Combining exchangeable p-values

Matteo Gasparin[1], Ruodu Wang[2], and Aaditya Ramdas[3,4]

[1]Department of Statistical Sciences, University of Padova
[2]Department of Statistics and Actuarial Science, University of Waterloo
[3]Department of Statistics and Data Science, Carnegie Mellon University
[4]Machine Learning Department, Carnegie Mellon University

February 7, 2025

Abstract

The problem of combining p-values is an old and fundamental one, and the classic assumption of independence is often violated or unverifiable in many applications. There are many well-known rules that can combine a set of arbitrarily dependent p-values (for the same hypothesis) into a single p-value. We show that essentially all these existing rules can be strictly improved when the p-values are exchangeable, or when external randomization is allowed (or both). For example, we derive randomized and/or exchangeable improvements of well known rules like "twice the median" and "twice the average", as well as geometric and harmonic means. Exchangeable p-values are often produced one at a time (for example, under repeated tests involving data splitting), and our rules can combine them sequentially as they are produced, stopping when the combined p-values stabilize. Our work also improves rules for combining arbitrarily dependent p-values, since the latter becomes exchangeable if they are presented to the analyst in a random order. The main technical advance is to show that all existing combination rules can be obtained by calibrating the p-values to e-values (using an α-dependent calibrator), averaging those e-values, converting to a level-α test using Markov's inequality, and finally obtaining p-values by combining this family of tests; the improvements are delivered via recent randomized and exchangeable variants of Markov's inequality.

Contents

1 Introduction

The combination of p-values represents a fundamental task frequently encountered in statistical inference and its applications in the natural sciences. Within the realm of multiple testing, for instance, the focus lies in testing whether all individual null hypotheses are simultaneously true. This particular challenge, often referred to as global null testing, can be addressed by merging multiple p-values into a single p-value. Potential solutions, assuming the p-values are statistically independent, are provided in Fisher (1934); Pearson (1934); Simes (1986), with the latter also working under a certain notion of positive dependence (Sarkar, 1998; Benjamini and Yekutieli, 2001). See Owen (2009) for a review of the methods. Recently, harmonic mean p-values (Wilson, 2019) and methods under negative dependence (Chi et al., 2024) have been developed in the context of multiple testing.

The assumption of independence (or positive/negative dependence) is often violated in many real-world applications (see e.g., Section 4.2 of Efron, 2010), and in certain scenarios, it may be preferable not to impose unverifiable conditions on the joint distribution of the p-values, beyond the minimum necessary assumption that each individual p-value is indeed (marginally) valid. Several methods are available for combining p-values with arbitrary dependencies; notably, the Bonferroni method is widely used, which involves multiplying the minimum of the p-values by the number of tests conducted. Other methods have been proposed in the literature, some based on order statistics (Rüger, 1978; Morgenstern, 1980; Hommel, 1983), while others rely on their arithmetic mean and other variants (Rüschendorf, 1982; Vovk and Wang, 2020; Vovk et al., 2022). Two prominent examples are that both 2 times the median of the p-values, and 2 times the average of the p-values, are valid combination rules, and the multiplicative factor of 2 cannot be reduced. Inevitably, these methods satisfying validity under arbitrary dependence come with a price to pay in terms of statistical power. Our work will improve all of these rules under the weaker assumption of validity under exchangeability of p-values.

To elaborate, the main objective of our work is to obtain simple new valid merging methods under the assumption of exchangeability of the input p-values, which are more powerful than methods that assume arbitrary dependence. Implied by the relative strength of the dependence assumptions, the new methods will be incomparable to methods assuming negative or positive dependence, and less powerful than methods assuming independence. However, specialized methods for handling exchangeable dependence are quite practically relevant. Such dependence is encountered, for example, in statistical testing via sample splitting, as we now elaborate.

There are at least two different reasons for which sample splitting is used: the first is to relax the assumptions needed to obtain theoretical guarantees, while the second is to reduce computational costs. Some examples of such procedures are Cox (1975); Wasserman and Roeder (2009); Banerjee et al. (2019); Wasserman et al. (2020); Shafer and Vovk (2008); Kim and Ramdas (2024). The drawback of these methods based on sample-splitting is that the obtained p-values are affected by the randomness of the split. Meinshausen et al. (2009) called this phenomenon as a *p-value lottery*. So one may instead repeat the same sample splitting procedure several times to obtain multiple p-values, which are exchangeable by design of the procedure.

One can of course combine such p-values by using the earlier mentioned rules (like twice the average) for arbitrarily dependent p-values. But we hope to do better by exploiting their exchangeability. In a paper that seemingly dampens that hope, Choi and Kim (2023) showed that in the aforementioned rule of "twice the average", the constant factor of 2 cannot be improved even under exchangeability. However, their result does not imply that "twice the average" cannot be improved; it simply states that any such improvement cannot proceed by attempting to lower the constant of 2. Indeed, our paper will improve on this well known rule (and many others), but it proceeds differently, not by lowering the constant. Instead, it calculates the twice the average of the first k p-values, and takes a minimum over all k (see Table 1).

We also show that no symmetric rule for merging arbitrarily dependent p-values (like the ones mentioned earlier) can be improved under exchangeability. To achieve any improvement, we must consider asymmetric rules that process the p-values in a particular order. This might initially appear paradoxical given the exchangeability of the p-values, but it is easily sorted out. In many practical settings, these exchangeable p-values can be generated one by one by repeating the same randomized procedure many times, generating a stream of p-values. In this case, our combination rules would simply process these p-values in the order that they are generated. This seems quite appropriate, and the main advantage of doing this is increased power over processing them symmetrically as a batch. In fact, as we discuss, we do not need to fix the number of p-values ahead of time, they can just be processed online, yielding a p-value whenever this procedure is stopped. This makes our merging rules particularly simple and practical.

The problem of combining such p-values from repeated sample splitting has been studied by other authors, such as DiCiccio et al. (2020), but our combination rules are more powerful, and also more general and systematic because they apply more broadly. The same problem has also been studied in Guo and Shah (2024), where the authors propose a combination method based on subsampling to merge test statistics based on different random splits. Unlike our nonasymptotic guarantees that work directly with the p-values and are cheap to compute, their work provides only asymptotic guarantees under certain additional assumptions (like an asymptotic pivotal null distribution for their test statistics) and they require access to the full dataset on which to perform expensive subsampling-based recomputations. However, when their additional assumptions hold, their procedure can be expected to be more powerful than ours because it estimates and exploits the joint distribution of the p-values. With a similar aim, Ritzwoller and Romano (2023) investigates a method to enhance the reproducibility of statistical results obtained through sample-splitting. Specifically, their algorithm sequentially aggregates statistics (not necessarily p-values or test statistics) across multiple sample splits until the variability induced by the different splits falls below a defined threshold. In particular, if the method is repeated twice on the same data, the probability of differing results remains close to a prespecified error rate.

It is perhaps interesting that all the aforementioned methods for merging under arbitrary dependence or under exchangeability are actually inadmissible when randomization is permitted. Randomization in the context of hypothesis testing is not new and is used, for example, in discrete tests; some examples are Fisher's exact test (Fisher, 1934) or the randomized test for a binomial proportion proposed in Stevens (1950). In our paper, we will see how the introduction of a simple external randomization (an independent uniform random variable and/or uniform permutation) can improve the existing merging rules for arbitrary dependence, as well as our new rules for exchangeable merging.

In terms of technical aspects, one of our main contributions is to point out explicitly how existing merging rules for arbitrary dependence are actually recovered in a unified manner: by transforming the p-values into e-values (Vovk and Wang, 2021; Wasserman et al., 2020; Grünwald et al., 2024) using different "calibrators", averaging the resulting e-values and finally applying Markov's inequality. This connection is particularly important, because then the improvements under exchangeability, or by randomization, are then achieved by invoking the recent "exchangeable Markov inequality" and "uniformly randomized Markov inequality" (Ramdas and Manole, 2024).

Paper outline and peek at results. The rest of this paper is organized as follows. In Section 2, we introduce the notation and tools necessary for the paper. In Section 3, the main results are presented in a general way, focusing on two distinct aspects: the first part addresses the case of exchangeable p-values, while the second part introduces novel findings under the assumption of arbitrarily dependent p-values when randomization is allowed. Subsequent sections investigate the implications of these results across various p-merging functions commonly found in the literature. Specifically, Section 4 and Section 5 delve into the combination proposed by Rüger (1978) and Hommel (1983), respectively. Section 6 examines the case of arithmetic mean. The following two sections address two additional scenarios within the family of generalized means: namely, the harmonic mean (Section 7) and the geometric mean (Section 8). Section 9 presents some simulation results, before we conclude in Section 10. All proofs are provided in Appendix A.

Before proceeding with the paper, Table 1 presents some notable combination rules introduced in the literature and their corresponding exchangeable and randomized versions introduced in the following sections. These results are first derived in a general form and then discussed case-by-case. Some of the rules in the table are not admissible, as will be explained in the following.

Combination rule	Arbitrary dependence (known)	Exchangeability (new)	Arbitrary dependence, randomized (new)
Rüger combination	$\frac{K}{k}p_{(k)}$	$\frac{K}{k}\bigwedge_{m=1}^{K} p^{m}_{(\lambda_m)}$	$\frac{K}{k}p_{(\lceil Uk\rceil)}$
Arithmetic mean	$2A(\mathbf{p})$	$2\bigwedge_{m=1}^{K} A(\mathbf{p}_m)$	$\frac{2}{2-U}A(\mathbf{p})$
Geometric mean	$eG(\mathbf{p})$	$e\bigwedge_{m=1}^{K} G(\mathbf{p}_m)$	$e^{U}G(\mathbf{p})$
Harmonic mean	$(T_K+1)H(\mathbf{p})$	$(T_K+1)\bigwedge_{m=1}^{K} H(\mathbf{p}_m)$	$(T_K U+1)H(\mathbf{p})$

Table 1: Some combination rules for arbitrarily dependent p-values documented in literature, along with their exchangeable and randomized improvements. If randomization is permitted, one can also improve the existing rules for combining arbitrarily dependent p-values by using the exchangeable combination rule applied to a random permutation of the p-values (this is not presented as a separate column). Here, $\mathbf{p} = (p_1, \dots, p_K)$ denotes the vector of p-values, and $\mathbf{p}_m$ represents the vector containing the first m values of $\mathbf{p}$. In the table, $p_{(k)}$ is the k-th smallest value of $\mathbf{p}$, while $p^{m}_{(\lambda_m)}$ is the $\lambda_m = \lceil m\frac{k}{K}\rceil$ ordered value of $\mathbf{p}_m$. The random variable U is uniformly distributed in the interval $[0,1]$. Additionally, A, G and H respectively denote the arithmetic mean, the geometric mean, and the harmonic mean. The value T_K is given by $T_K = \log K + \log\log K + 1$ for $K \geq 2$.

2 Problem setup and notation

Without loss of generality, let $(\Omega, \mathcal{F}, \mathbb{P})$ be an atomless probability space[1], and this is implicitly assumed in almost all papers in statistics; see Vovk and Wang (2021, Appendix D) for related results and discussions. Let $\mathcal{U}$ be the set of all uniform random variables on $[0,1]$ under $\mathbb{P}$. In the following, $K \geq 2$ is an integer. We use the shorthand notation $x \vee y = \max(x, y)$, $x \wedge y = \min(x, y)$, $\bigvee_{k=1}^{K} x_k = \max\{x_1, \dots, x_K\}$, and $\bigwedge_{k=1}^{K} x_k = \min\{x_1, \dots, x_K\}$.

A *p-variable* for testing $\mathbb{P}$ is a random variable $P : \Omega \to [0, \infty)$ satisfying

$$\mathbb{P}(P \leq \alpha) \leq \alpha,$$

for all $\alpha \in (0,1)$. Typically, of course, P will only take values in $[0,1]$, but nothing is lost by allowing the larger range above. For all results on validity of the methods in this paper, it suffices to consider p-variables in $\mathcal{U}$, i.e., $\mathbb{P}(P \leq \alpha) = \alpha$ for each $\alpha \in (0,1)$.

An *e-variable* for testing $\mathbb{P}$ is a non-negative extended random variable $E : \Omega \to [0, \infty]$ with $\mathbb{E}_{\mathbb{P}}[E] \leq 1$. A *calibrator* is a decreasing function $f : [0, \infty) \to [0, \infty]$ satisfying $f = 0$ on $(1, \infty)$ and $\int_0^1 f(p)\mathrm{d}p \leq 1$. Essentially, a calibrator transforms any p-variable to an e-variable. It is *admissible* if it is upper semicontinuous, $f(0) = \infty$, and $\int_0^1 f(p)\mathrm{d}p = 1$. Equivalently, a calibrator is admissible if it is not strictly dominated, in a natural sense, by any other calibrator (Proposition 2.1 and Proposition 2.2 in Vovk and Wang, 2021). We fix $\mathbb{P}$ throughout, and omit "for testing $\mathbb{P}$" when discussing p-variables and e-variables; we do not distinguish them from the commonly used terms "p-values" and "e-values", and this should create no confusion.

[1] A probability space is atomless if there exists a random variable on this space that is uniformly distributed on $[0,1]$.

Our starting point is a collection of K p-variables $\mathbf{P} = (P_1, \dots, P_K)$ and we denote their observed (realized) values by $\mathbf{p} = (p_1, \dots, p_K)$. Borrowing terminology from Vovk et al. (2022) and Vovk and Wang (2020), we have that a *p-merging function* is an increasing Borel function $F : [0, \infty)^{K+1} \to [0, \infty)$ such that $\mathbb{P}(F(\mathbf{P}) \leq \alpha) \leq \alpha$ whenever $P_1, \dots, P_K$ are p-variables. In other words, the function F, starting from K p-values, returns a p-value. A p-merging function is *symmetric* if $F(\mathbf{p})$ is invariant under any permutation of $\mathbf{p}$, and it is *homogeneous* if $F(\gamma \mathbf{p}) = \gamma F(\mathbf{p})$ for all $\mathbf{p}$ with $F(\mathbf{p}) \leq 1$ and $\gamma \in (0, 1]$. The class of homogeneous p-merging functions encompasses the *O-family* based on quantiles introduced in Rüger (1978), the Hommel's combination and the *M-family* introduced in Vovk and Wang (2020). We now introduce the notion of *domination* in the context of p-merging functions.

Definition 2.1. A function F dominates (interpreted as better being smaller) a function G if

$$F(\mathbf{p}) \leq G(\mathbf{p}), \quad \text{for all } \mathbf{p},$$

and the domination is strict if $F(\mathbf{p}) < G(\mathbf{p})$, for at least one $\mathbf{p}$. A p-merging function F is *admissible* if it is not strictly dominated by any other p-merging function.

Although we defined p-values and e-values for a single probability measure $\mathbb{P}$, all results hold for composite hypotheses. More precisely, if $\mathbf{P}$ is a vector of p-variables for a composite hypothesis and F is a p-merging function, then $F(\mathbf{P})$ is a p-variable for the same composite hypothesis. See Vovk and Wang (2021) for precise definitions and related discussions.

For any function $F : [0, \infty)^K \to [0, \infty)$ and $\alpha \in (0, 1)$, let its rejection region at level α be given by

$$R_\alpha(F) := \left\{ \mathbf{p} \in [0, \infty)^K : F(\mathbf{p}) \leq \alpha \right\}.$$

For any homogeneous F, $R_\alpha(F)$ for $\alpha \in (0, 1)$ takes the form $R_\alpha(F) = \alpha A$ for some $A \subseteq [0, \infty)^K$, where αA means the set $\{\alpha \, \mathbf{x} : \mathbf{x} \in A\}$.

Conversely, any increasing collection of Borel lower sets $\{R_\alpha \subseteq [0, \infty)^K : \alpha \in (0, 1)\}$ determines an increasing Borel function $F : [0, \infty)^K \to [0, 1]$ by the equation

$$F(\mathbf{p}) = \inf\{\alpha \in (0, 1) : \mathbf{p} \in R_\alpha\},$$

with the convention $\inf \varnothing = 1$ (throughout). It is immediate to see that F is a p-merging function if and only if $\mathbb{P}(\mathbf{P} \in R_\alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ and $\mathbf{P} \in \mathcal{U}^K$, where $\mathcal{U}^K$ is the set of all K-dimensional random vectors with components in $\mathcal{U}$.

Below, Δ_K is the standard K-simplex. Every admissible homogeneous p-merging function possesses a dual formulation expressed in terms of calibrators, as summarized below.

Theorem 2.2 (Vovk et al. (2022); Theorem 5.1)**.** *For any admissible homogeneous p-merging function F, there exist $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and admissible calibrators $f_1, \dots, f_K$ such that*

$$R_\alpha(F) = \left\{ \mathbf{p} \in [0, \infty)^K : \sum_{k=1}^{K} \lambda_k f_k\left(\frac{p_k}{\alpha}\right) \geq 1 \right\} \qquad \text{for each } \alpha \in (0, 1). \tag{1}$$

Conversely, for any $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and calibrators $f_1, \dots, f_K$, (1) *determines a homogeneous p-merging function.*

We will exploit this dual form to implement our “randomized” and “exchangeable” techniques, generating p-merging functions that consistently give smaller p-values (usually strictly) than those produced by their *original* counterparts. From (1), it is worth noting that $\sum_{k=1}^{K} \lambda_k f_k(P_k)$ is an e-value. We now present a useful lemma.

Lemma 2.3. *Let $f_1, \dots, f_K$ be K calibrators and $\mathbf{P} \in \mathcal{U}^K$. Then, for any $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and $\alpha \in (0, 1]$,*

$$\frac{1}{\alpha} \sum_{k=1}^{K} \lambda_k f_k \left(\frac{P_k}{\alpha} \right) \tag{2}$$

is an e-variable. If $f_1, \dots, f_K$ are admissible calibrators, then $\mathbb{E}\left[\frac{1}{\alpha} \sum_{k=1}^{K} \lambda_k f_k \left(\frac{P_k}{\alpha} \right) \right] = 1$.

Lemma 2.3 used the fact that a calibrator takes value 0 on $(1, \infty)$, which is not restrictive because the relevant range of p-values is $[0, 1]$. This condition is assumed when defining calibrators in Vovk et al. (2022).

In particular, choosing $\lambda_1 = 1$ and $\lambda_k = 0$, for $k \geq 2$, we have that $(1/\alpha) f_1(P_1/\alpha)$ is an e-value, for all $\alpha \in (0, 1]$.

Before introducing our results, we present some inequalities introduced in Ramdas and Manole (2024) that will be fundamental throughout the subsequent discussion. The following inequalities can be viewed as an extension of Markov's inequality.

Theorem 2.4 (Exchangeable Markov Inequality)**.** *Let $X_1, X_2, \dots$ form an exchangeable sequence of non-negative and integrable random variables. Then, for any $a > 0$,*

$$\mathbb{P}\left(\exists k \geq 1 : \frac{1}{k} \sum_{i=1}^{k} X_i \geq \frac{1}{a} \right) \leq a \mathbb{E}[X_1].$$

In addition, let $X_1, \dots, X_K$ be exchangeable, non-negative and integrable random variables. Then, for any $a > 0$,

$$\mathbb{P}\left(\exists k \leq K : \frac{1}{k} \sum_{i=1}^{k} X_i \geq \frac{1}{a} \right) \leq a \mathbb{E}[X_1].$$

The second inequality is based on an external randomization of the threshold of Markov's inequality.

Theorem 2.5 (Uniformly-randomized Markov Inequality)**.** *Let X be a non-negative random variable independent of $U \in \mathcal{U}$. Then, for any $a > 0$,*

$$\mathbb{P}(X \geq U/a) \leq a \mathbb{E}[X].$$

The third inequality combines the previous two theorems in the following way:

Theorem 2.6 (Exchangeable and uniformly-randomized Markov Inequality)**.** *Let $X_1, \dots, X_K$ be a set of exchangeable and non-negative random variables independent of $U \in \mathcal{U}$. Then, for any $a > 0$,*

$$\mathbb{P}\left(X_1 \geq U/a \text{ or } \exists k \leq K : \frac{1}{k} \sum_{i=1}^{k} X_i \geq \frac{1}{a} \right) \leq a \, \mathbb{E}[X_1].$$

These results will be used in the next section as technical tools to derive new combination rules in different situations.

3 New results on merging p-values

This section introduces our main results stated in abstract and general terms, which we instantiate in special cases (like arithmetic or geometric mean) in the sections that follow.

3.1 Exchangeable p-values

Assuming iid p-values is often overly stringent in numerous practical applications. A more pragmatic and less restrictive assumption is exchangeability of p-values, indicating that the distribution of the p-values is unchanged under a permutation of the indices. Formally,

$$(P_1, \dots, P_K) \overset{d}{=} (P_{\sigma(1)}, \dots, P_{\sigma(K)}),$$

where $\overset{d}{=}$ represents equality in distribution while $\sigma : \{1, \dots, K\} \to \{1, \dots, K\}$ is any permutation of the indices. As discussed in Section 1, this situation is encountered in statistical testing using repeated sample splitting (repeated K times on the same data in an identical fashion). In this section, we assume that the sequence of p-variables $\mathbf{P} = (P_1, \dots, P_K)$ is exchangeable and takes values in $[0, 1]^K$.

Remark 3.1. The reader may note that exchangeability can be induced by processing the (potentially non-exchangeable) sequence of p-values $(P_1, \dots, P_K)$ in a uniformly random order. As a consequence, this implies that if randomization is allowed, it is always possible to satisfy the exchangeability assumption even if the starting sequence has an arbitrary dependence. Stated alternatively, exchangeable combination rules can be safely applied to arbitrarily dependent p-values if the p-values are presented to the analyst in a random order.

We present an extension of the converse direction of Theorem 2.2, which is valid under exchangeability of the vector of p-values. In particular, to preserve exchangeability and use the result stated in Theorem 2.4, it is necessary that the same calibrator f is used for all p-values.

Theorem 3.2. *Let f be a calibrator, and $\mathbf{P} = (P_1, \dots, P_K) \in \mathcal{U}^K$ be exchangeable. For each $\alpha \in (0, 1)$, we have*

$$\mathbb{P}\left(\exists k \le K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\alpha}\right) \ge 1\right) \le \alpha.$$

We first give a formal definition of ex-p-merging function, which is a function that yields a p-value if its argument is a vector of exchangeable p-values.

Definition 3.3. An *ex-p-merging function* is an increasing Borel function $F : [0, 1]^K \to [0, 1]$ such that $\mathbb{P}(F(\mathbf{P}) \le \alpha) \le \alpha$ for all $\alpha \in (0, 1)$ and $\mathbf{P} \in \mathcal{U}^K$ that is exchangeable. It is *homogeneous* if $F(\gamma\mathbf{p}) = \gamma F(\mathbf{p})$ for all $\gamma \in (0, 1]$ and $\mathbf{p} \in [0, 1]^K$. An ex-p-merging function F is *admissible* if for any ex-p-merging function G, $G \le F$ implies $G = F$.

We now use the result given in Theorem 3.2 to derive better p-merging functions by exploiting the duality between rejection regions and p-merging functions. In particular, to derive an ex-p-merging the following steps are involved: Initially, the rejection regions at level α based on a given calibrator are determined. As explained in Section 2, the ex-p-merging is established by choosing the smallest α for which the p-values $\mathbf{p}$ falls within the rejection region R_α. In the first step, the result stated in Theorem 3.2 helps to derive functions that dominate their counterpart valid under arbitrary dependence. To elaborate, starting from a calibrator f and $\alpha \in (0, 1)$, we define the exchangeable rejection region

$$R_\alpha = \left\{\mathbf{p} \in [0, 1]^K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \ge 1 \text{ for some } k \le K\right\}.$$

Using R_α, we can define the function $F : [0,1]^K \to [0,1]$ by

$$\begin{aligned} F(\mathbf{p}) &= \inf\{\alpha \in (0,1) : \mathbf{p} \in R_\alpha\} \\ &= \inf\left\{\alpha \in (0,1) : \exists k \le K \text{ s.t. } \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \ge 1\right\} \\ &= \inf\left\{\alpha \in (0,1) : \bigvee_{k=1}^{K} \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \ge 1\right\}, \end{aligned} \tag{3}$$

where, and throughout, $\bigvee_{k=1}^{K} \frac{1}{k}\sum_{i=1}^{k} f(\frac{p_i}{\alpha})$ should be understood as $\bigvee_{k=1}^{K}(\frac{1}{k}\sum_{i=1}^{k} f(\frac{p_i}{\alpha}))$. Note that (3) is always smaller or equal than the p-merging function given by (1)

$$F'(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K} f\left(\frac{p_k}{\alpha}\right) \ge 1\right\},$$

which is valid for p-values with an arbitrary dependence. This is particularly important since all admissible homogeneous and symmetric p-merging functions have the form F' for some admissible calibrator (a symmetric version of Theorem 2.2; see Vovk et al. (2022)). This implies that the function defined (3) dominates the function F'. In the following theorem we prove that (3) is a homogeneous ex-p-merging function.

Theorem 3.4. *If f is a calibrator and $\mathbf{P} \in \mathcal{U}^K$ is an exchangeable sequence, then F in (3) is a homogeneous ex-p-merging function.*

It is clear from (3) that the function depends on the order of the values in $\mathbf{p}$, and hence $F(\mathbf{p})$ is not a symmetric function. This is not a coincidence: in the next result, we show that any symmetric ex-p-merging function is actually valid under arbitrary dependence, and hence they cannot improve over the admissible p-merging functions studied by Vovk et al. (2022). In particular, this implies that under exchangeability the multiplicative factor 2 for the arithmetic average cannot be improved, as earlier noted by Choi and Kim (2023), but it also extends their result to every other symmetric merging function.

Proposition 3.5. *A symmetric ex-p-merging function is necessarily a p-merging function. Hence, for an ex-p-merging function to strictly dominate an admissible p-merging function, it cannot be symmetric.*

Clearly, Proposition 3.5 implies that under symmetry, a function is a p-merging function if and only if it is an *ex-p-merging function*. Hence, to take advantage of the exchangeability of the p-values (over arbitrary dependence), one necessarily deviates from symmetric ways of merging p-values, as done in Theorem 3.4. More importantly, the proof of Proposition 3.5 illustrates the idea (mentioned in Remark 3.1) that for arbitrarily dependent p-values, we can first randomly permute them and then apply an ex-p-merging function (not necessarily symmetric) such as the one in Theorem 3.4, to obtain a p-value.

The next result gives a simple condition on the calibrator f that guarantees that the probability of rejection using F in (3) is sharp for some $\mathbf{P}$.

Proposition 3.6. *Suppose that f is a convex admissible calibrator with $f(0+) \le K$ and $f(1) = 0$, and F is in (3). For $\alpha \in (0,1)$, there exists an exchangeable $\mathbf{P} \in \mathcal{U}^K$ such that $\mathbb{P}(F(\mathbf{P}) \le \alpha) = \alpha$.*

Remark 3.7. Proposition 3.6 is not sufficient to justify admissibility of F in (3). In general, admissibility of *ex-p-merging functions* remains unclear. For instance, take F in (3) with $f(p) = (2-2p)_+$, corresponding to the arithmetic average, as in Section 6 below. If $K = 2$, then F is not admissible as it is strictly dominated by the Bonferroni p-merging function given by $F_{\text{Bonf}}(p_1, \dots, p_K) = K\min\{p_1, \dots, p_K\}$. For $K \ge 3$, F and F_{Bonf} are not comparable.

3.2 Homogeneity of p-merging functions

As seen from Theorem 3.4, the class of ex-p-merging functions we obtained in (3) are homogeneous. Indeed, all explicit p-merging functions in the literature are homogeneous; see Vovk et al. (2022) for many examples. In the next result, we show a very special feature of homogeneous p-merging functions, justifying their relevance in applications.

Theorem 3.8. *Let F be a p-merging function. For any $\alpha \in (0,1)$, there exists a homogeneous p-merging function G such that $R_\alpha(F) \subseteq R_\alpha(G)$.*

As a consequence of Theorem 3.8, if the level α is determined before choosing the p-merging function, then it suffices to consider homogeneous ones, since their rejection sets are at least as larger as those of other p-merging functions. Since all admissible homogeneous p-merging functions have the form in Theorem 2.2, the class of our ex-p-merging functions in (3), sharing a similar form to (1), is quite broad. Note that Theorem 3.8 does not imply that there exists a homogeneous p-merging function G dominating F in general, because the construction of G depends on the given α.

3.3 Sequentially combining a stream of exchangeable p-values

In Subsection 3.1 we have seen that, if our starting vector of p-values $\mathbf{P} = (P_1, \ldots, P_K)$ is exchangeable, then it is possible to derive new combination rules exploiting their exchangeability. The technique for developing these new rules involves converting the initial p-values into e-values through an α-dependent calibrator. Following this transformation, the exchangeable Markov inequality (Theorem 2.4) plays a crucial role in formulating these rules. In particular, notice that the inequality in Theorem 2.4 is uniformly valid; therefore, the exchangeable sequence of p-values need not be limited to a set with cardinality K, but it is possible to continue to add p-values and stop when the procedure is stable.

To provide a concrete example, in the case where p-values are obtained by applying a sample-splitting procedure to the same dataset, a researcher can obtain one p-value at a time simply by performing a new data split. The researcher would then aim to combine these p-values sequentially and potentially stop when the procedure appears to stabilize.

We first define the following lemma:

Lemma 3.9. *Let $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2) \in [0,1]^K$ be a vector with $\mathbf{p}_1 \in [0,1]^{K_1}, \mathbf{p}_2 \in [0,1]^{K_2}$ and $K = K_1 + K_2$. In addition, let F be the function defined in (3). Then*

$$F(\mathbf{p}) = F(\mathbf{p}_1, \mathbf{p}_2) \le F(\mathbf{p}_1).$$

The above result implies that the function F in (3) is non-increasing as more parameters (i.e., p-values) are added. In addition, the following holds.

Theorem 3.10. *Let $P_1, P_2, ... \in \mathcal{U}^\infty$ be an infinite exchangeable sequence and let F be the function defined in (3). Then,*

$$\mathbb{P}(\exists k \ge 1 : F(\mathbf{P}_k) \le \alpha) \le \alpha,$$

where $\mathbf{P}_k = (P_1, \ldots, P_k)$.

The result allows the analyst to either continue collecting new (exchangeable) p-values or to cease based on the outcome. In particular, in the example described at the beginning of the section, a researcher can stop the procedure when the result seems to stabilize. However, there are some issues to consider when applying such a procedure. In particular, some calibrators f depend on the number K of p-values. For example, this is the case for the Hommel combination (Section 5) and for the harmonic mean (Section 7). As a final caveat, exchangeability becomes more stringent when the number K grows; for instance, in general, for a given K-dimensional exchangeable vector $(P_1, \dots, P_K)$, there may not exist P_{K+1} such that $(P_1, \dots, P_{K+1})$ is exchangeable. Luckily, in many practical situations, the procedure that produced the K exchangeable p-values in the first place could also produce more of them; for example, this happens when the p-value was produced by sample splitting.

3.4 Randomized p-merging functions

In this subsection, we start with a collection of arbitrarily dependent p-values and we will show how it is possible to enhance existing merging rules using a simple randomization trick. In this case, we denote

$$\mathcal{U}^K \otimes \mathcal{U} = \{(\mathbf{P}, U) \in \mathcal{U}^K \times \mathcal{U} : U \text{ and } \mathbf{P} \text{ are independent}\},$$

and we state a randomized version of the converse direction of Theorem 2.2, by changing the constant 1 in (1) to a uniform random variable U.

Theorem 3.11. *Let $f_1, \dots, f_K$ be calibrators and $(P_1, \dots, P_K, U) \in \mathcal{U}^K \otimes \mathcal{U}$. For each $\alpha \in (0,1)$ and $(\lambda_1, \dots, \lambda_K) \in \Delta_K$, we have*

$$\mathbb{P}\left(\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\alpha}\right) \geq U\right) \leq \alpha.$$

If $f_1, \dots, f_K$ are admissible calibrators and $\mathbb{P}(\sum_{k=1}^{K} \lambda_k f_k(P_k/\alpha) \leq 1) = 1$, then equality holds

$$\mathbb{P}\left(\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\beta}\right) \geq U\right) = \beta \quad \textit{for all } \beta \in (0, \alpha]. \tag{4}$$

The result in Theorem 3.11 is a direct consequence of the uniformly randomized Markov inequality (UMI) introduced by Ramdas and Manole (2024); see Theorem 2.5.

Definition 3.12. A *randomized p-merging function* is an increasing Borel function $F : [0,1]^{K+1} \to [0,1]$ such that $\mathbb{P}(F(\mathbf{P}, U) \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$ and $(\mathbf{P}, U) \in \mathcal{U}^K \otimes \mathcal{U}$. It is *homogeneous* if $F(\gamma \mathbf{p}, u) = \gamma F(\mathbf{p}, u)$ for all $\gamma \in (0,1]$ and $(\mathbf{p}, u) \in [0,1]^{K+1}$. A randomized p-merging function F is admissible if for any randomized p-merging function G, $G \leq F$ implies $G = F$.

Let $f_1, \dots, f_K$ be calibrators and $(\lambda_1, \dots, \lambda_K) \in \Delta_K$. For $\alpha \in (0,1)$, define the randomized rejection region by

$$R_\alpha = \left\{(\mathbf{p}, u) \in [0,1]^{K+1} : \sum_{k=1}^{K} \lambda_k f_k\left(\frac{p_k}{\alpha}\right) \geq u\right\}$$

where we set $f_k(p_k/u) = 0$ if $u = 0$. Using R_α, we can define the function $F : [0,1]^{K+1} \to [0,1]$ by

$$\begin{aligned} F(\mathbf{p}, u) &= \inf\{\alpha \in (0,1) : (\mathbf{p}, u) \in R_\alpha\} \\ &= \inf\left\{\alpha \in (0,1) : \sum_{k=1}^{K} \lambda_k f_k\left(\frac{p_k}{\alpha}\right) \geq u\right\}, \end{aligned} \tag{5}$$

with the convention $0 \times \infty = \infty$ (this guarantees $F(\mathbf{p}, u) = 0$ when any component of $(\mathbf{p}, u)$ is 0).

Theorem 3.13. *If $f_1, \dots, f_K$ are calibrators and $(\lambda_1, \dots, \lambda_K) \in \Delta_K$, then F in (5) is a homogeneous randomized p-merging function. Moreover, F is lower semicontinuous.*

In case of symmetric p-merging functions (i.e., $F(\mathbf{p}, u) = F(\mathbf{q}, u)$ for any permutation $\mathbf{q}$ of $\mathbf{p}$), we have the following corollary, which directly follows from Theorem 3.13.

Corollary 3.14. *For any calibrator f,*

$$F(\mathbf{p}, u) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} f\left(\frac{p_k}{\alpha}\right) \geq u \right\},$$

is a homogeneous, symmetric randomized p-merging function.

A simple observation is that replacing $f(p)$ with $f(p) \wedge K$ does not change the function F. This observation allows us to only consider calibrators that are bounded above by K, which can improve some existing p-merging functions. This is similar to what was done in Vovk et al. (2022) in the context of deterministic p-merging functions.

Remark 3.15 (P-hacking via repeated derandomization)**.** Randomized methods may not cause a lack of reproducibility if they are part of a standard automated data analysis pipeline without a human in the loop. However, if a human is actively involved in the analysis, then there is a risk of "p-hacking", where a (malevolent) researcher re-runs the randomized method many times in order to obtain a desirable result according to their own utility. In our setting, the issue translates into sampling different U and picking the smallest one (which would be closer to 0 the more times the procedure is re-run). Clearly, this procedure is not valid and can be particularly problematic in the context of confirmatory analysis, where the end product is a binary decision.

Remark 3.16 (Internal randomization)**.** The use of *internal* (as opposed to *external*) randomization can be particularly useful to prevent and mitigate the risk of p-hacking mentioned above. We mention two such strategies below. First, note that F in (5) is increasing in each of its arguments. If one has prior information that one of the p-values, say P_1, is independent of the rest (but the rest can be arbitrarily dependent), then one can use P_1 for randomization and apply F (with one less input dimension for $\mathbf{p}$) to $(\mathbf{p}, u) = (P_2, \dots, P_K, U)$ with $U = P_1$ to obtain a p-value that does not depend on external randomization. Monotonicity of $u \mapsto F(\mathbf{p}, u)$ guarantees two things. First, a p-variable may be stochastically larger than a standard uniform one, so increasing monotonicity is needed for validity. Second, if P_1 is indeed very small, i.e., it carries signal against the null, then the combined p-value will benefit from this signal. This form of internal randomization has been discussed in Wang (2024, Section B.1).

An alternative method of internal randomization through data (instead of p-values) is discussed by Ramdas and Manole (2024, Section 10.6). To understand their proposal, assume that each p-value is calculated using a function of only the order statistics of iid data; for example, it could be based only on sums, like the t-statistic. In this case, the rank of the first data point (amongst all the data points) is a discrete uniform random variable, and can be used in place of U. This way, if the dataset is itself public (posted by a previous research paper, for example), the ordering itself is not in the hands of the researcher analyzing that data, reducing the risk of p-hacking. We refer interested readers to Ramdas and Manole (2024, Section 10) or Lei and Sudijono (2024) for further discussions.

It is feasible to combine the results presented in Subections 3.1 and 3.4 through the formulation of novel p-merging functions that exploit both the properties of exchangeability and randomization. These results are presented in Appendix B and are based on the exchangeable and uniformly randomized Markov inequality presented in Theorem 2.6.

3.5 Instantiating the above ideas

The ideas above were admittedly somewhat abstract, but provide us with the general tools to improve specific combination rules. The following sections do this for several rules, one by one. To elaborate, one of the most commonly employed p-merging functions is the Bonferroni method:

$$F_{\mathrm{Bonf}}(\mathbf{p}) = K p_{(1)},$$

where $p_{(1)}$ is the minimum of observed p-values. Rüger (1978) extended the aforementioned rule in a more general sense. In particular, it is possible to prove that

$$F_{\mathrm{R}}(\mathbf{p}) := \frac{K}{k} p_{(k)}, \quad k \in \{1, \ldots, K\}, \tag{6}$$

is a p-value, where $p_{(k)}$ represents the k-th smallest p-value among $(p_1, \ldots, p_K)$. In other words, the λ-quantile $p_{(\lceil \lambda K \rceil)}$ is a p-value if multiplied by the factor $1/\lambda$. In particular, the robust and widely used combination rule, twice the median, is part of this class. The next section improves on this combination rule.

Vovk and Wang (2020) introduced the class of p-merging functions based on the generalized mean, also called *M-family*. This general class takes the form

$$a_{r,K} \left(\frac{p_1^r + \cdots + p_K^r}{K} \right)^{1/r}, \tag{7}$$

where $r \in \mathbb{R} \setminus \{0\}$ and $a_{r,K}$ is the smallest constant making (7) a p-merging function. This class encompasses numerous well-known cases, each distinguished by different values of the parameter r. In particular, if $r = 1$ then (7) reduces to the simple average introduced in Rüschendorf (1982) and the value $a_{1,K} = 2$. Another important case is the harmonic mean obtained with $r = -1$. Among this class, the harmonic mean combination rule should be used when substantial dependence among the p-values is suspected. If the dependence is really strong, the arithmetic mean might be a safer option. In the following sections, we demonstrate that if p-values exhibit exchangeability or if randomization is allowed, then it becomes feasible to enhance most of these combination rules.

Before continuing, we provide a few simple examples to highlight the benefits of the proposed findings and demonstrate how our methods can enhance the existing approaches. In the examples we will use some rules proposed in Table 1, foreshadowing many results to come.

Example 3.17. Suppose that the vector $\mathbf{P}$ of 3 p-values is generated as follows. With 0.9 probability, $\mathbf{P} = (P_1, P_2, P_3)$ where P_1, P_2, P_3 are independent, and with 0.1 probability $\mathbf{P} = (P_4, P_4, P_4)$. We assume $P_i \in \mathcal{U}$ under the null, while under the alternative each P_i is distributed as Beta$(0.2, 1)$. The Beta distribution $(a, 1)$ with small $a > 0$ is a typical model for p-values under the alternative hypothesis; see Sellke et al. (2001). Clearly, the p-values are exchangeable under the null. Suppose that one want to use the rule $(3/2)p_{(2)}$ to combine the p-values. Then we can check numerically that, fixing the threshold α to 0.05, the probability of rejection is 0.5101 under the alternative. If we use the ex-p-merging derived from the median (see Theorem 4.1 below) the probability of rejection increases to 0.6207.

Example 3.18. Suppose that we want to test an hypothesis and we have that under the null $P_1 \in \mathcal{U}$ while under the alternative $P_1 \sim$ Beta$(0.2, 1)$. The vector of p-values is generated as follows $\mathbf{P} := (P_1, P_2) = (P_1, 1 - P_1)$, where P_2 is an exact p-value and $\mathbf{P}$ is exchangeable. In this scenario, the commonly applied twice the mean, even though controls the type I error, has no power under the alternative hypothesis because the result is always equal to 1. On the other hand, using the exchangeable rule derived from the arithmetic average (see Theorem 6.2), the probability of rejection is $\mathbb{P}(P_1 \leq \alpha/2) \approx 0.48$ under the alternative for $\alpha = 0.05$.

3.6 Combining asymptotic p-values

Before proceeding with the remainder of the paper and introducing new merging rules based on the results presented in the preceding sections, we want to examine the scenario wherein the p-values are asymptotically valid. Many of the results obtained in the literature rely on uniform or super-uniform p-values (see Section 2); however, in statistical applications, p-values are often asymptotic, and they are not necessarily valid p-values in finite sample. See, for example, Severini (2000) for an introduction to p-values obtained using likelihood-based methods.

All methods in our paper work also for asymptotic p-values, i.e., those that converge in distribution to p-values. Suppose that $(\mathbf{P}_n)_{n\in\mathbb{N}}$ is a sequence of nonnegative random vectors that converges to a vector $\mathbf{P}$ of p-values in distribution. With an upper semicontinuous calibrator f (recall that all admissible calibrators are upper semicontinuous), for each $\alpha \in (0,1)$ and $u \in (0,1)$, the rejection sets R_α given by

$$R_\alpha = \left\{ \mathbf{p} \in [0,\infty)^K : \bigvee_{k\le K} \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \ge 1 \right\}$$

or

$$R_\alpha = \left\{ \mathbf{p} \in [0,\infty)^K : \frac{1}{K}\sum_{k=1}^{K} f\left(\frac{p_k}{\alpha}\right) \ge u \right\}$$

are closed. As a consequence, by the Portmanteau Theorem,

$$\limsup_{n\to\infty} \mathbb{P}(\mathbf{P}_n \in R_\alpha) \le \mathbb{P}(\mathbf{P} \in R_\alpha).$$

Therefore, any methods in our paper that produce a p-value for the vector $\mathbf{P}$ of p-values (exchangeable or arbitrarily dependent) produce an asymptotic p-value for any $(\mathbf{P}_n)_{n\in\mathbb{N}}$ that converges to $\mathbf{P}$ in distribution.

4 Improving Rüger's combination rule

Vovk et al. (2022) showed that the p-merging function defined in (6), with a trivial modification (i.e., return 0 if $p_{(1)} = 0$; see Theorem 7.3 of Vovk et al. (2022)) is admissible for $k \ne K$, and it is admissible among symmetric p-merging functions when $k = K$. In particular, the corresponding calibrator that induces (6) is

$$f(p) = \frac{K}{k}\mathbb{1}\{p \in (0, k/K]\} + \infty\mathbb{1}\{p = 0\},$$

and this implies that we can exploit directly the duality between rejection regions and p-merging functions.

4.1 An exchangeable Rüger combination rule

If the exchangeability condition is satisfied, then we can obtain something better than the combination rule in (6). First, it is clear that $f(p) = \frac{K}{k}\mathbb{1}\{p \in (0, k/K]\} + \infty\mathbb{1}\{p = 0\}$ is an admissible calibrator. We now define the function

$$F_{\mathrm{ER}}(\mathbf{p}) = \inf\left\{ \alpha \in (0,1) : \bigvee_{\ell\le K} \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{K}{k}\mathbb{1}\left\{\frac{p_i}{\alpha} \le \frac{k}{K}\right\} \ge 1 \right\}.$$

Below, for fixed $k \in \{1,\dots,K\}$, we let $p^{\ell}_{(\lambda_\ell)}$ denote the $\lceil \ell\frac{k}{K}\rceil$-th ordered value obtained using the first ℓ values of $\mathbf{p}$. Essentially, $p^{\ell}_{(\lambda_\ell)}$ is the upper quantile of order k/K obtained using the first ℓ p-values.

Theorem 4.1. *For any fixed $k \in \{1,\dots,K\}$ the function F_{ER} satisfies*

$$F_{\mathrm{ER}}(\mathbf{p}) = \left(\frac{K}{k}\bigwedge_{\ell=1}^{K} p^{\ell}_{(\lambda_\ell)}\right)\mathbb{1}\{p_{(1)} > 0\} \text{ for } \mathbf{p} \in [0,1]^K,$$

where $\lambda_\ell := \lceil \ell\frac{k}{K}\rceil$, and it is an ex-p-merging function that strictly dominates F_{R} in (6).

It may be useful to note that the Bonferroni rule is not improved using this method. Indeed, fixing $k = 1$, we find that $p^{\ell}_{(\lambda_\ell)}$ reduces to the minimum of the first ℓ p-values subsequently taking the minimum of the obtained sequences coincides with the overall minimum. In addition, Rüger can be sharp, for some exchangeable $\mathbf{P}$, i.e., satisfying $F_{\mathrm{R}}(\mathbf{P}) \in \mathcal{U}$, and so is our proposal.

4.2 A randomized Rüger combination rule

In this part, we prove that if randomization is allowed, it becomes feasible to enhance the combination introduced by Rüger (1978), even if the sequence of p-values presents an arbitrary dependence and the obtained result has nice properties in terms of interpretability. As before, we define the merging function

$$F_{\mathrm{UR}}(\mathbf{p}, u) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} \frac{K}{k} \mathbb{1} \left\{ \frac{p_k}{\alpha} \leq \frac{k}{K} \right\} \geq u \right\}.$$

Theorem 4.2. *For any fixed $k \in \{1, \dots, K\}$, the function F_{UR} satisfies*

$$F_{\mathrm{UR}}(\mathbf{p}, u) = \frac{K}{k} p_{(\lceil uk \rceil)} \mathbb{1}\{p_{(1)} > 0\},$$

and it is a randomized p-merging function that strictly dominates F_{R} in (6).

The above theorem implies in particular that the Rüger combination rule is inadmissible if external randomization is allowed, despite it being admissible if randomization is not allowed (Vovk et al., 2022). The fact that $p_{(\lceil UK \rceil)}$ is a p-value is particularly interesting. It has a simple interpretation: sort the p-values and pick the one at a uniformly random index. In addition, for the latter combination rule (4) holds for all $\beta \in (0, 1]$.

It is worth noting that when $k = 1$ the Rüger combination rule reduces to the Bonferroni method; however, the introduction of a randomization does not yield any practical benefit since $\lceil U \rceil = 1$.

5 Improving Hommel's combination rule

The method proposed in Section 4 requires us to choose the value of k in advance; a solution that solves the problem is proposed by Hommel (1983). Hommel's combination rule is given by

$$F'_{\mathrm{Hom}}(\mathbf{p}) := h_K \bigwedge_{k=1}^{K} F_{\mathrm{R}}(\mathbf{p}; k) = \left(\sum_{k=1}^{K} \frac{1}{k} \right) \bigwedge_{k=1}^{K} \frac{K}{k} p_{(k)}, \tag{8}$$

with $h_K = \sum_{k=1}^{K} \frac{1}{k}$. This function allows selecting the minimum derived from the combinations based on ordered statistics with a multiplicative cost of $h_K \approx \log K$.

It is possible to prove that the Hommel combination rule is not admissible and is dominated by the *grid harmonic merging function* introduced in Vovk et al. (2022). For completeness, we state here a useful lemma.

Lemma 5.1. *Let f be a function defined by*

$$f(p) = \frac{K \mathbb{1}\{h_K p \leq 1\}}{\lceil K h_K p \rceil}. \tag{9}$$

Then f is an admissible calibrator. Moreover, the p-merging function induced by f is

$$F_{\mathrm{Hom}}(\mathbf{p}) := \inf \left\{ \alpha \in (0,1) : \sum_{k=1}^{K} \frac{\mathbb{1}\{h_K p_k / \alpha \leq 1\}}{\lceil K h_K p_k / \alpha \rceil} \geq 1 \right\},$$

is valid and it dominates the Hommel combination rule.

The calibrator in (9) coincides, with a slight adjustment, with the BY calibrator introduced in Xu et al. (2024). An interesting fact is that the function F_{Hom} is always admissible in the class of symmetric p-merging functions; while it is admissible in the class of p-merging function (not necessarily symmetric) if K is not a prime number (Vovk et al., 2022, Theorem 7.1). The Hommel function allows for the selection of the minimum among the K possible different quantiles of $\mathbf{p}$.

In reality, one can choose to select only certain quantiles among K (e.g., one can select the minimum between K times the minimum, 2 times the median and the maximum), hoping to pay a price less than $\log K$. We treat this problem in Appendix C, where we introduce a generalization of the Hommel combination rule.

5.1 An exchangeable Hommel's combination rule

Starting from the results in the previous sections, we can introduce a merging function, which improves Hommel's combination if the input p-values are exchangeable. In particular, we define the function

$$F_{\text{EHom}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{K\mathbb{1}\{h_K p_i/\alpha \le 1\}}{\lceil K h_K p_i/\alpha \rceil} \ge 1\right\}. \tag{10}$$

Theorem 5.2. *The function F_{EHom} is an ex-p-merging function and it strictly dominates the function F_{Hom}.*

The computation of a closed form for (10) is not straightforward, a possible solution to calculate the value of F_{EHom} is by using Algorithm 1 defined in Appendix E.

5.2 A randomized Hommel's combination rule

The *randomized* version of the function F_{Hom} has been proposed in Xu and Ramdas (2023, Appendix E) and takes the following form:

$$F_{\text{UHom}}(\mathbf{p}) := \inf\left\{\alpha \in (0,1) : \sum_{k=1}^{K} \frac{\mathbb{1}\{h_K p_k/\alpha \le 1\}}{\lceil K h_K p_k/\alpha \rceil} \ge u\right\}.$$

For completeness, we report here the following theorem.

Theorem 5.3. *The function F_{UHom} is a randomized p-merging function and it strictly dominates the function F_{Hom}.*

The value of the function F_{UHom} can be computed using Algorithm 1 in Vovk et al. (2022), substituting the value of 1 by u.

6 Improving the "twice the average" combination rule

We now study the case of $r = 1$ in (7), which corresponds to the arithmetic mean. The general case, which allows $r \in \mathbb{R}\backslash\{0\}$, is considered in Appendix D. In the following, let $A(\mathbf{p})$ denote the arithmetic average of any vector $\mathbf{p}$, let $\mathbf{p}_m := (p_1, \dots, p_m)$ denote the vector containing the first m p-values, and let $\mathbf{p}_{(m)}$ denote the vector containing the smallest m elements of $\mathbf{p}$: $\mathbf{p}_{(m)} = (p_{(1)}, \dots, p_{(m)})$ such that $p_{(1)} \le \dots \le p_{(m)}$. In addition, we denote by $\mathbf{p}^{\ell}_{(m)} = (p^{\ell}_{(1)}, \dots, p^{\ell}_{(m)})$, $m \in \{1, \dots, \ell\}$, the vector containing the smallest m elements of $(p_1, \dots, p_\ell)$. First, let us introduce a lemma that will be instrumental in subsequent discussions.

Lemma 6.1. *Let $f(p) = (2-2p)_+\mathbb{1}\{p \in (0,1]\} + \infty\mathbb{1}\{p = 0\}$. Then, f is an admissible calibrator.*

In particular, we have that the calibrator defined in Lemma 6.1 is larger or equal than $f'(p) := (2-2p)$, which is the function inducing the average combination rule

$$F'_{\mathrm{A}}(\mathbf{p}) := 2A(\mathbf{p}), \tag{11}$$

which is a valid p-merging function. Note that f' can take negative values (its arguments p/α can be larger than 1), so it is technically not a calibrator in the sense of our definitions.

6.1 Exchangeable average combination rule

We now define ex-p-merging functions

$$F_{\mathrm{EA}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(2 - 2\frac{p_i}{\alpha}\right)_+ \ge 1\right\};$$
$$F'_{\mathrm{EA}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(2 - 2\frac{p_i}{\alpha}\right) \ge 1\right\}.$$

Theorem 6.2. *The dominations among ex-p-merging functions $F_{\mathrm{EA}} \le F'_{\mathrm{EA}} \le F'_{\mathrm{A}}$ are strict. Moreover,*

$$F'_{\mathrm{EA}}(\mathbf{p}) = 2\left\{\bigwedge_{m=1}^{K} A(\mathbf{p}_m)\right\};$$
$$F_{\mathrm{EA}}(\mathbf{p}) = \left(\bigwedge_{\ell=1}^{K}\bigwedge_{m=1}^{\ell} \frac{2A(\mathbf{p}^{\ell}_{(m)})}{(2-\ell/m)_+}\right)\mathbb{1}\{p_{(1)} > 0\}.$$

Despite being strictly dominated, F'_{EA} is very interpretable: it is just the minimum (over m) of "twice the average" of the first m p-values.

6.2 Randomized average combination rule

We now derive an improvement for the "twice the average" rule using a simple randomization trick. In this case, we do not require exchangeability but we allow for an arbitrary dependence among the p-values. We define the randomized p-merging function F_{UA} by

$$F_{\mathrm{UA}}(\mathbf{p}, u) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(2 - \frac{2p_k}{\alpha}\right)_+ \ge u\right\}.$$

Clearly, $F_{\mathrm{UA}}(\mathbf{p}, u) \le F'_{\mathrm{UA}}(\mathbf{p}, u)$, where F'_{UA} is defined by

$$F'_{\mathrm{UA}}(\mathbf{p}, u) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(2 - \frac{2p_k}{\alpha}\right) \ge u\right\}.$$

In particular, if randomization is not allowed then u is replaced by 1 and $F'_{\mathrm{UA}}(\mathbf{p}, 1)$ coincides with F'_{A} in (11). A p-merging function (such as F'_{A}) can also be seen as a randomized p-merging function, for which the argument u does not affect its value.

Theorem 6.3. *The dominations among randomized p-merging functions $F_{\mathrm{UA}} \le F'_{\mathrm{UA}} \le F'_{\mathrm{A}}$ are strict. Moreover,*

$$F'_{\mathrm{UA}}(\mathbf{p}) = \frac{2A(\mathbf{p})}{2-u};$$
$$F_{\mathrm{UA}}(\mathbf{p}) = \left(\bigwedge_{m=1}^{K} \frac{2A(\mathbf{p}_{(m)})}{(2-Ku/m)_+} \right) \mathbb{1}\{p_{(1)} > 0\}.$$

A method that can be directly compared with Theorem 6.3 is to use $F^*_{\mathrm{UA}}(\mathbf{p}, u) := A(\mathbf{p})/(2-2u)$ proposed by Wang (2024, Section B.2). This function F^*_{UA} is also a randomized p-merging function. One can see that F^*_{UA} does not dominate and is not dominated by any of F_{A}, F_{UA} and F'_{UA}. Moreover, there is a simple relationship: $F'_{\mathrm{UA}}(\mathbf{p}, u) \le F^*_{\mathrm{UA}}(\mathbf{p}, u)$ if and only if $u \ge 2/3$ for every $\mathbf{p}$ that is not the zero vector.

7 Improving the harmonic mean combination rule

The harmonic mean p-value was studied by Wilson (2019). In our context, it corresponds to the merging function in (7) when $r = -1$. We first state a lemma on a calibrator that we will use later.

Lemma 7.1. *Define the function*

$$f(p) = \min\left\{ \frac{1}{T_K p} - \frac{1}{T_K}, K \right\} \mathbb{1}\{p \in [0,1]\},$$

with $T_K \ge 1$. Then f is a calibrator if T_K satisfies $KT_K + 1 - e^{T_K} \le 0$, and in particular, f is a calibrator if $T_K = \log K + \log\log K + 1$.

In the following, we fix $T_K = \log K + \log\log K + 1$ and denote by $H(\mathbf{p}) = K(\sum_{k=1}^K 1/p_k)^{-1}$ the harmonic mean of the vector $\mathbf{p}$ where K is the number of elements contained in $\mathbf{p}$. We begin with a result that is new even under arbitrary dependence.

Proposition 7.2. *$F'_{\mathrm{H}}(\mathbf{p}) := (T_K + 1)H(\mathbf{p})$ is a p-merging function.*

The above result differs from the formulation given in Vovk and Wang (2020, Proposition 9), which states that $e \log K H(\mathbf{p})$ is a p-merging function, thus sharpening their result for $K \ge 4$. Below we will further improve this result for exchangeable p-values. Moreover, a correction factor in the order of $\log K$ as $K \to \infty$ (although smaller than T_K) is needed even for independent p-values (see Proposition 6 of Chen et al. (2024)). The harmonic mean has some advantages over many other rules under certain dependence conditions; see Gui et al. (2023). It performs similarly to the Hommel combination; see Chen et al. (2023).

7.1 Exchangeable harmonic mean combination rule

We define homogeneous ex-p-merging functions as follows

$$F_{\mathrm{EH}}(\mathbf{p}) = \inf\left\{ \alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\frac{\alpha}{T_K p_i} - \frac{1}{T_K} \right)_+ \ge 1 \right\};$$
$$F'_{\mathrm{EH}}(\mathbf{p}) = \inf\left\{ \alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\frac{\alpha}{T_K p_i} - \frac{1}{T_K} \right) \ge 1 \right\}.$$

Theorem 7.3. *The dominations among ex-p-merging functions $F_{\mathrm{EH}} \le F'_{\mathrm{EH}} \le F'_{\mathrm{H}}$ are strict. Moreover,*

$$F'_{\mathrm{EH}}(\mathbf{p}) = \bigwedge_{m=1}^{K} (T_K + 1) H(\mathbf{p}_m);$$
$$F_{\mathrm{EH}}(\mathbf{p}) = \bigwedge_{\ell=1}^{K} \left(\bigwedge_{m=1}^{\ell} \left(\frac{\ell\, T_K}{m} + 1 \right) H(\mathbf{p}^{\ell}_{(m)}) \right).$$

7.2 Randomized harmonic mean combination rule

Similarly to Section 6, we derive an improvement for the harmonic mean using a randomization trick in the case of arbitrarily dependent p-values. Define

$$F_{\mathrm{UH}}(\mathbf{p}, u) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\alpha}{T_K p_k} - \frac{1}{T_K} \right)_+ \ge u \right\};$$
$$F'_{\mathrm{UH}}(\mathbf{p}, u) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\alpha}{T_K p_k} - \frac{1}{T_K} \right) \ge u \right\}.$$

Theorem 7.4. *The dominations among randomized p-merging functions $F_{\mathrm{UH}} \le F'_{\mathrm{UH}} \le F'_{\mathrm{H}}$ are strict. Moreover,*

$$F'_{\mathrm{UH}}(\mathbf{p}) = (T_K u + 1) H(\mathbf{p});$$
$$F_{\mathrm{UH}}(\mathbf{p}) = \bigwedge_{m=1}^{K} \left(\frac{uKT_K}{m} + 1 \right) H(\mathbf{p}_{(m)}).$$

A non-randomized improvement of F'_{H} can be achieved fixing $u = 1$ in F_{UH}. This coincides with the function $F_{\mathrm{H}}(\mathbf{p}) = \bigwedge_{m=1}^{K}((KT_K)/m + 1)H(\mathbf{p}_{(m)})$.

8 Improving the geometric mean combination rule

We now derive some new combination based on the geometric mean, a special case of (7) when $r \to 0$. Let $G(\mathbf{p}) = (\prod_{k=1}^{K} p_k)^{1/k}$ denote the geometric mean of the vector $\mathbf{p}$. The calibrator, in this case, is given by $f(p) = (-\log p)_+$, which is an admissible calibrator. Actually, a slightly improved calibrator is $f(p) = (-(\log p)/T)_+ \wedge K$ for some $T < 1$ satisfying $\int_0^1 f(p)\mathrm{d}p \le 1$. This condition is verified when $1 - e^{-KT} \le T$, which makes T very close to 1 for K moderately large (see Section 3.2 in Vovk and Wang, 2020). In the sequel, we denote by

$$F'_{\mathrm{G}}(\mathbf{p}) := eG(\mathbf{p}), \tag{12}$$

which is a valid p-merging function studied by Vovk and Wang (2020).

8.1 Exchangeable geometric mean combination rule

Following the same approach as in the preceding sections, we define

$$F_{\mathrm{EG}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\log\frac{\alpha}{p_i}\right)_+ \ge 1\right\};$$

$$F'_{\mathrm{EG}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell}\sum_{i=1}^{\ell}\log\frac{\alpha}{p_i} \ge 1\right\}.$$

Theorem 8.1. *The dominations among ex-p-merging functions $F_{\mathrm{EG}} \le F'_{\mathrm{EG}} \le F'_{\mathrm{G}}$ are strict. Moreover,*

$$F'_{\mathrm{EG}}(\mathbf{p}) = e\left\{\bigwedge_{m=1}^{K} G(\mathbf{p}_m)\right\};$$

$$F_{\mathrm{EG}}(\mathbf{p}) = \bigwedge_{\ell=1}^{K}\left(\bigwedge_{m=1}^{\ell} e^{\ell/m} G(\mathbf{p}^{\ell}_{(m)})\right).$$

8.2 Randomized geometric mean combination rule

As in the previous sections, we define the randomized p-merging functions as follows:

$$F_{\mathrm{UG}}(\mathbf{p}, u) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(\log\frac{\alpha}{p_k}\right)_+ \ge u\right\};$$

$$F'_{\mathrm{UG}}(\mathbf{p}, u) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\log\frac{\alpha}{p_k} \ge u\right\}.$$

Theorem 8.2. *The dominations among randomized p-merging functions $F_{\mathrm{UG}} \le F'_{\mathrm{UG}} \le F'_{\mathrm{G}}$ are strict. Moreover,*

$$F'_{\mathrm{UG}}(\mathbf{p}) = e^{u} G(\mathbf{p});$$

$$F_{\mathrm{UG}}(\mathbf{p}) = \bigwedge_{m=1}^{K}\left(e^{u\frac{K}{m}} G(\mathbf{p}_{(m)})\right).$$

A non-randomized improvement of the combination in (12) can be obtained fixing $u = 1$ in F_{UG}. This gives the combination rule $F_{\mathrm{G}}(\mathbf{p}) = \bigwedge_{m=1}^{K} e^{K/m} G(\mathbf{p}_{(m)})$.

9 Simulation study

In the previous sections, new p-merging functions have been introduced. These new rules are obtained using a randomization trick or they rely on exchangeability of p-values. Specifically, the introduced rules have been shown to dominate their original counterparts by utilizing randomness or exchangeability (or both). In this section, our aim is to investigate their performance using simulated data.

We consider the example described in Vovk and Wang (2020, Section 6) (a similar example is proposed in Chen et al. (2023)), where p-values are generated in the following way:

$$X_k = \rho Z + \sqrt{1-\rho^2} Z_k - \mu, \quad P_k = \Phi(X_k), \tag{13}$$

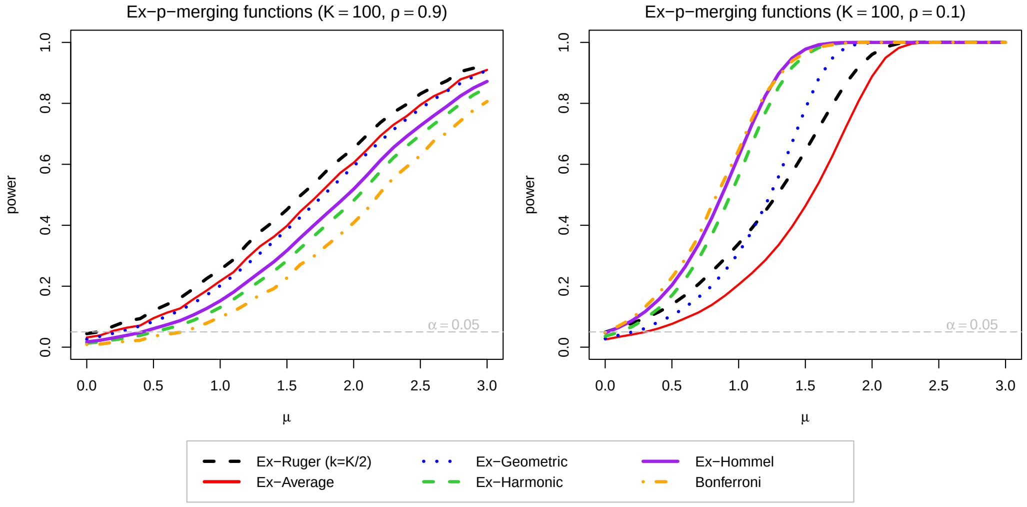

Figure 1: Combination of p-values using different ex-p-merging functions under high (left) and low (right) dependence. The performance of the different ex-p-merging functions is almost reversed in the two situations.

where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, $Z, Z_1, \dots, Z_K \overset{iid}{\sim} \mathcal{N}(0,1)$, and $\mu \geq 0$ and $\rho \in [0,1]$ are constants. It is simple to prove that $P_1, \dots, P_K$ are exchangeable and their marginal distribution does not depend on ρ. In addition, if $\rho = 0$ then $P_1, \dots, P_K$ are independent while if $\rho = 1$ then $P_1 = \cdots = P_K$. The value p_k, in this case, can be interpreted as the p-value resulting from a one-side z-test of the null hypothesis $\mu = 0$ against the alternative $\mu > 0$ from the statistic $X_k \sim \mathcal{N}(-\mu, 1)$ with unknown μ. We let the parameter μ vary in the interval $[0,3]$ (if $\mu = 0$ then H_0 is true) and fix the upper bound of the type I error to the nominal level 0.05.

We compare the different ex-p-merging functions introduced in the previous sections, with the addition of the Bonferroni method. The parameter k for the Rüger combination rule is set to $K/2$ (twice the median). The parameter ρ is set to the values $\rho = \{0.1, 0.9\}$, corresponding to weak and strong dependence among p-values. Each simulation is repeated for a total of $B = 10,000$ replications, and we report the observed empirical average. In Figure 1, we can notice that the error level is controlled at the nominal level 0.05 for all the proposed methods. Variations in terms of power are observed depending on whether the parameter ρ is set to 0.9 or 0.1. Specifically, ex-p-merging functions that exhibit strong performance in the left plot tend to show reduced effectiveness in the right plot, and the opposite is also true. As expected, the Bonferroni method shows a higher power near independence where also the ex-Hommel combination rule defined in (10) seems to perform well. A simulation study comparing the different randomized p-merging functions is reported in Appendix F.

In this last part, we aim to explore the scenario where p-values are exchangeable under the null hypothesis but not under the alternative. Indeed, the p-values generated as in (13) are exchangeable under both the null and the alternative hypotheses; however, for the results in Section 3, it is only necessary for the p-values to be exchangeable under H_0. Specifically, if the p-values are not exchangeable under the alternative, they could be arranged in a particular way to obtain a more powerful procedure. In other words, if it is suspected that some p-values are smaller under the alternative hypothesis, they could be placed at the beginning of the vector to enhance the power of the procedure since our rules that are valid under exchangeability process the vector of p-values sequentially. Clearly, the procedure of ordering p-values in a particular way must preserve the exchangeabilty under the null hypothesis (i.e., data-dependent ordering is usually not allowed since it violates exchangeability). In the simulation setting, suppose to have K independent studies, each with observations X_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$, that are iid from a normal distribution with mean μ and variance 1. In addition, $X_{0j}, j = 1, \ldots, n_0$, is assumed to be an additional sample from the same population and common for all studies. We define the quantity $\bar{X}_i$ as

$$\bar{X}_i = \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} X_{ij}, \quad i = 0, 1, \ldots, K,$$

that is distributed as $\mathcal{N}(\sqrt{n_i}\mu, 1)$. The interest is to test the hypothesis $\mu = 0$ under the alternative $\mu \neq 0$ and the test statistic used is

$$T_k := \frac{\bar{X}_k + \bar{X}_0}{\sqrt{2}}, \quad k = 1, \ldots, K,$$

where it is possible to see that each study use the common sample in the "same way". Under the null hypothesis, the test statistics have the same marginal distribution and the model $(T_1, \ldots, T_K)$ is exchangeable. Under the alternative, the mean of the test statistics depends on the sample size, so the model cannot be exchangeable. In particular, studies with a higher number of observations are more powerful. The p-values to test the null hypothesis are given by

$$P_k = 2\Phi(-|T_k|), \quad k = 1, \ldots, K.$$

Specifically, in the simulated scenario, $K = 10$, n_i, $i = 1, \ldots, K$, take values $10, 20, \ldots, 100$ and $n_0 = 25$. The number of replications is $B = 10,000$ and the ex-p-merging functions used are F_{EA} ("twice the mean") and F_{ER} with $k = K/2$ ("twice the median"). We let the parameter μ varies in the interval $[0, 0.5]$ and three different solutions are compared: (i) the p-values are ordered in increasing order with respect to the sample size, (ii) the p-values are ordered in decreasing order with respect to the sample size, and (iii) the p-values are randomly ordered. In addition, we compare the ex-p-merging functions with the "standard" rules valid under arbitrary dependence F_{A} and F_{R}. The results are reported in Figure 2, where we see that when the p-values are ordered in decreasing order with respect to the number of observations, the combined tests are more powerful. This is because the power of individual p-values increases with the sample size. Overall, the proposed ex-p-merging are more powerful than the rules valid under arbitrary dependence.

Results under other setups are reported in Appendix F, with similar qualitative conclusions.

10 Summary

In this paper, we derive novel p-merging functions for the scenario where the p-values are exchangeable. These new rules are demonstrated to dominate their original counterparts derived under the assumption of arbitrary dependence. Furthermore, we illustrate how a simple randomization trick (introduction of a uniform random variable or uniform permutation) can also be employed in the case of arbitrarily dependent p-values to yield more powerful rules than existing ones. These results are proposed in a fully general form, addressing the relationship between p-merging functions and e-values introduced by their respective calibrators. In particular, once the corresponding e-value is obtained, it becomes feasible to utilize Markov's inequality or its randomized and exchangeable generalizations from Ramdas and Manole (2024).

As a practical recommendation, we suggest the exchangeable improvement of the Hommel combination if we have no apriori idea how strong the dependence is likely to be, but to use the exchangeable improvement of "twice the median" if the p-values are thought to be strongly dependent.

As an open question, it remains unclear whether our methods are further improvable, stemming from the unknown admissibility or tightness of exchangeable Markov's inequality beyond extreme cases.

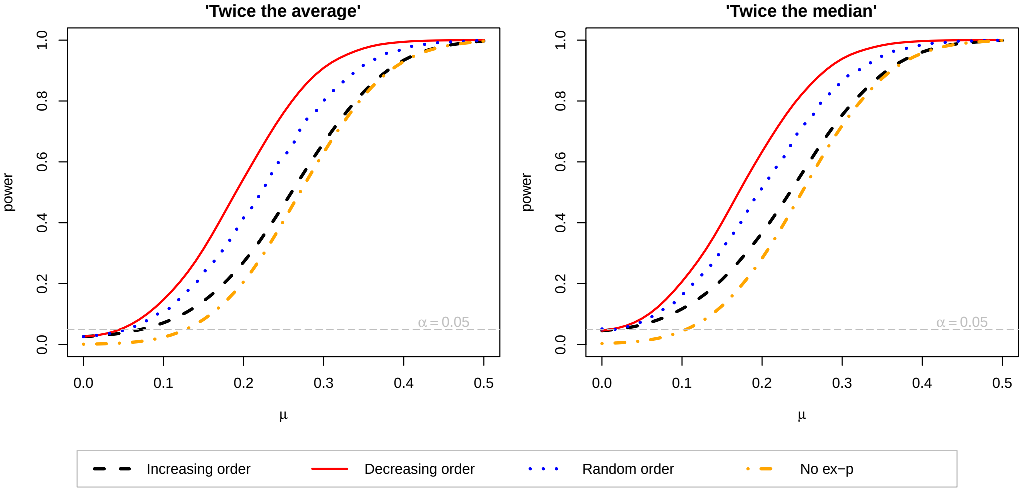

Figure 2: Combination of p-values using different ex-p-merging functions and different ordering based on the sample size. Non ex-p-merging functions valid under arbitrary dependence are added for comparison. The ex-p-merging rules are more powerful if p-values are ordered in decreasing order with respect to the sample size.

Acknowledgments

The authors thank Ilmun Kim for a suggestion on asymptotic p-values.

References

Banerjee, M., Durot, C., and Sen, B. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2):720 – 757.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(4):377–380.

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

Chen, Y., Liu, P., Tan, K. S., and Wang, R. (2023). Trade-off between validity and efficiency of merging p-values under arbitrary dependence. *Statistica Sinica*, 33(2):851–872.

Chen, Y., Wang, R., Wang, Y., and Zhu, W. (2024). Sub-uniformity of harmonic mean p-values. *arXiv preprint arXiv:2405.01368*.

Chi, Z., Ramdas, A., and Wang, R. (2024). Multiple testing under negative dependence. *Bernoulli*, forthcoming.

Choi, W. and Kim, I. (2023). Averaging p-values under exchangeability. *Statistics & Probability Letters*, 194:109748.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

DiCiccio, C. J., DiCiccio, T. J., and Romano, J. P. (2020). Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166:108865.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction.* Cambridge University Press.

Fisher, R. A. (1934). *Statistical methods for research workers.* Number 5. Oliver and Boyd.

Grünwald, P., de Heide, R., and Koolen, W. M. (2024). Safe testing. *Journal of the Royal Statistical Society, Series B (to appear with discussion).*

Gui, L., Jiang, Y., and Wang, J. (2023). Aggregating dependent signals with heavy-tailed combination tests. *arXiv preprint arXiv:2310.20460.*

Guo, F. R. and Shah, R. D. (2024). Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology.*

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

Ignatiadis, N., Wang, R., and Ramdas, A. (2024). E-values as unnormalized weights in multiple testing. *Biometrika*, 111(2):417–439.

Kim, I. and Ramdas, A. (2024). Dimension-agnostic inference using cross U-statistics. *Bernoulli*, 30(1):683 – 711.

Lei, L. and Sudijono, T. (2024). Inference for synthetic controls via refined placebo tests. *arXiv preprint arXiv:2401.07152.*

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Morgenstern, D. (1980). Berechnung des maximalen signifikanzniveaus des testes "Lehne H_0 ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika*, 27:285–286.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *The Annals of Statistics*, 37(6B):3867 – 3892.

Pearson, K. (1934). On a new method of determining "goodness of fit". *Biometrika*, 26(4):425–442.

Ramdas, A. and Manole, T. (2024). Randomized and Exchangeable Improvements of Markov's, Chebyshev's and Chernoff's inequalities. *Statistical Science*, forthcoming.

Ritzwoller, D. M. and Romano, J. P. (2023). Reproducible aggregation of sample-split statistics. *arXiv preprint arXiv:2311.14204.*

Rüger, B. (1978). Das maximale signifikanzniveau des tests: "Lehne H_0 ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika*, 25:171–178.

Rüschendorf, L. (1982). Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: A proof of the simes conjecture. *The Annals of Statistics*, 26(2):494–504.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71.

Severini, T. A. (2000). *Likelihood methods in statistics.* Oxford University Press.

Shafer, G. and Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3).

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Stevens, W. L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika*, 37(1/2):117–129.

Vovk, V., Wang, B., and Wang, R. (2022). Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

Wang, B. and Wang, R. (2016). Joint mixability. *Mathematics of Operations Research*, 41(3):808–826.

Wang, R. (2024). Testing with p*-values: Between p-values, mid p-values, and e-values. *Bernoulli*, 30(2):1313–1346.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201.

Westfall, P. H., Kropf, S., and Finos, L. (2004). Weighted FWE-controlling methods in high-dimensional situations. *Lecture Notes-Monograph Series*, pages 143–154.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200.

Xu, Z. and Ramdas, A. (2023). More powerful multiple testing under dependence via randomization. *arXiv preprint arXiv:2305.11126.*

Xu, Z., Wang, R., and Ramdas, A. (2024). Post-selection inference for e-value based confidence intervals. *Electronic Journal of Statistics*, 18(1):2292–2338.

A Proofs of the results

A.1 Proof of Section 2

Proof of Lemma 2.3. By definition, the quantity in (2) is non-negative. In addition, for any $\alpha \in (0,1]$,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\alpha}\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\alpha}\right)\right] &= \frac{1}{\alpha}\sum_{k=1}^{K}\lambda_k \mathbb{E}\left[f_k\left(\frac{P_k}{\alpha}\right)\right] = \frac{1}{\alpha}\sum_{k=1}^{K}\lambda_k \int_0^{\alpha} f_k\left(\frac{p}{\alpha}\right)\mathrm{d}p \\ &= \sum_{k=1}^{K}\lambda_k \int_0^{1} f_k\left(p\right)\mathrm{d}p \le 1.\end{aligned}$$

If the calibrators are admissible, one can see that the equality holds since $\int_0^1 f_k(p)\,\mathrm{d}p = 1$ for each k. □

A.2 Proofs of Section 3

Proof of Theorem 3.2. The proof involves the use of the exchangeable Markov inequality (EMI) recalled in Theorem 2.4 for finite sequences:

$$\mathbb{P}\left(\exists k \le K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\alpha}\right) \ge 1\right) \overset{(i)}{\le} \mathbb{E}\left[f\left(\frac{P_1}{\alpha}\right)\right] = \alpha\mathbb{E}\left[\frac{1}{\alpha}f\left(\frac{P_1}{\alpha}\right)\right] \overset{(ii)}{\le} \alpha,$$

where (i) is due to EMI while (ii) holds due to Lemma 2.3. □

Proof of Theorem 3.4. It is clear that F is increasing and Borel since R_α is a lower set. For an exchangeable sequence $\mathbf{P} \in \mathcal{U}^K$ and $\alpha \in (0,1)$, using Theorem 3.2 and the fact that $(R_\beta)_{\beta\in(0,1)}$ is nested, we have

$$\begin{aligned}\mathbb{P}\left(F(\mathbf{P}) \le \alpha\right) &= \mathbb{P}\big(\inf\{\beta \in (0,1) : \mathbf{P} \in R_\beta\} \le \alpha\big) \\ &= \mathbb{P}\left(\inf\left\{\beta \in (0,1) : \exists k \le K \text{ such that } \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right) \ge 1\right\} \le \alpha\right) \\ &= \mathbb{P}\left(\bigcap_{\beta>\alpha}\left\{\exists k \le K : \left(\frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right)\right) \ge 1\right\}\right) \\ &= \inf_{\beta>\alpha}\mathbb{P}\left(\exists k \le K : \left(\frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right)\right) \ge 1\right) \le \inf_{\beta>\alpha}\beta = \alpha.\end{aligned}$$

Therefore F is a valid p-merging function. Homogeneity comes directly from the definition of (3). □

Proof of Proposition 3.5. Let $\mathbf{P} \in \mathcal{U}^K$, and let σ be a random permutation of $\{1,\dots,K\}$, uniformly drawn from all permutations of $\{1,\dots,K\}$ and independent of $\mathbf{P}$. Let $\mathbf{P}^\sigma = (P_{\sigma(1)},\dots,P_{\sigma(K)})$. Note that $\mathbf{P}^\sigma$ is exchangeable by construction. If F is a symmetric ex-p-merging function, it must satisfy $F(\mathbf{P}^\sigma) = F(\mathbf{P})$. Because $F(\mathbf{P}^\sigma)$ is a p-variable, so is $F(\mathbf{P})$, showing that F is a p-merging function. □

Proof of Proposition 3.6. Take $U \in \mathcal{U}$ and an event A with $\mathbb{P}(A) = \alpha$ independent of U. Let $b = f(0+) \le K$. The condition on f guarantees that $f(U)$ is a random variable with support $[0,b]$, mean 1 and a decreasing density. The above conditions, using Theorem 3.2 of Wang and Wang

(2016), guarantee that there exists $\mathbf{U} = (U_1, \dots, U_K) \in \mathcal{U}^K$ such that $\mathbb{P}(\sum_{i=1}^K f(U_i) = K) = 1$. We assume that $U, A, \mathbf{U}$ are mutually independent; this is possible as we are only concerned with distributions. Taking a uniformly drawn random permutation further allows us to assume that $\mathbf{U}$ is exchangeable. Let $P_i = \alpha U_i \mathbb{1}_A + (\alpha + (1-\alpha)U)\mathbb{1}_{A^c}$ for $i = 1, \dots, K$. It is clear that each $P_i \in \mathcal{U}$ and $\mathbf{P} = (P_1, \dots, P_K)$ is exchangeable. Moreover, by the definition of F,

$$\mathbb{P}(F(\mathbf{P}) \le \alpha) \ge \mathbb{P}\left(\frac{1}{K}\sum_{i=1}^K f(P_i/\alpha) \ge 1\right) = \mathbb{P}(A)\mathbb{P}\left(\sum_{i=1}^K f(U_i) = K\right) = \alpha.$$

The other inequality $\mathbb{P}(F(\mathbf{P}) \le \alpha) \le \alpha$ follows from Theorem 3.2. □

Proof of Theorem 3.8. We say that a set $R \subseteq [0,\infty)^K$ is a decreasing set if $\mathbf{x} \in R$ implies $\mathbf{y} \in R$ for all $\mathbf{y} \in [0,\infty)^K$ with $\mathbf{y} \le \mathbf{x}$ (componentwise).

Fix $\alpha \in (0,1)$, and note that $R_\alpha(F) = \{\mathbf{p} \in [0,\infty)^K : F(\mathbf{p}) \le \alpha\}$ is a decreasing set. Define

$$G(\mathbf{p}) = \inf\left\{\beta \in (0,1) : \mathbf{p} \in \frac{\beta}{\alpha} R_\alpha(F)\right\}, \quad \mathbf{p} \in [0,\infty)^K.$$

First, G is homogeneous, which follows from its definition. Second, $\mathbf{p} \in R_\alpha(F)$ implies $G(\mathbf{p}) \le \alpha$, and hence $R_\alpha(F) \subseteq R_\alpha(G)$. Third, G is increasing because $R_\alpha(F)$ is a decreasing set.

It remains to show that G is a p-merging function. The definition of G gives

$$G(\mathbf{p}) \le \beta \iff \mathbf{p} \in \bigcap_{\gamma > \beta} \frac{\gamma}{\alpha} R_\alpha(F).$$

If we can show

$$\mathbb{P}\left(\mathbf{P} \in \frac{\beta}{\alpha} R_\alpha(F)\right) \le \beta \quad \text{for all } \mathbf{P} \in \mathcal{U}^K \text{ and } \beta \in (0,1), \tag{14}$$

then

$$\mathbb{P}(G(\mathbf{P}) \le \beta) \le \inf_{\gamma > \beta} \mathbb{P}\left(\mathbf{P} \in \frac{\gamma}{\alpha} R_\alpha(F)\right) \le \inf_{\gamma > \beta} \gamma = \beta,$$

Lemma A.1. *Let $R \subseteq [0,\infty)^K$ be a decreasing Borel set. For any $\beta \in (0,1)$, we have*

$$\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \beta R) \ge \beta \iff \sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in R) = 1.$$

Proof of the lemma. Let $\mathcal{V}$ be the set of p-variables for $\mathbb{P}$. For any decreasing set L, we have

$$\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in L) = \sup_{\mathbf{P} \in \mathcal{V}^K} \mathbb{P}(\mathbf{P} \in L). \tag{15}$$

This fact will be repeatedly used in the proof below.

We first prove the ⇐ direction by contraposition. Suppose

$$\gamma := \sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \beta R) < \beta.$$

Take an event A with probability β and any $\mathbf{P} \in \mathcal{U}^K$ independent of A. Define $\mathbf{P}^*$ by

$$\mathbf{P}^* = \beta \mathbf{P} \times \mathbb{1}_A + \mathbf{1} \times \mathbb{1}_{A^c}.$$

It is straightforward to check $\mathbf{P}^* \in \mathcal{V}^K$. Hence, by (15),

$$\beta \mathbb{P}(\mathbf{P} \in R) = \mathbb{P}(A)\mathbb{P}(\mathbf{P} \in R) \le \mathbb{P}(\mathbf{P}^* \in \beta R) \le \gamma,$$

and thus $\mathbb{P}(\mathbf{P} \in R) \le \gamma/\beta$. Since $\gamma/\beta < 1$, this yields $\sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}(\mathbf{P} \in R) < 1$ and completes the $\Leftarrow$ direction.

Next we show the $\Rightarrow$ direction. Suppose $\sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \beta R) \ge \beta$. For any $\epsilon \in (0, \beta)$, there exists $\mathbf{P} = (P_1, \dots, P_K) \in \mathcal{U}^K$ such that $\mathbb{P}(\mathbf{P} \in \beta R) > \beta - \epsilon$. Let $A = \{\mathbf{P} \in \beta R\}$, $\gamma = \mathbb{P}(A)$, and B be an event containing A with $\mathbb{P}(B) = \beta \vee \gamma$. Let $\mathbf{P}^* = (P_1^*, \dots, P_K^*)$ follow the conditional distribution of $\mathbf{P}/\beta$ given B. We have

$$\mathbb{P}(\mathbf{P}^* \in R) = \mathbb{P}(\mathbf{P} \in \beta R \mid B) = \mathbb{P}(A \mid B) = \frac{\gamma}{\beta \vee \gamma}.$$

Note that for $k \in \{1, \dots, K\}$,

$$\mathbb{P}(P_k^* \le p) = \mathbb{P}(P_k/\beta \le p \mid B) \le \frac{\mathbb{P}(P_k \le \beta p)}{\mathbb{P}(B)} = \frac{\beta p}{\beta \vee \gamma} \le p,$$

and hence $\mathbf{P}^* \in \mathcal{V}^K$. Since $\gamma > \beta - \epsilon$ and $\epsilon \in (0, \beta)$ is arbitrary, we can conclude $\sup_{\mathbf{P}\in\mathcal{V}^K} \mathbb{P}(\mathbf{P} \in R) = 1$, yielding $\sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}(\mathbf{P} \in R) = 1$ via (15). □

Now we resume the proof of Theorem 3.8. Fix $\beta \in (0, 1)$. Take any $\lambda > 1$ such that $\lambda(\beta \vee \alpha) < 1$. Lemma A.1 yields, for any decreasing set R,

$$\sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \lambda\beta R) < \lambda\beta \iff \sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \lambda\alpha R) < \lambda\alpha. \tag{16}$$

Let $R = R_\alpha(F)/(\alpha\lambda)$, which is a decreasing set. Note that

$$\sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \lambda\alpha R) = \sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}(\mathbf{P} \in R_\alpha(F)) \le \alpha < \lambda\alpha$$

and this leads to, by using (16),

$$\sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}\left(\mathbf{P} \in \frac{\beta}{\alpha} R_\alpha(F)\right) = \sup_{\mathbf{P}\in\mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \lambda\beta R) < \lambda\beta.$$

Since $\lambda > 1$ can be arbitrarily close to 1, we conclude that (14) holds true, and this completes the proof. □

Proof of Lemma 3.9. Fix any $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2) \in [0, 1]^K$ and suppose that $F(\mathbf{p}_1) = \alpha$. By definition, we have

$$\exists k \le K_1 : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \ge 1 \implies \exists k \le K_1 + K_2 : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \ge 1.$$

This implies

$$F(\mathbf{p}) = \inf\left\{\beta \in (0, 1) : \exists k \le K \text{ s.t. } \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\beta}\right) \ge 1\right\} \le \alpha,$$

due to the fact that f is increasing in β. □

Proof of Theorem 3.10. From Lemma 3.9, we have that

$$\exists k \le K : F(\mathbf{p}_k) \le \alpha \iff F(\mathbf{p}_K) \le \alpha,$$

where $\mathbf{p}_k = (p_1, \dots, p_k)$ is the sequence containing the first k values of $\mathbf{p}$. Then we can write,

$$\begin{aligned}
\mathbb{P}\left(\exists k \ge 1 : F(\mathbf{P}_k) \le \alpha\right) &= \mathbb{P}\left(\bigcup_{K\in\mathbb{N}} \{\exists k \le K : F(\mathbf{P}_k) \le \alpha\}\right) \\
&= \lim_{K\to\infty} \mathbb{P}\left(\exists k \le K : F(\mathbf{P}_k) \le \alpha\right) \\
&= \lim_{K\to\infty} \mathbb{P}\left(F(\mathbf{P}_K) \le \alpha\right) \le \alpha,
\end{aligned}$$

where the last inequality is due to Theorem 3.4. □

Proof of Theorem 3.11. From direct calculation and using Lemma 2.3,

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\alpha}\right)\geq U\right) &= \mathbb{E}\left[\mathbb{P}\left(\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\alpha}\right)\geq U\right)\mid \mathbf{P}\right]\\ &= \mathbb{E}\left[\left(\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\alpha}\right)\right)\wedge 1\right]\\ &\leq \mathbb{E}\left[\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\alpha}\right)\right] = \alpha\mathbb{E}\left[\frac{1}{\alpha}\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\alpha}\right)\right]\leq \alpha,\end{aligned}$$

The equality for $\beta = \alpha$ follows because $\sum_{k=1}^{K}\lambda_k f_k(P_k/\alpha) \leq 1$ and $\int_0^1 f_k(p)\mathrm{d}p = 1$ for each k guarantee the inequalities in the above set of equations are equalities. For $\beta < \alpha$, it suffices to notice that $\sum_{k=1}^{K}\lambda_k f_k(P_k/\alpha)$ is increasing in α. □

Proof of Theorem 3.13. It is clear that F is increasing and Borel since R_α is a lower set. For $(\mathbf{P}, U) = (P_1, \dots, P_K, U) \in \mathcal{U}^K \otimes \mathcal{U}$ and $\alpha \in (0, 1)$, using Theorem 3.11 and the fact that $(R_\beta)_{\beta\in(0,1)}$ is nested, we have

$$\begin{aligned}\mathbb{P}(F(\mathbf{P}, U) \leq \alpha) &= \mathbb{P}\left(\inf\left\{\beta\in(0,1):\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\beta}\right)\geq U\right\}\leq\alpha\right)\\ &= \mathbb{P}\left(\bigcap_{\beta>\alpha}\left\{\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\beta}\right)\geq U\right\}\right)\\ &= \inf_{\beta>\alpha}\mathbb{P}\left(\sum_{k=1}^{K}\lambda_k f_k\left(\frac{P_k}{\beta}\right)\geq U\right)\leq\inf_{\beta>\alpha}\beta=\alpha.\end{aligned}$$

Therefore, F is a randomized p-merging function. Homogeneity of F follows from (5).

Since F is homogeneous and increasing, it is continuous in $\mathbf{p}$. Moreover, for fixed $\mathbf{p} \in [0, 1]^K$, since

$$\bigcap_{v<u}\left\{\alpha\in(0,1):\sum_{k=1}^{K}\lambda_k f_k\left(\frac{p_k}{\alpha}\right)\geq v\right\}=\left\{\alpha\in(0,1):\sum_{k=1}^{K}\lambda_k f_k\left(\frac{p_k}{\alpha}\right)\geq u\right\},$$

we have

$$\begin{aligned}\lim_{v\uparrow u}F(\mathbf{p},v) &= \lim_{v\uparrow u}\inf\left\{\alpha\in(0,1):\sum_{k=1}^{K}\lambda_k f_k\left(\frac{p_k}{\alpha}\right)\geq v\right\}\\ &= \inf\bigcap_{v<u}\left\{\alpha\in(0,1):\sum_{k=1}^{K}\lambda_k f_k\left(\frac{p_k}{\alpha}\right)\geq v\right\}=F(\mathbf{p},u).\end{aligned}$$

Therefore, $u \mapsto F(\mathbf{p}, u)$ is lower semi-continuous. □

A.3 Proofs of Section 4

Proof of Theorem 4.1. According to Theorem 3.4, it follows that the function F_{ER} is a valid ex-p-merging function. Fix any $\alpha \in (0, 1)$ and $\mathbf{p} \in (0, 1]^K$. Note that $F_{\mathrm{ER}}(\mathbf{p}) \leq \alpha$ if and only if

$$\exists\ell\leq K:\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{K}{k}\mathbb{1}\left\{\frac{p_i}{\alpha}\leq\frac{k}{K}\right\}\geq 1\implies\exists\ell\leq K:\sum_{i=1}^{\ell}\mathbb{1}\left\{p_i\leq\alpha\frac{k}{K}\right\}\geq\left\lceil\ell\frac{k}{K}\right\rceil.$$

Where rounding up is due to the fact that the summation takes values in positive integers. This holds true if and only if

$$\exists \ell \leq K : p^{\ell}_{(\lambda_\ell)} \leq \alpha \frac{k}{K},$$

where $p^{\ell}_{(\lambda_\ell)}$ is the $\lceil \ell \frac{k}{K} \rceil$-th ordered value of $(p_1, \dots, p_\ell)$. Rearranging the terms, we obtain that it is verified when

$$\frac{K}{k} \bigwedge_{\ell=1}^{K} p^{\ell}_{(\lambda_\ell)} \leq \alpha,$$

which complete the first part of the proof. For the second statement, it is possible to note that the element $p^{\ell}_{(\lambda_\ell)}$ in the sequence coincides with $p_{(k)}$ when $\ell = K$. □

Proof of Theorem 4.2. According to Corollary 3.14, it follows that the function F_{UR} is a valid randomized p-merging function. Fix any $\alpha \in (0,1)$ and $(\mathbf{p}, u) \in (0,1]^{K+1}$, then it is possible to note that $F(\mathbf{p}, u) \leq \alpha$ if and only if

$$\frac{1}{K} \sum_{i=1}^{K} \frac{K}{k} \mathbb{1} \left\{ \frac{p_i}{\alpha} \leq \frac{k}{K} \right\} \geq u \implies \sum_{i=1}^{K} \mathbb{1} \left\{ p_i \leq \alpha \frac{k}{K} \right\} \geq \lceil uk \rceil.$$

Rearranging the terms this holds true only if

$$p_{(\lceil uk \rceil)} \leq \alpha \frac{k}{K} \implies \frac{K}{k} p_{(\lceil uk \rceil)} \leq \alpha,$$

which concludes the claim. Since $u \leq 1$ almost surely, we have $p_{(\lceil uk \rceil)} \leq p_{(k)}$. □

A.4 Proofs of Section 5

Proof of Lemma 5.1. It is simple to see that f is decreasing and it is upper semicontinuous (the function f has discontinuity points in $i/(Kh_K)$, $i = 1, \dots, K$, but it is simple to prove that $\lim_{x \to x_0} f(x) \leq f(x_0)$). In addition,

$$\int_0^1 f(p) \mathrm{d}p = \int_0^1 \frac{K \mathbb{1}\{h_K p \leq 1\}}{\lceil K h_K p \rceil} \mathrm{d}p = \sum_{j=1}^{K} \frac{K}{j} \frac{1}{K h_K} = \frac{1}{h_K} \sum_{j=1}^{K} \frac{1}{j} = 1.$$

Due to Theorem 2.2 we have that F_{Hom} is admissible.

To prove the last part, we see that for any $\mathbf{p} \in (0,1]^K$ and $\alpha \in (0,1)$ we have that $F_{\text{Hom}}(\mathbf{p}) \leq \alpha$, if and only if

$$\bigwedge_{k=1}^{K} \frac{K}{k} p_{(k)} \leq \frac{\alpha}{h_K}.$$

This implies that, for some $m \in \{1, \dots, K\}$, we have that

$$\sum_{j=1}^{K} \mathbb{1} \left\{ \frac{K}{m} h_K p_j \leq \alpha \right\} \geq m.$$

We now define this chain of inequalities,

$$1 \leq \sum_{j=1}^{K} \frac{1}{m} \mathbb{1} \left\{ \frac{K}{m} h_K p_j \leq \alpha \right\} \overset{(i)}{\leq} \sum_{j=1}^{K} \frac{1}{\lceil K h_K p_j / \alpha \rceil} \mathbb{1} \left\{ \frac{K}{m} h_K p_j \leq \alpha \right\} \overset{(ii)}{\leq} \sum_{j=1}^{K} \frac{1}{\lceil K h_K p_j / \alpha \rceil} \mathbb{1} \left\{ h_K p_j \leq \alpha \right\},$$

where (i) is a consequence of $(1/m)\mathbb{1}\{K p_j h_K \leq \alpha m\} \leq (1/\lceil K h_K p_j / \alpha \rceil) \mathbb{1}\{K p_j h_K \leq \alpha m\}$, for all $j = 1, \dots, K$, while (ii) to the fact that $K/m \geq 1$. This concludes the proof. □

Proof of Theorem 5.2. According to Theorem 3.4 and Lemma 5.1, it follows that F_{EHom} is a ex-p-merging function. It is simple to see that $F_{\mathrm{EHom}} \le F_{\mathrm{Hom}}$. □

Proof of Theorem 5.3. According to Corollary 3.14 and Lemma 5.1, it follows that F_{UHom} is a randomized p-merging function. It is simple to see that $F_{\mathrm{UHom}} \le F_{\mathrm{Hom}}$. □

A.5 Proofs of Section 6

Proof of Lemma 6.1. It is easy to see that f is decreasing and $\int_0^1 (2-2p)\mathrm{d}p = 1$. In addition, it is upper semi-continuous and $f(0) = \infty$. □

The statements on domination in Theorems 6.2, 6.3, and other similar results are straightforward from definitions and we omit the proof.

Proof of Theorem 6.2. According to Theorem 3.4 and Lemma 6.1, it follows that the function F_{EA} is an ex-p-merging function. Fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$. We note that $F_{\mathrm{EA}}(\mathbf{p}) \le \alpha$ if and only if, for some $\ell \in \{1,\dots,K\}$, we have

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(2-\frac{2p_i}{\alpha}\right)_+ \ge 1. \tag{17}$$

This implies that there exists $\ell \le K$ such that

$$\frac{1}{\ell}\sum_{j=1}^{m}\left(2-\frac{2p^{\ell}_{(j)}}{\alpha}\right) \ge 1 \quad \text{for some } m \in \{1,\dots,\ell\},$$

where we recall that $p^{\ell}_{(j)}$ is the j-th ordered value of the vector $(p_1,\dots,p_\ell)$. This is due to the fact that the contribution of p_i in the left-hand side of (17) vanishes for large values of p_i. Rearranging the terms, it is possible to obtain that exists $\ell \le K$ such that

$$\frac{2A(\mathbf{p}^{\ell}_{(m)})}{2-\ell/m} \le \alpha \quad \text{for some } m \in \{1,\dots,\ell\}.$$

Taking an infimum over m yields

$$\left(\bigwedge_{m=1}^{\ell}\frac{2A(\mathbf{p}^{\ell}_{(m)})}{(2-\ell/m)_+}\right) \le \alpha \quad \text{for some } \ell \in \{1,\dots,K\}.$$

Actually, the index m can start from $\lceil \ell/2 \rceil$ since the first $\lceil \ell/2 \rceil - 1$ terms in the denominator are smaller than zero. Taking an infimum also over ℓ gives the desired result. □

Proof of Theorem 6.3. According to Corollary 3.14 and Lemma 6.1, it follows that the function F_{UA} is a randomized p-merging function. Fix any $\alpha \in (0,1)$ and $(\mathbf{p},u) \in (0,1]^{K+1}$. Note that $F_{\mathrm{UA}}(\mathbf{p},u) \le \alpha$ if and only if

$$\frac{1}{K}\sum_{k=1}^{m}\left(2-\frac{2p_{(k)}}{\alpha}\right) \ge u \quad \text{for some } m \in \{1,\dots,K\}.$$

Rearranging terms, it is

$$\frac{\sum_{k=1}^{m} 2p_{(k)}}{2m-Ku} \le \alpha \quad \text{for some } m \in \{1,\dots,K\}.$$

Taking an infimum over m yields the desired formula. □

A.6 Proofs of Section 7

Proof of Lemma 7.1. Let $K \geq 2$ and $T_K \geq 1$. Define the function $f : [0, \infty) \to [0, \infty]$ as

$$p \mapsto \min\left\{\frac{1}{T_K p} - \frac{1}{T_K}, K\right\} \mathbb{1}\{p \in [0,1]\},$$

that is decreasing in p. Then,

$$\begin{aligned}
\int_0^1 f(p)\mathrm{d}p &= \int_0^1 \min\left\{\frac{1}{T_K p} - \frac{1}{T_K}, K\right\} \mathbb{1}\{p \in [0,1]\}\mathrm{d}p \\
&= \int_0^{(T_K K+1)^{-1}} K\mathrm{d}p + \int_{(T_K K+1)^{-1}}^1 \left(\frac{1}{T_K p} - \frac{1}{T_K}\right)\mathrm{d}p \\
&= \frac{K}{T_K K + 1} - \frac{K}{T_K K + 1} + \frac{\log(T_K K + 1)}{T_K} = \frac{\log(K T_K + 1)}{T_K}.
\end{aligned}$$

This implies that $\int_0^1 f(p)\mathrm{d}p \leq 1$ if and only if

$$KT_K + 1 - e_K^T \leq 0. \tag{18}$$

We would like to choose T_K as small as possible. One possible candidate is $T_K = \log K + \log\log K + 1$. Indeed, plugging $T_K = \log K + \log\log K + 1$ into the left-hand side of (18) we find that it is verified when

$$\log\log K + 1 + \frac{1}{K} \leq (e-1)\log K,$$

and this holds if $K \geq 2$ by checking $K = 2, 3$ and using the derivative of both sides for $K \geq 4$. □

Proof of Proposition 7.2. It is straightforward to see that F is an increasing function. By direct calculation,

$$\begin{aligned}
\mathbb{P}(F(\mathbf{P}) \leq \alpha) &= \mathbb{P}((T_K + 1)H(\mathbf{P}) \leq \alpha) \\
&= \mathbb{P}\left((T_K+1)K\left(\sum_{k=1}^K \frac{1}{P_k}\right)^{-1} \leq \alpha\right) \\
&= \mathbb{P}\left(\frac{1}{K}\sum_{k=1}^K\left(\frac{\alpha}{T_K P_k} - \frac{1}{T_K}\right) \geq 1\right) \\
&\overset{(i)}{\leq} \mathbb{P}\left(\frac{1}{K}\sum_{k=1}^K\left(\frac{1}{T_K P_k/\alpha} - \frac{1}{T_K}\right)\mathbb{1}\left\{\frac{P_k}{\alpha} \in [0,1]\right\} \geq 1\right) \\
&= \mathbb{P}\left(\frac{1}{K}\sum_{k=1}^K \min\left\{\left(\frac{1}{T_K P_k/\alpha} - \frac{1}{T_K}\right), K\right\}\mathbb{1}\left\{\frac{P_k}{\alpha} \in [0,1]\right\} \geq 1\right) \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \min\left\{\left(\frac{1}{T_K P_k/\alpha} - \frac{1}{T_K}\right), K\right\}\mathbb{1}\left\{\frac{P_k}{\alpha} \in [0,1]\right\}\right] \\
&= \alpha\mathbb{E}\left[\frac{1}{\alpha}\frac{1}{K}\sum_{k=1}^K \min\left\{\left(\frac{1}{T_K P_k/\alpha} - \frac{1}{T_K}\right), K\right\}\mathbb{1}\left\{\frac{P_k}{\alpha} \in [0,1]\right\}\right] \leq \alpha,
\end{aligned}$$

where (i) is due to the fact that $(1/(T_K x) - 1/T_K)$ is negative for $x > 1$, and (ii) holds due to Markov's inequality. The last inequality is a consequence of Lemma 2.3 and Lemma 7.1. □

Proof of Theorem 7.3. According to Theorem 3.4 and Lemma 7.1, it follows that F_{EH} is an ex-p-merging function. Fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$. We note that $F_{\mathrm{EH}}(\mathbf{p}) \le \alpha$ if and only if, for some $\ell \in \{1,\ldots,K\}$, we have

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{\alpha}{T_K p_i} - \frac{1}{T_K}\right)_+ \ge 1.$$

This implies that exists $\ell \le K$ such that

$$\frac{1}{\ell}\sum_{j=1}^{m}\left(\frac{\alpha}{T_K p^{\ell}_{(j)}} - \frac{1}{T_K}\right) \ge 1 \quad \text{for some } m \in \{1,\ldots,\ell\},$$

where we recall that $p^{\ell}_{(j)}$ is the j-th ordered value of the vector $(p_1,\ldots,p_\ell)$. Rearranging the terms it is possible to obtain that exist $\ell \le K$ such that

$$\left(\frac{\ell T_K}{m} + 1\right) H(\mathbf{p}^{\ell}_{(m)}) \le \alpha \quad \text{for some } m \in \{1,\ldots,\ell\},$$

Taking an infimum over m, we get

$$\bigwedge_{m=1}^{\ell}\left(\frac{\ell T_K}{m} + 1\right) H(\mathbf{p}^{\ell}_{(m)}) \le \alpha \quad \text{for some } \ell \in \{1,\ldots,K\}.$$

Taking an infimum over ℓ yields the desired result. □

Proof of Thereom 7.4. According to Corollary 3.14 and Lemma 7.1, it follows that F_{UH} is a randomized p-merging function. Fix any $\alpha \in (0,1)$ and $(\mathbf{p},u) \in (0,1]^{K+1}$. Then $F_{\mathrm{UH}}(\mathbf{p},u) \le \alpha$ if and only if

$$\frac{1}{K}\sum_{k=1}^{m}\left(\frac{\alpha}{T_K p_{(k)}} - \frac{1}{T_K}\right) \ge u \quad \text{for some } m \in \{1,\ldots,K\}.$$

Rearranging the terms, it is

$$(uKT_K + m)\left(\sum_{k=1}^{m}\frac{1}{p_{(k)}}\right)^{-1} \le \alpha \quad \text{for some } m \in \{1,\ldots,K\}.$$

Taking a minimum over m yields the desired formula. □

A.7 Proofs of Section 8

Proof of Theorem 8.1. According to Theorem 3.4, it follows that the function F_{EG} is an ex-p-merging function. Fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$. Then $F_{\mathrm{EG}}(\mathbf{p}) \le \alpha$, if and only if exists $\ell \in \{1,\ldots,K\}$ such that

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(-\log p_i + \log\alpha)_+ \ge 1.$$

This is verified when exists $\ell \le K$ such that

$$\frac{1}{\ell}\sum_{j=1}^{m}(-\log p^{\ell}_{(j)} + \log\alpha) \ge 1 \quad \text{for some } m \in \{1,\ldots,\ell\},$$

where we recall that $p^{\ell}_{(j)}$ is the j-th ordered value of the vector $(p_1,\ldots,p_\ell)$. Rearranging the terms it is possible to obtain that exists $\ell \le K$ such that

$$e^{\ell/m} G(\mathbf{p}^{\ell}_{(m)}) \le \alpha \quad \text{for some } m \in \{1,\ldots,\ell\}.$$

Taking an infimum over m, we get

$$\bigwedge_{m=1}^{\ell}\left(e^{\ell/m}G(\mathbf{p}^{\ell}_{(m)})\right) \le \alpha \quad \text{for some } \ell \in \{1,\dots,K\}.$$

Taking an infimum over ℓ yields the desired result. □

Proof of Theorem 8.2. According to Corollary 3.14, it follows that F_{UG} is a randomized p-merging function. Fix any $\alpha \in (0,1)$ and $(\mathbf{p},u) \in (0,1]^{K+1}$. We have that $F_{\mathrm{UG}}(\mathbf{p},u) \le \alpha$ if and only if

$$\frac{1}{K}\sum_{k=1}^{m}(-\log p_{(k)} + \log\alpha) \ge u \quad \text{for some } m \in \{1,\dots,K\}.$$

Rearranging the terms, it is

$$e^{u\frac{K}{m}}G(\mathbf{p}_{(m)}) \le \alpha \quad \text{for some } m \in \{1,\dots,K\}.$$

Taking a minimum over m yields the desired formula. □

B Exchangeable and randomized p-merging function

In this part, we will integrate the results in Section 3, using both exchangeability and randomization. In fact, starting from exchangeable p-values, it is possible to prove that if randomization is allowed, then it is possible to improve some of the results obtained in Section 3. We start by defining a "randomized" version of Theorem 3.2.

Theorem B.1. *Let f be a calibrator, and $(\mathbf{P},U) = (P_1,\dots,P_K,U) \in \mathcal{U}^K \otimes \mathcal{U}$ such that $\mathbf{P}$ is exchangeable. For each $\alpha \in (0,1)$, we have*

$$\mathbb{P}\left(f\left(\frac{P_1}{\alpha}\right) \ge U \text{ or } \exists k \le K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\alpha}\right) \ge 1\right) \le \alpha.$$

Proof. The proof invokes the exchangeable and uniformly-randomized Markov inequality (EUMI) introduced in Ramdas and Manole (2024); see Theorem 2.6. In particular,

$$\mathbb{P}\left(f\left(\frac{P_1}{\alpha}\right) \ge U \text{ or } \exists k \le K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\alpha}\right) \ge 1\right) \overset{(i)}{\le} \mathbb{E}\left[f\left(\frac{P_1}{\alpha}\right)\right] = \alpha\mathbb{E}\left[\frac{1}{\alpha}f\left(\frac{P_1}{\alpha}\right)\right] \overset{(ii)}{\le} \alpha,$$

where (i) is due to EUMI and (ii) is due to Lemma 2.3. □

Similarly to how it was done in the preceding sections, let us now define a randomized ex-p-merging function.

Definition B.2. A randomized ex-p-merging function is an increasing Borel function $F : [0,1]^{K+1} \to [0,1]$ such that $\mathbb{P}(F(\mathbf{P},U) \le \alpha) \le \alpha$ for all $\alpha \in (0,1)$ and $(\mathbf{P},U) \in \mathcal{U}^K \otimes \mathcal{U}$ with $\mathbf{P}$ exchangeable. It is homogeneous if $F(\gamma\mathbf{p},u) = \gamma F(\mathbf{p},u)$ for all $\gamma \in (0,1]$ and $(\mathbf{p},u) \in [0,1]^{K+1}$. A randomized ex-p-merging function is admissible if for any randomized ex-p-merging function G, $G \le F$ implies $G = F$.

Let f be a calibrator; then for $\alpha \in (0,1)$, we define the exchangeable and randomized rejection region by

$$R_\alpha = \left\{(\mathbf{p},u) \in [0,1]^{K+1} : f\left(\frac{p_1}{\alpha}\right) \ge u \text{ or } \exists k \le K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \ge 1\right\},$$

where we set $f(p_i/u) = 0$ if $u = 0$. Starting from R_α, we can define the function $F : [0,1]^{K+1} \to [0,1]$ by

$$\begin{aligned} F(\mathbf{p}, u) &= \inf\left\{\alpha \in (0,1) : (\mathbf{p}, u) \in R_\alpha\right\} \\ &= \inf\left\{\alpha \in (0,1) : f\left(\frac{p_1}{\alpha}\right) \geq u \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1\right\}, \end{aligned} \tag{19}$$

with the convention $\inf \varnothing = 1$ and $0 \times \infty = \infty$.

Theorem B.3. *If f is a calibrator and $(\mathbf{P}, U) \in \mathcal{U}^K \otimes \mathcal{U}$ with $\mathbf{P}$ exchangeable, then F in* (19) *is a homogeneous randomized ex-p-merging function.*

Proof. It is clear that F is increasing and Borel since R_α is a lower set. For an exchangeable $\mathbf{P} \in \mathcal{U}^K$ and $\alpha \in (0,1)$, using Theorem B.1 and the fact that $(R_\beta)_\beta \in (0,1)$ is nested, we have

$$\begin{aligned} \mathbb{P}(F(\mathbf{P}, U) \leq \alpha) &= \mathbb{P}\left(\inf\left\{\beta \in (0,1) : f\left(\frac{P_1}{\beta}\right) \geq U \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right) \geq 1\right\} \leq \alpha\right) \\ &= \mathbb{P}\left(\bigcap_{\beta > \alpha}\left\{f\left(\frac{P_1}{\beta}\right) \geq U \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right) \geq 1\right\}\right) \\ &= \inf_{\beta > \alpha} \mathbb{P}\left(f\left(\frac{P_1}{\beta}\right) \geq U \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right) \geq 1\right) \\ &\leq \inf_{\beta > \alpha} \beta = \alpha. \end{aligned}$$

therefore F is a valid randomized ex-p-merging function. Homogeneity comes directly from the definition of (19). □

C Generalized Hommel combination rule

We can generalize the Hommel combination by allowing the selection of certain quantiles from the possible K different quantiles of $\mathbf{p} = (p_1, \ldots, p_K)$, for example, one can select the minimum between K times the minimum, 2 times the median and the maximum. This can be obtained by noting that the Hommel combination rule can be rewritten in terms of quantiles. In particular, the following holds:

$$F'_{\text{Hom}}(\mathbf{p}) = h_K \bigwedge_{k=1}^{K} \frac{1}{\lambda_k} p_{(\lceil \lambda_k K \rceil)}, \qquad \text{with } h_K = \sum_{j=1}^{K} \frac{\lambda_j - \lambda_{j-1}}{\lambda_j},$$

where $(\lambda_0, \lambda_1, \ldots, \lambda_K)$ is the vector of quantiles such that $\lambda_j = j/K$, $j = 0, 1, \ldots, K$. This gives the intuition to define a generalization of the aforementioned rule.

Let us define the vector of quantiles $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_M)$, such that $\lambda_0 = 0$, $\lambda_j \in (0,1]$, if $j = 1, \ldots, M$ and $\lambda_j < \lambda_{j+1}$. Then, we can define

$$F'_{\text{GHom}}(\mathbf{p}) := h_M \bigwedge_{k=1}^{M} \frac{1}{\lambda_k} p_{(\lceil \lambda_k K \rceil)}, \qquad \text{with } h_M = \sum_{j=1}^{M} \frac{\lambda_j - \lambda_{j-1}}{\lambda_j}. \tag{20}$$

Lemma C.1. *Let f be a function defined by*

$$f(p) = \sum_{j=1}^{M} \frac{1}{\lambda_j} \mathbb{1}\left\{p \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M}\right]\right\} + \infty \mathbb{1}\{p = 0\}.$$

Then f is an admissible calibrator. Moreover, the p-merging function induced by f is

$$F_{\text{GHom}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{M}\frac{1}{\lambda_j}\mathbb{1}\left\{\frac{p_k}{\alpha} \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M}\right]\right\} \geq 1\right\},$$

that is valid and dominates (20)*.*

Proof. It is simple to see that f is decreasing and upper semicontinuous. In addition,

$$\int_0^1 f(p)\mathrm{d}p = \sum_{j=1}^{M}\frac{1}{\lambda_j}\left(\frac{\lambda_j - \lambda_{j-1}}{h_M}\right) = \frac{1}{h_M}\sum_{j=1}^{M}\frac{\lambda_j - \lambda_{j-1}}{\lambda_j} = 1.$$

This implies that F_{GHom} is admissible. To prove the last part, we see that for any $\mathbf{p} \in (0,1]^K$ and $\alpha \in (0,1)$ we have that $F_{\text{GHom}}(\mathbf{p}) \leq \alpha$, if and only if,

$$\bigwedge_{k=1}^{M}\frac{1}{\lambda_k}p_{(\lceil\lambda_k K\rceil)} \leq \frac{\alpha}{h_M}.$$

This implies that, for some $m \in \{1,\ldots,K\}$, we have that

$$\sum_{i=1}^{K}\frac{1}{K}\mathbb{1}\left\{p_i \leq \alpha\frac{\lambda_m}{h_M}\right\} \geq \lambda_m.$$

We now define this chain of inequalities,

$$\begin{aligned}1 &\leq \sum_{i=1}^{K}\frac{1}{K\lambda_m}\mathbb{1}\left\{p_i \leq \alpha\frac{\lambda_m}{h_M}\right\} = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{\lambda_m}\sum_{j=1}^{m}\mathbb{1}\left\{\frac{p_i}{\alpha} \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M}\right]\right\}\\ &\leq \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{m}\frac{1}{\lambda_j}\mathbb{1}\left\{\frac{p_i}{\alpha} \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M}\right]\right\} \leq \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{M}\frac{1}{\lambda_j}\mathbb{1}\left\{\frac{p_i}{\alpha} \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M}\right]\right\}.\end{aligned}$$

□

It is possible to see that (20) is a special case of the Rüger combination when $M = 1$, while it coincides with the Hommel combination rule when $\lambda = (0, 1/K, \ldots, (K-1)/K, 1)$.

D Improving generalized mean

In this section, we discuss the generalized mean combination rule, for $r \in \mathbb{R}\setminus\{0\}$. This combination rule, introduced in Vovk and Wang (2020), is quite broad and contains some important cases well known in the literature. In particular, if $r = 1$ it reduces to the sample average (Section 6), while if $r = -1$ it coincides with the harmonic mean described in Section 7. We introduce a lemma characterizing the calibrator used in the context of the generalized mean combination rule.

Lemma D.1. *Let $r \in \mathbb{R}\setminus\{0\}$ and $f : [0,\infty) \to [0,\infty]$ be given by*

$$f(p) = \min\left\{\frac{r(1-p^r)}{T_{r,K}}, K\right\}\mathbb{1}\{p \in [0,1]\}, \tag{21}$$

where $T_{r,K} > 0$ is any constant, possibly dependent on K, such that $\int_0^1 f(p)\mathrm{d}p \leq 1$. Then f is a calibrator.

Proof. Since $f(p)$ is continuously decreasing in $T_{r,K}$, it can be verified that there exists $T_{r,K} > 0$ such that $\int_0^1 f(p)\mathrm{d}p = 1$. Moreover, f is decreasing and non-negative which completes the claim. □

It is simple to see that for $r > 0$, we have that $T_{r,K} = r^2/(r+1)$ satisfies $\int_0^1 f(p)\mathrm{d}p \leq 1$. From previous sections, we obtain $T_{1,K} = 1/2$ while a more complex result appears for $T_{-1,K}$. We now see how the calibrator defined in (21) is related to the rule defined in Section 3. First, we define the function $F'_{\mathrm{M_r}}$ as

$$F'_{\mathrm{M_r}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K} \frac{r(1-(p_k/\alpha)^r)}{T_{r,K}} \geq 1\right\} = \frac{M_{\mathrm{r}}(\mathbf{p})}{(1-T_{r,K}/r)^{1/r}}, \tag{22}$$

where $M_{\mathrm{r}}(\mathbf{p})$ is the r-generalized mean of $\mathbf{p}$, defined by $M_{\mathrm{r}}(\mathbf{p}) = (\sum_{k=1}^{K} p_k^r/K)^{1/r}$. In particular, $F'_{\mathrm{M_r}}(\mathbf{p})$ coincides with the generalized mean combination rule where $a_{r,K} = (1-T_{r,K}/r)^{-1/r}$. In addition, if $r > 0$ then $F'_{\mathrm{M_r}}(\mathbf{p}) = (r+1)^{1/r} M_r(\mathbf{p})$ which coincides with the asymptotically precise merging function studied by Vovk and Wang (2020). It is possible to prove that $F'_{\mathrm{M_r}}(\mathbf{p}) \leq F_{\mathrm{M_r}}(\mathbf{p})$, where

$$F_{\mathrm{M_r}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K} \left(\frac{r(1-(p_k/\alpha)^r)}{T_{r,K}}\right)_+ \geq 1\right\}, \tag{23}$$

which is the p-merging function induced by the calibrator defined in (21). In particular, according to Theorem 2.2 we have that $F_{\mathrm{M_r}}(\mathbf{p})$ is a p-merging function.

D.1 Exchangeable generalized mean

We now define the function $F_{\mathrm{EM_r}}$ in the following way:

$$F_{\mathrm{EM_r}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell} \left(\frac{r(1-(p_i/\alpha)^r)}{T_{r,K}}\right)_+ \geq 1\right\}. \tag{24}$$

If we define $F'_{\mathrm{EM_r}}$ by

$$F'_{\mathrm{EM_r}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell} \left(\frac{r(1-(p_i/\alpha)^r)}{T_{r,K}}\right) \geq 1\right\}, \tag{25}$$

then it holds that $F_{\mathrm{EM_r}} \leq F'_{\mathrm{EM_r}}$.

Theorem D.2. *Let $r \in \mathbb{R} \setminus \{0\}$, then the function $F'_{\mathrm{EM_r}}$ defined in (25) equals*

$$\left(1 - \frac{T_{r,K}}{r}\right)^{-1/r} \left(\bigwedge_{m=1}^{K} M_r(\mathbf{p}_m)\right),$$

is an ex-p-merging function, and it strictly dominates the function $F_{\mathrm{M_r}}$ in (23). However, it is strictly dominated by the function $F_{\mathrm{EM_r}}$ defined in (24) that is also an ex-p-merging function.

Proof. According to Theorem 3.4 and Lemma D.1, it follows that the function $F_{\mathrm{EM_r}}$ is an ex-p-merging function. In addition, fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$, then $F'_{\mathrm{EM_r}} \leq \alpha$ if and only if

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \frac{r(1-(p_i/\alpha)^r)}{T_{r,K}} \geq 1 \text{ for some } \ell \leq K \implies \left(1 - \frac{T_{r,K}}{r}\right)^{-1/r} \left(\frac{1}{\ell}\sum_{i=1}^{\ell} p_i^r\right)^{1/r} \leq \alpha \text{ for some } \ell \leq K.$$

Taking a minimum over ℓ yields the desired formula. □

D.2 Randomized generalized mean

According to the the previous sections, we define the randomized p-merging function $F_{\mathrm{UM_r}}$ as follows

$$F_{\mathrm{UM_r}}(\mathbf{p}, u) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(\frac{r(1-(p_k/\alpha)^r)}{T_{r,K}}\right)_+ \geq u\right\}. \tag{26}$$

The function $F_{\mathrm{UM_r}} \leq F'_{\mathrm{UM_r}}$ where the function $F'_{\mathrm{UM_r}}$ is defined by

$$F'_{\mathrm{UM_r}}(\mathbf{p}, u) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(\frac{r(1-(p_k/\alpha)^r)}{T_{r,K}}\right) \geq u\right\}, \tag{27}$$

and can be considered as the randomized version of (22).

Theorem D.3. *Let $r \in \mathbb{R}\setminus\{0\}$, then the function $F'_{\mathrm{EM_r}}$ defined in* (27) *equals*

$$\frac{M_r(\mathbf{p})}{(1-uT_{r,K}/r)^{1/r}},$$

is a randomized p-merging function, and it strictly dominates the function $F_{\mathrm{M_r}}$ in (23)*. However, it is strictly dominated by the function $F_{\mathrm{UM_r}}$ defined in* (26) *that is also a randomized p-merging function.*

Proof. According to Corollary 3.14 and Lemma D.1, it follows that $F_{\mathrm{UM_r}}$ is a valid randomized p-merging function. In addition, fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$, then $F'_{\mathrm{UM_r}} \leq \alpha$ if and only if

$$\frac{1}{K}\sum_{k=1}^{K}\left(\frac{r(1-(p_k/\alpha)^r)}{T_{r,K}}\right) \geq u \implies \frac{M_r(\mathbf{p})}{(1-uT_{r,K}/r)^{1/r}} \leq \alpha.$$

□

E General algorithm

One general algorithm to compute the ex-p-merging function defined in (3) induced by a calibrator f is proposed in Algorithm 1. The algorithm employs the bisection method and it consistently yields a p-value exceeding that of the induced ex-p-merging function by at most 2^{-B}.

Algorithm 1 Ex-p-merging function

```
Require: A calibrator $f$, $B \in \mathbb{N}$, and a sequence of p-values $(p_1, \dots, p_K)$
  $L := 0$ and $U := 1$
  for $m = 1, \dots, B$ do
    $\alpha := (L+U)/2$
    if $\bigvee_{k=1}^{K}\left(\frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right)\right) \geq 1$ then
      $U := \alpha$
    else
      $L := \alpha$
    end if
  end for
Ensure: $U$
```

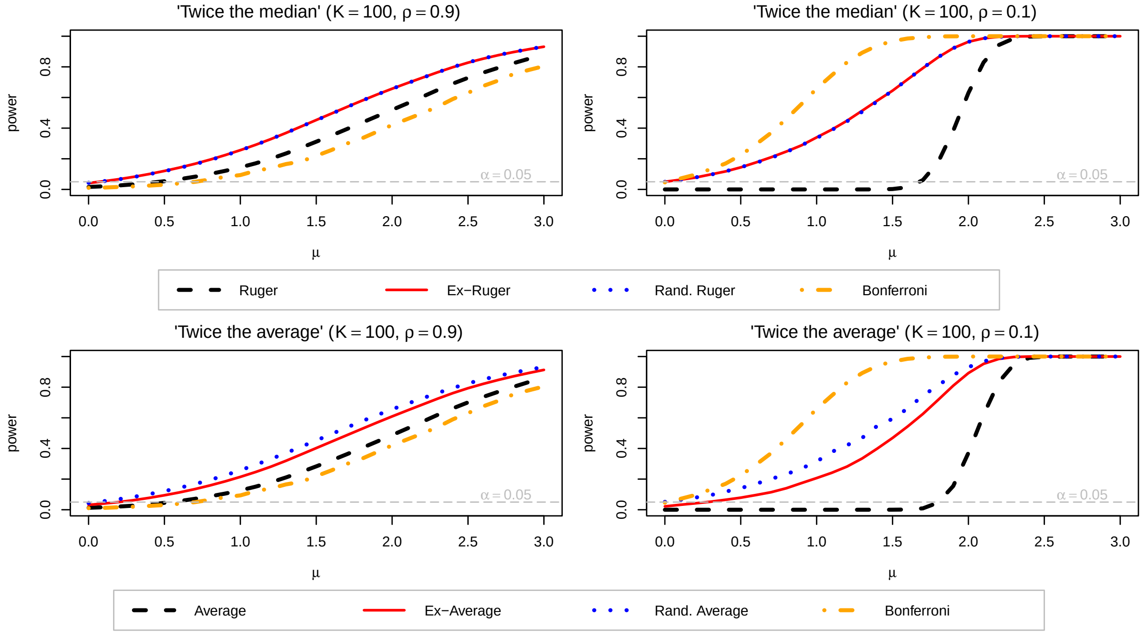

Figure 3: Combination of p-values using different rules. Every subplot illustrates power against μ. The left endpoint of $\mu = 0$ actually represents the empirical type I error, which is controlled at the nominal level $\alpha = 0.05$ for all methods proposed. The first column has $\rho = 0.9$, while the second column has $\rho = 0.1$ — as expected, the Bonferroni correction is more powerful near independence, but is less powerful under strong dependence. Further, our exchangeable and randomized improvements offer sizeable increases in power over the original variants in all settings.

F Additional simulation results

In this section we report some additional simulation results. We first compare the merging rules introduced in Section 4 and in Section 6, specifically the rules: "twice the median" and "twice the average". In particular, $K = 100$ p-values are generated as in (13), and two different values of ρ are chosen, 0.9 and 0.1, respectively. The parameter k for the randomized Rüger combination rule is set to $K/2$ and μ varies in the interval $[0, 3]$. The results are computed by averaging the outcomes obtained in $10,000$ replications.

In Figure 3, we can see that the type I error is controlled at the nominal level 0.05 for all proposed methods. In the case of the Rüger combination, in both dependence scenarios, the power of the combinations obtained by exploiting exchangeability or employing external randomization is highly similar. In the case of the combination based on the arithmetic mean, it appears that the rules obtained using randomization exhibit greater power than those based on exchangeability (and, naturally, than the original rules). In general, across all observed scenarios, the new rules demonstrate a quite significant improvement in terms of power.

In addition, we compare all the randomized combination rules reported in the paper. We omit the randomized functions that are dominated by other randomized p-merging functions. As before, $K = 100$ p-values are generated as in (13), and $\rho = \{0.1, 0.9\}$. The parameter k for the randomized Rüger combination rule is set to $K/2$ and μ varies in the interval $[0, 3]$. The results are obtained by repeating the procedure $10,000$ times and reporting the average.

The results of the randomized versions of "twice" the median and (improved) "twice" the average

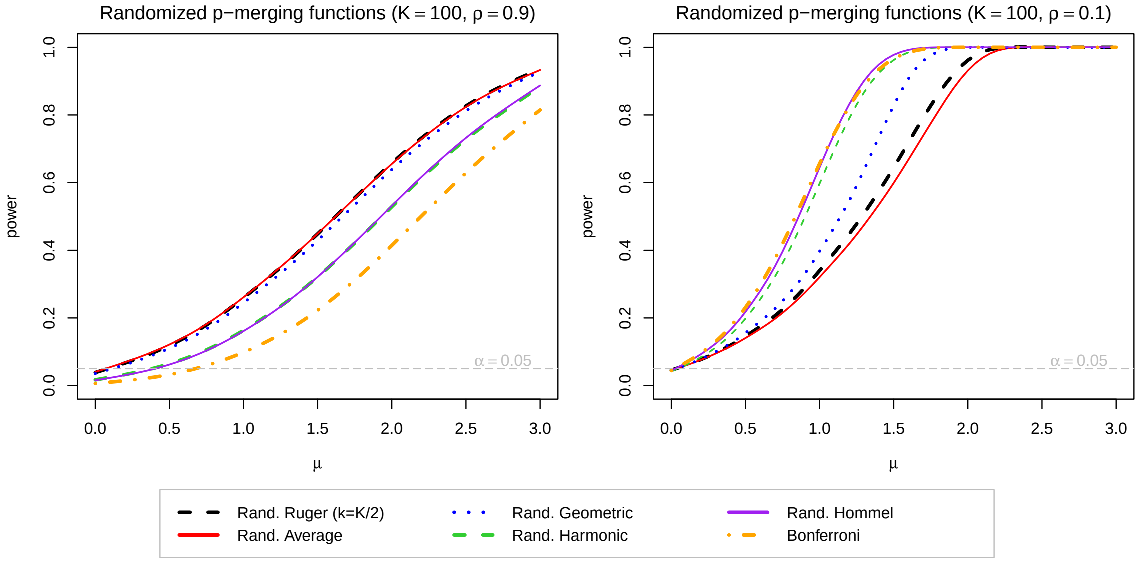

Figure 4: Combination of p-values using different randomized p-merging functions. The order of the performance of the different ex-p-merging functions is almost the opposite in the two situations.

are similar in both scenario. In addition, F_{UHom} and F_{UH} are quite similar in terms of power (Figure 4). From Figure 4, we can observe an opposite behavior of the functions in the case where $\rho = 0.9$ or $\rho = 0.1$. In general, the most powerful functions in the left graph tend to be the least powerful in the right graph, and vice versa.

At the end, we examine the 3 different randomized combination rules defined in Section 6, respectively, $F_{\mathrm{UA}}, F'_{\mathrm{UA}}, F^*_{\mathrm{UA}}$. In particular, we recall that, F_{UA} dominates F'_{UA}, while the combination F^*_{UA} (introduced in Wang (2024, Appendix B.2)) neither dominates nor is dominated by either of the two.

The set of $K = 100$ p-values is generated as in (13), and two different values of ρ are chosen, 0.9 and 0.1, respectively. The results are obtained by repeating the procedure $10,000$ times and reporting the average. From Figure 5, we can see that the power F_{UA} is always higher than the power of the other two combination rules. The function F^*_{UA} is more powerful than F'_{UA} only in the first part of the y-axis in both cases (for $\mu \leq 2.5$, indicatively), so there is no clear preference between the two. As expected, the Bonferroni correction is more powerful near independence ($\rho = 0.1$).

Before concluding, let us see what our procedures give for testing a global null. The goal of this last part is to explore a case where the order of the p-values in our proposed ex-p-merging functions is chosen in a data-driven manner. In the considered scenario, we will see how a particular statistics can be chosen to order the p-values with the aim of improving the statistical power under the alternative. In particular, this data-driven ordering does not alter the exchangeability under the null that is required in our ex-p-merging functions. Specifically, we investigate the issue of performing simultaneous one-sample t-tests using n observations for each hypothesis. Let us suppose to have K samples (one for each hypothesis) from a normal distribution,

$$X_{ki} \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad i = 1, \dots, n, \quad \mu_k \in \mathbb{R}, \quad \sigma_k > 0,$$

where the observations X_{ki}, $k \in \{1, \dots, K\}$, $i \in \{1, \dots, n\}$, are mutually independent. Our goal is to test the hypothesis $H_k : \mu_k = 0$ and to do so we use t-test. In addition, we define the global null as $H_0 : \bigcap_{k=1}^{K} H_k$, that can be tested by merging the different p-values into a single p-value. The

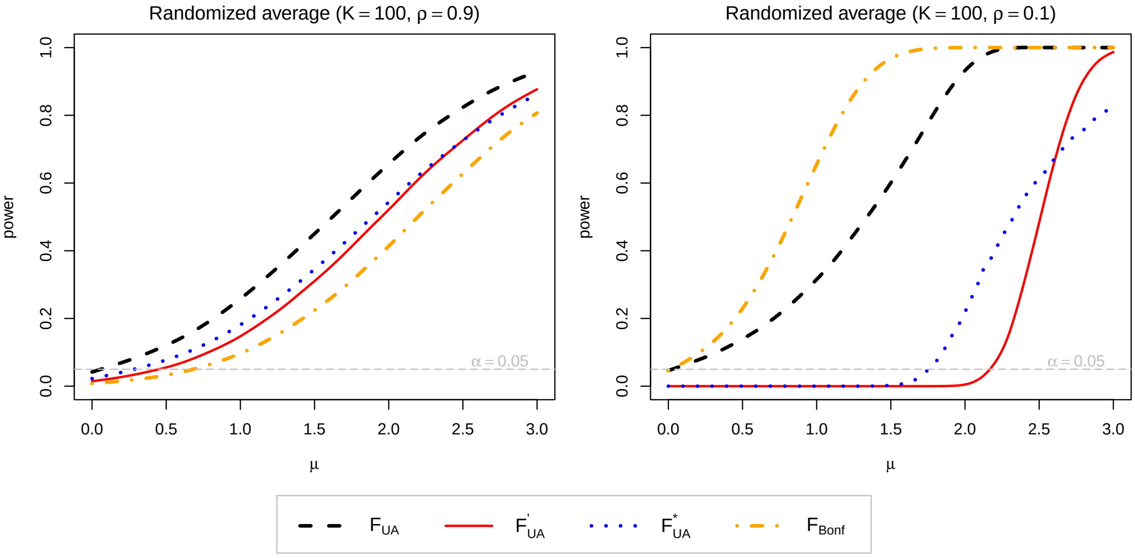

Figure 5: Combination of p-values using different randomized combination rules based on the average of p-values (and the Bonferroni method, for comparison). F^*_{UA} is more powerful than F'_{UA} only when $\mu \lesssim 2.5$.

same problem has been studied, for example, in Westfall et al. (2004) and Ignatiadis et al. (2024). Let us define

$$\bar{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ki}, \quad \hat{\sigma}_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ki} - \bar{X}_k)^2, \quad T_k = \frac{\sqrt{n}\bar{X}_k}{\hat{\sigma}_k^2},$$

and p-values are $P_k := 2G_{n-1}(-|T_k|)$, where G_{n-1} is the t-distribution with $n-1$ degrees of freedom.

Our p-values will be ordered using the statistics $S_k^2 = \sum_{i=1}^{n} X_{ki}^2$, and this can be done since the proposed statistics are independent from T_k (and so P_k). The key observation is that, under H_0 (i.e., when $\mu_k = 0$ for all $k = 1, \ldots, K$), then the statistic S_k^2 is sufficient and complete for the inference on the parameter σ_k^2. On the other hand, the test statistic T_k is constant in distribution with respect to σ_k^2. By Basu's Theorem (Basu, 1955), S_k^2 and T_k are independent and, in addition, the test T_k and the statistics S_k^2 are independent among themselves since they are functions of independent random variables.

We assume $\sigma_k = \sigma = 1$ for all $k = 1, \ldots, K$, and $\mu_k = k \cdot \mu$, where μ is a parameter that varies in $[0, 0.2]$. The parameter α is set to 0.05, while $K = 20$ and $n = 10$. As can be seen in Figure 6, in all situations the type I error is controlled at the level α under H_0 (i.e., when $\mu = 0$). If p-values are ordered in decreasing order with respect to the statistics S_k^2 then we have a rule that has a power comparable (or higher) than the Bonferroni rule. It is expected that the descending order is more effective, as a large S_k^2 indicates that H_k is likely false. The Fisher combination, which requires the strong assumption of independence, has the largest power.

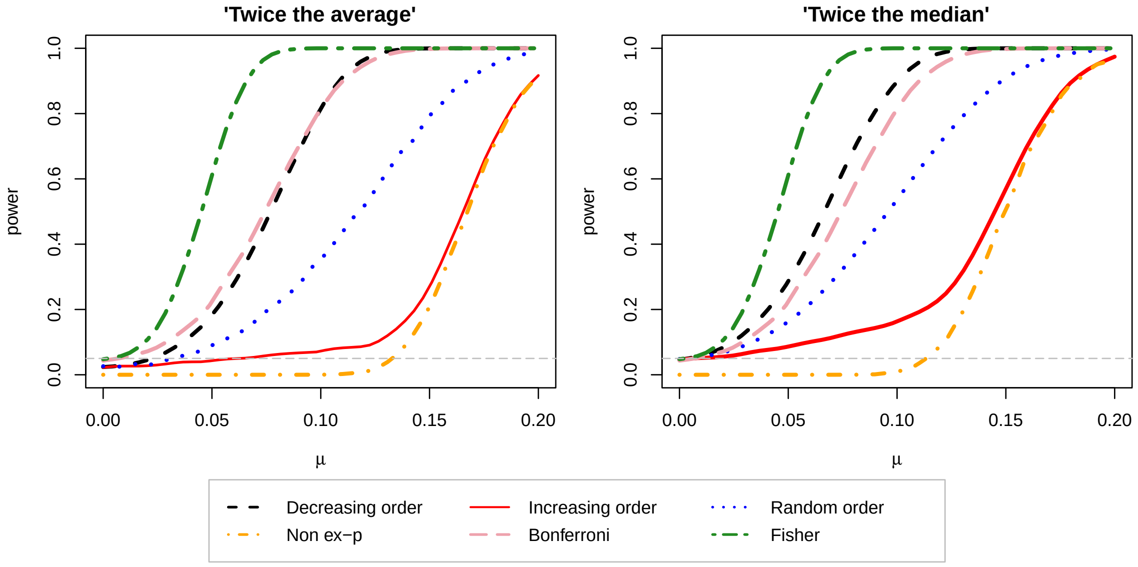

Figure 6: Combination of p-values for testing a global null. If p-values are ordered in decreasing order with respect to S_k^2 we have a power that is comparable with the Bonferroni rule. However, in all situations Fisher's combination is the most powerful.