Statistica Sinica **33** (2023), 851-872
doi:https://doi.org/10.5705/ss.202021.0071

TRADE-OFF BETWEEN VALIDITY AND EFFICIENCY OF MERGING p-VALUES UNDER ARBITRARY DEPENDENCE

Yuyu Chen[1], Peng Liu[2], Ken Seng Tan[3] and Ruodu Wang[1]

[1] *University of Waterloo,* [2] *University of Essex and* [3] *Nanyang Technological University*

Abstract: Various methods are widely used to combine individual p-values into one p-value in many areas of statistical applications. We say that a combining method is valid for arbitrary dependence if it does not require any assumption on the dependence structure of the p-values, whereas it is valid for some dependence if it requires some specific, perhaps realistic, but unjustifiable, dependence structures. The trade-off between the validity and efficiency of these methods is studied by analyzing the choices of critical values under different dependence assumptions. We introduce the notions of independence-comonotonicity balance (IC-balance) and the price for validity. In particular, IC-balanced methods always produce an identical critical value for independent and perfectly positively dependent p-values, a specific type of insensitivity to a family of dependence assumptions. We show that, among two very general classes of merging methods commonly used in practice, the Cauchy combination method and the Simes method are the only IC-balanced ones. Simulation studies and a real-data analysis are conducted to analyze the size and power of various combining methods in the presence of weak and strong dependence.

Key words and phrases: Efficiency, hypothesis testing, multiple hypothesis testing, validity.

1. Introduction

In many statistical applications in which multiple hypothesis testing is involved, the task of merging several p-values into one naturally arises. Depending on the specific application, these p-values may be from a single hypothesis or from multiple hypotheses, in small or large numbers, independent or correlated, and with sparse or dense signals, leading to different considerations when choosing merging procedures.

Let K be a positive integer, and $F : [0,1]^K \to [0,\infty)$ be an increasing Borel function used to combine K p-values, which we refer to as a *combining function*. In general, the combined value may not be a valid p-value itself, and

Corresponding author: Yuyu Chen, Department of Statistics and Actuarial Science, University of Waterloo, Canada. E-mail: y937chen@uwaterloo.ca.

a critical point needs to be specified. Different dependence assumptions on the p-values lead to significantly different critical points, and thus different statistical decisions. The problem of merging p-values has a long history, and early results can be found in Tippett (1931), Pearson (1933), and Fisher (1948), where p-values are assumed to be independent. Based on an idea of Tukey, Donoho and Jin (2004) developed the higher criticism statistics to detect weak and sparse signals effectively using independent p-values. Certainly, these methods do not always produce a valid p-value if the assumption of independence is violated. On the other hand, the independence assumption is often difficult or impossible to verify when only one set of p-values is available.

However, some methods produce valid p-values without any dependence assumption. A classic example is the Bonferroni method, which takes the minimum of the p-values times K (we allow combined p-values to be greater than one and they can be treated as one) or, equivalently, dividing the critical value by K. Other methods that are valid without assumptions include those based on order statistics by Rüger (1978) and Hommel (1983), and those based on averaging by Vovk and Wang (2020); details of these merging methods are presented in Section 3.

Other methods work under weak or moderate dependence assumptions, such as the method of Simes (1986), which uses the minimum of $Kp_{(i)}/i$ over $i = 1, \ldots, K$, where $p_{(i)}$ is the ith smallest order statistic of $p_1, \ldots, p_K$. The validity of the Simes method is shown under a large class of dependence structures (e.g., Sarkar (1998, 2008); Benjamini and Yekutieli (2001); Rødland (2006)), although even such dependence assumptions are unlikely to hold in practice (see, e.g., Efron (2010, p.51)). Two recent methods include the Cauchy combination test proposed by Liu and Xie (2020), which uses the weighted average of Cauchy transformed p-values, and the harmonic mean p-value of Wilson (2019), which uses the harmonic mean of the p-values. Under mild dependence assumptions, these two methods are asymptotically valid as the significance level goes to zero (see Theorem 2).

Here, we present a comprehensive and unifying treatment of p-value merging methods under various dependence assumptions. Some methods are valid without any assumption on the interdependence of the p-values. We refer to such methods as valid for arbitrary dependence (VAD). On the other hand, methods that are valid for some specific, but realistic dependence assumption (e.g., independence, positive dependence, or joint normality dependence) are referred to as valid for some dependence (VSD). Our main goal is to understand the difference and the trade-off between these methods.

For a fixed combining function F, using a VAD method means choosing a smaller critical value (threshold) for making rejections than when using a VSD method. Thus, the gain of validity comes at the price of a loss of detection power. Because it is often difficult to make a valid statistical inference on the dependence structure of p-values, our analysis also helps to understand the relative performance of VSD combining methods in the presence of model misspecification. As a byproduct, we obtain several new theoretical results on the popular Simes, harmonic, and Cauchy merging methods.

In the next section, we collect some basic definitions of VAD and VSD merging methods and their corresponding threshold functions. We focus on symmetric merging functions for the tractability in their comparison. In Section 3, we introduce two general classes of combining functions, which include all methods mentioned above. We also derive formulae for their VAD and VSD threshold functions, some based on results from robust risk aggregation; see, Wang, Peng and Yang (2013). In Section 4, we introduce independence-comonotonicity balanced (IC-balanced) combining functions, which are indifferent between the two dependence assumptions. We show that the Cauchy combination method and the Simes method are the only IC-balanced ones among two general classes of combining methods, thus highlighting their unique roles. In Section 5, we establish the strong similarity between the Cauchy combination and the harmonic averaging methods, and obtain an algebraic relationship between the harmonic averaging and the Simes functions. In Section 6, the price for validity is introduced to assess the loss of power of VAD methods compared to their VSD versions. Simulation studies and a real-data analysis are conducted to analyze the relative performance of these methods. Owing to space constraints, we present all the results and observations of the numerical studies in Section S1 of the Supplementary Material. Proofs of all technical results are also provided in the Supplementary Material.

We conclude this section by providing additional notation and terminology adopted in this paper. All random variables are defined on an atomless probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Random variables $X_1, \dots, X_n$ are comonotonic if there exist increasing functions $f_1, \dots, f_n$ and a random variable Z such that $X_i = f_i(Z)$, for each $i = 1, \dots, n$. For $\alpha \in (0, 1]$, $q_\alpha(X)$ is the left α-quantile of a random variable X, defined as

$$q_\alpha(X) = \inf\{x \in \mathbb{R} \mid \mathbb{P}(X \le x) \ge \alpha\}.$$

We also use $F^{-1}(\alpha)$ for $q_\alpha(X)$ if X follows the distribution F. The set $\mathcal{U}$ is the set of all standard uniform random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ (i.e., the set of all

measurable functions on $(\Omega, \mathcal{F})$ whose distribution under $\mathbb{P}$ is uniform on $[0,1]$) and $\mathbb{1}$ is the indicator function. The equality $\overset{\mathrm{d}}{=}$ represents equality in distribution. For given $p_1, \dots, p_K$, the order statistics $p_{(1)}, \dots, p_{(K)}$ are ordered from smallest to largest. The equivalence $A_x \sim B_x$ as $x \to x_0$ means that $A_x/B_x \to 1$ as $x \to x_0$. All terms of "increasing" and "decreasing" are in the non-strict sense.

2. Merging Methods and Thresholds

Following the terminology of Vovk and Wang (2020), a *p-variable* is a random variable P such that $\mathbb{P}(P \le \varepsilon) \le \varepsilon$, for all $\varepsilon \in (0,1)$ (such random variables are called *superuniform* by Ramdas et al. (2019)). Values realized by p-variables are p-values. In the Introduction, p-values are used loosely for p-variables, which should be clear from the context.

Let $P_1, \dots, P_K$ be K p-variables for testing a common hypothesis. A *combining function* is an increasing Borel measurable function $F : [0,1]^K \to [0,\infty)$ that transforms $P_1, \dots, P_K$ into a single random variable $F(P_1, \dots, P_K)$. The choice of combining function depends on how one integrates information, and some common options are mentioned in the Introduction. In general, $F(P_1, \dots, P_K)$ may not be a valid p-variable. For different choices of F and assumptions on $P_1, \dots, P_K$, one needs to assign a critical value $g(\varepsilon)$ so that the hypothesis can be rejected with significance level $\varepsilon \in (0,1)$ if $F(P_1, \dots, P_K) < g(\varepsilon)$. We call g a *threshold (function)* for F and $P_1, \dots, P_K$. Clearly, $g(\varepsilon)$ is increasing in ε. In case g is strictly increasing, which is the most common situation, the above specification of g is equivalent to requiring $g^{-1} \circ F(P_1, \dots, P_K)$ to be a p-variable. To objectively compare various combining methods, one should compare the corresponding values of the function $g^{-1} \circ F$.

In some situations, it might be convenient and practical to assume additional information on the dependence structure of the p-variables, for example, independence, comonotonicity (i.e., perfectly positive dependence), and specific copulas. The choice of the threshold g certainly depends on such assumptions. If no assumption is made on the interdependence of the p-variables, the corresponding threshold function is called a *VAD threshold*; otherwise, it is a *VSD threshold*. A testing procedure based on a VAD threshold always produces a size less than or equal to the significance level, regardless of the dependence structure of the p-variables.

We denote the VAD threshold of a combining function F by a_F. If a merging method is valid for independent (resp. comonotonic) dependence of the p-variables, we use b_F (resp. c_F) to denote the corresponding valid threshold func-

tion, and call it the *VI* (resp. *VC*) *threshold*. More precisely, for the equation

$$\mathbb{P}(F(P_1,\dots,P_K) < g(\varepsilon)) \le \varepsilon, \quad \varepsilon \in (0,1), \tag{2.1}$$

a VAD threshold $g = a_F$ satisfies (2.1) for all p-variables $P_1,\dots,P_K$, a VI threshold $g = b_F$ satisfies (2.1) for all independent p-variables $P_1,\dots,P_K$, and a VC threshold $g = c_F$ satisfies (2.1) for all comonotonic p-variables $P_1,\dots,P_K$.

The comonotonicity assumption on the p-variables to be combined (actually they are identical if they are uniform on $[0,1]$) is not interesting by itself in practice. Nevertheless, comonotonicity is a benchmark for (extreme) positive dependence, and we analyze c_F for the purpose of comparison. It helps us to understand how valid thresholds for different methods vary as the dependence assumption gradually shifts from independence to extreme positive dependence. This point is clarified in Sections 4–6.

An immediate observation is that the p-variables can be equivalently replaced by uniform random variables on $[0,1]$, because for each p-variable P, we can find $U \in \mathcal{U}$ with $U \le P$; see, for example, Vovk and Wang (2020). Therefore, it suffices to consider p-variables in $\mathcal{U}$. Moreover, if g satisfies (2.1), then any function that is smaller than g is also valid. Hence, for the sake of power, it is natural to use the largest functions that satisfy (2.1). Putting these considerations together, we formally define the thresholds of interest as follows.

Definition 1. The thresholds a_F, b_F, and c_F of a combining function F are given by, for $\varepsilon \in (0,1)$,

$$a_F(\varepsilon) = \inf\{q_\varepsilon(F(U_1,\dots,U_K)) \mid U_1,\dots,U_K \in \mathcal{U}\}, \tag{2.2}$$

$$b_F(\varepsilon) = q_\varepsilon(F(V_1,\dots,V_K)), \tag{2.3}$$

$$c_F(\varepsilon) = q_\varepsilon(F(U,\dots,U)), \tag{2.4}$$

where $U, V_1, \dots, V_K$ are independent standard uniform random variables.

In what follows, we focus on the thresholds in Definition 1. It is clear that $g = a_F$, b_F, or c_F in Definition 1 satisfies (2.1) under the respective dependence assumptions.

Remark 1. While the objects b_F and c_F in (2.3)–(2.4) can often be explicitly calculated, the object a_F in (2.2) is generally difficult to calculate for a chosen function F, owing to the infimum taken over all possible dependence structures. Techniques in the field of robust risk aggregation, particularly those of Wang, Peng and Yang (2013), Embrechts, Puccetti and Rüschendorf (2013),

Embrechts, Wang and Wang (2015), and Wang and Wang (2016), are designed for such calculation, as shown by Vovk and Wang (2020). By definition, for any threshold $g(\varepsilon) > a_F(\varepsilon)$, there exists some dependence structure of $(P_1, \dots, P_K)$ such that validity is lost, that is, (2.1) is violated. Moreover, if the combining function F is continuous, the infimum in (2.2) is attainable; the proof of this statement is similar to that of Lemma 4.2 of Bernard, Jiang and Wang (2014).

3. Combining Functions

3.1. Two general classes of combining functions

We first introduce two general classes of combining functions, the generalized mean class and the order statistics class. Let $p_1, \dots, p_K \in [0, 1]$ be K realized p-values. The first class of combining functions is the generalized mean, that is,

$$M_{\phi,K}(p_1, \dots, p_K) = \phi^{-1}\left(\frac{1}{K}\sum_{i=1}^{K}\phi(p_i)\right),$$

where $\phi : [0, 1] \to [-\infty, \infty]$ is a continuous and strictly monotone function and ϕ^{-1} is its inverse on the domain $\phi([0, 1])$. Many combining functions used in the statistical literature are included in this class. For example, the Fisher method (Fisher (1948)) corresponds to the geometric mean with $\phi(p) = \log(p)$; the averaging methods of Vovk and Wang (2020) and Wilson (2019) correspond to the functions $\phi(p) = p^r$ and $r \in [-\infty, \infty]$ (including limit cases), and the Cauchy combination method of Liu and Xie (2020) corresponds to $\phi(p) = \tan(\pi(p - 1/2))$.

The second class of combining functions is built on order statistics. Let $\alpha = (\alpha_1, \dots, \alpha_K) \in \mathbb{R}_+^K$, where $\mathbb{R}_+ = [0, \infty)$. We define the combining function

$$S_{\alpha,K}(p_1, \dots, p_K) = \min_{i\in\{1,\dots,K\}} \frac{p_{(i)}}{\alpha_i},$$

where the convention is $p_{(i)}/\alpha = \infty$ if $\alpha = 0$. If $\alpha_1 = 1/K$ and all the other components of α are zero, then using $S_{\alpha,K}$ yields the Bonferroni method based on the minimum of p-values. The VAD method via order statistics of Rüger (1978) uses $S_{\alpha,K}$ by setting $\alpha_i = i/K$, for a fixed $i \in \{1, \dots, K\}$, and all other components of α to zero. On the other hand, if $\alpha_i = i/K$, for each $i = 1, \dots, K$, then we arrive at the method of Simes (1986); in this case, we simply denote $S_{\alpha,K}$ by S_K, that is,

$$S_K(p_1, \dots, p_K) := \min_{i\in\{1,\dots,K\}} \frac{Kp_{(i)}}{i},$$

and S_K is called the *Simes function*. The method of Hommel (1983) uses $\ell_K S_K$, which is S_K adjusted via the VAD threshold, where

$$\ell_K = \sum_{k=1}^{K} \frac{1}{k}. \tag{3.1}$$

If $\alpha_{i+1} \le \alpha_i$, then the term $p_{(i+1)}/\alpha_{i+1}$ does not contribute to the calculation of $S_{\alpha,K}(p_1, \ldots, p_K)$. Hence, we can safely replace α_{i+1} with α_i without changing the function $S_{\alpha,K}$. Thus, we assume, without loss of generality, that $\alpha_1 \le \cdots \le \alpha_K$. The admissibility of VAD merging methods in the above two classes is studied by Vovk, Wang and Wang (2022).

Recall that a function $F : \mathbb{R}_+^K \to \mathbb{R}$ is homogeneous if $F(\lambda \mathbf{x}) = \lambda F(\mathbf{x})$, for all $\lambda > 0$ and $\mathbf{x} \in \mathbb{R}_+^K$. It is clear that the function $S_{\alpha,K}$ is homogeneous, and so are the averaging methods of Vovk and Wang (2020). In such cases, we can show that the VAD threshold a_F is a linear function.

Proposition 1. *If the combination function F is homogeneous, then the VAD threshold $a_F(x)$ is a constant times x on $(0, 1)$.*

In the subsections below, we discuss several special cases of the above two classes of combining functions, and analyze their corresponding threshold functions. As the first example, note that the functions a_F, b_F, and c_F for the Bonferroni method can be easily verified.

Proposition 2. *Let $F(p_1, \ldots, p_K) = \min\{p_1, \ldots, p_K\}$, for $p_1, \ldots, p_K \in [0, 1]$. Then, $a_F(\varepsilon) = \varepsilon/K$, $b_F(\varepsilon) = 1 - (1 - \varepsilon)^{1/K}$, and $c_F(\varepsilon) = \varepsilon$, for $\varepsilon \in (0, 1)$.*

3.2. The averaging methods

The aforementioned averaging methods of Vovk and Wang (2020) use the combining functions given by

$$M_{r,K}(p_1, \ldots, p_K) = \left(\frac{p_1^r + \cdots + p_K^r}{K} \right)^{1/r},$$

for $r \in \mathbb{R} \setminus \{0\}$, together with its limit cases

$$\begin{aligned} M_{-\infty,K}(p_1, \ldots, p_K) &= \min\{p_1, \ldots, p_K\}; \\ M_{0,K}(p_1, \ldots, p_K) &= \left(\prod_{i=1}^{K} p_i \right)^{1/K}; \\ M_{\infty,K}(p_1, \ldots, p_K) &= \max\{p_1, \ldots, p_K\}. \end{aligned}$$

Some special cases of the combining functions above are $r = -\infty$ (minimum), $r = -1$ (harmonic mean), $r = 0$ (geometric mean), $r = 1$ (arithmetic mean), and $r = \infty$ (maximum); the cases $r \in \{-1, 0, 1\}$ are known as Platonic means. Note that $M_{-\infty,K}$ gives rise to the Bonferroni method, and the geometric mean yields Fisher's method (Fisher (1948)) under the independence assumption. The harmonic mean p-value of Wilson (2019) is a VSD method using the harmonic mean.

Because the mean function $M_{r,K}$ is homogeneous, by Proposition 1, the VAD threshold is a linear function $a_F(x) = a_r x$, $x \in (0,1)$, for some $a_r > 0$. The multipliers a_r are well studied in Vovk and Wang (2020), and here we focus mainly on the Platonic means and the Bonferroni method. It is known that $a_{-\infty} = 1/K$ and $a_1 = 1/2$. For $r = 0$ or $r = -1$, the values of a_r and their asymptotic formulae are calculated by Propositions 4 and 6 of Vovk and Wang (2020), summarized below for $K \geq 3$.

(i) For $F = M_{0,K}$,

$$a_F(x) = a_0 x = c_K \exp\left(\frac{K-1}{1-Kc_K}\right) \times x, \quad x \in (0,1), \tag{3.2}$$

where c_K is the unique solution to the equation $\log(1/c-(K-1)) = K-K^2c$, for $c \in (0, 1/K)$. Moreover, $a_0 \geq 1/e$, and $a_0 \to 1/e$ as $K \to \infty$.

(ii) For $F = M_{-1,K}$,

$$a_F(x) = a_{-1} x = \frac{(y_K+1)K}{(y_K+K)^2} \times x, \quad x \in (0,1), \tag{3.3}$$

where y_K is the unique solution to the equation $y^2 = K((y+1)\log(y+1)-y)$, for $y \in (0,\infty)$. Moreover, $a_{-1} \geq (e \log K)^{-1}$, and $a_{-1} \log K \to 1$ as $K \to \infty$.

To determine the VC threshold, it is easy to check that $c_{M_{r,K}}(x) = x$, $x \in (0,1)$ for all $r \in [-\infty, \infty]$, because the generalized mean of identical objects is equal to themselves; this obviously holds for all functions in the family of $M_{\phi,K}$.

Next, we study $b_r := b_{M_{r,K}}$, or its approximate form. For this, we use stable distributions (e.g., Uchaikin and Zolotarev (2011) and Samorodnitsky (2017)). Let F_α be the stable distribution with stability parameter $\alpha \in (0,2)$, skewness parameter $\beta = 1$, scale parameter $\sigma = 1$, and shift parameter $\mu = 0$. The characteristic function of F_α is given by, for $\theta \in \mathbb{R}$,

$$\int \exp(i\theta x)\, \mathrm{d}F_\alpha(x) = \begin{cases} \exp\left(-|\theta|^\alpha(1 - i\,\mathrm{sgn}(\theta)\tan(\pi\alpha/2))\right) & \text{if } \alpha \neq 1, \\ \exp\left(-|\theta|(1 + i(2/\pi)\,\mathrm{sgn}(\theta)\log|\theta|)\right) & \text{if } \alpha = 1, \end{cases}$$

Table 1. Coefficients C_α and b_K for $r = -1/\alpha < 0$.

$r = -1/\alpha$	C_α	b_K
$-1/2 < r < 0$	$(K(\alpha/(\alpha-2) - (\alpha/(\alpha-1))^2))^{1/2}$	$K\alpha/(\alpha-1)$
$r = -1/2$	$\sqrt{K \log K}$	$K\alpha/(\alpha-1)$
$-1 < r < -1/2$	$K^{1/\alpha}\left(\Gamma(1-\alpha)\cos(\pi\alpha/2)\right)^{1/\alpha}$	$K\alpha/(\alpha-1)$
$r = -1$	$K\pi/2$	$(\pi K^2/2)\int_1^\infty \sin(2x/K\pi)\alpha x^{-\alpha-1}\,\mathrm{d}x$
$r < -1$	$K^{1/\alpha}\left(\Gamma(1-\alpha)\cos(\pi\alpha/2)\right)^{1/\alpha}$	0

where $\mathrm{sgn}(\cdot)$ is the sign function. For $\alpha \geq 2$, let F_α stand for the standard normal distribution.

Proposition 3. *Let b_r be the VI threshold of $M_{r,K}$, $r \in \mathbb{R}$.*

(i) If $r < 0$, then for $K \in \mathbb{N}_+$,

$$b_r(\varepsilon) \sim K^{-1-1/r}\varepsilon, \qquad \text{as } \varepsilon \downarrow 0, \tag{3.4}$$

and for $\varepsilon \in (0,1)$,

$$b_r(\varepsilon) \sim \left(\frac{C_\alpha F_\alpha^{-1}(1-\varepsilon) + b_K}{K}\right)^{1/r}, \qquad \text{as } K \to \infty,$$

where $\alpha = -1/r > 0$ and the constants C_α and b_K are given in Table 1.

(ii) If $r = 0$, then

$$b_r(\varepsilon) = \exp\left(-\frac{1}{2K} q_{1-\varepsilon}\left(\chi^2_{2K}\right)\right). \tag{3.5}$$

(iii) If $r > 0$, then for $K \in \mathbb{N}_+$,

$$b_r(\varepsilon) = \frac{(\Gamma(1+K/p))^{1/K}\varepsilon^{1/K}}{K^{1/r}\Gamma(1+1/p)}, \quad \text{if} \quad \varepsilon \leq \frac{(\Gamma(1+1/p))^K}{\Gamma(1+K/p)},$$

where Γ is the Gamma function. For $\varepsilon \in (0,1)$,

$$b_r(\varepsilon) \sim \left(\frac{\sigma}{\sqrt{K}}\Phi^{-1}(\varepsilon) + \mu\right)^{1/r}, \qquad \text{as } K \to \infty,$$

where $\mu = (r+1)^{-1}$ and $\sigma^2 = r^2(1+2r)^{-1}(1+r)^{-2}$.

3.3. The Cauchy combination method

The Cauchy combination method was recently proposed by Liu and Xie (2020), and relies on a special case of the generalized mean via $\phi = \mathcal{C}^{-1}$, where

$\mathcal{C}$ is the standard Cauchy cdf, that is,

$$\mathcal{C}(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2},\ x \in \mathbb{R}; \qquad \mathcal{C}^{-1}(p) = \tan\left(\pi\left(p - \frac{1}{2}\right)\right),\ p \in (0,1).$$

We denote this combining function by $M_{\mathcal{C},K}$ (instead of $M_{\mathcal{C}^{-1},K}$, for simplicity),

$$M_{\mathcal{C},K}(p_1,\dots,p_K) := \mathcal{C}\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(p_i)\right).$$

It is well known that the arithmetic average of either independent or comonotonic standard Cauchy random variables follows again the standard Cauchy distribution. This feature allows us to use such a combination method to combine p-values under uncertain dependence assumptions. In addition, Liu and Xie (2020) showed that under a bivariate normality assumption of the individual test statistics (i.e., a normal copula), the combined p-value has the same asymptotic behavior as that under the assumption of independence (see Theorem 2 (ii) below).

Because $(1/K)\sum_{i=1}^{K}\mathcal{C}^{-1}(U_i)$ follows a standard Cauchy distribution if $U_1, \dots, U_K \in \mathcal{U}$ are either independent or comonotonic, we have $b_F(x) = c_F(x) = x$, for all $x \in (0,1)$. This convenient feature is examined in greater detail in Section 4.

By Definition 1, we get, for $F = M_{\mathcal{C},K}$,

$$a_F(\varepsilon) = \mathcal{C}\left(\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(U_i)\right) \mid U_1,\dots,U_K \in \mathcal{U}\right\}\right). \tag{3.6}$$

The function a_F does not admit an explicit formula, but it can be calculated using results from robust risk aggregation (Corollary 3.7 in Wang, Peng and Yang (2013)), as in the following proposition.

Proposition 4. *For $\varepsilon \in (0,1/2)$, we have*

$$a_F(\varepsilon) = \mathcal{C}\left(-\frac{H_\varepsilon(x_K)}{K}\right), \tag{3.7}$$

where $H_\varepsilon(x) = (K-1)\mathcal{C}^{-1}(1-\varepsilon+(K-1)x) + \mathcal{C}^{-1}(1-x)$, $x \in (0,\varepsilon/K)$, and x_K is the unique solution $x \in (0,\varepsilon/K)$ to the equation

$$K\int_x^{\varepsilon/K} H_\varepsilon(t)\,\mathrm{d}t = (\varepsilon - Kx)H(x).$$

3.4. The Simes method

The method of Simes (1986) uses the Simes function S_K in the order statistics family, given by $S_K(p_1,\dots,p_K) = \min_{i\in\{1,\dots,K\}}(K/i)p_{(i)}$. For $F = S_K$, the results in Hommel (1983), together with Proposition 1, suggest that $a_F(x) = x/\ell_K$, for $x \in (0,1)$. For independent p-variables $P_1,\dots,P_K \in \mathcal{U}$, Simes (1986) obtained

$$\mathbb{P}\left(\min_{i\in\{1,\dots,K\}} \frac{K}{i}P_{(i)} > \varepsilon\right) = 1-\varepsilon, \quad \varepsilon \in (0,1),$$

which gives $b_F(x) = x$, for $x \in (0,1)$. For comonotonic p-variables $P_1,\dots,P_K \in \mathcal{U}$, it is clear that $S_K(P_1,\dots,P_K) = P_{(K)}$, which follows a standard uniform distribution, and hence we again have $c_F(x) = x$, for $x \in (0,1)$. The validity of the Simes function using the VI (VC) threshold (called the Simes inequality) holds under many positive dependence structures; see, for example, Sarkar (1998, 2008).

In the context of testing multiple hypotheses, if the p-variables for several hypotheses are independent, the Benjamini–Hochberg procedure for controlling the false discovery rate (FDR) (Benjamini and Hochberg (1995)) also relies on the Simes function (in case all hypotheses are null). Although the Benjamini–Hochberg procedure is valid for many practical models, to control the FDR under an arbitrary dependence structure of p-variables, one needs to multiply the p-values by ℓ_K, resulting in the Benjamini–Yekutieli procedure (Benjamini and Yekutieli (2001)). This constant is exactly $x/a_F(x)$, and the function a_F is called a reshaping function by Ramdas et al. (2019) in the FDR context.

4. IC-Balance

As we have seen above, the Cauchy function and the Simes function both satisfy $b_F = c_F$, and hence the corresponding merging methods are invariant under independence or comonotonicity assumptions, an arguably convenient feature. Inspired by this observation, we introduce the property of IC-balance for combining functions in this section. This property distinguishes the Cauchy combination method and the Simes method from their corresponding classes $M_{\phi,K}$ and $S_{\alpha,K}$, respectively.

A combining function is said to be balanced between two different dependence structures of p-variables if the combined random variable under the two dependence assumptions coincide in distribution. Recall that $U, V_1,\dots,V_K$ are independent standard uniform random variables.

Definition 2. A combining function $F : [0,1]^K \to [0,\infty)$ is IC-balanced if $F(V_1,\dots,V_K) \overset{\mathrm{d}}{=} F(U,\dots,U)$.

Because the VI and VC thresholds are the corresponding quantile functions of $F(P_1,\dots,P_K)$, we immediately conclude that a combining function $F : [0,1]^K \to [0,\infty)$ is IC-balanced if and only if $b_F = c_F$ on $(0,1]$; recall that c_F is the identity for all functions in Section 3.

IC-balanced methods have the same threshold $b_F = c_F$ if the dependence structure of the p-variables is a mixture of independence and comonotonicity, that is, with the copula

$$\lambda \prod_{i=1}^{n} x_i + (1-\lambda) \min_{i=1,\dots,n} x_i, \quad (x_1,\dots,x_n) \in [0,1]^n, \tag{4.1}$$

where $\lambda \in [0,1]$. This is because $\mathbb{P}(F(U_1,\dots,U_K) \le b_F(\varepsilon))$ is linear in the distribution of $(U_1,\dots,U_K)$.

For any combining function F, VI and VC thresholds generally yield more power to the test than does the corresponding VAD threshold, but the gain in power may come with invalidity owing to model misspecification. If a combining function F is IC-balanced, the validity is preserved under independence, comonotonicity, and their mixtures, and we may expect (without mathematical justification) that, to some extent, the size of the test can be controlled properly, even if mild model misspecification exists. Therefore, the notion of IC-balance can be interpreted as insensitivity to some specific type of model misspecification (e.g., dependence structure given in (4.1)) for VSD merging methods.

We have already seen in Section 3 that the Cauchy combination method and the Simes method are IC-balanced. Below, we show that they are the only IC-balanced methods among the two classes of combining functions based on generalized mean and order statistics.

Theorem 1. *For a generalized mean function $M_{\phi,K}$ and an order statistics function $S_{\alpha,K}$,*

(i) $M_{\phi,K}$ is IC-balanced for all $K \in \mathbb{N}$ if and only if it is the Cauchy combining function, that is, $\phi(p)$ is a linear transform of $\tan(\pi(p-(1/2)))$, $p \in (0,1)$;

(ii) $S_{\alpha,K}$ is IC-balanced if and only if it is a positive constant times the Simes function.

The IC-balance of $M_{\phi,K}$ for some fixed K (instead of all $K \in \mathbb{N}$) does not imply that ϕ is the quantile function of a Cauchy distribution; see the counterexample (Example 1) in the Supplementary Material. As a direct consequence

of Theorem 1, if $S_{\alpha,K}$ is IC-balanced, then $S_{\alpha,k}$ for $k = 2, \dots, K-1$ are also IC-balanced (here, we use the first k components of α); a similar statement does not hold in general for the generalized mean functions, also shown by Example 1.

5. Connecting the Simes, the Harmonic Averaging and the Cauchy Combination Methods

As we have seen from Theorem 1, the Cauchy and Simes combining functions are the only IC-balanced ones among the two classes considered in Section 3. Although the harmonic combining function does not satisfy $b_F = c_F$, we observe empirically that the harmonic averaging method and the Cauchy combination method report very similar results in all simulations; see Section S1 of the Supplementary Material.

In this section, we explore the relationship between the three methods based on S_K, $M_{-1,K}$, and $M_{\mathcal{C},K}$. We first show that the harmonic averaging method is equivalent to the Cauchy combination method asymptotically in a few senses. Second, we show the Simes function S_K and the harmonic averaging function $M_{-1,K}$ are closely connected via $M_{-1,K} \le S_K \le \ell_K M_{-1,K}$, where ℓ_K is given in (3.1). Throughout this section, for fixed $K \in \mathbb{N}$, we write $a_{\mathcal{C}} = a_{M_{\mathcal{C},K}}$, $a_{\mathcal{S}} = a_{S_K}$, and $a_{\mathcal{H}} = a_{M_{-1,K}}$, and similarly for $b_{\mathcal{C}}$, $b_{\mathcal{S}}$, and $b_{\mathcal{H}}$.

We use the following assumption on the p-variables $U_1, \dots, U_K \in \mathcal{U}$:

(G) For each $1 \le i < j \le K$, (U_i, U_j) follows a bivariate Gaussian copula (which can be different for each pair).

The assumption (G) is mild and is imposed by Liu and Xie (2020, Condition C.1). Note that condition (G) includes independence and comonotonicity as special cases. The following theorem confirms the close relationship between the harmonic averaging method and the Cauchy combination method. Recall that the VC thresholds for both methods are the identity function, and thus it suffices to consider VAD and VI thresholds.

Theorem 2. *For fixed $K \in \mathbb{N}$, the harmonic averaging and the Cauchy combination methods are asymptotically equivalent in the following senses:*

(i) If $\min_{i\in\{1,\dots,K\}} p_i \downarrow 0$ and $\max_{i\in\{1,\dots,K\}} p_i \le c$ for some fixed $c \in (0,1)$, then

$$\frac{M_{\mathcal{C},K}(p_1, \dots, p_K)}{M_{-1,K}(p_1, \dots, p_K)} \to 1.$$

(ii) For K standard uniform random variables $U_1, \dots, U_K$ satisfying condition

(G),

$$\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\dots,U_K)<\varepsilon\right)\sim\mathbb{P}\left(M_{-1,K}(U_1,\dots,U_K)<\varepsilon\right)\sim\varepsilon,\ \text{as } \varepsilon\downarrow 0. \tag{5.1}$$

In particular, $b_{\mathcal{C}}(\varepsilon)\sim b_{\mathcal{H}}(\varepsilon)$ *as* $\varepsilon\downarrow 0$.

(iii) $a_{\mathcal{C}}(\varepsilon)\sim a_{\mathcal{H}}(\varepsilon)$ *as* $\varepsilon\downarrow 0$.

(iv) For $r\neq -1$,

$$\frac{M_{\mathcal{C},K}(p_1,\dots,p_K)}{M_{r,K}(p_1,\dots,p_K)}\not\to 1,\ \text{as}\max_{i\in\{1,\dots,K\}}p_i\downarrow 0.$$

Remark 2. The statement $\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\dots,U_K)<\varepsilon\right)\sim\varepsilon$ in Theorem 2 (ii) is implied by Theorem 1 of Liu and Xie (2020), which gives the same convergence rate for the weighted Cauchy combination method. For the weighted harmonic averaging method, we have a similar result (see (S3.13) in the Supplementary Material): For standard uniform random variables $U_1,\dots,U_K$ satisfying condition (G) and any $(w_1,\dots,w_K)\in[0,1]^K$ with $\sum_{i=1}^K w_i=1$, we have

$$\mathbb{P}\left(\sum_{i=1}^K w_i U_i^{-1}>\frac{1}{\varepsilon}\right)\sim\varepsilon,\ \text{as } \varepsilon\downarrow 0.$$

We omit a discussion on weighted merging methods because our focus is on comparing symmetric combination functions.

The first statement of Theorem 2 means that if at least one of the realized p-values is close to zero, the harmonic averaging and Cauchy combining functions will produce very close numerical results. This case is likely to happen in high-dimensional situations, where the number of p-variables is very large. Because condition (G) for (ii) in Theorem 2 is arguably mild, the thresholds of the two methods are similar for a small significance level under a wide range of dependence structures of p-variables (including independence and comonotonicity). Therefore, if the significance level is small, one likely arrives at the same statistical conclusions on the hypothesis testing by using either method. The third result in Theorem 2 illustrates the equivalence between the VAD thresholds of the harmonic averaging and Cauchy combination methods as the significance level goes to zero. The final result in Theorem 2 shows that among all averaging methods, the harmonic averaging method is the only one that is asymptotically equivalent to the Cauchy combination method.

The next result reveals a close relationship between the Simes and harmonic averaging methods.

Theorem 3. *For* $p_1, \dots, p_K \in [0,1]$,

$$M_{-1,K}(p_1, \dots, p_K) \le S_K(p_1, \dots, p_K) \le \ell_K M_{-1,K}(p_1, \dots, p_K).$$

The first inequality holds as an equality if $p_1 = \cdots = p_K$. *The second inequality holds as an equality if* $p_1 = p_k/k$, *for* $k = 2, \dots, K$. *As a result,* $a_{\mathcal{S}}/a_{\mathcal{H}} \in [1, \ell_K]$ *and* $b_{\mathcal{S}}/b_{\mathcal{H}} \in [1, \ell_K]$.

By Proposition 3 (i), the VI threshold of the harmonic averaging method satisfies $b_{\mathcal{H}}(\varepsilon) \sim \varepsilon = b_{\mathcal{S}}$ as $\varepsilon \downarrow 0$. Using Theorem 3, we further know that $b_{\mathcal{H}}(\varepsilon) < \varepsilon$ (the inequality is strict because $M_{-1,K} < S_K$ has probability one for independent p-variables). Therefore, we cannot directly use the asymptotic VI threshold ε of the harmonic averaging method, which needs to be corrected; see Wilson (2019).

To summarize the results in this section, the Cauchy combining function and the harmonic averaging function are very similar in several senses, and the Simes function is more conservative than the harmonic averaging function. Empirically, we find that the Simes function is only slightly more conservative; see Section S1 of the Supplementary Material.

6. Price for Validity

For a set of realized p-values, the decision on the hypothesis testing for some specific combining function is determined by the corresponding threshold. The VAD method can always control the size below the significance level; VSD methods may not have the correct size, but they yield more power than that of the VAD method. Therefore, there is always a trade-off between validity and efficiency, and thus, a price for validity.

For a combining function F and K standard uniform random variables $U_1, \dots, U_K$ with some specific dependence assumption (e.g., independence, comonotonicity, or condition (G)), let g_F be the VSD threshold, that is, $g_F(\varepsilon) = q_\varepsilon(F(U_1, \dots, U_K))$. Let a_F be defined as in (2.2). For some fixed $\varepsilon \in (0,1)$, the ratio $g_F(\varepsilon)/a_F(\varepsilon)$ is called the *price for validity* under the corresponding dependence assumption of the p-variables. For instance, $b_F(\varepsilon)/a_F(\varepsilon)$ is the price paid for validity under the independence assumption, and $c_F(\varepsilon)/a_F(\varepsilon)$ is the corresponding price under the comonotonicity assumption. For a specific application, one may consider the price for validity under other dependence assumptions. The

calculation of the price for validity serves two purposes:

i (Power gain/loss): On the one hand, if additional information on the dependence structure of the p-values is available, the price for validity can be used as a measure for the gain of power from the dependence information. On the other hand, if the dependence information is not available or credible, the price can be used to measure the power loss by switching to the VAD threshold.

ii (Sensitivity to model misspecification): If the dependence structure is ambiguous, VAD thresholds should be used. A small price for validity indicates a relatively small change of threshold due to the model ambiguity. Hence, the price for validity can be used as a tool to assess the sensitivity of VSD methods to model misspecification.

Remark 3. Instead of using the price for validity, a more direct way to assess the trade-off between using VSD and VAD methods is to compare the sizes, $\mathbb{P}(F(P_1,\dots,P_K) < g_F(\varepsilon))/\mathbb{P}(F(P_1,\dots,P_K) < a_F(\varepsilon))$, where the dependence of the p-variables $P_1,\dots,P_K$ corresponds to the VSD method. More precisely, for a fixed $\varepsilon \in (0,1)$, the ratio of the sizes is $\varepsilon/g_F^{-1}(a_F(\varepsilon))$, where g_F^{-1} is the (generalized) inverse of g_F. The connection between the price for validity and the ratio of sizes is explained below.

(i) For the Simes and Cauchy combination methods, the ratios of the sizes under independence and comonotonicity are identical to the corresponding price for validity, because $b_F(\varepsilon) = c_F(\varepsilon)$, for $\varepsilon \in (0,1)$.

(ii) For the averaging methods, the ratios of the sizes under comonotonicity are identical to the price for validity, because c_F is an identity function. The ratios of the sizes under independence may differ from $b_F(\varepsilon)/a_F(\varepsilon)$; however, by letting $\delta = a_F(\varepsilon)$, we have ($a_F$ is strictly increasing in all cases we consider)

$$\frac{\varepsilon}{b_F^{-1}(a_F(\varepsilon))} = \frac{a_F^{-1}(\delta)}{b_F^{-1}(\delta)}.$$

This is very similar to $b_F(\varepsilon)/a_F(\varepsilon)$; it is a matter of examing the ratio of the threshold functions or that of their inverses. In fact, if $r < 0$, by Proposition 3, we have,

$$\frac{\varepsilon}{b_F^{-1}(a_F(\varepsilon))} \sim \frac{b_F(\varepsilon)}{a_F(\varepsilon)},\ \varepsilon \downarrow 0,$$

which suggests that the ratio of the sizes is almost the same as the price for validity under independence for small significance levels.

Table 2. Thresholds for K p-variables at significance level $\varepsilon \in (0,1)$.

	Bonferroni	Negative-quartic	Simes	Cauchy	Harmonic	Geometric
$a_F(\varepsilon)$	ε/K	$(3/4)K^{-3/4}\varepsilon$	ε/ℓ_K	(3.7)	(3.3)	(3.2)
$b_F(\varepsilon)$	$1-(1-\varepsilon)^{1/K}$	(3.4)	ε	ε	(3.4)	(3.5)

Table 3. $b_F(\varepsilon)/a_F(\varepsilon)$ and $c_F(\varepsilon)/a_F(\varepsilon)$ for $\varepsilon = 0.01$ and $K \in \{50, 100, 200, 400\}$.

	$K=50$		$K=100$		$K=200$		$K=400$	
	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F
Bonferroni	1.005	50.000	1.005	100.000	1.005	200.000	1.005	400.000
Negative-quartic	1.340	25.071	1.340	42.164	1.340	70.911	1.340	119.257
Simes	4.499	4.499	5.187	5.187	5.878	5.878	6.570	6.570
Cauchy	6.625	6.625	7.465	7.465	8.277	8.277	9.058	9.058
Harmonic	6.658	6.625	7.496	7.459	8.314	8.273	9.117	9.072
Geometric	69.903	2.718	78.096	2.718	84.214	2.718	88.694	2.718

We use the Bonferroni method based on the combining function $F = M_{-\infty,K}$ as an example to illustrate the above idea. Using Proposition 2 and noting that $K(1-(1-\varepsilon)^{1/K}) \sim \varepsilon$ as $\varepsilon \downarrow 0$, we obtain that the prices for validity of the Bonferroni method satisfy $c_F(\varepsilon)/a_F(\varepsilon) = K$ for $\varepsilon \in (0,1)$ and $b_F(\varepsilon)/a_F(\varepsilon) \to 1$ as $\varepsilon \downarrow 0$. Therefore, for a small ε close to zero, the price for validity under the independence assumption is close to one, and the price for validity under the comotonicity assumption increases linearly as the number of p-variables increases. This means a model misspecification of independence does not affect the Bonferroni method significantly, whereas a model misspecification of comonotonicity greatly affects the statistical conclusion of the Bonferroni method.

Next, we numerically calculate the prices for validity under the independence and comonotonicity assumptions for various merging methods using the results in Section 3. We consider the Bonferroni, harmonic averaging, geometric averaging, Cauchy combination, Simes, and negative-quartic (using $M_{-4,K}$, a compromise between the Bonferroni and harmonic averaging) methods. The (asymptotic) VAD and VI thresholds of these methods are summarized in Table 2. The VC threshold is an identity function for all these methods. The VAD threshold of the negative-quartic method is given by Proposition 5 of Vovk and Wang (2021). Numerical results on the prices for validity are reported in Table 3 for $\varepsilon = 0.01$. Although some of the VAD thresholds in Table 2 do not have explicit forms, the numerical computation is very fast. The results for $\varepsilon = 0.05$ and $\varepsilon = 0.0001$ are similar and reported in Tables 1 and 2, respectively, in the Supplementary Material.

The Bonferroni and negative-quartic methods pay a much lower price under the independence assumption than they do under the comonotonicity assumption, and the geometric averaging method is the opposite. On the other hand, the harmonic averaging, Simes, and Cauchy combination methods have relatively small prices under both independence and comonotonicity assumptions, and their prices increase at moderate rates as K increases, compared with those of the other methods. In particular, the harmonic averaging and Cauchy combination methods exhibit very similar performance (cf. Theorem 2), and their prices are slightly larger than that of the Simes method. If mild model misspecification exists, it may be safer to choose one of the harmonic averaging, Simes, or Cauchy combination methods and use the corresponding VAD threshold without losing much power. The prices for validity in Table 3 can also be interpreted as inflations of sizes by using the VSD threshold against the VAD threshold, except for the geometric averaging method (see Remark 3).

Next, we show that the prices for validity of the harmonic averaging, Cauchy combination, and Simes methods behave like $\log K$ for K large enough and ε small enough.

Proposition 5. *For $\varepsilon \in (0,1)$, the prices for validity satisfy the following:*

(i) For the harmonic averaging method, $F = M_{-1,K}$,

$$\lim_{\delta \downarrow 0} \frac{b_F(\delta)}{a_F(\delta)} = \frac{c_F(\varepsilon)}{a_F(\varepsilon)} \sim \log K, \ \text{as } K \to \infty.$$

(ii) For the Cauchy combination method, $F = M_{\mathcal{C},K}$,

$$\lim_{\delta \downarrow 0} \frac{b_F(\delta)}{a_F(\delta)} = \lim_{\delta \downarrow 0} \frac{c_F(\delta)}{a_F(\delta)} \sim \log K, \ \text{as } K \to \infty.$$

(iii) For the Simes method, $F = S_K$,

$$\frac{b_F(\varepsilon)}{a_F(\varepsilon)} = \frac{c_F(\varepsilon)}{a_F(\varepsilon)} \sim \log K, \ \text{as } K \to \infty.$$

Numerical values of the ratios of the prices for validity under the independence assumption and $\log K$ are reported in Table 4; the results for the corresponding ratios under the comonotonicity assumption are similar for these methods. The Simes method has the fastest convergence rate among the three methods. The ratios for the harmonic averaging and Cauchy combination methods converge quite slowly, and have similar rates. This fact can also be explained by

Table 4. Numerical values of $(1/\log(K))(b_F(\varepsilon)/a_F(\varepsilon))$ for the Simes, Cauchy combination, and harmonic averaging methods.

	ε	$K = 10$	20	50	100	200	500
Simes	0.05	1.272035	1.200955	1.150097	1.126425	1.109415	1.093041
	0.01	1.272035	1.200955	1.150097	1.126425	1.109415	1.093041
Cauchy	0.05	1.979572	1.82826	1.693025	1.620527	1.561670	1.511264
	0.01	1.980144	1.828822	1.693562	1.621011	1.562121	1.504288
Harmonic	0.05	2.026308	1.873762	1.73641	1.661098	1.601539	1.539448
	0.01	1.989255	1.837605	1.701851	1.627702	1.569179	1.508248

Theorem 3, where we see that the Simes function is, in general, larger than the harmonic averaging function.

Based on Proposition 5, one may be tempted to use $b_F/\log K$ as the corrected critical value under a model misspecification; however, for the harmonic averaging and Cauchy combination methods, the asymptotic rate of $\log K$ can only be expected for very large K (instead, $1.7 \log K$ works for $K \geq 100$).

7. Conclusions

We have discussed two aspects of merging p-values: the impact of the dependence structure on the critical thresholds, and the trade-off between validity and efficiency. The Cauchy combination method and Simes method are shown to be the only IC-balanced members among the generalized mean class and the order statistics class of combining functions. The harmonic averaging and Cauchy combination methods are asymptotically equivalent, and the Simes and harmonic averaging methods have a simple algebraic relationship. For the above three methods, the prices for validity under the independence (comonotonicity) assumption all behave like $\log K$ for large K. Moreover, numerical studies in the Supplementary Material suggest that these methods lose a moderate amount of power if VAD thresholds are used, and their performance against model misspecification is better than that of other methods. This explains the wide application of these methods in various statistical procedures.

Merging p-values is not only useful for testing a single hypothesis, but is also important for testing multiple hypotheses, controlling the FDR (Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001)), and exploratory research (Goeman and Solari (2011), Goeman et al. (2019)). In many situations, especially those involving a large number of hypotheses and tests, dependence information is rarely available. Our results offer some insights, especially in terms of the gain/loss of validity and power, into how the absence of such information influ-

ences different statistical procedures that merge p-values.

In many practical applications, p-values arrive sequentially over time, and the existence of the nth p-variable may depend on previously observed p-values (only promising experiments may be continued); thus, the number of experiments to combine is a stopping time. Unfortunately, the merging method of p-values discussed in this paper cannot be used to sequentially update p-values with an arbitrary stopping rule. To deal with such a situation, one has to rely on anytime-valid methods, typically using a test supermartingale (see Howard et al. (2021) and Ramdas et al. (2020)) or e-values (see Shafer (2021) and Vovk and Wang (2021)). Moreover, e-values are nicer to combine (e.g., using the average or product, as in Vovk and Wang (2021)), especially under arbitrary dependence, in contrast to the complicated methods of merging p-values.

Supplementary Material

The online Supplementary Material contains simulation studies and a real-data analysis. An `R` package `pmerge` for the various merging methods discussed in this paper is available at `https://github.com/YuyuChen-UW/pmerge`. All technical proofs, additional remarks, and tables are also provided in the Supplementary Material.

Acknowledgments

The authors thank Aaditya Ramdas and Vladimir Vovk for their helpful advice on an earlier version of the paper. The authors thank the associate editor and two anonymous reviewers for their valuable comments and suggestions. Ruodu Wang acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-03823, RGPAS-2018-522590) and the University of Waterloo CAE Research Grant from the Society of Actuaries.

References

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

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

Bernard, C., Jiang, X. and Wang, R. (2014). Risk aggregation with dependence uncertainty. *Insurance: Mathematics and Economics* **54**, 93–108.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32**, 962–994.

Embrechts, P., Puccetti, G. and Rüschendorf, L. (2013). Model uncertainty and var aggregation. *Journal of Banking and Finance* **37**, 2750–2764.

Embrechts, P., Wang, B. and Wang, R. (2015). Aggregation-robustness and model uncertainty of regulatory risk measures. *Finance and Stochastics* **19**, 763–790.

Efron, B. (2010). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge.

Fisher, R. A. (1948). Combining independent tests of significance. *American Statistician* **2**, 30.

Goeman, J. J., Meijer, R. J., Krebs, T. J. and Solari, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika* **106**, 841–856.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science* **26**, 584–597.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal* **25**, 423–430.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* **49**, 1055–1080.

Liu, Y. and Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association* **115**, 393–402.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika* **25**, 379–410.

Ramdas, A. K., Barber, R. F., Wainwright, M. J. and Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics* **47**, 2790–2821.

Ramdas, A., Ruf, J., Larsson, M. and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint*, arXiv:2009.03167.

Rødland, E. A. (2006). Simes' procedure is 'valid on average'. *Biometrika* **93**, 742–746.

Rüger, B. (1978). Das maximale signifikanzniveau des tests: "lehneh o ab, wennk untern gegebenen tests zur ablehnung führen". *Metrika* **25**, 171–178.

Samorodnitsky, G. (2017). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Routledge, New York.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *The Annals of Statistics* **26**, 494–504.

Sarkar, S. K. (2008). On the Simes inequality and its generalization. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (Edited by N. Balakrishnan, E. A. Pena and M. J. Silvapulle), 231–242. Institute of Mathematical Statistics, Beachwood.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **184**, 407–431.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

Tippett, L. H. C. (1931). *The Methods of Statistics: An Introduction Mainly for Experimentalists*. Williams and Norgate, London.

Uchaikin, V. V. and Zolotarev, V. M. (2011). *Chance and Stability: Stable Distributions and their Applications*. Walter de Gruyter, Berlin.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika* **107**, 791–808.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *The Annals of Statistics* **49**, 1736–1754.

Vovk, V., Wang, B. and Wang, R. (2022). Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics* **50**, 351–375.

Wang, B. and Wang, R. (2016). Joint mixability. *Mathematics of Operations Research* **41**, 808–826.

Wang, R., Peng, L. and Yang, J. (2013). Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities. *Finance and Stochastics* **17**, 395–417.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences* **116**, 1195–1200.

Yuyu Chen

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

E-mail: y937chen@uwaterloo.ca

Peng Liu

Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK.

E-mail: peng.liu@essex.ac.uk

Ken Seng Tan

Nanyang Business School, Nanyang Technological University, Singapore.

E-mail: kenseng.tan@ntu.edu.sg

Ruodu Wang

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

E-mail: wang@uwaterloo.ca

(Received February 2021; accepted August 2021)