statements in Theorem 2.

For every fixed $t \in [0, 1]$ and fixed $\vartheta \in \Theta$, we define the function $q : [0, 1] \to [0, 1]$ by

$$q(c) = \mathbb{P}_\vartheta(p_j^{rand}(X, U_j, c) \leq t) = t\mathbb{P}_\vartheta(p_j^{LFC}(X) > c) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq ct).$$

Furthermore, we denote by f_ϑ the Lebesgue density of $p_j^{LFC}(X)$ under ϑ, such that it holds $q'(c) = -tf_\vartheta(c) + tf_\vartheta(ct)$, which is not positive if f_ϑ is non-decreasing and not negative if f_ϑ is non-increasing.

Let R_j and $\tilde{R}_j$ be random variables fulfilling the assumptions of the theorem. If q is non-decreasing, then it holds that $\mathbb{E}[q(R_j)] \leq \mathbb{E}[q(\tilde{R}_j)]$, and if q is non-increasing it holds that $\mathbb{E}[q(R_j)] \geq \mathbb{E}[q(\tilde{R}_j)]$, where $\mathbb{E}$ refers to the joint distribution of R_j and $\tilde{R}_j$. Since $\mathbb{E}[q(R_j)] = \mathbb{P}_\vartheta(p_j^{rand}(X, U_j, R_j) \leq t)$ and $\mathbb{E}[q(\tilde{R}_j)] = \mathbb{P}_\vartheta(p_j^{rand}(X, U_j, \tilde{R}_j) \leq t)$, the proof is completed. □

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-021-00797-0.

Acknowledgements Financial support by the Deutsche Forschungsgemeinschaft via Grant No. DI 1723/5-1 is gratefully acknowledged.

References

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 57*(1), 289–300.

Benjamini, Y., Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics, 25*(1), 60–83.

Blanchard, G., Roquain, E. (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research, 10,* 2837–2871.
Chen, X. (2019). Uniformly consistently estimating the proportion of false null hypotheses via Lebesgue-Stieltjes integral equations. *Journal of Multivariate Analysis, 173,* 724–744. https://doi.org/10.1016/j.jmva.2019.06.003.
Dickhaus, T. (2013). Randomized p-values for multiple testing of composite null hypotheses. *Journal of Statistical Planning and Inference, 143*(11), 1968–1979.
Dickhaus, T. (2014). *Simultaneous statistical inference with applications in the life sciences*. Berlin, Heidelberg: Springer.
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(4):Art. 12, front matter+31, https://doi.org/10.1515/1544-6115.1776.
Finner, H., Gontscharuk, V. (2009). Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71*(5), 1031–1048.
Finner, H., Strassburger, K. (2007). A note on *P*-values for two-sided tests. *Biometrical Journal, 49*(6), 941–943. https://doi.org/10.1002/bimj.200710382.
Ghosal, S., Roy, A. (2011). Identifiability of the proportion of null hypotheses in skew-mixture models for the *p*-value distribution. *Electronic Journal of Statistics, 5,* 329–341. https://doi.org/10.1214/11-EJS609.
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.
Habiger, J. D., Pena, E. A. (2011). Randomised p-values and nonparametric procedures in multiple testing. *Journal of Nonparametric Statistics, 23*(3), 583–604.
Heesen, P., Janssen, A. (2015). Inequalities for the false discovery rate (FDR) under dependence. *Electronic Journal of Statistics, 9*(1), 679–716. https://doi.org/10.1214/15-EJS1016.
Heesen, P., Janssen, A. (2016). Dynamic adaptive multiple tests with finite sample FDR control. *Journal of Statistical Planning and Inference, 168,* 38–51. https://doi.org/10.1016/j.jspi.2015.06.007.
Hoang, A. T., Dickhaus, T. (2021). Randomized p-values for multiple testing and their application in replicability analysis. *Biometrical Journal early view*. https://doi.org/10.1002/bimj.202000155.
Karlin, S. (1956). Decision theory for Pólya type distributions. Case of two actions, I. In: *Proceedings of the third berkeley symposium on mathematical statistics and probability, Volume 1: contributions to the theory of statistics* pp. 115–128. Berkeley and Los Angeles: University of California Press.
Karlin, S., Rubin, H. (1956). Distributions possessing a monotone likelihood ratio. *Journal of the American Statistical Association, 51*(276), 637–643.
Lehmann, E. L., Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.
MacDonald, P. W., Liang, K., Janssen, A. (2019). Dynamic adaptive procedures that control the false discovery rate. *Electronic Journal of Statistics, 13*(2), 3009–3024. https://doi.org/10.1214/19-ejs1589.
Meinshausen, N., Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics, 34*(1), 373–393. https://doi.org/10.1214/009053605000000741.
Neumann, A., Bodnar, T., & Dickhaus, T. (2021). Estimating the proportion of true null hypotheses under dependency: a marginal bootstrap approach. *Journal of Statistical Planning and Inference, 210,* 76–86. https://doi.org/10.1016/j.jspi.2020.04.011.
Schweder, T., Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika, 69*(3), 493–502.
Stange, J., Dickhaus, T., Navarro, A., & Schunk, D. (2016). Multiplicity- and dependency-adjusted *p*-values for control of the family-wise error rate. *Statistics & Probability Letters, 111,* 32–40.
Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(3), 479–498.
Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the *q*-value. *The Annals of Statistics, 31*(6), 2013–2035. https://doi.org/10.1214/aos/1074290335.
Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66*(1), 187–205.

Tian, J., Ramdas, A. (2019). ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. *Advances in Neural Information Processing Systems (NeurIPS 2019), 32,* 9388–9396.

Zhao, Q., Small, D. S., & Su, W. (2019). Multiple testing when many *p*-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *Journal of the American Statistical Association, 114*(527), 1291–1304. https://doi.org/10.1080/01621459.2018.1497499.