Combining p-values via averaging

Vladimir Vovk and Ruodu Wang

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #21

First posted 20 December 2012 (on arXiv) and 21 November 2017
(as Working Paper). Last revised April 25, 2018.

Project web site:
http://alrw.net

Abstract

This paper proposes general methods for the problem of multiple testing of a single hypothesis, with a standard goal of combining a number of p-values without making any assumptions about their dependence structure. An old result by Rüschendorf and, independently, Meng implies that the p-values can be combined by scaling up their arithmetic mean by a factor of 2 (and no smaller factor is sufficient in general). Based on more recent developments in mathematical finance, specifically, robust risk aggregation techniques, we show that K p-values can be combined by scaling up their geometric mean by a factor of e (for all K) and by scaling up their harmonic mean by a factor of $\ln K$ (asymptotically as $K \to \infty$). These and other results lead to a generalized version of the Bonferroni–Holm method. A simulation study compares the performance of various averaging methods.

Contents

1 Introduction

Suppose we are testing the same hypothesis using $K \geq 2$ different statistical tests and obtain p-values $p_1, \dots, p_K$. How can we combine them into a single p-value?

One of the earliest papers answering this question was Fisher's [8]. However, Fisher's paper assumes that the p-values are independent (in practice, obtained from independent test statistics), whereas we would like to avoid any assumptions besides all $p_1, \dots, p_K$ being bona fide p-values. Fisher's method has been extended to dependent p-values in, e.g., [5, 19], but the combined p-values obtained in those papers are approximate; in this paper we are interested only in precise or conservative p-values.

Without assuming any particular dependence structure among p-values, the simplest way of combining them is the Bonferroni method:

$$F(p_1, \dots, p_K) := K \min(p_1, \dots, p_K) \tag{1}$$

(when $F(p_1, \dots, p_K)$ exceeds 1 it can be replaced by 1, but we usually ignore this trivial step). Albeit $F(p_1, \dots, p_K)$ is a p-value (see Section 2 for a precise definition of a *p-value*), it has been argued that in some cases it is overly conservative. Rüger [29] extends the Bonferroni method by showing that, for any fixed $k \in \{1, \dots, K\}$,

$$F(p_1, \dots, p_K) := \frac{K}{k} p_{(k)} \tag{2}$$

is a p-value, where $p_{(k)}$ is the kth smallest p-value among $p_1, \dots, p_K$; see [27] for a simpler exposition. Hommel [14] develops this by showing that

$$F(p_1, \dots, p_K) := \left(1 + \frac{1}{2} + \dots + \frac{1}{K}\right) \min_{k=1,\dots,K} \frac{K}{k} p_{(k)} \tag{3}$$

is also a p-value. Simes [32] improves (3) by removing the first factor on the right-hand side of (3), but he assumes the independence of $p_1, \dots, p_K$.

To us, intuitively, the most natural way to combine K p-values is to average them, by using $\bar{p} := (p_1 + \dots + p_K)/K$. Unfortunately, $\bar{p}$ is not necessarily a p-value. An old result by Rüschendorf [30, Theorem 1] shows that $2\bar{p}$ is a p-value; moreover, the factor of 2 cannot be improved in general. In the statistical literature this result was rediscovered by Meng [26, Theorem 1].

In this paper (see Section 3) we move on to a generalized notion of the mean as axiomatized by Andrei Kolmogorov [18] and adapt various results of *robust risk aggregation* [6, 2, 7, 36, 17] to combining p-values by averaging them in Kolmogorov's wider sense. In particular, to obtain a p-value from given p-values $p_1, \dots, p_K$, it is sufficient to multiply their geometric mean by e (as noticed by Mattner [24]) and to multiply their harmonic mean by $e \ln K$ (for $K > 2$). More generally, we consider the mean $M_{r,K}(p_1, \dots, p_K)$ defined by $((p_1^r + \dots + p_K^r)/K)^{1/r}$ for $r \in [-\infty, \infty]$; in particular, our results cover the Bonferroni method (1), which corresponds to $M_{-\infty,K}(p_1, \dots, p_K) = K \min(p_1, \dots, p_K)$ (see, e.g., [12, (2.3.1)]).

Median is also sometimes regarded as a kind of average. Rüger's (2), applied to $k := \lceil K/2 \rceil$, says that p-values can be combined by scaling up their median by a factor of 2 (exactly for even K and approximately for large odd K). Therefore, we have the same factor of 2 as in Rüschendorf's [30] result. (Taking $k = \lfloor (K+1)/2 \rfloor = \lceil K/2 \rceil$ is suggested in [23, Section 1.1].) More generally, the α quantile $p_{(\lceil \alpha K \rceil)}$ becomes a p-value if multiplied by $1/\alpha$.

It is often possible to automatically transform results about multiple testing of a single hypothesis into results about testing multiple hypotheses; the standard procedures are Marcus et al.'s [22] closed testing procedure and its modification by Hommel [15]. In particular, when applied to the Bonferroni method the closed testing procedure gives the well-known method due to Holm [13], which we will refer to as the Bonferroni–Holm method; see, e.g., [15, 16] for its further applications. In Section 4 we briefly discuss a similar application to one of the procedures of Section 3.

Section 5 discusses weighted averaging, and Section 6 contains a simulation study which compares various methods of averaging. Section 7 concludes the paper. As this paper targets a statistical audience, in the main text, we shall omit proofs and techniques based on results from the literature of robust risk aggregation; the details can be found in the Appendix.

Combining p-values by arithmetic averaging is used in the method of cross-conformal prediction [33, (11)], as discussed in detail in Appendix B.

Some notation and terminology

If E is a property of elements of a set X, $\mathbf{1}_E : X \to [0, \infty)$ is the indicator function of E: $\mathbf{1}_E(x) = 1$ if x satisfies E and $\mathbf{1}_E(x) = 0$ if not. A function $F : [0,1] \to [0, \infty)$ is *increasing* (resp. *decreasing*) if $F(x_1) \le F(x_2)$ (resp. $F(x_1) \ge F(x_2)$) whenever $x_1 \le x_2$. A function $F : [0,1]^K \to [0, \infty)$ is *increasing* (resp. *decreasing*) if it is increasing (resp. decreasing) in each of its arguments. A function is *strictly increasing* or *strictly decreasing* when these conditions hold with strict inequalities.

2 Merging functions

A *p-value function* is a random variable P that satisfies

$$\mathbb{P}(P \le \epsilon) \le \epsilon, \quad \forall \epsilon \in (0,1). \tag{4}$$

The values taken by a p-value function are *p-values* (allowed to be conservative). In Section 1 the expression "p-value" was loosely used to refer to p-value functions as well. A *merging function* is an increasing Borel function $F : [0,1]^K \to [0, \infty)$ such that $F(U_1, \dots, U_K)$ is a p-value function for any choices of random variables $U_1, \dots, U_K$ (on the same probability space, which can be arbitrary) distributed uniformly on $[0,1]$. Without loss of generality we can assume that $U_1, \dots, U_K$ are defined on the same atomless probability space, which is fixed throughout the paper (cf. [9, Proposition A.27]). Let $\mathcal{U}$ be the set

of all uniformly distributed random variables (on our probability space). Using the notation $\mathcal{U}$, an increasing Borel function $F : [0,1]^K \to [0,\infty)$ is a merging function if, for each $\epsilon \in (0,1)$,

$$\sup\{\mathbb{P}(F(U_1,\dots,U_K) \le \epsilon) \,|\, U_1,\dots,U_K \in \mathcal{U}\} \le \epsilon \tag{5}$$

We say that a merging function F is *precise* if, for each $\epsilon \in (0,1)$,

$$\sup\{\mathbb{P}(F(U_1,\dots,U_K) \le \epsilon) \,|\, U_1,\dots,U_K \in \mathcal{U}\} = \epsilon. \tag{6}$$

Remark. The requirement that a merging function be Borel does not follow automatically from the requirement that it be increasing: see the remark after Theorem 4.4 in [11] (Theorem 4.4 itself says that every increasing function on $[0,1]^K$ is Lebesgue measurable).

It may be practically relevant to notice that, for any merging function F, $F(P_1,\dots,P_K)$ is a p-value function whenever $P_1,\dots,P_K$ are p-value functions (on the same probability space). Indeed, for each $k \in \{1,\dots,K\}$ we can define a uniformly distributed random variable $U_k \le P_k$ by

$$U_k(\omega) := \mathbb{P}(P_k < P_k(\omega)) + \theta\,\mathbb{P}(P_k = P_k(\omega)), \;\; \omega \in \Omega,$$

where θ is a random variable distributed uniformly on $[0,1]$ and independent of $P_1,\dots,P_K$, and Ω is the underlying probability space extended (if required) to carry such a θ; we then have

$$\mathbb{P}(F(P_1,\dots,P_K) \le \epsilon) \le \mathbb{P}(F(U_1,\dots,U_K) \le \epsilon) \le \epsilon, \quad \forall \epsilon \in (0,1).$$

Therefore, the procedure of merging can be carried out in multiple layers (although it may make the resulting p-value overly conservative).

3 Combining p-values by symmetric averaging

In this section we present our methods of combining p-values via averaging. A general notion of averaging, axiomatized by Kolmogorov [18], is

$$M_{\phi,K}(p_1,\dots,p_K) := \psi\left(\frac{\phi(p_1)+\dots+\phi(p_K)}{K}\right), \tag{7}$$

where $\phi : [0,1] \to [-\infty,\infty]$ is a continuous strictly monotonic function and ψ is its inverse (with the domain $\phi([0,1])$). For example, arithmetic mean corresponds to the identity function $\phi(p) = p$, geometric mean corresponds to $\phi(p) = \ln p$, and harmonic mean corresponds to $\phi(p) = 1/p$.

The two general results of this section, Theorems 1 and 7, will be consequences of known results in the field of mathematical finance dealing with robust risk aggregation. In Appendix A, with a few auxiliary results, we establish a connection between the p-value merging problem and recent results on robust risk aggregation, where one also finds the proofs of the vast majority of other results of this paper.

Theorem 1. *Suppose a continuous strictly monotonic $\phi : [0,1] \to [-\infty, \infty]$ is integrable, i.e., $\int_0^1 |\phi(u)| \, \mathrm{d}u < \infty$. Then, for any $K \in \{2, 3, \ldots\}$ and any $\epsilon > 0$,*

$$\mathbb{P}\left(M_{\phi,K}(p_1, \ldots, p_K) \le \psi\left(\frac{1}{\epsilon} \int_0^\epsilon \phi(u) \mathrm{d}u \right) \right) \le \epsilon. \tag{8}$$

As we stated it, Theorem 1 gives a critical region of size ϵ. An alternative statement is that $\Psi^{-1}(M_{\phi,K})$ is a merging function, where the strictly increasing function Ψ is defined by

$$\Psi(\epsilon) := \psi\left(\frac{1}{\epsilon} \int_0^\epsilon \phi(u) \mathrm{d}u \right), \quad \epsilon \in (0,1). \tag{9}$$

We focus on the most important special case of (7), namely,

$$M_{r,K}(p_1, \ldots, p_K) := \left(\frac{p_1^r + \cdots + p_K^r}{K} \right)^{1/r}, \tag{10}$$

where $r \in \mathbb{R} \setminus \{0\}$ and the following standard conventions are used: $0^c := \infty$ for $c < 0$, $0^c := 0$ for $c > 0$, $\infty + c := \infty$ for $c \in \mathbb{R} \cup \{\infty\}$, and $\infty^c := 0$ for $c < 0$. The case $r = 0$ (considered in [24]) is treated separately (as the limit as $r \to 0$):

$$M_{0,K}(p_1, \ldots, p_K) := \exp\left(\frac{\ln p_1 + \cdots + \ln p_K}{K} \right) = \left(\prod_{k=1}^K p_k \right)^{1/K},$$

where, as usual, $\ln 0 := -\infty$, $-\infty + c := -\infty$ for $c \in \mathbb{R} \cup \{-\infty\}$, and $\exp(-\infty) := 0$. It is also natural to set

$$M_{\infty,K}(p_1, \ldots, p_K) := \max(p_1, \ldots, p_K),$$
$$M_{-\infty,K}(p_1, \ldots, p_K) := \min(p_1, \ldots, p_K).$$

The most important special cases of $M_{r,K}$ are perhaps those corresponding to $r = -\infty$ (minimum), $r = -1$ (harmonic mean), $r = 0$ (geometric mean), $r = 1$ (arithmetic mean), and $r = \infty$ (maximum); the cases $r \in \{-1, 0, 1\}$ are known as Platonic means.

The main results of this section are summarized in Table 1, where a family F_K, $K = 2, 3, \ldots$, of merging functions on $[0,1]^K$ is called *asymptotically precise* if, for any $a \in (0,1)$, aF_K is not a merging function for a large enough K; in other words, this family of merging functions cannot be improved by a constant multiplier. It is well known [12, Theorem 16] that $M_{r_1,K} \le M_{r_2,K}$ on $[0,1]^K$ if $r_1 \le r_2$; therefore, the constant in front of the precise merging functions should be decreasing in r.

We start by presenting results for $r > -1$. The following corollary is a direct consequence of Theorem 1.

Corollary 2. *Let $r \in (-1, \infty]$. Then $(r+1)^{1/r} M_{r,K}$ is a merging function.*

Table 1: The main results of Section 3: examples of merging functions, all of them precise or asymptotically precise (except for the case $r = -1$ where the asymptotic formula is not a merging function for finite K); the column "claimed in" contains the number(s) of the relevant proposition and/or corollary (if not obvious)

range of r	merging function		special case	claimed in
$r = \infty$	$M_{r,K}$	precise	maximum	
$r \in [K-1, \infty)$	$K^{1/r} M_{r,K}$	precise		2 and 5
$r \in [\frac{1}{K-1}, K-1]$	$(r+1)^{1/r} M_{r,K}$	precise	arithmetic	2 and 4
$r \in (-1, \infty]$	$(r+1)^{1/r} M_{r,K}$	asymptotically precise		2 and 3
$r = 0$	$e M_{r,K}$	asymptotically precise	geometric	2 and 3
$r = -1$	$e(\ln K) M_{r,K}$	$K > 2$; not precise	harmonic	10
	$(\ln K) M_{r,K}$	(asymptotic formula)		11
$r \in (-\infty, -1)$	$\frac{r}{r+1} K^{1+1/r} M_{r,K}$	asymptotically precise		8 and 9
$r = -\infty$	$K M_{r,K}$	precise	Bonferroni	

The expression $(r+1)^{1/r}$ is understood to be $e = \lim_{r \to 0}(r+1)^{1/r}$ when $r = 0$ and $1 = \lim_{r \to \infty}(r+1)^{1/r}$ when $r = \infty$.

Proof. Evaluating the term (9) in (8), we obtain:

$$\psi\left(\frac{1}{\epsilon}\int_0^\epsilon \phi(u)\mathrm{d}u\right) = \begin{cases} \epsilon/e & \text{if } r = 0 \\ (r+1)^{-1/r}\epsilon & \text{otherwise.} \end{cases} \quad \square$$

Next we show that the constant $(r+1)^{1/r}$ in Corollary 2 cannot be improved in general.

Proposition 3. *Let $r \in (-1, \infty)$. The family of merging functions $(r+1)^{1/r} M_{r,K}$, $K \in \{2, 3 \ldots\}$, is asymptotically precise.*

In particular, Proposition 3 implies that the constant factor e for the geometric mean cannot be improved in general for large K.

Proposition 4. *For $r \in (-1, \infty)$ and $K \in \{2, 3, \ldots\}$, the merging function $M := (r+1)^{1/r} M_{r,K}$ is precise if and only if $r \in [\frac{1}{K-1}, K-1]$.*

The most straightforward yet relevant example of Proposition 4, the arithmetic average multiplied by 2, namely,

$$M_{1,K}(p_1, \ldots, p_K) := \frac{2}{K}\sum_{k=1}^{K} p_k,$$

is a precise merging function for all $K \geq 2$. As another special case of Proposition 4, the scaled quadratic average multiplied by $\sqrt{3}$, namely $\sqrt{3}M_{2,K}$, is a merging function, and it is precise if and only if $K \geq 3$.

In the case $r \geq 1$, the merging function in Proposition 4 can be modified in such a way that it remains precise even for $r > K - 1$:

Proposition 5. *For $K \in \{2, 3, \ldots\}$ and $r \in [1, \infty)$,*

$$\min(r + 1, K)^{1/r} M_{r,K}$$

is a precise merging function.

Because of the importance of geometric mean as one of the Platonic means, the following result gives a precise (albeit somewhat implicit) expression for the corresponding precise merging function.

Proposition 6. *For each $K \in \{2, 3, \ldots\}$, $a_K M_{0,K}$ is a precise merging function, where*

$$a_K := \frac{1}{c_K} \exp\left(-(K-1)(1 - Kc_K)\right)$$

and c_K is the unique solution to the equation

$$\ln(1/c - (K-1)) = K - K^2 c \tag{11}$$

over $c \in (0, 1/K)$.

Proposition 3 already suggests that $a_K \to e$ as $K \to \infty$. Table 2 reports several values of a_K/e numerically calculated in R. It suggests that in practice there is no point in improving the factor e for $K \geq 5$.

Table 2: Numeric values of a_K/e for geometric mean

K	a_K/e	K	a_K/e	K	a_K/e
2	0.7357589	5	0.9925858	10	0.9999545
3	0.9286392	6	0.9974005	15	0.9999997
4	0.9779033	7	0.9990669	20	1.0000000

The condition $r > -1$ in Corollary 2 ensures that the term (9) is finite, and also that the condition $\int_0^1 |\phi(u)| \, \mathrm{d}u < \infty$ in Theorem 1 is satisfied. However, the condition rules out the harmonic mean (for which $r = -1$) and the minimum ($r = -\infty$). The next simple corollary of another known result will cover those cases as well.

Theorem 7. *Suppose $\phi : [0, 1] \to [-\infty, \infty]$ is a strictly decreasing continuous function satisfying $\phi(0) = \infty$. Then, for any $\epsilon \in (0, 1)$ such that $\phi(\epsilon) \geq 0$,*

$$\mathbb{P}\left(M_{\phi,K}(p_1, \ldots, p_K) \leq \epsilon\right) \leq \inf_{t \in (0, \phi(\epsilon)]} \frac{\int_{\phi(\epsilon)-t}^{\phi(\epsilon)+(K-1)t} \psi(u) \mathrm{d}u}{t}. \tag{12}$$

As $t \to 0$, the upper bound in (12) is not informative since, for $t \approx 0$,

$$\frac{\int_{\phi(\epsilon)-t}^{\phi(\epsilon)+(K-1)t} \psi(u)\mathrm{d}u}{t} \approx \frac{Kt\psi(\phi(\epsilon))}{t} = K\epsilon,$$

which is dominated by the Bonferroni bound. On the other hand, the upper bound is informative when $t = \phi(\epsilon)$ provided the integral is convergent. For example, we have the following corollary.

Corollary 8. *For $r < -1$, $\frac{r}{r+1}K^{1+1/r}M_{r,K}$ is a merging function.*

Proof. By Theorem 7 applied to $\phi(u) := u^r$, $r < -1$, we have:

$$\mathbb{P}\left(M_{\phi,K}(p_1,\dots,p_K) \le \epsilon\right) \le \frac{\int_0^{K\phi(\epsilon)} \psi(u)\mathrm{d}u}{\phi(\epsilon)} = \frac{r}{r+1}K^{1+1/r}\epsilon. \qquad \square$$

Corollary 8 includes the Bonferroni bound (1) as special case: for $r := -\infty$, we obtain that $KM_{-\infty,K}$ is a merging function.

Proposition 9. *Let $r \in (-\infty, -1)$. The family of merging functions $\frac{r}{r+1}K^{1+1/r}M_{r,K}$ is asymptotically precise.*

Corollary 8 does not cover the case $r = -1$ of harmonic mean directly, but easily implies a bound for this case that turns out to be not so crude, as we will see later.

Corollary 10. *For $K > 2$, $(e \ln K)M_{-1,K}$ is a merging function.*

Proof. Let us find the smallest value of the coefficient $\frac{r}{r+1}K^{1+1/r}$ in Corollary 8 and Proposition 9. Setting the derivative in r of the logarithm of this coefficient to 0, we obtain a linear equation whose solution is

$$r = \frac{\ln K}{1 - \ln K}. \tag{13}$$

Plugging this into the coefficient gives $e \ln K$. It remains to notice that r defined by (13) satisfies $r < -1$ and apply the inequality $M_{r,K} \le M_{-1,K}$ [12, Theorem 16]. $\square$

Tighter, albeit more complicated, versions of Corollary 10 can be derived from other known results in robust risk aggregation.

Proposition 11. *Set $a_K := \frac{(y_K+K)^2}{(y_K+1)K}$, $K > 2$, where y_K is the unique solution to the equation*

$$y^2 = K((y+1)\ln(y+1) - y), \qquad y \in (0,\infty).$$

Then $a_K M_{-1,K}$ is a precise merging function. Moreover, $a_K / \ln K \to 1$ as $K \to \infty$.

Table 3: Numeric values of a_K for the harmonic mean

K	a_K	K	a_K	K	a_K
3	2.499192	10	1.980287	100	1.619631
4	2.321831	20	1.828861	200	1.561359
5	2.214749	50	1.693497	400	1.514096

Even though $a_K/\ln K \to 1$, the rate of convergence is very slow, and $a_K/\ln K > 1$ for moderate values of K. In practice, it might be better to use the conservative merging function $(e \ln K)M_{-1,K}$ of Corollary 10. Table 3 reports several values of a_K numerically calculated in R. For instance, for $K \geq 10$, one may use $(2 \ln K)M_{-1,K}$, and for $K \geq 50$, one may use $(1.7 \ln K)M_{-1,K}$.

The main emphasis of this section has been on characterizing $a > 0$ such that $F := aM_{r,K}$ is a merging function, or a precise merging function. Recall that $F : [0,1]^K \to [0,\infty)$ is a merging function if and only if (5) holds for all $\epsilon \in (0,1)$, and that F is a precise merging function if and only if (6) holds for all $\epsilon \in (0,1)$. The next proposition shows that in both statements "for all" can be replaced by "for some" if $F = aM_{r,K}$. A practical implication is that even if an applied statistician is interested in the property of validity (4) only for specific values of ϵ (such as 0.01 or 0.05) and would like to use $aM_{r,K}$ as a merging function, she is still forced to ensure that (4) folds for all ϵ.

Proposition 12. *For any $a > 0$, $r \in [-\infty, \infty]$, and $K \in \{2, 3, \ldots\}$:*

(a) $F := aM_{r,K}$ is a merging function if and only if (5) holds for some $\epsilon \in (0,1)$;

(b) $F := aM_{r,K}$ is a precise merging function if and only if (6) holds for some $\epsilon \in (0,1)$.

4 Application to testing multiple hypotheses

In this section we apply the results of the previous section, concerning multiple testing of a single hypothesis, to testing multiple hypotheses. Namely, we will arrive at a generalization of the Bonferroni–Holm method [13]. Fix a parameter

$$r \leq \frac{\ln K}{1 - \ln K} \tag{14}$$

(cf. (13)); the Bonferroni–Holm case is $r := -\infty$.

Suppose p_k is a p-value for testing a composite null hypothesis H_k (meaning that, for any $\epsilon \in (0,1)$, $\mathbb{P}(p_k \leq \epsilon) \leq \epsilon$ under H_k). For $I \subseteq \{1, \ldots, K\}$, let H_I be the hypothesis

$$H_I := (\cap_{k \in I} H_k) \cap \left(\cap_{k \in \{1,\ldots,K\} \setminus I} H_k^c\right),$$

Algorithm 1 Generalized Bonferroni–Holm procedure

Require: A significance level $\epsilon > 0$ and parameter $r < -1$ (or, w.l.o.g., (14)).
Require: A sequence of p-values $p_1, \dots, p_K$ ordered as $p_{k_1} \le \dots \le p_{k_K}$.
  for $k = 1, \dots, K$ **do**
    `reject` := `true`
    $I := \{k\}$
    for $i = K, \dots, 1, 0$ **do**
      if $\frac{r}{r+1} |I|^{1+1/r} M_{r,|I|}(p_I) > \epsilon$ **then**
        `reject` := `false`
      end if
      $I := I \cup \{k_i\}$
    end for
    if `reject` = `true` **then**
      reject H_k
    end if
  end for

where H_k^c is the complement of H_k.

Fix a significance level ϵ. Let us reject H_I when

$$\frac{r}{r+1} |I|^{1+1/r} M_{r,|I|}(p_I) \le \epsilon,$$

where p_I is the vector of p_k for $k \in I$; by Corollary 8, the probability of error will be at most ϵ. If we now reject H_k when all H_I with $I \supseteq \{k\}$ are rejected, the family-wise error rate (FWER) will be at most ϵ. This gives the procedure described as Algorithm 1, in which $(k_1, \dots, k_K)$ stands for a fixed permutation of $\{1, \dots, K\}$ such that $p_{k_1} \le \dots \le p_{k_K}$.

An alternative representation of the generalized Bonferroni–Holm procedure given as Algorithm 1 is in terms of adjusting the p-values $p_1, \dots, p_K$ to new p-values $p_1^*, \dots, p_K^*$ that are valid in the sense of the FWER: we are guaranteed to have $\mathbb{P}(\min_{k \in I} p_k^* \le \epsilon) \le \epsilon$ for all $\epsilon \in (0,1)$, where I is the set of the indices k of the true hypotheses H_k. The adjusted p-values can be defined as

$$p_k^* := \max_{k \in I \subseteq \{1,\dots,K\}} \frac{r}{r+1} |I|^{1+1/r} M_{r,|I|}(p_I)$$

and computed using Algorithm 2.

If we do not insist on controlling the FWER, we can still use our ways of merging p-values instead of Bonferroni's in more flexible procedures for testing multiple hypotheses, such as those described in [10].

5 Combining p-values by weighted averaging

In this section we will briefly consider a more general notion of averaging:

$$M_{\phi,\mathbf{w}}(p_1, \dots, p_K) := \psi\left(w_1 \phi(p_1) + \dots + w_K \phi(p_K)\right)$$

Algorithm 2 Generalized Bonferroni–Holm procedure for adjusting p-values

Require: A parameter $r < -1$ (or, w.l.o.g., (14)).
Require: A sequence of p-values $p_1, \dots, p_K$ ordered as $p_{k_1} \le \dots \le p_{k_K}$.
  for $k = 1, \dots, K$ **do**
    $p_k^* := 0$
    $I := \{k\}$
    for $i = K, \dots, 1, 0$ **do**
      if $\frac{r}{r+1} |I|^{1+1/r} M_{r,|I|}(p_I) > p_k^*$ **then**
        $p_k^* := \frac{r}{r+1} |I|^{1+1/r} M_{r,|I|}(p_I)$
      end if
      $I := I \cup \{k_i\}$
    end for
  end for

in the notation of (7), where $\mathbf{w} = (w_1, \dots, w_K) \in \Delta_K$ is an element of the standard K-simplex

$$\Delta_K := \left\{ (w_1, \dots, w_K) \in [0,1]^K \,\middle|\, w_1 + \dots + w_K = 1 \right\}.$$

One might want to use a weighted average in a situation where some of p-values are based, e.g., on bigger experiments, and then we might want to take them with bigger weights. Intuitively, the weights reflect the prior importance of the p-values (see, e.g., [4, p. 5] for further details).

Much fewer mathematical results in the literature are available for asymmetric risk aggregation. For this reason, we will concentrate on the easier integrable case, namely, $r > -1$. Theorem 1 can be generalized as follows (proofs of all results in this section are given in Appendix A).

Theorem 1w. *Suppose a continuous strictly monotonic $\phi : [0,1] \to [-\infty, \infty]$ is integrable and $\mathbf{w} \in \Delta_K$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(M_{\phi,\mathbf{w}}(p_1, \dots, p_K) \le \psi\left(\frac{1}{\epsilon} \int_0^\epsilon \phi(u) \mathrm{d}u \right) \right) \le \epsilon.$$

Similarly to (10), we set

$$M_{r,\mathbf{w}}(p_1, \dots, p_K) := (w_1 p_1^r + \dots + w_K p_K^r)^{1/r}$$

for $r \in \mathbb{R}$ and $\mathbf{w} = (w_1, \dots, w_K) \in \Delta_K$. We can see that Corollary 2 still holds when $M_{r,K}$ is replaced by $M_{r,\mathbf{w}}$, for any $r \in \mathbb{R}$ and $\mathbf{w} \in \Delta_K$. This is complemented by the following proposition.

Proposition 4w. *For $\mathbf{w} = (w_1, \dots, w_K) \in \Delta_K$ and $r \in (-1, \infty)$, the merging function $(r+1)^{1/r} M_{r,\mathbf{w}}$ is precise if and only if $w \le 1/2$ and $r \in [\frac{w}{1-w}, \frac{1-w}{w}]$, where $w := \max_{k=1,\dots,K} w_k$.*

Next we generalize Proposition 5 to non-uniform weights.

Proposition 5w. *For* $\mathbf{w} = (w_1, \ldots, w_K) \in \Delta_K$ *and* $r \in [1, \infty)$*, the function* $\min(r+1, \frac{1}{w})^{1/r} M_{r,\mathbf{w}}$ *is a precise merging function, where* $w := \max_{k=1,\ldots,K} w_k$*.*

An interesting special case of Proposition 5w is for $r = 1$ (weighted arithmetic mean). If $w \leq 1/2$, i.e., no single experiment outweighs the total of all the other experiments, the optimal multiplier for the weighted average is 2, exactly as in the case of the arithmetic average. If $w > 1/2$, i.e., there is a single experiment that outweighs all the other experiments, our merging function is, assuming $w_1 = w$,

$$\frac{1}{w} M_{1,\mathbf{w}}(p_1, \ldots, p_K) = p_1 + \sum_{k=2}^{K} \frac{w_k}{w} p_k.$$

It is obtained by adding weighted adjustments to the p-value obtained from the most important experiment.

6 Simulation study

The merging functions $M_{r,K}$ for $r \in [-\infty, \infty]$ provide different ways of combining p-values via averaging. It is in general unclear which way of merging is more efficient in a particular situation. In this section we present a simplistic simulation study to compare different ways of averaging.

We assume a normal population of variance 1, and test whether a given set of data has zero mean. There are K tests conducted based on different samples which are not necessarily independent. Each sample in a test contains n iid normal observations with mean $d/\sqrt{n}$ and known variance 1. We merge the p-values obtained from a two-sided z-test on each sample. In addition to the number of tests K, the sample size n, and the deviation d, we specify the following parameters:

1. o is the percentage of overlap of observations between any two tests, between 0 and 1; $o = 1$ corresponds to identical tests.

2. c is the correlation coefficient of non-overlapping observations among tests (not within a test); $c = 1$ also corresponds to identical tests. More precisely, the kth sample is $\{X_{k,1}, \ldots, X_{k,n}\}$, and we assume $X_{k_1,i} = X_{k_2,i}$ for $i = 1, \ldots, \lfloor on \rfloor$, and $\operatorname{corr}(X_{k_i,j}, X_{k_2,j}) = c$ for $k_1 \neq k_2$ and $j = \lfloor on \rfloor + 1, \ldots, n$. All other correlation coefficients are zero.

3. N is the number of trials of the procedure.

In Figures 1 and 2, we report the percentage of rejecting the null hypothesis (the sample has mean zero) at the 5% significance level for each of the following averaging methods: Bonferroni ($r = -\infty$), harmonic ($r = -1$), geometric ($r = 0$), arithmetic ($r = 1$), quadratic ($r = 2$), and maximum ($r = \infty$), under several representative choices of parameters. The percentage, interpreted as the power of the combined test, is represented as the height of the corresponding column.

The sample size n does not affect our results since we standardized the deviation by dividing it by $\sqrt{n}$.

For a practical implementation, the constant in front of the geometric mean is chosen as e, the constant in front of the harmonic mean is chosen as $2\ln K$, which is valid for $K \geq 10$, and for all other methods we use precise merging functions.

The average combined p-values (i.e., the averages of the realizations of $M_{r,K}(p_1, \ldots, p_K)$) for each method across the N trials are also reported. The sizes (i.e., the probabilities of rejecting the null hypothesis when it is true) of our combined tests are not reported as the merging methods are all conservative. A bigger power (i.e., taller column) or a smaller average combined p-value indicates a more efficient merging method.

From Figures 1 and 2, we make the following observations.

1. The Bonferroni method performs very well in cases where tests are independent or moderately overlapping and correlated.

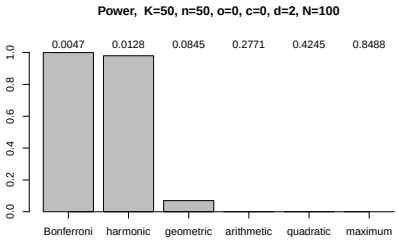
2. When tests are heavily overlapping and correlated, the arithmetic and quadratic methods perform well. The arithmetic method has a relatively stable average p-value across all examples in Figure 2 ($d = 3$ in all examples there), since it is simply twice the average p-value of the individual tests.

3. The geometric method seems to perform quite nicely in most cases, especially for large K, in terms of both the power and the average combined p-value. On the other hand, the harmonic and maximum methods almost always perform poorly.

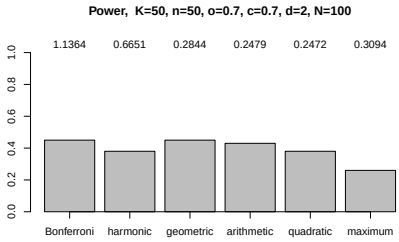
4. The average combined p-value of the Bonferroni method is often the largest among all methods. It seems that Bonferroni merging is unstable in the sense that, for the same setting, it sometimes gives very small p-values and sometimes very big ones.

5. In the case of independent tests, the Bonferroni method gives an average combined p-value that is very close to zero, even when the deviation is small. It seems that the Bonferroni method is able to utilize the extra information produced by multiple tests if they are independent. This is not the case for the arithmetic mean, as whether the tests are independent is irrelevant for the mean of its combined p-value.

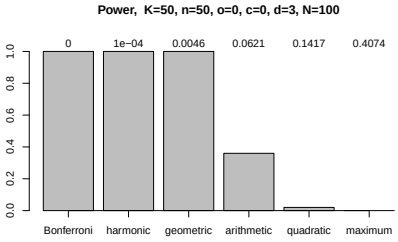
To summarize the observations into a rule of thumb, if the tests are highly similar (i.e., using similar data sets), then the arithmetic and geometric averaging methods are more powerful; if the tests are almost independent, then the Bonferroni method performs the best. The geometric mean seems to have the advantages of both incorporating independent information sources and being efficient when the tests are similar.



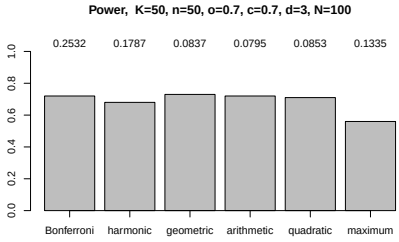
(a) independent tests, $d = 2$



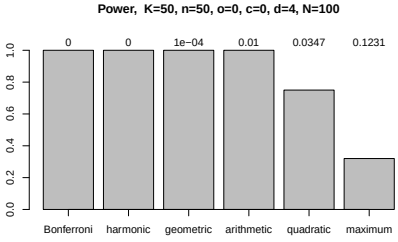
(b) overlap & correlation 70%, $d = 2$



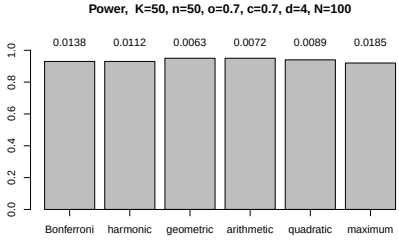
(c) independent tests, $d = 3$



(d) overlap & correlation 70%, $d = 3$

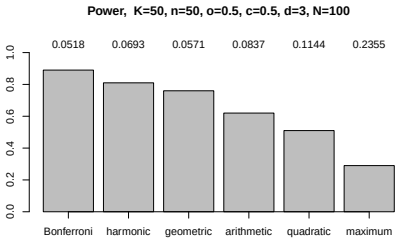


(e) independent tests, $d = 4$

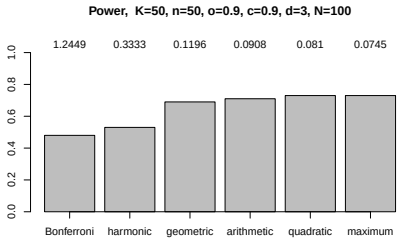


(f) overlap & correlation 70%, $d = 4$

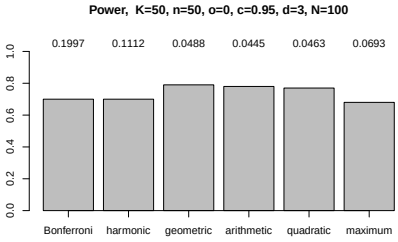
Figure 1: The power of each merging method under different settings. The average combined p-values are reported on top of each column.



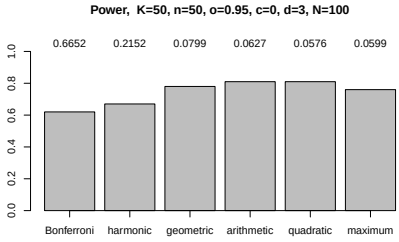
(a) overlap 50%, correlation 50%



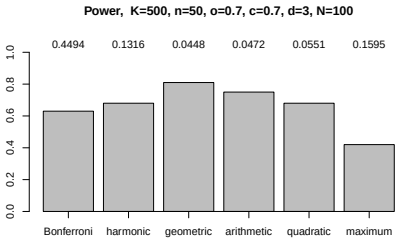
(b) overlap 90%, correlation 90%



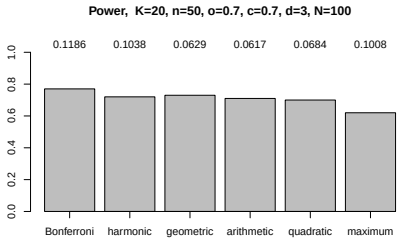
(c) overlap 0%, correlation 95%



(d) overlap 95%, correlation 0%



(e) large number of tests ($K = 500$)



(f) small number of tests ($K = 20$)

Figure 2: The power of each merging method under different settings. The average combined p-values are reported on top of each column.

7 Conclusion

These are examples of specific mathematical questions for further research:

- Explore how far our merging functions are from being precise for $r \in (-\infty, -1)$, complementing Proposition 9 by results for small K.
- How far is our merging function $(r+1)^{1/r} M_{r,K}$ from being precise for $r \in (-1, \frac{1}{K-1})$?

We note that these questions have corresponding versions in the context of robust risk aggregation, which are also unanswered. But perhaps the most important direction of research is to find practically useful applications, in multiple testing of a single hypothesis or testing multiple hypotheses, for our methods of merging p-values. The Bonferroni method of merging a set of p-values works very well when experiments are almost independent, while it produces unsatisfactory results if all p-values are approximately equal. Our methods are designed to work for intermediate situations; in particular, the arithmetic and the geometric averaging methods perform very well in many cases.

Acknowledgments

We are grateful to Dave Cohen, Alessio Sancetta, Wouter Koolen, and Lutz Mattner for their advice. Our special thanks go to Paul Embrechts, who, in addition to illuminating discussions, has been instrumental in starting our collaboration. The first author's work was partially supported by the Cyprus Research Promotion Foundation and the European Union's Horizon 2020 Research and Innovation programme (671555), and the second author's by the Natural Sciences and Engineering Research Council of Canada (RGPIN-435844-2013).

References

[1] Grace E. Bates. Joint distributions of time intervals for the occurrence of successive accidents in a generalized Polya scheme. *Annals of Mathematical Statistics*, 26:705–720, 1955.

[2] Carole Bernard, Xiao Jiang, and Ruodu Wang. Risk aggregation with dependence uncertainty. *Insurance: Mathematics and Economics*, 54:93–108, 2014.

[3] Valeria Bignozzi, Tiantian Mao, Bin Wang, and Ruodu Wang. Diversification limit of quantiles under dependence uncertainty. *Extremes*, 19:143–170, 2016.

[4] Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. *Introduction to Meta-Analysis*. Wiley, Chichester, 2009.

[5] Morton B. Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31:987–992, 1975.

[6] Paul Embrechts and Giovanni Puccetti. Bounds for functions of dependent risks. *Finance and Stochastics*, 10:341–352, 2006.

[7] Paul Embrechts, Bin Wang, and Ruodu Wang. Aggregation-robustness and model uncertainty of regulatory risk measures. *Finance and Stochastics*, 19:763–790, 2015.

[8] Ronald A. Fisher. Combining independent tests of significance. *American Statistician*, 2:30, 1948.

[9] Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. De Gruyter, Berlin, third edition, 2011.

[10] Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26:584–597, 2011.

[11] Benjamin T. Graham and Geoffrey R. Grimmett. Influence and sharp-threshold theorems for monotonic measures. *Annals of Probability*, 34:1726–1745, 2006.

[12] G. H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, Cambridge, England, second edition, 1952.

[13] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[14] Gerhard Hommel. Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal*, 25:423–430, 1983.

[15] Gerhard Hommel. Multiple test procedures for arbitrary dependence structures. *Metrika*, 33:321–336, 1986.

[16] Gerhard Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75:383–386, 1988.

[17] Edwards Jakobsons, Xiaoying Han, and Ruodu Wang. General convex order on risk aggregation. *Scandinavian Actuarial Journal*, 2016(8):713–740, 2016.

[18] Andrei N. Kolmogorov. Sur la notion de la moyenne. *Atti della Reale Accademia Nazionale dei Lincei. Classe di scienze fisiche, matematiche, e naturali. Rendiconti Serie VI*, 12(9):388–391, 1930.

[19] James T. Kost and Michael P. McDermott. Combining dependent p-values. *Statistics and Probability Letters*, 60:183–190, 2002.

[20] Henrik Linusson, Ulf Norinder, Henrik Boström, Ulf Johansson, and Tuve Löfström. On the calibration of aggregated conformal predictors. *Proceedings of Machine Learning Research*, 60:154–173, 2017.

[21] G. D. Makarov. Estimates for the distribution function of the sum of two random variables with given marginal distributions. *Theory of Probability and its Applications*, 26:803–806, 1981.

[22] Ruth Marcus, Eric Peritz, and K. Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660, 1976.

[23] Lutz Mattner. Combining individually valid and conditionally i.i.d. P-variables. Technical Report `arXiv:1008.5143 [stat.ME]`, `arXiv.org` e-Print archive, August 2011.

[24] Lutz Mattner. Combining individually valid and arbitrarily dependent P-variables. In *Abstract Book of the Tenth German Probability and Statistics Days,*, page p. 104, Mainz, Germany, March 2012. Institut für Mathematik, Johannes Gutenberg-Universität Mainz.

[25] Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ, 2015.

[26] Xiao-Li Meng. Posterior predictive p-values. *Annals of Statistics*, 22:1142–1160, 1993.

[27] Dietrich Morgenstern. Berechnung des maximalen Signifikanzniveaus des Testes "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika*, 27:285–286, 1980.

[28] Svetlozar T. Rachev and Ludger Rüschendorf. *Mass Transportation Problems*. Springer, New York, 1998. Volume I: Theory; Volume II: Applications.

[29] Bernhard Rüger. Das maximale Signifikanzniveau des Tests "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika*, 25:171–178, 1978.

[30] Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14:623–632, 1982.

[31] Ludger Rüschendorf. *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios*. Springer, Heidelberg, 2013.

[32] R. John Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.

[33] Vladimir Vovk. Cross-conformal predictors, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 6, August 2012.

[34] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[35] Bin Wang and Ruodu Wang. The complete mixability and convex minimization problems with monotone marginal densities. *Journal of Multivariate Analysis*, 102:1344–1360, 2011.

[36] Bin Wang and Ruodu Wang. Joint mixability. *Mathematics of Operations Research*, 41:808–826, 2016.

[37] Ruodu Wang, Liang Peng, and Jingping Yang. Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities. *Finance and Stochastics*, 17:395–417, 2013.

A Robust risk aggregation and proofs

The main topic of this paper is closely connected to robust risk aggregation. The origin of this field lies in a problem posed by Kolmogorov (see, e.g., [21]). This appendix is devoted to the proofs that depend on known results in robust risk aggregation, in particular, many results in [6, 2, 7, 36, 17].

Merging functions and quantiles

We start from a simple result (Lemma 13 below) that translates probability statements of a merging function into corresponding quantile statements. This result will allow us to freely use some recent results in the literature on robust risk aggregation.

Define the left α-quantile of a random variable X for $\alpha \in (0,1]$,

$$q_\alpha(X) := \sup\{x \in \mathbb{R} : \mathbb{P}(X \le x) < \alpha\},$$

and the right α-quantile of X for $\alpha \in [0,1)$,

$$q^+_\alpha(X) := \sup\{x \in \mathbb{R} : \mathbb{P}(X \le x) \le \alpha\}.$$

Notice that $q_1(X)$ is the essential supremum of X and $q^+_0(X)$ is the essential infimum of X. For $a > 0$, let $\mathcal{U}(a)$ be the set of all random variables distributed uniformly over the interval $[0,a]$, $a \ge 0$; we can regard $\mathcal{U}$ as an abbreviation for $\mathcal{U}(1)$. For a function $F : [0,1]^K \to [0,\infty)$ and $\alpha \in (0,1)$, write

$$\underline{q}_\alpha(F) := \inf\left\{q_\alpha(F(U_1,\dots,U_K)) \,|\, U_1,\dots,U_K \in \mathcal{U}\right\}.$$

Lemma 13. *For an increasing Borel function $F : [0,1]^K \to [0,\infty)$:*

(a) F is a merging function if and only if $\underline{q}_\epsilon(F) \ge \epsilon$ for all $\epsilon \in (0,1)$;

(b) F is a precise merging function if and only if $\underline{q}_\epsilon(F) = \epsilon$ for all $\epsilon \in (0,1)$.

Proof. Part "if" of (a): Suppose $\underline{q}_\epsilon(F) \ge \epsilon$ for all $\epsilon \in (0,1)$. Consider arbitrary $U_1,\dots,U_K \in \mathcal{U}$. We have $q_\epsilon(F(U_1,\dots,U_K)) \ge \epsilon$ for all $\epsilon \in (0,1)$. By

the definition of left quantiles, $\mathbb{P}(F(U_1,\dots,U_K)<\epsilon)\le\epsilon$. It follows that, for all $\delta\in(0,1-\epsilon)$,

$$\mathbb{P}(F(U_1,\dots,U_K)\le\epsilon)\le\mathbb{P}(F(U_1,\dots,U_K)<\epsilon+\delta)\le\epsilon+\delta,$$

which implies

$$\mathbb{P}(F(U_1,\dots,U_K)\le\epsilon)\le\epsilon,$$

since δ is arbitrary. Therefore, F is a merging function.

PART "ONLY IF" OF (A): Suppose F is a merging function. Let $U_1,\dots,U_K\in\mathcal{U}$ and $\epsilon\in(0,1)$. We have $\mathbb{P}(F(U_1,\dots,U_K)\le\epsilon)\le\epsilon$. By the definition of right quantiles, $q_\epsilon^+(F(U_1,\dots,U_K))\ge\epsilon$. It follows that, for all $\delta\in(0,\epsilon)$,

$$q_\epsilon(F(U_1,\dots,U_K))\ge q^+_{\epsilon-\delta}(F(U_1,\dots,U_K))\ge\epsilon-\delta,$$

which implies $q_\epsilon(F(U_1,\dots,U_K))\ge\epsilon$ since δ is arbitrary.

PART "IF" OF (B): Suppose $\underline{q}_\epsilon(F)=\epsilon$ for all $\epsilon\in(0,1)$. By (a), F is a merging function. For all $\epsilon,\delta\in(0,1)$, there exist $U_1,\dots,U_K\in\mathcal{U}$ such that $q_\epsilon(F(U_1,\dots,U_K))\in[\epsilon,\epsilon+\delta)$, which implies $\mathbb{P}(F(U_1,\dots,U_K)\le\epsilon+\delta)\ge\epsilon$. Since δ is arbitrary, we have

$$\sup\{\mathbb{P}(F(U_1,\dots,U_K)\le\epsilon)\,|\,U_1,\dots,U_K\in\mathcal{U}\}=\epsilon,$$

and thus F is precise.

PART "ONLY IF" OF (B): Suppose F is a precise merging function. Since F is a merging function, by (a) we have $\underline{q}_\epsilon(M)\ge\epsilon$ for all $\epsilon\in(0,1)$. Suppose, for the purpose of contradiction, that $\underline{q}_\epsilon(M)>\epsilon$ for some $\epsilon\in(0,1)$. Then, there exists $\delta\in(0,1-\epsilon)$ such that $q_\epsilon(F(U_1,\dots,U_K))>\epsilon+\delta$ for all $U_1,\dots,U_K\in\mathcal{U}$. As a consequence, we have

$$\mathbb{P}(F(U_1,\dots,U_K)\le\epsilon+\delta/2)\le\mathbb{P}(F(U_1,\dots,U_K)<\epsilon+\delta)\le\epsilon.$$

Therefore,

$$\sup\{\mathbb{P}(F(U_1,\dots,U_K)\le\epsilon+\delta/2)\,|\,U_1,\dots,U_K\in\mathcal{U}\}<\epsilon+\delta/2,$$

contradicting F being precise. □

Remark. This remark discusses briefly how the problem of merging p-values is related to robust risk aggregation. In quantitative risk management, the term *robust risk aggregation* refers to evaluating the value of a *risk measure* of an aggregation of risks $X_1,\dots,X_K$ with specified marginal distributions and unspecified dependence structure. More specifically, if the risk measure is chosen as a quantile q_α, known as *Value-at-Risk* and very popular in finance, the quantities of interest are typically

$$\overline{q}:=\sup\{q_\alpha(X_1+\cdots+X_n)\,|\,X_1\sim F_1,\dots,X_n\sim F_n\}$$

and

$$\underline{q}:=\inf\{q_\alpha(X_1+\cdots+X_n)\,|\,X_1\sim F_1,\dots,X_n\sim F_n\},$$

where $F_1, \dots, F_n$ denote the pre-specified marginal distributions of the risks. The motivation behind this problem is that, in practical applications of banking and insurance, the dependence structure among risks to aggregate is very difficult to accurately model, as compared with the corresponding marginal distributions. The interval $[\underline{q}, \overline{q}]$ thus represents all possible values of the aggregate risk measure given the marginal information. A more detailed introduction to this topic can be found in [25, Section 8.4.4] and [31, Chapter 4]. Via Lemma 13, the quantities $\overline{q}$ and $\underline{q}$ are obviously closely related to the problem of merging p-values. There are few explicit formulas for $\overline{q}$ and $\underline{q}$ but fortunately some do exist in the literature, and we will rely on them in our study of merging functions.

In all proofs below, for statements that have a weighted version in Section 5, namely Theorem 1 and Propositions 4 and 5, we present a proof of the corresponding weighted version, which is stronger.

Proof of Theorem 1w

Without loss of generality we can, and will, assume that ϕ is strictly increasing. Indeed, if ϕ is strictly decreasing, we can redefine $\phi := -\phi$ and $\psi(u) := \psi(-u)$ and notice that the statement of the theorem for new ϕ and ψ will imply the analogous statement for the original ϕ and ψ.

Define an accessory function $\Phi : (0,1) \to [-\infty, \infty]$ by $\Phi(\epsilon) = \frac{1}{\epsilon} \int_0^\epsilon \phi(u) \mathrm{d}u$. Fix $\epsilon \in (0,1)$. Since ϕ is integrable, $\Phi(\epsilon)$ is finite.

Known results from the literature on robust risk aggregation can be applied to random variables $X_k := \phi(U_k)$, where $U_k \in \mathcal{U}$; notice that the distribution function of X_k is ψ:

$$\mathbb{P}(X_k \le x) = \mathbb{P}(\phi(U_k) \le x) = \mathbb{P}(U_k \le \psi(x)) = \psi(x).$$

Theorem 4.6 of [2] gives the following relation:

$$\underline{q}_\epsilon(M_{\phi,\mathbf{w}}) = \inf \left\{ q_1 \left(\psi \left(\sum_{k=1}^K w_k \phi(V_k) \right) \right) \,\middle|\, V_1, \dots, V_K \in \mathcal{U}(\epsilon) \right\}. \tag{15}$$

Since

$$q_1 \left(w_1 \phi(V_1) + \cdots + w_K \phi(V_K) \right) \ge \mathbb{E} \left[w_1 \phi(V_1) + \cdots + w_K \phi(V_K) \right] = \Phi(\epsilon)$$

for $V_1, \dots, V_K \in \mathcal{U}(\epsilon)$, we have $\underline{q}_\epsilon(M_{\phi,\mathbf{w}}) \ge \psi(\Phi(\epsilon))$. □

Proof of Proposition 3

Now $\phi(u) = u^r$, which gives $\Phi(\epsilon) = \epsilon^r/(r+1)$, in the notation of the previous proof. Using Corollary 3.4 of [7], we have

$$\lim_{K \to \infty} \frac{\phi(\underline{q}_\epsilon(M_{r,K}))}{\Phi(\epsilon)} = 1,$$

leading to $\lim_{K\to\infty} \underline{q}_\epsilon(M_{r,K}) = \epsilon(r+1)^{-1/r}$. It follows that, for $a < (r+1)^{1/r}$,

$$\lim_{K\to\infty} \underline{q}_\epsilon(aM_{r,K}) < \epsilon$$

and so, by Lemma 13, $aM_{r,K}$ is not a merging function for K large enough. □

Proof of Proposition 4w

Using (15) and Theorem 1w, we have, for $\epsilon \in (0,1)$:

$$(\underline{q}_\epsilon(M_{r,K}))^r = \inf\left\{ q_1\left(\sum_{k=1}^K w_k V_k^r\right) \,\middle|\, V_1,\dots,V_K \in \mathcal{U}(\epsilon)\right\} \geq \frac{\epsilon^r}{1+r} \quad (16)$$

if $r > 0$,

$$(\underline{q}_\epsilon(M_{r,K}))^r = \sup\left\{ q_0^+\left(\sum_{k=1}^K w_k V_k^r\right) \,\middle|\, V_1,\dots,V_K \in \mathcal{U}(\epsilon)\right\} \leq \frac{\epsilon^r}{1+r} \quad (17)$$

if $r < 0$, and

$$\underline{q}_\epsilon(M_{r,K}) = \exp\left(\inf\left\{ q_1\left(\sum_{k=1}^K w_k \ln V_k\right) \,\middle|\, V_1,\dots,V_K \in \mathcal{U}(\epsilon)\right\}\right) \geq \frac{\epsilon}{e} \quad (18)$$

if $r = 0$. By Lemma 13, M is a precise merging function if and only if the inequality in (16)–(18) is an equality for all $\epsilon \in (0,1)$.

Fix $\epsilon \in (0,1)$ and $r \in (-1,\infty)$. For $k = 1,\dots,K$, let F_k be the distribution of $w_k V_k^r$ where $V_k \in \mathcal{U}(\epsilon)$. Using the terminology of [36], notice that the inequality in (16)–(18) is an equality if and only if $(F_1,\dots,F_K)$ is jointly mixable due to a standard compactness argument (see [36, Proposition 2.3]). Therefore, we can first settle the cases $r = 0$ and $r < 0$, as in these cases the supports of $F_1,\dots,F_K$ are unbounded on one side, and $(F_1,\dots,F_K)$ is not jointly mixable (see [36, Remark 2.2]).

Next assume $r > 0$. Since $F_1,\dots,F_K$ have monotone densities on their respective supports, by Theorem 3.2 of [36], $(F_1,\dots,F_K)$ is jointly mixable if and only if the "mean condition"

$$w\epsilon^r \leq \frac{\epsilon^r}{1+r} \leq \epsilon^r - w\epsilon^r$$

is satisfied. This is equivalent to $w \leq \frac{1}{1+r} \leq 1-w$ and, therefore, to the conjunction of $w \leq 1/2$ and $r \in [\frac{w}{1-w}, \frac{1-w}{w}]$. This completes the proof. □

Proof of Proposition 5w

Notice that, for each $k = 1,\dots,K$, the distribution of $w_k U_k^r$, where $U_k \in \mathcal{U}$, has a decreasing density on its support. Therefore, we can apply Corollary 4.7 of

[17], which gives

$$\inf\left\{q_\epsilon\left(w_1U_1^r+\cdots+w_KU_K^r\right)\middle|\,U_1,\ldots,U_K\in\mathcal{U}\right\}=\max\left(w\epsilon^r,\frac{\epsilon^r}{1+r}\right).$$

Simple algebra leads to

$$\underline{q}_\epsilon(M_{r,K})=\max\left(w,\frac{1}{1+r}\right)^{1/r}\epsilon,$$

and by Lemma 13, M is a precise merging function. □

Proof of Proposition 6

Our goal is to obtain the precise value of $\underline{q}_\epsilon(M_{0,K})$. Set

$$\begin{aligned}b_K&:=\sup\left\{q_0^+\left(-(\ln U_1+\cdots+\ln U_K)\right)\middle|\,U_1,\ldots,U_K\in\mathcal{U}\right\}\\&=\sup\left\{q_0^+\left(-(\ln V_1+\cdots+\ln V_K)+K\ln\epsilon\right)\middle|\,V_1,\ldots,V_K\in\mathcal{U}(\epsilon)\right\}.\end{aligned}$$

It is easy to see that

$$\begin{aligned}\underline{q}_\epsilon(M_{0,K})&=\exp\left(\inf\left\{q_1\left(\frac{\ln V_1+\cdots+\ln V_K}{K}\right)\middle|\,V_1,\ldots,V_K\in\mathcal{U}(\epsilon)\right\}\right)\\&=\exp\left(-\sup\left\{q_0^+\left(-\frac{\ln V_1+\cdots+\ln V_K}{K}\right)\middle|\,V_1,\ldots,V_K\in\mathcal{U}(\epsilon)\right\}\right)\\&=\exp\left(-b_K/K+\ln\epsilon\right)\\&=\epsilon\exp(-b_K/K).\end{aligned}$$

It is clear that $e^{b_K/K}M_{0,K}$ is a precise merging function. In the proof of Proposition 3, we have already seen that, as $K\to\infty$, $b_K/K\to 1$ since $\underline{q}_\epsilon(M_{0,K})\to\epsilon/e$. Next we focus on b_K for finite K.

Since $-\ln U$ has the standard exponential distribution for $U\in\mathcal{U}$ and, therefore, a decreasing density on $\mathbb{R}$, we can apply Theorem 3.2 of [2] (essentially Theorem 3.5 of [35]) to arrive at

$$b_K=-(K-1)\ln(1-(K-1)c_K)-\ln c_K,$$

where c_K is the unique solution to (11) (see [35, Corollary 4.1]). Using (11), we can write

$$b_K/K=-\ln c_K-(K-1)(1-Kc_K).$$

Using $a_K=e^{b_K/K}$ one obtains the desired result. □

Proof of Theorem 7

We will apply Theorem 4.2 of [6] in our situation where the function ϕ (and, therefore, ψ as well) in (7) is decreasing. Letting $X_k:=\phi(p_k)$ and using the

notation m_+ (used in Theorem 4.2 of [6]), we have, by the definition of m_+,

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^{K} X_k < s\right) &\geq m_+(s),\\
\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} \phi(p_k) < s/K\right) &\geq m_+(s),\\
\mathbb{P}\left(M_{\phi,K}(p_1,\ldots,p_K) > \psi(s/K)\right) &\geq m_+(s),\\
\mathbb{P}\left(M_{\phi,K}(p_1,\ldots,p_K) \leq \psi(s/K)\right) &\leq 1 - m_+(s).
\end{aligned}$$

The lower bound on $m_+(s)$ given in Theorem 4.2 of [6] involves $1-F(x)$, where F is the common distribution function of X_k, and in our current context we have:

$$1 - F(x) = \mathbb{P}(X_k > x) = \mathbb{P}(\phi(p_k) > x) = \mathbb{P}(p_k < \psi(x)) = \psi(x).$$

The last inequality and chain of equalities in combination with Theorem 4.2 of [6] give

$$\mathbb{P}\left(M_{\phi,K}(p_1,\ldots,p_K) \leq \psi(s/K)\right) \leq K \inf_{r\in[0,s/K)} \frac{\int_r^{s-(K-1)r} \psi(x)\mathrm{d}x}{s-Kr}.$$

Setting $\epsilon := \psi(s/K) \in [\psi(\infty), \psi(0)]$ (so that it is essential that $\psi(\infty) = 0$), we obtain, using $s = K\phi(\epsilon)$,

$$\mathbb{P}\left(M_{\phi,K}(p_1,\ldots,p_K) \leq \epsilon\right) \leq K \inf_{r\in[0,\phi(\epsilon))} \frac{\int_r^{K\phi(\epsilon)-(K-1)r} \psi(x)\mathrm{d}x}{(\phi(\epsilon)-r)K}.$$

Setting $t := \phi(\epsilon) - r$ and renaming x to u, this can be rewritten as (12). □

Proof of Proposition 9

We first show a simple property of a precise merging function via general averaging. Define the following constant:

$$a_{r,K} := \left(\frac{1}{K}\sup\{q_0^+(U_1^r + \cdots + U_K^r) \;\big|\; U_1,\ldots,U_K \in \mathcal{U}\}\right)^{-1/r}.$$

It is clear that $a_{r,K} \geq 1$ for $r < 0$.

Lemma 14. *For $r < 0$, the function $a_{r,K}M_{r,K}$ is a precise merging function.*

Proof. By straightforward algebra and Theorem 4.6 of [2],

$$\begin{aligned}
\underline{q}_\epsilon(a_{r,K}M_{r,K}) &= a_{r,K}\inf\{q_1(M_{r,K}(V_1,\ldots,V_K)) \,|\, V_1,\ldots,V_K \in \mathcal{U}(\epsilon)\}\\
&= a_{r,K}\inf\{\epsilon q_1(M_{r,K}(U_1,\ldots,U_K)) \,|\, U_1,\ldots,U_K \in \mathcal{U}\}
\end{aligned}$$

$$= a_{r,K}\epsilon \left(\frac{1}{K} \sup \left\{ q_0^+(U_1^r + \cdots + U_K^r) \,\middle|\, U_1, \ldots, U_K \in \mathcal{U} \right\} \right)^{1/r}$$
$$= \epsilon.$$

By Lemma 13, $a_{r,K} M_{r,K}$ is a precise merging function. □

To construct precise merging functions, it remains to find values of $a_{r,K}$. Unfortunately, for $r < 0$ no analytical formula for $a_{r,K}$ is available. There is an asymptotic result available in [3], which leads to the following proposition.

Proposition 15. *For $r \in (-\infty, -1)$,*

$$\lim_{K\to\infty} \frac{a_{r,K}}{K^{1+1/r}} = \frac{r}{r+1}.$$

Proof. The quantity $\overline{\Delta}^{\mathcal{F}_d}$ in [3], defined as

$$\overline{\Delta}^{\mathcal{F}_d} := \lim_{\alpha\to 1} \frac{\sup \left\{ q_\alpha(U_1^r + \cdots + U_K^r) \,|\, U_1, \ldots, U_K \in \mathcal{U}(\alpha) \right\}}{K(1-\alpha)^r},$$

satisfies

$$\overline{\Delta}^{\mathcal{F}_d} = \frac{1}{K} \sup \left\{ q_0^+(U_1^r + \cdots + U_K^r) \,\middle|\, U_1, \ldots, U_K \in \mathcal{U} \right\} = a_{r,K}^{-r}.$$

Using Proposition 3.5 of [3], we have, for $r < -1$, by substituting $\beta := -1/r$ in (3.25) of [3] and $\overline{\Delta}^{\mathcal{F}_d} = a_{r,K}^{-r}$,

$$\lim_{K\to\infty} \frac{a_{r,K}^{-r}}{K^{-r-1}} = \left(\frac{r}{r+1} \right)^{-r},$$

and this gives the desired result. □

Proof of Proposition 9

This proposition immediately follows from Proposition 15. □

Proof of Proposition 11

Let us check that $a_K = a_{-1,K}$. For this we use Corollary 3.7 of [37]. Write

$$H(t) := \frac{K-1}{1-(K-1)t} + \frac{1}{t} = \frac{1}{t(1-(K-1)t)}, \quad t \in [0, 1/K].$$

By Corollary 3.7 of [37], we have

$$a_{-1,K} = \frac{1}{K} \sup \left\{ q_0^+(U_1^{-1} + \cdots + U_K^{-1}) \,\middle|\, U_1, \ldots, U_K \in \mathcal{U} \right\}$$
$$= \frac{1}{K} H(x_K) = \frac{1}{K x_K (1-(K-1)x_K)}$$

where x_K solves the equation

$$\int_x^{1/K} H(t)\mathrm{d}t = \left(\frac{1}{K} - x\right) H(x), \quad x \in [0, 1/K).$$

Plugging in the expression for H and rearranging the above equation, we obtain

$$\begin{aligned}\frac{1-Kx}{Kx(1-(K-1)x)} &= \int_x^{1/K} \left(\frac{K-1}{1-(K-1)t} + \frac{1}{t}\right) \mathrm{d}t \\ &= \int_{(K-1)x}^{(K-1)/K} \frac{1}{1-y}\mathrm{d}y + \int_x^{1/K} \frac{1}{t}\mathrm{d}t \\ &= \ln(1-(K-1)x) - \ln x. \end{aligned} \tag{19}$$

The uniqueness of the solution x_K to (19) can be easily checked, and it is a special case of Lemma 3.1 of [17]. Writing $y = \frac{1-Kx}{x} > 0$, (19) reads as $\frac{y}{(y+1)K/(y+K)} = \ln(y+1)$. Rearranging the terms gives

$$y^2 = K((y+1)\ln(y+1) - y), \tag{20}$$

which admits a unique solution, $y_K = \frac{1-Kx_K}{x_K}$. Therefore,

$$a_{-1,K} = \frac{1}{Kx_K(1-(K-1)x_K)} = \frac{(y_K+K)^2}{(y_K+1)K},$$

and the first part of the proposition is shown.

Next we analyze the asymptotic behaviour of $a_{-1,K}$ as $K \to \infty$. Using $\ln(y+1) \geq y - y^2/2$ for $y \geq 0$, we can see that (20) implies the inequality

$$y^2 \geq \frac{K}{2}y^2 - \frac{K}{2}y^3,$$

which leads to $2 \geq K(1-y)$. Hence, we have $\liminf_{K\to\infty} y_K \geq 1$.

Notice that $(y+1)\ln(y+1) - y$ is a strictly increasing function of $y \in (0, \infty)$. Using $\liminf_{K\to\infty} y_K \geq 1$, we obtain that

$$\liminf_{K\to\infty} y_K^2 \geq K(2\ln 2 - 1).$$

Therefore, $\lim_{K\to\infty} y_K = \infty$. Applying logarithms to both sides of (20) and taking a limit in their ratio, we obtain

$$1 = \lim_{K\to\infty} \frac{2\ln(y_K)}{\ln K + \ln(y_K+1) + \ln\left(\ln(y_K+1) - \frac{y_K}{y_K+1}\right)} = \lim_{K\to\infty} \frac{2\ln(y_K)}{\ln K + \ln y_K},$$

and hence $\ln y_K / \ln K \to 1$ as $K \to \infty$. Using (20) again, we have

$$1 = \lim_{K\to\infty} \frac{y_K^2}{K((y_K+1)\ln(y_K+1) - y_K)} = \lim_{K\to\infty} \frac{y_K^2}{K(y_K \ln y_K)} = \lim_{K\to\infty} \frac{y_K}{K\ln K}.$$

Therefore, we have

$$\lim_{K\to\infty} \frac{a_{-1,K}}{\ln K} = \lim_{K\to\infty} \frac{(y_K+K)^2}{(y_K+1)K\ln K} = \lim_{K\to\infty} \frac{(K\ln K)^2}{(K\ln K)K\ln K} = 1.$$

This completes the proof. □

Proof of Proposition 12

Let us check that for $F := aM_{r,K}$ the following statements are equivalent:

(a) F is a merging function;

(b) (5) holds for some $\epsilon \in (0,1)$;

(c) $\underline{q}_\epsilon(F) \geq \epsilon$ for some $\epsilon \in (0,1)$.

The implication (a) $\Rightarrow$ (b) holds by definition.

To check (b) $\Rightarrow$ (c), let us assume (b). Since $\mathbb{P}(X \leq \epsilon) \leq \epsilon$ implies $q^+_\epsilon(X) \geq \epsilon$ for any random variable X,

$$\inf\left\{q^+_\epsilon(F(U_1,\dots,U_K)) \,\middle|\, U_1,\dots,U_K \in \mathcal{U}\right\} \geq \epsilon.$$

Using Lemma 4.5 of [2],

$$\underline{q}_\epsilon(F) = \inf\left\{q^+_\epsilon(F(U_1,\dots,U_K)) \,\middle|\, U_1,\dots,U_K \in \mathcal{U}\right\},$$

and hence (c) follows.

It remains to check (c) $\Rightarrow$ (a). For any $\epsilon \in (0,1)$, by straightforward algebra and Theorem 4.6 of [2],

$$\begin{aligned}\underline{q}_\epsilon(F) &= \inf\left\{q_1(F(V_1,\dots,V_K)) \,\middle|\, V_1,\dots,V_K \in \mathcal{U}(\epsilon)\right\} \\ &= \epsilon \inf\left\{q_1(F(U_1,\dots,U_K)) \,\middle|\, U_1,\dots,U_K \in \mathcal{U}\right\}.\end{aligned}$$

Therefore, to check $\underline{q}_\epsilon(F) \geq \epsilon$ for all $\epsilon \in (0,1)$, one only needs to check the inequality for one $\epsilon \in (0,1)$. By Lemma 13, F is a merging function.

This completes the proof of the first part of Proposition 12, and the second part can be proved similarly. □

B Connections with conformal prediction

This section assumes the knowledge of basic definitions of conformal prediction (see, e.g., [34]). The method of cross-conformal prediction, already mentioned in Section 1, defines putative p-values as, essentially, the arithmetic means of the p-values computed from K folds (cf. [33, (11)]). The experiments reported in [33] confirm the empirical validity of the arithmetic means in the sense of approximately satisfying (4), although [33, Appendix A] gives examples showing that in general (4) may be violated. In their recent paper [20] Linusson et al. give examples of particularly unstable randomized underlying algorithms which make the arithmetic means used in the method of conformal prediction violate (4) in their experimental studies. This makes the phenomenon discussed theoretically in [33, Appendix A] a practical issue. Corollary 2 with $r = 1$ (due to Rüschendorf and Meng) shows that there is a limit to the degree to which the validity of cross-conformal predictors can be violated, even for the most unstable underlying algorithms: namely, we always have

$$\mathbb{P}(P \leq \epsilon) \leq 2\epsilon, \quad \forall \epsilon \in (0,1), \tag{21}$$

where P is the arithmetic mean of p-values.

In fact, cross-conformal predictors output only approximate arithmetic means of p-values; the precise expression [33, (11)] for their putative p-values p is

$$p = \bar{p} + \frac{K-1}{l+1}(\bar{p} - 1), \tag{22}$$

where l is the size of the training set, K is the number of equal-sized folds, and $\bar{p}$ is the arithmetic mean of the p-values computed from the different folds. Expressing $\bar{p}$ via p,

$$\bar{p} = \frac{l+1}{l+K}\left(p + \frac{K-1}{l+1}\right)$$

and using $2\bar{p}$ being a p-value function, we obtain

$$\mathbb{P}\left(\frac{l+1}{l+K}\left(p + \frac{K-1}{l+1}\right) \le \delta\right) \le 2\delta$$

for any $\delta > 0$. Rewriting the last inequality as

$$\mathbb{P}\left(p \le \frac{l+K}{l+1}\delta - \frac{K-1}{l+1}\right) \le 2\delta$$

and setting

$$\epsilon := \frac{l+K}{l+1}\delta - \frac{K-1}{l+1},$$

we finally obtain

$$\mathbb{P}(p \le \epsilon) \le 2\epsilon + 2\frac{K-1}{l+K}(1-\epsilon) \tag{23}$$

for any $\epsilon > 0$. The last equation, (23), is the precise version of (21) in the case where p is the putative p-value output by a cross-conformal predictor; we can see that (23) is marginally weaker.

In the rest of this section we will ignore the difference between the expression (22) and the arithmetic mean $\bar{p}$, and it will be convenient to generalize the definition of $\bar{p}$ as $\bar{p} := \int p_\theta P(\mathrm{d}\theta)$, where p_θ is now a family of uniformly distributed random variables indexed by, and measurable in, $\theta \in \Theta$, where Θ is a measurable space and P is a probability measure on Θ. We will refer to the uniform distribution on $[0,1]$ as U. Let us say that a random variable $\xi \in [0,1]$ is *stochastically less variable than* U if $\mathbb{E}(h(\xi)) \le \int_0^1 h(u)\mathrm{d}u$ for any convex function h on $[0,1]$. The following result is a restatement of Theorem 1 in [26].

Proposition 16. *It is always true that $\bar{p}$ is stochastically less variable than U.*

Proof. For any convex h,

$$\begin{aligned}
\mathbb{E}(h(\bar{p})) &= \mathbb{E}\left(h\left(\int p_\theta P(\mathrm{d}\theta)\right)\right) && \text{by definition} \\
&\le \mathbb{E}\left(\int h(p_\theta) P(\mathrm{d}\theta)\right) && \text{by Jensen's inequality}
\end{aligned}$$

$$
\begin{aligned}
&= \int \mathbb{E}(h(p_\theta))P(\mathrm{d}\theta) && \text{by Fubini's theorem} \\
&= \int \left(\int_0^1 h(u)\mathrm{d}u \right) P(d\theta) && \text{as } p_\theta \in \mathcal{U} \\
&= \int_0^1 h(u)\mathrm{d}u. && \square
\end{aligned}
$$

Meng [26, Corollary 1] proves the following stronger version of Rüschendorf's result, applicable to any random variable that is stochastically less variable than U, not just $\bar{p}$.

Proposition 17. *If G is the distribution function of a random variable that is stochastically less variable than U, then, for all $\epsilon \in (0,1)$,*

$$
\epsilon - \sqrt{\epsilon^2 - 2\int_0^\epsilon G(u)\mathrm{d}u} \le G(\epsilon) \le \epsilon + \sqrt{\epsilon^2 - 2\int_0^\epsilon G(u)\mathrm{d}u} \le 2\epsilon \tag{24}
$$

(in particular, the expression under the square root signs is always nonnegative).

For example, (24) says that if at some ϵ the property of validity of $\bar{p}$ as a p-value is violated to the greatest degree, $\mathbb{P}(\bar{p} \le \epsilon) = 2\epsilon$, then, at all smaller significance levels $\epsilon' < \epsilon$, $\bar{p}$ must be extremely conservative, $\mathbb{P}(\bar{p} \le \epsilon') = 0$.

Applying Proposition 16 to $h(u) := u$ and $h(u) := -u$, we obtain $\mathbb{E}\,\bar{p} = 1/2$. Therefore, $\bar{p}$ has the right expectation but is more concentrated around it than U is. There are two particularly interesting special cases that have been discussed in literature. One is where $\bar{p}$ is a constant (necessarily $1/2$): see, e.g., [28, Example 3.6.4]. The other is where $\bar{p}$ is distributed as the sum K independent random variables each distributed uniformly on $[0,1]$; the resulting distribution of $\bar{p}$, which we now again assume to be $(p_1 + \cdots + p_K)/K$, is then known as the Bates distribution [1] and is concentrated around $1/2$ for large K. In the context of cross-conformal prediction, this corresponds to the extreme case where the p-value for each fold is computed using a randomized algorithm that does not pay any attention to the data. In general, we need to be careful when using randomized algorithms as underlying algorithms in cross-conformal prediction [20].