Trade-off between validity and efficiency of merging p-values under arbitrary dependence

Yuyu Chen[*] Peng Liu[†] Ken Seng Tan[‡] Ruodu Wang[§]

Abstract

Various methods of combining individual p-values into one p-value are widely used in many areas of statistical applications. We say that a combining method is valid for arbitrary dependence (VAD) if it does not require any assumption on the dependence structure of the p-values, whereas it is valid for some dependence (VSD) if it requires some specific, perhaps realistic but unjustifiable, dependence structures. The trade-off between validity and efficiency of these methods is studied via 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 sizes and powers of various combining methods in the presence of weak and strong dependence.

Keywords: Hypothesis testing; multiple hypothesis testing; validity; efficiency.

[*] Department of Statistics and Actuarial Science, University of Waterloo. E-mail: y937chen@uwaterloo.ca.

[†] Department of Mathematical Sciences, University of Essex. E-mail: peng.liu@essex.ac.uk.

[‡] Nanyang Business School, Nanyang Technological University. E-mail: kenseng.tan@ntu.edu.sg.

[§] Department of Statistics and Actuarial Science, University of Waterloo. E-mail: wang@uwaterloo.ca.

1 Introduction

In many areas of statistical applications where 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 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 shall refer to as a *combining function*. Generally, the combined value may not be a valid p-value itself, and 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. 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 very difficult or impossible to verify in many applications where only one set of p-values is available.

There are, however, some methods that produce valid p-values without any dependence assumption. A classic one is the Bonferroni method by taking the minimum of the p-values times K (we allow combined p-values to be greater than 1 and they can be treated as 1) or equivalently, dividing the critical value by K. Other methods that are valid without assumptions include the ones based on order statistics by Rüger (1978) and Hommel (1983), and the ones based on averaging by Vovk and Wang (2020a); details of these merging methods are presented in Section 3.

Some 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 $i = 1, \dots, K$, where $p_{(i)}$ is the i-th smallest order statistic of $p_1, \dots, p_K$. The validity of the Simes method is shown under a large class of dependence structures (e.g., Sarkar (1998, 2008); Benjamini and Yekutieli (2001) and Rødland (2006)), although even such dependence assumptions are unlikely to hold in practice (see e.g., Efron (2010, p.51)). Two more recent methods include the Cauchy combination test proposed by Liu and Xie (2020) using the weighted average of Cauchy transformed p-values, and the harmonic mean p-value of Wilson (2019) using the harmonic mean of p-values. Under mild dependence assumptions, these two methods are asymptotically valid as the significance level goes to 0 (see Theorem 2).

This paper is dedicated to 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 p-values, and they will be referred to as *VAD methods*. On the other hand, methods that are valid for some specific but realistic dependence assumption (e.g., independence, positive dependence, or joint normality dependence) will be referred to as *VSD methods*. 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 compared to a VSD method. Thus, the gain of validity comes at the price of a loss of detection power. As it is often difficult to make valid statistical inference on the dependence structure of p-values, our analysis also helps to understand the relative performance of VSD combining methods under 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. Formulas for their VAD and VSD threshold functions are derived, some based on results from robust risk aggregation, e.g., Wang et al. (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 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 presented in Section 7 to analyze the relative performance of these methods. Proofs of all technical results are put in the supplementary material.

We conclude the section by providing additional notation and terminology that will be 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

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

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 the smallest to the 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 (2020a), a *p-variable* is a random variable P such that $\mathbb{P}(P \leq \varepsilon) \leq \varepsilon$, 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 $F : [0,1]^K \to [0, \infty)$ which 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. Generally, $F(P_1, \dots, P_K)$ 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 dependence structure of p-variables, e.g., 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 p-variables, we use b_F (resp. c_F) to denote the corresponding valid threshold function, and we call it the *VI* (resp. *VC*) *threshold*. More precisely, for the equation

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

a VAD threshold $g = a_F$ satisfies (1) for all p-variables $P_1,\dots,P_K$; a VI threshold $g = b_F$ satisfies (1) for all independent p-variables $P_1,\dots,P_K$, and a VC threshold $g = c_F$ satisfies (1) for all comonotonic p-variables $P_1,\dots,P_K$.

The comonotonicity assumption on the p-variables to combine (actually they are identical if they are uniform on $[0,1]$) is not interesting by itself for statistical 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 will be made more clear in Sections 4-7.

An immediate observation is that the p-variables can be equivalently replaced by uniform random variables on $[0,1]$ as for each p-variable P, we can find $U \in \mathcal{U}$ with $U \leqslant P$; see e.g., Vovk and Wang (2020a). Therefore, it suffices to consider p-variables in $\mathcal{U}$. Moreover, if g satisfies (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 (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 $\varepsilon \in (0,1)$,

$$a_F(\varepsilon) = \inf\{q_\varepsilon(F(U_1,\dots,U_K)) \mid U_1,\dots,U_K \in \mathcal{U}\}, \tag{2}$$

$$b_F(\varepsilon) = q_\varepsilon(F(V_1,\dots,V_K)), \tag{3}$$

$$c_F(\varepsilon) = q_\varepsilon(F(U,\dots,U)), \tag{4}$$

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

It is clear that $g = a_F$, b_F or c_F in Definition 1 satisfies (1) under the respective dependence assumptions.

Remark 1. While the objects b_F and c_F in (3)-(4) can often be explicitly calculated, the object a_F in (2) is generally difficult to calculate for a chosen function F due to the infimum taken over all possible dependence structures. Techniques in the field of robust risk aggregation, in particular, results in Wang et al. (2013), Embrechts et al. (2013, 2015) and Wang and Wang (2016), are designed for such calculation, as illustrated by Vovk and Wang (2020a).

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 the 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 (2020a) 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 $\phi(p) = \tan\left(\pi\left(p - \frac{1}{2}\right)\right)$.

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 0, 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 the other components of α to be 0. 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 will simply denote $S_{\alpha,K}$ by S_K, namely,

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

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

$$\ell_K = \sum_{k=1}^{K} \frac{1}{k}. \tag{5}$$

If $\alpha_{i+1} \leqslant \alpha_i$, then the term $p_{(i+1)}/\alpha_{i+1}$ does not contribute to the calculation of $S_{\alpha,K}(p_1, \dots, p_K)$. Hence, we can safely replace α_{i+1} by α_i without changing the function $S_{\alpha,K}$. Thus, we shall assume, without loss of generality, that $\alpha_1 \leqslant \ldots \leqslant \alpha_K$. Admissibility of VAD merging methods in the above two classes are studied by Vovk et al. (2020).

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 (2020a). 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 will discuss several special cases of the above two classes of combining functions, and analyze their corresponding threshold functions. As the first example, we note that the functions a_F, b_F and c_F for the Bonferroni method can be easily verified.

Proposition 2. *Let $F(p_1, \dots, p_K) = \min\{p_1, \dots, p_K\}$ for $p_1, \dots, p_K \in [0, 1]$. 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 (2020a) use the combining functions given by

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

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

$$M_{-\infty,K}(p_1, \dots, p_K) = \min\{p_1, \dots, p_K\};$$

$$M_{0,K}(p_1, \dots, p_K) = \left(\prod_{i=1}^K p_i \right)^{\frac{1}{K}};$$

$$M_{\infty,K}(p_1, \dots, p_K) = \max\{p_1, \dots, p_K\}.$$

Some special cases of the combining function 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.

Since 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 have been well studied in Vovk and Wang (2020a), and here we mainly focus on the cases of Platonic means and the Bonferroni method. It is known that $a_{-\infty} = K$ and $a_1 = 2$. For $r = 0$ or $r = -1$, the values of a_r

and their asymptotic formulas are calculated by Propositions 4 and 6 of Vovk and Wang (2020a), summarized below for $K \geqslant 3$.

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

(ii) $a_{-1} = \frac{(y_K+1)K}{(y_K+K)^2}$ 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 will use stable distributions (e.g., Samorodnitsky (2017)) below. 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 $\theta \in \mathbb{R}$,

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

where $\mathrm{sgn}(\cdot)$ is the sign function. For $\alpha \geqslant 2$, F_α stands 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 $K \in \mathbb{N}_+$

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

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

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

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

(ii) If $r = 0$, then

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

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

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

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

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

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

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

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

3.3 The Cauchy combination method

The Cauchy combination method is recently proposed by Liu and Xie (2020) which 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), namely,

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

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 the use of 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 behaviour as the one under the assumption of independence (see Theorem 2 (ii) below).

Since $\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(U_i)$ 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 will be studied in more detail in Section 4.

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

$$a_F(\varepsilon) = \mathcal{C}\left(\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(U_i)\right) \mid U_1,\dots,U_K \in \mathcal{U}\right\}\right). \tag{6}$$

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

Proposition 4. *For $\varepsilon \in (0,1/2)$, we have $a_F(\varepsilon) = \mathcal{C}\left(-H_\varepsilon(x_K)/K\right)$, where*

$$H_\varepsilon(x) = (K-1)\mathcal{C}^{-1}(1-\varepsilon+(K-1)x) + \mathcal{C}^{-1}(1-x), \quad x \in (0, \varepsilon/K),$$

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 $S_K(p_1,\dots,p_K) = \min_{i\in\{1,\dots,K\}} \frac{K}{i}p_{(i)}$. 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 e.g., Sarkar (1998, 2008).

In the context of testing multiple hypotheses, if p-variables for several hypothesis 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 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 Independence-comonotonicity 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 assumption, an arguably convenient feature. Inspired by this observation, we introduce the property of *independence-comonotonicity 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 resulting combined 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 *independence-comonotonicity balanced* (*IC-balanced*) if $F(V_1,\dots,V_K) \overset{\mathrm{d}}{=} F(U,\dots,U)$.

As 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 p-variables is a mixture of independence and comonotonicity, i.e., with the copula

$$\lambda \prod_{i=1}^{n} x_i + (1-\lambda) \min_{i=1,\dots,n} x_i, \quad (x_1,\dots,x_n) \in [0,1]^n, \tag{7}$$

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

For any combining function F, VI (VC) thresholds generally yield more power to the test compared with the corresponding VAD threshold, but the gain of power may come with the invalidity due 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 (7)) 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, i.e., $\phi(p)$ is a linear transform of $\tan\left(\pi\left(p - \frac{1}{2}\right)\right)$, *$p \in (0,1)$;*

(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 counter-example (Example A.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 component of α); a similar statement does not hold in general for the generalized mean functions, also shown by Example A.1.

Remark 2. The property of IC-balance should be seen as a necessary but not sufficient condition for a merging method to be insensitive to dependence between independence and comonotonicity. As shown by Sarkar (1998), the Simes method is valid for positive regression dependence, which is a large spectrum of dependence structures connecting independence and comonotonicity (larger than (7)); on the other hand, the Cauchy combination method using VI threshold is valid under a bivariate Gaussian assumption asymptotically but not precisely (Liu and Xie (2020)); see Theorem 2 below and the simulation studies in Section 7. Instead of arguing for the practical usefulness of IC-balance, we emphasize it as a necessary condition for insensitivity to dependence. The main aim of Theorem 1 is, via this necessary condition, to pin down the unique role of the Simes and the Cauchy combination methods among their respective generalized classes, thus justifying their advantages with respect to dependence.

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 7.

In this section, we explore the relationship among the three methods based on S_K, $M_{-1,K}$ and $M_{\mathcal{C},K}$. We first show that 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 $M_{-1,K} \leqslant S_K \leqslant \ell_K M_{-1,K}$, where ℓ_K is given in (5). Throughout this section, for fixed $K \in \mathbb{N}$, we write $a_{\mathcal{C}} = a_{M_{\mathcal{C},K}}$, $a_{\mathcal{S}} = a_{S_K}$, $a_{\mathcal{H}} = a_{M_{-1,K}}$ and similarly for $b_{\mathcal{C}}$, $b_{\mathcal{S}}$ and $b_{\mathcal{H}}$.

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

(G) For each $1 \leq i < j \leq K$, (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 look at 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 $\min_{i\in\{1,\ldots,K\}} p_i \downarrow 0$ and $\max_{i\in\{1,\ldots,K\}} p_i \leq c$ for some fixed $c \in (0,1)$, then

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

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

$$\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\ldots,U_K) < \varepsilon\right) \sim \mathbb{P}\left(M_{-1,K}(U_1,\ldots,U_K) < \varepsilon\right) \sim \varepsilon, \ \text{ as } \varepsilon \downarrow 0. \tag{8}$$

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 $r \neq -1$,

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

Remark 3. The statement $\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\ldots,U_K) < \varepsilon\right) \sim \varepsilon$ 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 (A.20) in the supplementary material): For standard uniform random variables $U_1,\ldots,U_K$ satisfying condition (G) and any $(w_1,\ldots,w_K) \in [0,1]^K$ with $\sum_{i=1}^K w_i = 1$, we have

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

We omit a discussion on weighted merging methods as the focus of this paper is comparing symmetric combination functions.

The first statement of Theorem 2 means that, if at least one of realized p-values are close to 0, the harmonic averaging and the 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. As the 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 method and the Cauchy combination method as the significance level goes to 0. 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.

Remark 4. We note that the equivalence

$$\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\dots,U_K)<\varepsilon\right)\sim\mathbb{P}\left(M_{-1,K}(U_1,\dots,U_K)<\varepsilon\right)$$

in (8) does not always hold under arbitrary dependence structures. Since the Cauchy distribution is symmetric, it is possible that $\mathbb{P}(\mathcal{C}^{-1}(U_1)+\cdots+\mathcal{C}^{-1}(U_K)=0)=1$ for some $U_1,\dots,U_K\in\mathcal{U}$, implying $\mathbb{P}(M_{\mathcal{C},K}(U_1,\dots,U_K)<1/2)=0$. Indeed, Theorem 4.2 of Puccetti et al. (2019) implies that there exist K standard Cauchy random variables whose sum is a constant c, for each $c\in[-K\log(K-1)/\pi, K\log(K-1)/\pi]$. On the other hand, $\mathbb{P}(M_{-1,K}(U_1,\dots,U_K)<\varepsilon)>0$ for all $\varepsilon>0$ and all $U_1,\dots,U_K\in\mathcal{U}$. Thus, $\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\dots,U_K)<\varepsilon\right)\sim\mathbb{P}\left(M_{-1,K}(U_1,\dots,U_K)<\varepsilon\right)$ does not hold.

Remark 5. The equivalence in Theorem 2 (ii) relies on the p-variables being uniform on $[0,1]$. For p-variables that are stochastically larger than uniform, the behaviour of the Cauchy combination method and that of the harmonic averaging method may diverge; nevertheless, by Theorem 2 (i), for a realized vector of p-values with at least one very small component, the two methods would produce similar values.

The next result reveals an intimate relationship between the Simes and the harmonic averaging methods.

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

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

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 $k = 2, \ldots, K$. As a consequence, $a_{\mathcal{S}}/a_{\mathcal{H}} \in [1, \ell_K]$ 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 since $M_{-1,K} < S_K$ has probability 1 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 see that the Simes function is only slightly more conservative; see Section 7.

6 Prices for validity

For a given set of realized p-values, the decision to the hypothesis testing for some specific combining function will be 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 the VAD method. Therefore, there is always a trade-off between validity and efficiency, thus a price for validity.

For a combining function F, let g_F be the VSD threshold under some specific dependence assumption of the p-variables, e.g., independence, comonotonicity, or condition (G). 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 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 for 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 that 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.

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 0, the price for validity under the independence assumption is close to 1 while the price for validity under the comotonicity assumption increases linearly as the number of p-variables increases. This means a model misspecification of independence is not affecting the Bonferroni method much, whereas a model misspecification of comonotonicity greatly affects the statistical conclusion of the Bonferroni method.

Next we numerically calculate the prices for validity under independence and comonotonicity assumptions for various merging methods using results in Section 3. We consider the Bonferroni, the harmonic averaging, the geometric averaging, the Cauchy combination, the Simes, and the negative-quartic (using $M_{-4,K}$, a compromise between Bonferroni and harmonic averaging) methods. Numerical results on the prices for validity are reported in Table 2 for $\varepsilon = 0.01$. The results for $\varepsilon = 0.05$ are similar and reported in Table B.4 in the supplementary material.

	$K = 50$		$K = 100$		$K = 200$		$K = 400$	
	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

Table 2: $b_F(\varepsilon)/a_F(\varepsilon)$ and $c_F(\varepsilon)/a_F(\varepsilon)$ for $\varepsilon = 0.01$ and $K \in \{50, 100, 200, 400\}$

The Bonferroni and the negative-quartic methods pay much lower price under the independence assumption than the comonotonicity assumption, and the geometric averaging method is the absolute opposite. On the other hand, the harmonic averaging, the Simes and the Cauchy combination methods have relatively small prices under both independence and comonotonicity assumptions and their prices increase at moderate rates as K increases, compared to other methods. In particular, the harmonic averaging and the Cauchy combination methods have 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, the Simes and the Cauchy combination methods and use the corresponding VAD threshold without losing much power.

Next, we show that the prices for validity of the harmonic averaging, the Cauchy combination and the 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:*

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

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

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

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

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

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

Numerical values of the ratios between the price for validity under independence assumption and $\log K$ are reported in Table 3; the results for the corresponding ratios under 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 the Cauchy combination methods converge quite slowly and have similar rates. This fact can also be explained by Theorem 3, where we see that the Simes function is generally 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 model misspecification; however, for the harmonic averaging and the Cauchy combination methods, the asymptotic rate of $\log K$ can only be expected for very large K (instead, $1.7 \log K$ works for $K \geqslant 100$).

Remark 6. Instead of using $g_F(\varepsilon)/a_F(\varepsilon)$, an alternative way to define the price for validity is the ratio of the type-I errors, $\mathbb{P}(F(P_1, \dots, P_K) < g_F(\varepsilon))/\mathbb{P}(F(P_1, \dots, P_K) < a_F(\varepsilon))$, where the dependence of p-variables $P_1, \dots, P_K$ corresponds to the VSD method. More precisely, for a fixed $\varepsilon \in (0,1)$, the price for validity is $\varepsilon / g_F^{-1}(a_F(\varepsilon))$, where g_F^{-1} is the (generalized) inverse of g_F. This alternative formulation is similar to our $g_F(\varepsilon)/a_F(\varepsilon)$, as we explain below.

	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

Table 3: Numerical values of $\frac{1}{\log(K)}\frac{b_F(\varepsilon)}{a_F(\varepsilon)}$ for the Simes, the Cauchy combination and the harmonic averaging methods.

(i) For the averaging, the Simes and the Cauchy combination methods, the alternative prices under comonotonicity are identical to our definition $c_F(\varepsilon)/a_F(\varepsilon)$ since c_F is the identity.

(ii) Similarly, the alternative prices for the Simes and the Cauchy combination methods under independence are identical to our definition $b_F(\varepsilon)/a_F(\varepsilon)$ since b_F is the identity.

(iii) For the averaging methods, $\varepsilon/b_F^{-1}(a_F(\varepsilon))$ may be different 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 our definition of prices, $b_F(\varepsilon)/a_F(\varepsilon)$; it is a matter of looking at the ratio of threshold functions or that of their inverses.

7 Simulations and a real data example

7.1 Simulation studies

We conduct K one-sided z-tests of the null hypothesis: $\mu_i = 0$ against the alternative hypothesis $\mu_i > 0$, $i = 1, \dots, K$, using the test statistic X_i and the p-value p_i from the ith test, $i = 1, \dots, K$. The tests are formulated as the following:

$$p_i = \Phi(X_i), \quad X_i = \rho Z + \sqrt{1-\rho^2} Z_i - \mu_i, \quad i = 1, \dots, K.$$

where Φ is the standard normal distribution function, $Z, Z_1, \dots, Z_K$ are iid standard normal random variables, $\mu_i \geq 0$, $i = 1, \dots, K$, and ρ is a parameter in $[0, 1]$. Note that for $\rho = 0$, the p-variables are independent, and $\rho = 1$ corresponds to the case where p-variables are comonotonic.

Let $K \in \{50, 200\}$ and set the significance level $\varepsilon = 0.01$. To see how different dependence structures and signals affect the size and the power for various methods using both VAD and VSD thresholds, the rejection probabilities (RPs) are computed over $\rho \in [0,1]$ under the following four cases:

(i) (no signal) 100% of μ_i's are 0;

(ii) (needle in a haystack) 98% of μ_i's are 0 and 2% of μ_i's are 4;

(iii) (sparse signal) 90% of μ_i's are 0 and 10% of μ_i's are 3;

(iv) (dense signal) 100% of μ_i's are 2.

The RP corresponds to the size under case (i), and it corresponds to the power under (ii), (iii) and (iv). The RP is computed as the ratio between the number of the combined values which are less than the critical threshold and the number of simulations for some $\rho \in [0,1]$, that is,

$$\mathrm{RP} = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{F_i < g(\varepsilon)\}}}{N},$$

where N is the number of simulations and is equal to 15000 in our study, F_i is the realized value of the combining function for the i-th simulation, $i = 1, \dots, N$, and $g(\varepsilon)$ is the corresponding critical value. For $\rho \in [0,1]$, graphs of RPs for different combining methods are drawn using VAD thresholds and VSD thresholds. Some observations from Figures 1-4 are made below, and those on the averaging methods using $M_{r,K}$ are consistent with the observations in Vovk and Wang (2020a).

(a) All VAD methods give sizes less than $\varepsilon = 0.01$ as expected. Using VAD thresholds, the Bonferroni, the harmonic averaging, the Cauchy combination and the Simes methods have good powers.

(b) The Simes method using thresholds b_F or c_F reports the right size for all values of ρ. Sarkar (1998) showed the validity of the Simes method in the so-called MTP_2 class including multivariate normal distributions with nonnegative correlations (the setting of our simulation).

(c) Using thresholds b_F or c_F, the harmonic averaging and Cauchy combination methods perform similarly with sizes possibly larger than 0.01 (see Theorems 2 and 3).

(d) The geometric averaging method using b_F and the Bonferroni and negative-quartic methods using c_F do not yield correct sizes under model misspecification, and the sizes increase rapidly as the misspecification gets bigger.

(e) Using b_F or c_F, the harmonic averaging, the Cauchy combination and the Simes methods have good performances on capturing the signals.

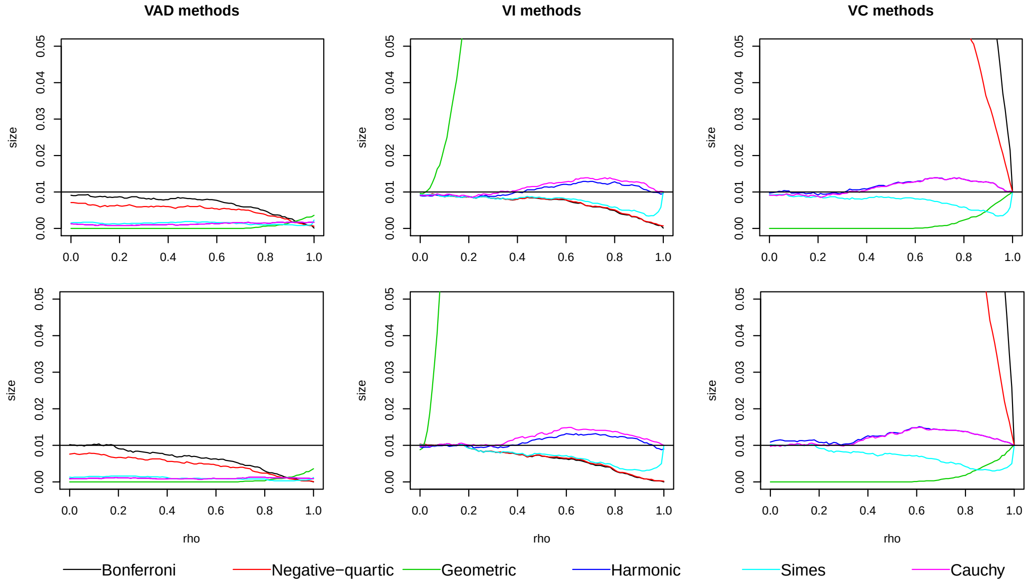

Figure 1: Case (i): size (top: $K = 50$, bottom: $K = 200$)

7.2 Real data analysis

We apply several merging methods to a genomewide study to compare their performances. We use the dataset of p-values of Storey and Tibshirani (2003) which contains 3170 p-values computed based on the data from Hedenfalk et al. (2001) for testing whether genes are differentially expressed between BRCA1- and BRCA2-mutation-positive tumors. As mentioned in Section 2, $g^{-1} \circ F(P_1, \dots, P_K)$ is a p-variable if the threshold g is strictly increasing, and it is the quantity we choose to compare combined p-values for different methods.

For each method, we calculate the combined p-value, and remove the smallest p-value from the dataset. Repeat this procedure until the resulting combined p-value loses significance. Using the Bonferroni combining function, this leads to the Bonferroni-Holm (BH) procedure (Holm (1979)); thus we mimic the BH procedure for other methods in a naive manner. The rough interpretation is to report the number of significant discoveries (this procedure generally does not control the family-wise error rate (FWER); to control FWER one needs to use a generalized BH procedure as in Vovk and Wang (2020a) or Goeman et al. (2019). This procedure can be seen as a lower confidence bound from a closed testing perspective). For a visual comparison of detection power,

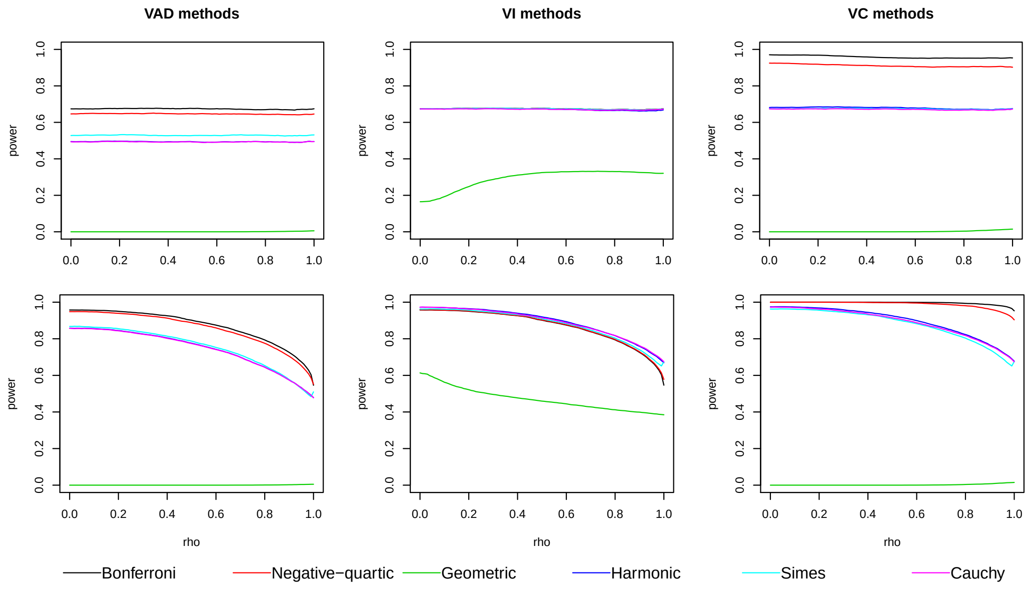

Figure 2: Case (ii): needle in a haystack (top: $K = 50$, bottom: $K = 200$)

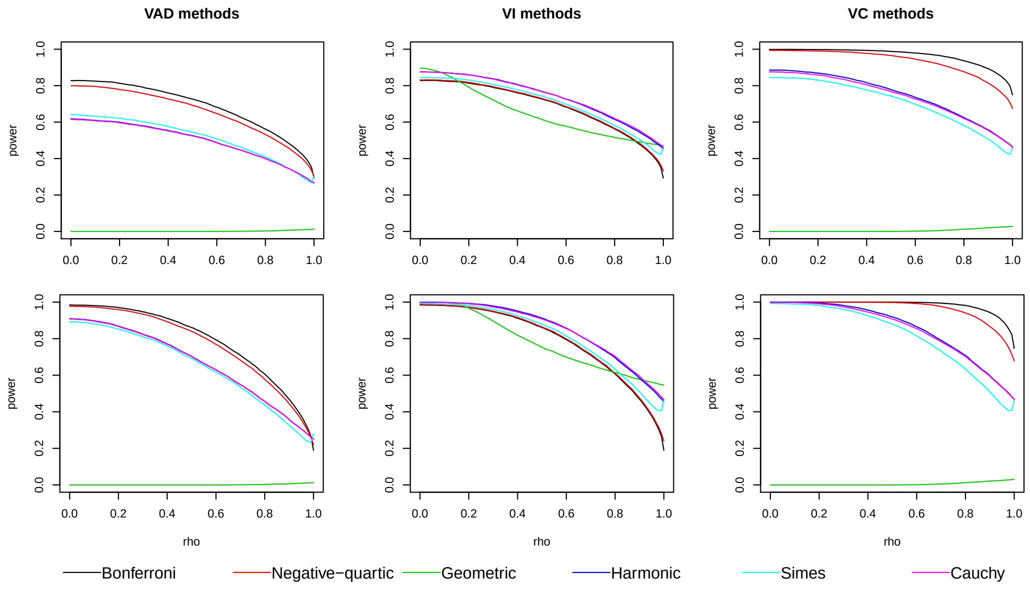

Figure 3: Case (iii): sparse signal (top: $K = 50$, bottom: $K = 200$)

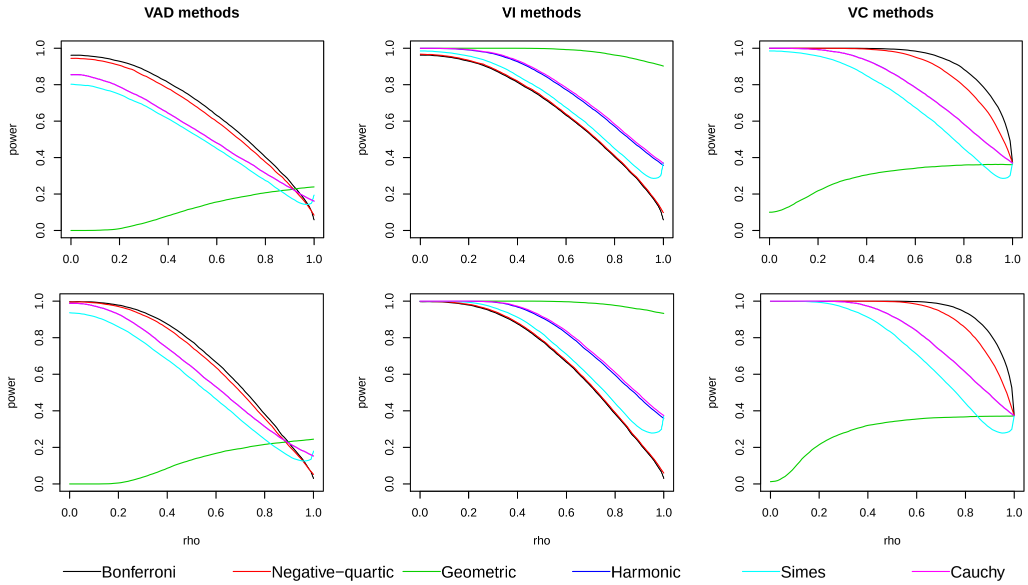

Figure 4: Case (iv): dense signal (top: $K = 50$, bottom: $K = 200$)

the combined p-values against the numbers of removed p-values are plotted in Figure 5, where we use both the VAD and the VI thresholds (comonotonicity is obviously unrealistic here).

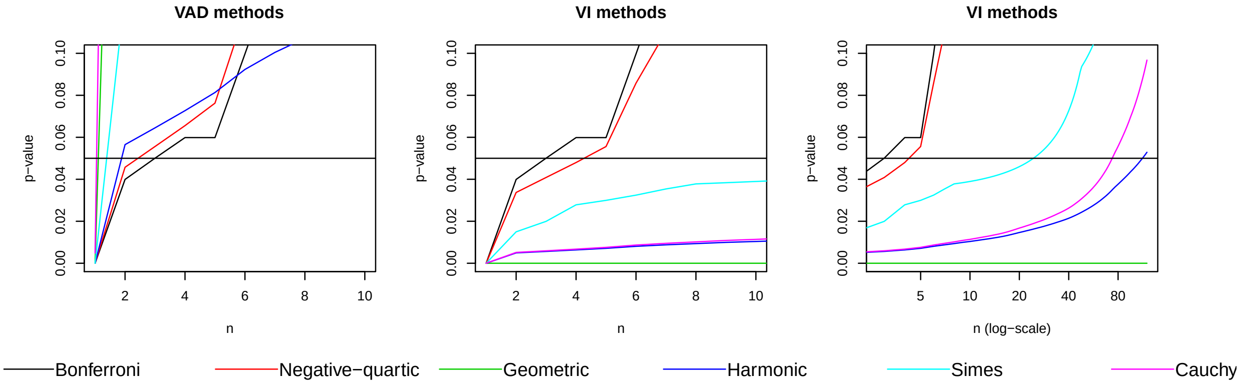

Figure 5: Combined p-value after removing n smallest p-values

All VAD methods lose significance at $\varepsilon = 0.05$ after omitting the first or the second smallest p-value (the smallest p-value is 0 and the second smallest is 1.26×10^{-5}). Using thresholds b_F for independence, the Bonferroni and the negative quartic methods behave similarly to their VAD versions (as their price for validity is close to 1). In contrast, the Simes, the Cauchy combination and the harmonic averaging methods lose significance at $\varepsilon = 0.05$ after removing around 20, 70 and 110 p-values respectively. The geometric averaging method (Fisher's) exceeds 0.05 only after removing around 400 p-values. However, this method relies heavily on the independence assumption, which is impossible to verify from just one set of p-values.

8 Concluding remarks

We 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 the 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 the Cauchy combination methods are asymptotically equivalent, and the Simes and the harmonic averaging methods have simple algebraic relationship. For the above three methods, the prices for validity under independence (comonotonicity) assumption all behaves like $\log K$ for large K. Moreover, these methods lose moderate amount of power if VAD thresholds are used, and their performance against model misspecification is better than other methods. This explains the wide applications of these methods in different statistical procedures.

Merging p-values is not only useful for testing a single hypothesis, but also important in testing multiple hypotheses, controlling false discovery rate (Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001)), and exploratory research (Goeman and Solari (2011), Goeman et al. (2019)). In many situations especially involving a large number of hypotheses and tests, dependence information is hardly available. The results in our paper offer some insights, especially in terms of gain/loss of validity and power, on how the absence of such information influences different statistical procedures of merging p-values.

In many practical applications, p-values arrive sequentially in time, and the existence of the n-th 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 current merging method of p-values discussed in this paper cannot be used to sequentially update p-values with arbitrary stopping rule. To deal with such a situation, one has to rely on anytime-valid methods, typically through the use of a test supermartingale (see Howard et al. (2020) and Ramdas et al. (2020)) or through e-values (see Shafer (2020) and Vovk and Wang (2020b)). Moreover, e-values are nicer to combine (e.g., using average and product as in Vovk and Wang (2020b)) especially under arbitrary dependence, in contrast to the complicated methods of merging p-values.

Acknowledgements

The authors thank Aaditya Ramdas and Vladimir Vovk for helpful comments on an earlier version of the paper. 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*, *57*(1), 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*(4), 1165–1188.

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

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling Extremal Events for Insurance and Finance.* Springer Science & Business Media.

Embrechts, P., Puccetti, G., and Rüschendorf, L. (2013). Model uncertainty and var aggregation. *Journal of Banking and Finance*, *37*(8), 2750–2764.

Embrechts, P., Wang, B. and Wang, R. (2015). Aggregation-robustness and model uncertainty of regulatory risk measures. *Finance and Stochastics*, *19*(4), 763–790.

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

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

Föllmer, H. and Schied, A. (2016). *Stochastic Finance. An Introduction in Discrete Time.* Walter de Gruyter, Berlin, Fourth Edition.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal*, *25*(5), 423–430.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, *26*(4), 584–597.

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*(4), 841–856.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., et al. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, *344*(8), 539–548.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, forthcoming.

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.

McNeil, A. J., Frey, R. and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools*. Revised Edition. Princeton, NJ: Princeton University Press.

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*(3), 379–410.

Puccetti, G., Rigo, P., Wang, B., and Wang, R. (2019). Centers of probability measures without the mean. *Journal of Theoretical Probability*, *32*(3), 1482–1501.

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. *Annals of Statistics*, *47*(5), 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*(3), 742–746.

Rüger, B. (1978). Das maximale signifikanzniveau des tests:"lehneh o ab, wennk untern gegebenen tests zur ablehnung führen". *Metrika*, *25*(1), 171–178.

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

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. *Annals of Statistics*, *26*(2), 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* (pp. 231–242). Institute of Mathematical Statistics.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, Cambridge.

Shafer, G. (2020). The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, forthcoming.

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

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440–9445.

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

Vovk, V., Wang, B. and Wang, R. (2020). Admissible ways of merging p-values under arbitrary dependence. *arXiv preprint*, arXiv:2007.14208.

Vovk, V. and Wang, R. (2020a). Combining p-values via averaging. *Biometrika*, forthcoming.

Vovk, V. and Wang, R. (2020b). E-values: Calibration, combination, and applications. *Annals of Statistics*, forthcoming.

Wang, B. and Wang, R. (2016). Joint mixability. *Mathematics of Operations Research*, *41*(3), 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*(2), 395–417.

Wang, X. (2005). Volumes of generalized unit balls. *Mathematics Magazine*, *78*(5), 390–395.

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

Supplementary Material for

Trade-off between validity and efficiency of merging p-values under arbitrary dependence

A Proofs of theorems and propositions

A.1 Proof of Proposition 1

By definition, we have

$$a_F(\varepsilon) = \inf\{q_\varepsilon(F(U_1,\dots,U_K)) \mid U_1,\dots,U_K \in \mathcal{U}\},\ \varepsilon \in (0,1).$$

We shall show

$$a_F(\varepsilon) = \inf\{q_1(F(V_1,\dots,V_K)) \mid V_1,\dots,V_K \in \mathcal{U}_\varepsilon\},\ \varepsilon \in (0,1), \tag{A.9}$$

where $\mathcal{U}_\varepsilon$ denotes the collection of all uniform random variables distributed on $[0,\varepsilon]$. Denote by $S = F(U_1,\dots,U_K)$ and $G_S^{-1}(t) = q_t(S)$, $t \in (0,1]$. We can find $U_S \in \mathcal{U}$ such that $G_S^{-1}(U_S) = S$ a.s. (e.g., Lemma A.32 of Föllmer and Schied (2016)). Let $f_i(t) = \mathbb{P}\left(U_i \le t | U_S < \varepsilon\right)$, $t \in [0,1]$. Then $f_i(U_i)$ conditionally on $U_S < \varepsilon$ is a uniform random variable on $[0,1]$ and $V_i^\varepsilon := \varepsilon f_i(U_i)$ conditionally on $U_S < \varepsilon$ is a uniform random variable on $[0,\varepsilon]$. We construct the following two random variables:

$$S_1 = S\mathbb{1}_{\{U_S<\varepsilon\}} + d\mathbb{1}_{\{U_S\ge\varepsilon\}},\ S_2 = F(V_1^\varepsilon,\dots,V_n^\varepsilon)\mathbb{1}_{\{U_S<\varepsilon\}} + d\mathbb{1}_{\{U_S\ge\varepsilon\}}, \tag{A.10}$$

where $d > F(\varepsilon,\dots,\varepsilon)$. Noting the fact that $\varepsilon f_i(t) = \mathbb{P}(U_i \le t, U_S < \varepsilon) \le t$, $t \in [0,1]$ and F is increasing, we have $S_1 \ge S_2$. Hence $q_\varepsilon(S_1) \ge q_\varepsilon(S_2)$. Moreover, direct calculation shows $q_\varepsilon(S) = q_\varepsilon(S_1)$. Thus $q_\varepsilon(S) \ge q_\varepsilon(S_2)$. Let $\hat{V}_1,\dots,\hat{V}_n$ be uniform random variables on $[0,\varepsilon]$ such that $(\hat{V}_1,\dots,\hat{V}_n)$ has the joint distribution identical to the conditional distribution of $(V_1^\varepsilon,\dots,V_n^\varepsilon)$ on $U_S < \varepsilon$. Hence, for $x < d$,

$$\begin{aligned}\mathbb{P}(S_2 \le x) &= \mathbb{P}(F(V_1^\varepsilon,\dots,V_n^\varepsilon) \le x, U_S < \varepsilon)\\ &= \varepsilon\mathbb{P}(F(V_1^\varepsilon,\dots,V_n^\varepsilon) \le x | U_S < \varepsilon)\\ &= \varepsilon\mathbb{P}(F(\hat{V}_1,\dots,\hat{V}_n) \le x).\end{aligned}$$

This implies $q_\varepsilon(S_2) = q_1(F(\hat{V}_1,\dots,\hat{V}_n))$. Thus we have

$$a_F(\varepsilon) \ge \inf\{q_1(F(V_1,\dots,V_K)) \mid V_1,\dots,V_K \in \mathcal{U}_\varepsilon\}.$$

We next show "$\le$" in (A.9). Take $V_1,\dots,V_n \in \mathcal{U}_\varepsilon$ and $U \in \mathcal{U}$ such that U is independent of $V_1,\dots,V_n$. Let $\hat{U}_i = V_i \mathbb{1}_{\{U<\varepsilon\}} + U\mathbb{1}_{\{U\ge\varepsilon\}}$, $i=1,2,\dots,n$. It is clear that $\hat{U}_i \in \mathcal{U}$, $i=1,2,\dots,n$ and $F(\hat{U}_1,\dots,\hat{U}_n) = F(V_1,\dots,V_n)\mathbb{1}_{\{U<\varepsilon\}} + F(U,\dots,U)\mathbb{1}_{\{U\ge\varepsilon\}}$. Noting that F is increasing, we have $q_1(F(V_1,\dots,V_n)) = q_\varepsilon(F(\hat{U}_1,\dots,\hat{U}_n))$. This implies

$$a_F(\varepsilon) \le \inf\{q_1(F(V_1,\dots,V_K)) \mid V_1,\dots,V_K \in \mathcal{U}_\varepsilon\}.$$

Therefore, (A.9) holds. By (A.9) and the homogeneity of F we have that for $\varepsilon \in (0,1)$,

$$\begin{aligned} a_F(\varepsilon) &= \inf\{q_1(F(V_1,\dots,V_K)) \mid V_1,\dots,V_K \in \mathcal{U}_\varepsilon\} \\ &= \inf\{q_1(F(\varepsilon U_1,\dots,\varepsilon U_K)) \mid U_1,\dots,U_K \in \mathcal{U}\} \\ &= \varepsilon \inf\{q_1(F(U_1,\dots,U_K)) \mid U_1,\dots,U_K \in \mathcal{U}\}. \end{aligned}$$

This completes the proof. □

A.2 Proof of Proposition 2

It is well known that the Bonferroni correction yields $a_F(\varepsilon) = \varepsilon/K$. Also, since the average of identical objects is itself, $c_F(\varepsilon) = \varepsilon$ for any averaging method, including the Bonferroni method. For iid standard uniform random variables $V_1,\dots,V_K$, we have $\mathbb{P}(\min\{V_1,\dots,V_K\} \le x) = 1-(1-x)^K$. Therefore, $b_F(\varepsilon) = 1-(1-\varepsilon)^{1/K}$ for $\varepsilon \in (0,1)$. □

A.3 Proof of Proposition 3

(a) Suppose $r<0$. We first fix K and find the asymptotic of b_r as $\varepsilon \downarrow 0$ satisfying

$$\mathbb{P}\left(\sum_{i=1}^{K} P_i^r \ge K\left(b_r(\varepsilon)\right)^r\right) = \varepsilon.$$

Observe that the random variables P_i^r, $i=1,\dots,K$, follow a common Pareto distribution with cdf $\mathbb{P}(P_i^r \le x) = 1-x^{1/r}$, $x \in (1,\infty)$, $i=1,\dots,K$. Note that the tail probability of the sum of iid Pareto random variables is asymptotically the same as that of the maximum of the iid Pareto random variables (e.g., Embrechts et al. (2013), Corollary 1.3.2). Hence

$$\lim_{\varepsilon\downarrow 0} \frac{\mathbb{P}\left(\sum_{i=1}^{K} P_i^r \ge K\left(b_r(\varepsilon)\right)^r\right)}{\mathbb{P}\left(\max\{P_1^r,\dots,P_K^r\} > K\left(b_r(\varepsilon)\right)^r\right)} = \lim_{\varepsilon\downarrow 0} \frac{\varepsilon}{1-\left(1-K^{\frac{1}{r}}b_r(\varepsilon)\right)^K} = 1.$$

This implies

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

The case $K \to \infty$ follows directly from the generalized central limit theorem (e.g., Theorem 1.8.1 of Samorodnitsky (2017)).

(b) If $r = 0$, in a similar way, we first have,

$$\mathbb{P}\left(2\sum_{i=1}^{K}\log\frac{1}{P_i} \geq 2K\log\frac{1}{b_r(\varepsilon)}\right) = \varepsilon.$$

The random variable $\log\frac{1}{P_i}$, $i = 1, \ldots, K$, follows exponential distribution with parameter 1. Thus $2\sum_{i=1}^{K}\log\frac{1}{P_i}$ follows a chi-square distribution with parameter $2K$. We denote $q_\alpha(\chi^2_\nu)$ the α-quantile of the chi-square distribution with ν degrees of freedom. Hence

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

(c) If $r > 0$, using the result of Wang (2005), we have for $0 \leq x \leq K^{-r}$,

$$\begin{aligned}\mathbb{P}\left(M_{r,K}(U_1, \ldots, U_K) \leq x\right) &= \mathbb{P}\left(\sum_{i=1}^{K}U_i^r \leq Kx^r\right)\\ &= \lambda\left\{(x_1, \ldots, x_K) : \sum_{i=1}^{K}x_i^r \leq Kx^r,\ x_1, \ldots, x_K \geq 0\right\}\\ &= \frac{(\Gamma(1+1/p))^K}{\Gamma(1+K/p)}K^{K/r}x^K,\end{aligned}$$

where λ is the Lebesgue measure. This implies that if $\varepsilon \leq \frac{(\Gamma(1+1/p))^K}{\Gamma(1+K/p)}$,

$$b_r(\varepsilon) = \frac{(\Gamma(1+K/p))^{1/K}\varepsilon^{1/K}}{K^{1/r}\Gamma(1+1/p)}. \tag{A.11}$$

The asymptotic behaviour of $b_r(\varepsilon)$ for fixed $\varepsilon \in (0,1)$ as $K \to \infty$ can be obtained by the Central Limit Theorem. Note that the random variables P_i^r, $i = 1, \ldots, K$, follow a common Beta distribution with mean and variance given by, respectively,

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

The Central Limit Theorem gives $(\sum_{i=1}^{K}P_i^r - K\mu)/\sqrt{K}\sigma \xrightarrow{\text{d}} \mathrm{N}(0,1)$. Hence

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

where Φ^{-1} is the inverse of the standard normal distribution function. □

A.4 Proof of Proposition 4

By symmetry of the standard Cauchy distribution,

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

Moreover, $\mathcal{C}^{-1}(U_i),\ i = 1,\dots,K$, follow the standard Cauchy distribution with decreasing density on $[\mathcal{C}^{-1}(1-\varepsilon),\infty]$ for $\varepsilon \in (0,1/2)$. The proposition follows directly from applying Corollary 3.7 of Wang et al. (2013). □

A.5 Proof of Theorem 1

(i) IC-balance of $M_{\phi,K}$ for all $K \in \{2,3,\dots\}$ is equivalent to $\frac{1}{K}\sum_{i=1}^{K}\phi(V_i) \overset{\mathrm{d}}{=} \phi(U)$ for all $K \in \{2,3,\dots\}$, which is further equivalent to the fact that $\phi(U)$ follows a strictly 1-stable distribution. We know that strictly 1-stable distributions are Cauchy distributions (see, e.g., Theorem 14.15 of Sato (1999)). This proves the statement of part (i).

(ii) For the Simes function $S_{\alpha,K} = S_K$, $\alpha_i = i$ for $i \in \{1,\dots,K\}$ and $b_F(x) = c_F(x) = x$ for $x \in [0,1]$. Therefore, $S_{\alpha,K}$ is IC-balanced.

Below we show the opposite direction of the statement. For $n \in \{2,\dots,K\}$, let $V_{(1)},\dots,V_{(n)}$ be the order statistics for n independent standard uniform random variables $V_1,\dots,V_n$. Let $(X_1,\dots,X_{n-1}) = (V_{(1)}/V_{(n)},\dots,V_{(n-1)}/V_{(n)})$ which is identically distributed as the order statistics for $n-1$ independent standard uniform random variables, independent of $V_{(n)}$. Hence, for $x \in (0,1/\alpha_n)$,

$$
\begin{aligned}
&\mathbb{P}\left(S_{\alpha,n}(V_1,\dots,V_n) > x\right)\\
&= \mathbb{P}\left(V_{(1)} > x\alpha_1,\dots,V_{(n-1)} > x\alpha_{n-1}, V_{(n)} > x\alpha_n\right)\\
&= \mathbb{P}\left(X_1 > x\alpha_1/V_{(n)},\dots,X_{n-1} > x\alpha_{n-1}/V_{(n)}, V_{(n)} > x\alpha_1\right)\\
&= \int_{x\alpha_n}^{1}\mathbb{P}\left(X_1 > x\alpha_1/p,\dots,X_{n-1} > x\alpha_{n-1}/p\right) np^{n-1}\,\mathrm{d}p\\
&= \int_{x\alpha_n}^{1}\mathbb{P}\left(S_{\alpha,n-1}(V_1,\dots,V_{n-1}) > x/p\right) np^{n-1}\,\mathrm{d}p, \qquad \text{(A.12)}
\end{aligned}
$$

where for simplicity we use $S_{\alpha,n-1}$ for $S_{(\alpha_1,\dots,\alpha_{n-1}),n-1}$. Note that

$$
\mathbb{P}\left(S_{\alpha,1}(V_1) > x\right) = 1 - \alpha_1 x, \quad x \in (0,1/\alpha_1). \qquad \text{(A.13)}
$$

Plugging (A.13) in (A.12), we obtain that $\mathbb{P}\left(S_{\alpha,2}(V_1,V_2)>x\right)$ is a polynomial function of x of degree less than or equal to 2. Recursively, using (A.12) we are able to show that the function $\mathbb{P}\left(S_{\alpha,n}(V_1,\dots,V_n)>x\right)$ for $x\in(0,1/\alpha_n)$ is a polynomial of x of degree less than or equal to n for $n=2,\dots,K$. Hence, there exist K constants $\beta_0,\dots,\beta_{K-1}$ such that

$$\mathbb{P}\left(S_{\alpha,K-1}(V_1,\dots,V_{K-1})>x\right)=\sum_{i=0}^{K-1}\beta_i x^i,\quad x\in(0,1/\alpha_{K-1}).$$

Moreover, noting that $S_{\alpha,K}$ is IC-balanced, we have

$$\int_{x\alpha_K}^{1}\mathbb{P}\left(S_{\alpha,K-1}(V_1,\dots,V_{K-1})>x/p\right)Kp^{K-1}\,\mathrm{d}p=\mathbb{P}\left(S_{\alpha,K}(U,\dots,U)>x\right)=1-x\alpha_K,$$

for $x\in(0,1/\alpha_K)$. Therefore, we have

$$\int_{x\alpha_K}^{1}\left(\sum_{i=0}^{K-1}\beta_i x^i p^{-i}\right)Kp^{K-1}\,\mathrm{d}p=1-x\alpha_K,$$

which implies that for $x\in(0,1/\alpha_K)$,

$$\sum_{i=0}^{K-1}\frac{K\beta_i}{K-i}x^i-\left(\sum_{i=0}^{K-1}\frac{K\beta_i}{K-i}\alpha_K^{K-i}\right)x^K=1-x\alpha_K.$$

Solving the above equation, we get $\beta_0=1$, $\beta_1=-\frac{K-1}{K}\alpha_K$ and $\beta_2=\cdots=\beta_{K-1}=0$. Consequently,

$$\mathbb{P}\left(S_{\alpha,K-1}(V_1,\dots,V_{K-1})>x\right)=1-\frac{K-1}{K}\alpha_K x,\quad x\in(0,1/\alpha_{K-1}).$$

Recursively, using (A.12) we have

$$\mathbb{P}\left(S_{\alpha,n}(V_1,\dots,V_n)>x\right)=1-\frac{n}{K}\alpha_K x,\quad x\in(0,1/\alpha_n)\tag{A.14}$$

for $n=1,\dots,K$, which gives, using (A.13),

$$\alpha_K=K\alpha_1.\tag{A.15}$$

Inserting (A.14) into (A.12), we obtain, for $x\in(0,1/\alpha_n)$ and $n=2,\dots,K$,

$$\begin{aligned}1-\frac{n}{K}\alpha_K x&=\int_{x\alpha_n}^{1}\left(1-\frac{n-1}{K}\alpha_K x p^{-1}\right)np^{n-1}\,\mathrm{d}p\\&=1-\frac{n}{K}\alpha_K x+\left(\frac{n}{K}\alpha_K\alpha_n^{n-1}-\alpha_n^n\right)x^n.\end{aligned}$$

Consequently,

$$\alpha_n=\frac{n}{K}\alpha_K,\quad n=2,\dots,K,$$

which together with (A.15) implies $\alpha_n=n\alpha_1,\ k=1,\dots,K$. This gives the desired statement.

□

Example A.1 (IC-balanced generalized mean for a finite K)**.** We show that IC-balance of $M_{\phi,K}$ for a finite K does not imply $M_{\phi,K}$ that ϕ is the Cauchy quantile function (up to an affine transform). For this purpose, we construct a continuous distribution μ such that

$$\frac{1}{K}\sum_{i=1}^{K} X_i \overset{\mathrm{d}}{=} X, \tag{A.16}$$

where X and $X_i, i = 1, \ldots, K$ are iid random variables with distribution μ, but μ is not a Cauchy distribution. Define

$$\hat{\mu}(z) = \exp\left(\int_{\mathbb{R}} \left(e^{izx} - 1 - \mathbb{1}_{[-1,1]}(x)\right) \nu(\mathrm{d}x)\right), \ z \in \mathbb{R},$$

where $\mathbb{1}_{[-1,1]}(\cdot)$ is the indicator function, $i^2 = -1$ and ν is a symmetric measure on $\mathbb{R}\setminus\{0\}$ satisfying

$$\nu(\{K^n\}) = \nu(\{-K^n\}) = K^{-n}, \ n \in \mathbb{Z}, \text{ and } \nu\left(\mathbb{R}\setminus\left(\{0\} \cup \bigcup_{n\in\mathbb{Z}} \{K^n, -K^n\}\right)\right) = 0.$$

It follows from Theorem 8.1 of Sato (1999) that $\hat{\mu}$ is the characterization function of some infinitely divisible distribution μ. Also noting that $\nu(\mathbb{R}\setminus\{0\}) = \infty$, by Theorem 27.16 of Sato (1999) we know that μ is a continuous distribution. By Theorem 14.7 of Sato (1999), $(\mu(z))^b = \mu(bz), \ z \in \mathbb{R}, b > 0$ holds if and only if

$$T_b\nu(B) = b\nu(B), \text{ and } \int_{1<|x|\le b} x\nu(\mathrm{d}x) = 0,$$

where $T_b\nu(B) = \nu(b^{-1}B)$ for all Borel sets $B \subset \mathbb{R}$. By symmetry of ν, $\int_{1<|x|\le b} x\nu(\mathrm{d}x) = 0$ holds for any $b > 0$. However, $T_b\nu(B) = b\nu(B)$ holds only for $b \in \{K^n, n \in \mathbb{Z}\}$. Consequently, $(\mu(z))^b = \mu(bz), \ z \in \mathbb{R}$ if and only if $b \in \{K^n, n \in \mathbb{Z}\}$. This implies that μ is not a Cauchy distribution (strictly 1-stable distribution) but (A.16) holds.

A.6 Proof of Theorem 2

(i) Recall that

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

Note that $\mathcal{C}^{-1}(x) \sim -1/(\pi x)$ as $x \downarrow 0$ and $\mathcal{C}(y) \sim -1/(\pi y)$ as $y \to -\infty$. For any $\delta_1, \delta_2 \in (0, 1/K)$, there exists $0 < \varepsilon < 1$ and $m < 0$ such that for all $x \in (0, \varepsilon)$ and $y \in (-\infty, m)$,

$$-\frac{(1+\delta_1)}{\pi x} \le \mathcal{C}^{-1}(x) \le -\frac{(1-\delta_1)}{\pi x}; \tag{A.17}$$

$$-\frac{(1-\delta_2)}{\pi y} \leq \mathcal{C}(y) \leq -\frac{(1+\delta_2)}{\pi y}. \tag{A.18}$$

For $0 < c < 1$, there exists $0 < \varepsilon' < \varepsilon$ such that

$$\sup_{x \in [\varepsilon, c]} \left| \tan\left(-\frac{\pi}{2} + \pi x\right) + \frac{1}{\pi x} \right| \leq \frac{\delta_1}{\pi \varepsilon'}. \tag{A.19}$$

Take $(p_1, \dots, p_K)$ such that $p_{(1)} < \varepsilon'$ and $p_{(K)} \leq c < 1$. Let $l = \max\{i = 1, \dots, K : p_{(i)} < \varepsilon\}$. As a consequence of (A.17), we have

$$-\sum_{i=1}^{l} \frac{(1+\delta_1)}{\pi p_{(i)}} \leq \sum_{i=1}^{l} \tan\left(-\frac{\pi}{2} + \pi p_{(i)}\right) \leq -\sum_{i=1}^{l} \frac{(1-\delta_1)}{\pi p_{(i)}}.$$

For $j > l$, (A.19) implies

$$\left| \tan\left(-\frac{\pi}{2} + \pi p_{(j)}\right) + \frac{1}{\pi p_{(j)}} \right| \leq \frac{\delta_1}{\pi \varepsilon'} \leq \frac{\delta_1}{\pi p_{(1)}}.$$

Therefore,

$$\begin{aligned} \sum_{i=1}^{K} \tan\left(-\frac{\pi}{2} + \pi p_i\right) &\leq -\sum_{i=1}^{l} \frac{(1-\delta_1)}{\pi p_{(i)}} - \sum_{i=l+1}^{K} \frac{1}{\pi p_{(i)}} + \frac{(K-l)\delta_1}{\pi p_{(1)}} \\ &\leq -\sum_{i=1}^{K} \frac{(1-K\delta_1)}{\pi p_{(i)}} \\ &= -\sum_{i=1}^{K} \frac{(1-K\delta_1)}{\pi p_i}. \end{aligned}$$

Similarly, we can show

$$\sum_{i=1}^{K} \tan\left(-\frac{\pi}{2} + \pi p_i\right) \geq \sum_{i=1}^{K} -\frac{(1+K\delta_1)}{\pi p_i}.$$

Using (A.18), for any $(p_1, \dots, p_K)$ satisfying $p_{(1)} < \min(\varepsilon', \frac{K\delta_1 - 1}{K\pi m})$ and $p_{(K)} \leq c < 1$,

$$\frac{1-\delta_2}{1+K\delta_1} M_{-1,K}(p_1, \dots, p_K) \leq M_{\mathcal{C},K}(p_1, \dots, p_K) \leq \frac{1+\delta_2}{1-K\delta_1} M_{-1,K}(p_1, \dots, p_K).$$

We establish the claim by letting $\delta_1, \delta_2 \downarrow 0$, and the above inequalities hold as long as $p_{(1)}$ is sufficiently small.

(ii) The statement

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

follows directly from Theorem 1 of Liu and Xie (2020) by noting that standard Cauchy distribution is symmetric at 0. Below we show $\mathbb{P}\left(M_{-1,K}(U_1,\dots,U_K)<\varepsilon\right)\sim\varepsilon$ as $\varepsilon\downarrow 0$, based on similar techniques as in Theorem 1 of Liu and Xie (2020). Observe that

$$\mathbb{P}\left(M_{-1,K}(U_1,\dots,U_K)<\varepsilon\right)=\mathbb{P}\left(\frac{1}{K}\sum_{i=1}^{K}U_i^{-1}>1/\varepsilon\right).$$

Condition (G) means that for any $1\le i<j\le K$, $(\Phi^{-1}(U_i),\Phi^{-1}(U_j))$ is a bivariate normal random variable with $\mathrm{cov}(\Phi^{-1}(U_i),\Phi^{-1}(U_j))=\sigma_{ij}$, where Φ is the standard normal distribution function and Φ^{-1} is its inverse. Clearly, $\sigma_{ij}=1$ implies that $U_i=U_j$ a.s. In this case we can combine them in one and the corresponding coefficient becomes $2/K$. Thus, it suffices to prove the stronger statement

$$\mathbb{P}\left(\sum_{i=1}^{K}w_iU_i^{-1}>1/\varepsilon\right)\sim\varepsilon,\ \text{as }\varepsilon\downarrow 0,\tag{A.20}$$

where $w_i>0,\ i=1,\dots,K,\ \sum_{i=1}^{K}w_i=1$ and $\sigma_{ij}<1,\ i,j=1,\dots,K$. We choose some positive constant δ_ε depending on ε, such that $\delta_\varepsilon\to 0$ and $\delta_\varepsilon/\varepsilon\to\infty$ as $\varepsilon\downarrow 0$. Denote by $S=\sum_{i=1}^{K}w_iU_i^{-1}$, and define the following events: for $i\in\{1,\dots,K\}$,

$$A_{i,\varepsilon}=\left\{U_i^{-1}>\frac{1+\delta_\varepsilon}{w_i\varepsilon}\right\},\quad B_{i,\varepsilon}=\left\{U_i^{-1}\le\frac{1+\delta_\varepsilon}{w_i\varepsilon},\ S>1/\varepsilon\right\}.$$

Let $A_\varepsilon=\bigcup_{i=1}^{K}A_{i,\varepsilon}$ and $B_\varepsilon=\bigcap_{i=1}^{K}B_{i,\varepsilon}$ and thus we have

$$\mathbb{P}\left(S>1/\varepsilon\right)=\mathbb{P}(A_\varepsilon)+\mathbb{P}(B_\varepsilon).$$

First we show $\mathbb{P}(B_\varepsilon)=o(\varepsilon)$. Note that $S>1/\varepsilon$ implies that there exists $i\in\{1,\dots,K\}$ such that $U_i^{-1}>\frac{1}{w_iK\varepsilon}$. Hence,

$$\begin{aligned}\mathbb{P}\left(B_\varepsilon\right)&\le\sum_{i=1}^{K}\mathbb{P}\left(\frac{1}{w_iK\varepsilon}<U_i^{-1}\le\frac{1+\delta_\varepsilon}{w_i\varepsilon},S>1/\varepsilon\right)\\&\le\sum_{i=1}^{K}\mathbb{P}\left(\frac{1}{w_iK\varepsilon}<U_i^{-1}\le\frac{1-\delta_\varepsilon}{w_i\varepsilon},S>1/\varepsilon\right)+\sum_{i=1}^{K}\mathbb{P}\left(\frac{1-\delta_\varepsilon}{w_i\varepsilon}<U_i^{-1}\le\frac{1+\delta_\varepsilon}{w_i\varepsilon}\right)\\&\le\sum_{i=1}^{K}\mathbb{P}\left(\frac{1}{w_iK\varepsilon}<U_i^{-1}\le\frac{1-\delta_\varepsilon}{w_i\varepsilon},S>1/\varepsilon\right)+\sum_{i=1}^{K}w_i\varepsilon\left(\frac{1}{1-\delta_\varepsilon}-\frac{1}{1+\delta_\varepsilon}\right)\\&=:I_1+I_2.\end{aligned}$$

Noting that $\delta_\varepsilon \downarrow 0$ as $\varepsilon \downarrow 0$, we have $I_2 = o(\varepsilon)$. We next focus on I_1. Observe

$$\begin{aligned} I_1 &\le \sum_{i=1}^K \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \le \frac{1-\delta_\varepsilon}{w_i\varepsilon}, \sum_{j\ne i}^K w_j U_j^{-1} > \delta_\varepsilon/\varepsilon\right) \\ &\le \sum_{i=1}^K \sum_{j\ne i}^K \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \le \frac{1-\delta_\varepsilon}{w_i\varepsilon}, U_j^{-1} > \frac{\delta_\varepsilon}{w_j K\varepsilon}\right). \end{aligned}$$

It remains to show for $1 \le i \ne j \le K$,

$$I_{i,j} := \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \le \frac{1-\delta_\varepsilon}{w_i\varepsilon}, U_j^{-1} > \frac{\delta_\varepsilon}{w_j K\varepsilon}\right) = o(\varepsilon).$$

Condition (G) implies that there exist $Z_{i,j}$ and $\delta_{i,j}$ such that

$$\Phi^{-1}(U_j) = \sigma_{ij}\Phi^{-1}(U_i) + \delta_{ij} Z_{ij}, \tag{A.21}$$

where Z_{ij} is a standard normal random variable that is independent of U_i and $\sigma_{ij}^2 + \delta_{ij}^2 = 1$. If $\sigma_{ij} = -1$, we have $U_i = 1 - U_j$. This implies that $I_{i,j} = 0$ for $\varepsilon > 0$ sufficiently small. Next, assume $|\sigma_{ij}| < 1$, and write $\gamma_{ij} = \Phi^{-1}(w_i K\varepsilon)$ if $-1 < \sigma_{ij} \le 0$ and $\gamma_{ij} = \Phi^{-1}\left(\frac{w_i\varepsilon}{1-\delta_\varepsilon}\right)$ if $0 < \sigma_{ij} < 1$. We have

$$\begin{aligned} I_{i,j} &= \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \le \frac{1-\delta_\varepsilon}{w_i\varepsilon}, \sigma_{ij}\Phi^{-1}(U_i) + \delta_{ij} Z_{ij} < \Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right)\right) \\ &\le \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \le \frac{1-\delta_\varepsilon}{w_i\varepsilon}, \delta_{ij} Z_{ij} < \Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right) - \sigma_{ij}\gamma_{ij}\right) \\ &= \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \le \frac{1-\delta_\varepsilon}{w_i\varepsilon}\right) \mathbb{P}\left(\delta_{ij} Z_{ij} < \Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right) - \sigma_{ij}\gamma_{ij}\right). \end{aligned}$$

Note that $\Phi^{-1}(\varepsilon) \sim -\sqrt{-2\ln\varepsilon}$, as $\varepsilon \downarrow 0$, which is a slowly varying function. Taking $\delta_\varepsilon = -1/\log\varepsilon$, we have

$$\Phi^{-1}\left(\frac{w_i\varepsilon}{1-\delta_\varepsilon}\right) \sim \Phi^{-1}(w_i K\varepsilon) \sim \Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right) \quad \text{as } \varepsilon \downarrow 0.$$

This implies

$$\Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right) - \sigma_{ij}\gamma_{ij} \to -\infty, \text{ as } \varepsilon \downarrow 0.$$

Hence $I_{i,j} = o(\varepsilon)$. Consequently, $I_1 = o(\varepsilon)$ and further $\mathbb{P}(B_\varepsilon) = o(\varepsilon)$. Next, we show $\mathbb{P}(A_\varepsilon) \sim \varepsilon$. By the Bonferroni inequality, we have,

$$\sum_{i=1}^K \mathbb{P}(A_{i,\varepsilon}) - \sum_{1\le i<j\le K} \mathbb{P}(A_{i,\varepsilon} \cap A_{j,\varepsilon}) \le \mathbb{P}(A_\varepsilon) \le \sum_{i=1}^K \mathbb{P}(A_{i,\varepsilon}).$$

Direct calculation gives

$$\sum_{i=1}^{K} \mathbb{P}(A_{i,\varepsilon}) = \sum_{k=1}^{K} \frac{w_i \varepsilon}{1+\delta_\varepsilon} \sim \varepsilon.$$

For any $1 \le i < j \le K$, since the Gaussian copula is tail independent (e.g., Example 7.38 of McNeil et al. (2015)), we have, writing $w = \max\{w_i, w_j\}$,

$$\begin{aligned}\mathbb{P}(A_{i,\varepsilon} \cap A_{j,\varepsilon}) &= \mathbb{P}\left(U_i^{-1} > \frac{1+\delta_\varepsilon}{w_i\varepsilon}, U_j^{-1} > \frac{1+\delta_\varepsilon}{w_j\varepsilon}\right) \\ &\leqslant \mathbb{P}\left(U_i < \frac{w\varepsilon}{1+\delta_\varepsilon}, U_j < \frac{w\varepsilon}{1+\delta_\varepsilon}\right) = o(1)\mathbb{P}\left(U_1 < \frac{w\varepsilon}{1+\delta_\varepsilon}\right) = o(1)\varepsilon.\end{aligned}$$

Hence $\mathbb{P}(A_{i,\varepsilon} \cap A_{j,\varepsilon}) = o(\varepsilon)$. This implies $\mathbb{P}(A_\varepsilon) \sim \varepsilon$, and we establish (A.20).

(iii) By Lemma A.1 of Vovk and Wang (2020a), we have

$$a_{\mathcal{H}}(\varepsilon) = \varepsilon \left(\sup \left\{ q_0^+ \left(\frac{1}{K} \sum_{i=1}^{K} P_i^{-1} \right) \mid P_1, \dots, P_K \in \mathcal{U} \right\} \right)^{-1}, \quad \varepsilon \in (0,1),$$

where $q_0^+(X) = \sup\{x \in \mathbb{R} \mid \mathbb{P}(X \le x) = 0\}$. Note that for any $\delta > 0$, there exists $0 < \varepsilon_\delta < 1$ such that for all $x \in (0, \varepsilon_\delta)$

$$-\frac{(1+\delta)}{x} < \tan\left(-\frac{\pi}{2} + x\right) < -\frac{(1-\delta)}{x}.$$

For $\delta > 0$, letting $0 < \varepsilon < \varepsilon_\delta/\pi$ and using Theorem 4.6 in Bernard et al. (2014), we have

$$\begin{aligned}&\inf\left\{ q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K} \mathcal{C}^{-1}(P_i)\right) \mid P_1,\dots,P_K \in \mathcal{U}\right\} \\ &= \inf\left\{ q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K} \tan\left(\pi\left(P_i - \frac{1}{2}\right)\right)\right) \mid P_1,\dots,P_K \in \mathcal{U}\right\} \\ &= \inf\left\{ q_1\left(\frac{1}{K}\sum_{i=1}^{K} \tan\left(\pi\left(\varepsilon P_i - \frac{1}{2}\right)\right)\right) \mid P_1,\dots,P_K \in \mathcal{U}\right\} \\ &\le \inf\left\{ q_1\left(\frac{1}{K}\sum_{i=1}^{K} -\frac{1-\delta}{\varepsilon\pi P_i}\right) \mid P_1,\dots,P_K \in \mathcal{U}\right\} \\ &= -\frac{1-\delta}{\varepsilon\pi} \sup\left\{ q_0^+\left(\frac{1}{K}\sum_{i=1}^{K} P_i^{-1}\right) \mid P_1,\dots,P_K \in \mathcal{U}\right\} = -\frac{1-\delta}{a_{\mathcal{H}}(\varepsilon)\pi}.\end{aligned}$$

Similarly, we obtain, for $0 < \varepsilon < \varepsilon_\delta/\pi$,

$$\inf\left\{ q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K} \mathcal{C}^{-1}(P_i)\right)\right\} \ge -\frac{1+\delta}{a_{\mathcal{H}}(\varepsilon)\pi}.$$

Consequently,

$$\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(P_i)\right)\right\}\sim-\frac{1}{a_{\mathcal{H}}(\varepsilon)\pi}\quad\text{as } \varepsilon\downarrow 0.$$

Plugging the above result in the formula for $a_{\mathcal{C}}$ in (6), and using $\mathcal{C}(y)\sim -1/(\pi y)$ as $y\to-\infty$, we have, as $\varepsilon\downarrow 0$,

$$\begin{aligned}a_{\mathcal{C}}(\varepsilon)&=\mathcal{C}\left(\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(P_i)\right)\right\}\right)\\&\sim-\frac{1}{\pi}\left(\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(P_i)\right)\right\}\right)^{-1}\sim a_{\mathcal{H}}(\varepsilon).\end{aligned}$$

This completes the proof.

(iv) By (i), it suffices to show that for $r\neq -1$

$$\frac{M_{-1,K}(p_1,\dots,p_K)}{M_{r,K}(p_1,\dots,p_K)}\nrightarrow 1,\ \text{ as } \max_{i\in\{1,\dots,K\}}p_i\downarrow 0.$$

Take $p_1=p^2$ and $p_i=x_ip$ with $x_i>0$ and $p>0$ for $i=2,\dots,K$. By homogeneity of M_r, for $r\le -1$,

$$\frac{M_{-1,K}(p_1,\dots,p_K)}{M_{r,K}(p_1,\dots,p_K)}=\frac{M_{-1,K}(p,x_2,\dots,x_K)}{M_{r,K}(p,x_2,\dots,x_K)}.$$

Hence

$$\lim_{p\downarrow 0}\frac{M_{-1,K}(p_1,\dots,p_K)}{M_{r,K}(p_1,\dots,p_K)}=K^{1/r+1}\neq 1,\quad r<-1.$$

This proves the claim of (iv) for $r<-1$. The case for $r>-1$ can be argued similarly. □

A.7 Proof of Theorem 3

Take arbitrary $p_1,\dots,p_K\in(0,1]$, and let $j\in\{1,\dots,K\}$ be such that $\min_{k\in\{1,\dots,K\}}\frac{1}{k}p_{(k)}=\frac{1}{j}p_{(j)}$. We have

$$\frac{S_K(p_1,\dots,p_K)}{M_{-1,K}(p_1,\dots,p_K)}=\sum_{i=1}^{K}\frac{1}{j}p_{(j)}\frac{1}{p_i}\le\sum_{i=1}^{K}\frac{1}{i}p_i\frac{1}{p_i}=\sum_{i=1}^{K}\frac{1}{i}=\ell_K.$$

Moreover,

$$\frac{S_K(p_1,\dots,p_K)}{M_{-1,K}(p_1,\dots,p_K)}=\frac{1}{j}p_{(j)}\left(\sum_{i=1}^{K}\frac{1}{p_{(i)}}\right)\ge\frac{1}{j}p_{(j)}\left(\sum_{i=1}^{j}\frac{1}{p_{(j)}}+\sum_{i=j+1}^{K}\frac{1}{p_{(i)}}\right)\ge 1.$$

Therefore, $M_{-1,K}\leqslant S_K\leqslant \ell_K M_{-1,K}$. The two special cases of equalities are straightforward to check. □

A.8 Proof of Proposition 5

(i) Recall that $a_F(x) = a_F x$ for $x \in (0,1)$. By (i) of Proposition 3, we have $b_F(\delta) \sim \delta$ as $\delta \downarrow 0$. Hence $\lim_{\delta\downarrow 0} b_F(\delta)/a_F(\delta) = 1/a_F$. By Proposition 6 of Vovk and Wang (2020a), we have $a_F \sim 1/\log K$, as $K \to \infty$. Consequently,

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

Moreover, for the harmonic averaging method, $c_F(\varepsilon) = \varepsilon$. This implies $c_F(\varepsilon)a_F(\varepsilon) = 1/a_F$. We establish the claim by the fact $a_F \sim \frac{1}{\log K}$, as $K \to \infty$.

(ii) By Theorem 2, we have $a_{\mathcal{C}}(\delta) \sim a_{\mathcal{H}}(\delta)$ and $b_{\mathcal{C}}(\delta) \sim b_{\mathcal{H}}(\delta)$ as $\delta \downarrow 0$, which together with (i) leads to

$$\lim_{\delta\downarrow 0} \frac{b_{\mathcal{C}}(\delta)}{a_{\mathcal{C}}(\delta)} \sim \log K, \text{ as } K \to \infty.$$

The rest of the statement follows by noting that $c_{\mathcal{C}}(\delta) = b_{\mathcal{C}}(\delta)$.

(iii) For the Simes method, recall that $a_F(x) = x/\ell_K$ and $b_F(x) = c_F(x) = x$. The claim follows directly from the fact that $\ell_K = \sum_{k=1}^{K} \frac{1}{k} \sim \log K$, as $K \to \infty$. □

B Additional tables

In Table B.4 we report numerical results of prices for validity for $\varepsilon = 0.05$.

	$K = 50$		$K = 100$		$K = 200$		$K = 400$	
	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	1.025	50.000	1.026	100.000	1.026	200.000	1.026	400.000
Negative-quartic	1.367	25.071	1.367	42.164	1.368	70.911	1.368	119.257
Simes	4.499	4.499	5.187	5.187	5.878	5.878	6.570	6.570
Cauchy	6.623	6.623	7.463	7.463	8.274	8.274	9.055	9.055
Harmonic	6.793	6.625	7.650	7.459	8.485	8.273	9.306	9.072
Geometric	15.679	2.718	16.874	2.718	17.755	2.718	18.395	2.718

Table B.4: $b_F(\varepsilon)/a_F(\varepsilon)$ and $c_F(\varepsilon)/a_F(\varepsilon)$ for $\varepsilon = 0.05$ and $K \in \{50, 100, 200, 400\}$