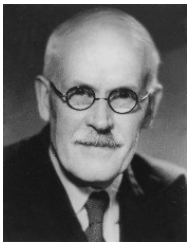
E-values: Calibration, combination, and applications

Vladimir Vovk Ruodu Wang



Users of these tests speak of the
5 per cent. point [p-value of 5%]
in much the same way as I should
speak of the $K = 10^{-1/2}$ point
[e-value of $10^{1/2}$], and of the 1
per cent. point [p-value of 1%]
as I should speak of the
$K = 10^{-1}$ point [e-value of 10].

Project "Hypothesis testing with e-values"

Working Paper #2

First posted 13 December 2019 (on arXiv) and 1 March 2020 (as Working Paper). Last revised March 14, 2021.

Project web site:
http://alrw.net/e

Abstract

Multiple testing of a single hypothesis and testing multiple hypotheses are usually done in terms of p-values. In this paper we replace p-values with their natural competitor, e-values, which are closely related to betting, Bayes factors, and likelihood ratios. We demonstrate that e-values are often mathematically more tractable; in particular, in multiple testing of a single hypothesis, e-values can be merged simply by averaging them. This allows us to develop efficient procedures using e-values for testing multiple hypotheses.

Contents

1 Introduction

The problem of multiple testing of a single hypothesis (also known as testing a global null) is usually formalized as that of combining a set of p-values. The notion of p-values, however, has a strong competitor, which we refer to as e-values in this paper. E-values can be traced back to various old ideas, but they have started being widely discussed in their pure form only recently: see, e.g., Shafer [2019], who uses the term "betting score" in the sense very similar to our "e-value", Shafer and Vovk [2019, Section 11.5], who use "Skeptic's capital", and Grünwald et al. [2020]. The power and intuitive appeal of e-values stem from their interpretation as results of bets against the null hypothesis [Shafer, 2019, Section 1].

Formally, an *e-variable* is a nonnegative extended random variable whose expected value under the null hypothesis is at most 1, and an *e-value* is a value taken by an e-variable. Whereas p-values are defined in terms of probabilities, e-values are defined in terms of expectations. As we regard an e-variable E as a bet against the null hypothesis, its realized value $e := E(\omega)$ shows how successful our bet is (it is successful if it multiplies the money it risks by a large factor). Under the null hypothesis, it can be larger than a constant $c > 1$ with probability at most $1/c$ (by Markov's inequality). If we are very successful (i.e., e is very large), we have reasons to doubt that the null hypothesis is true, and e can be interpreted as the amount of evidence we have found against it. In textbook statistics e-variables typically appear under the guise of likelihood ratios and Bayes factors.

The main focus of this paper is on combining e-values and multiple hypothesis testing using e-values. The picture that arises for these two fields is remarkably different from, and much simpler than, its counterpart for p-values. To clarify connections between e-values and p-values, we discuss how to transform p-values into e-values, or *calibrate* them, and how to move in the opposite direction.

We start the main part of the paper by defining the notion of e-values in Section 2 and reviewing known results about connections between e-values and p-values; we will discuss how the former can be turned into the latter and vice versa (with very different domination structures for the two directions). In Section 3 we show that the problem of merging e-values is more or less trivial: a convex mixture of e-values is an e-value, and symmetric merging functions are essentially dominated by the arithmetic mean. For example, when several analyses are conducted on a common (e.g., public) dataset each reporting an e-value, it is natural to summarize them as a single e-value equal to their weighted average (the same cannot be said for p-values). In Section 4 we assume, additionally, that the e-variables being merged are independent and show that the domination structure is much richer; for example, now the product of e-values is an e-value. The assumption of independence can be replaced by the weaker assumption of being *sequential*, and we discuss connections with the popular topic of using martingales in statistical hypothesis testing: see, e.g., Duan et al. [2019] and Shafer and Vovk [2019]. In Section 5 we apply these results to multiple hy-

pothesis testing. In the next section, Section 6, we briefly review known results on merging p-values (e.g., the two classes of merging methods in Rüger [1978] and Vovk and Wang [2019a]) and draw parallels with merging e-values; in the last subsection we discuss the case where p-values are independent. Section 7 is devoted to experimental results; one finding in this section is that, for multiple testing of a single hypothesis in independent experiments, a simple method based on e-values outperforms standard methods based on p-values. Section 8 concludes the main part of the paper.

Appendix A describes numerous connections with the existing literature, including Bayes factors and multiple hypothesis testing. Appendix B describes the origins of the problem of calibrating p-values and gives interesting examples of calibrators. A short Appendix C deals with merging infinite e-values. Appendix D explores the foundations of calibration and merging of e-values and p-values; in particular, whether the universal quantifiers over probability spaces in the definitions given in the main paper are really necessary. Appendix E proves Theorem 3.2 in the main paper characterizing the domination structure of the e-merging functions. Appendix F presents an informative minimax view of essential and weak domination. Appendix G discusses "cross-merging": how do we merge several p-values into one e-value and several e-values into one p-value? Appendix H contains additional experimental results. Finally, Appendix I briefly describes the procedure that we use for multiple hypothesis testing in combination with Fisher's [1932] method of combining p-values.

The journal version of this paper is to appear in the *Annals of Statistics*.

2 Definition of e-values and connections with p-values

For a probability space $(\Omega, \mathcal{A}, Q)$, an *e-variable* is an extended random variable $E : \Omega \to [0, \infty]$ satisfying $\mathbb{E}^Q[E] \le 1$; we refer to it as "extended" since its values are allowed to be ∞, and we let $\mathbb{E}^Q[X]$ (or $\mathbb{E}[X]$ when Q is clear from context) stand for $\int X \, \mathrm{d}Q$ for any extended random variable X. The values taken by e-variables will be referred to as *e-values*, and we denote the set of e-variables by $\mathcal{E}_Q$. It is important to allow E to take value ∞; in the context of testing Q, observing $E = \infty$ for an *a priori* chosen e-variable E means that we are entitled to reject Q as null hypothesis.

Remark 2.1. Until Section 5 we will concentrate on the case of a simple null hypothesis Q. A composite null hypothesis $\mathcal{Q}$ is a set of probability measures on $(\Omega, \mathcal{A})$, and an e-variable for $\mathcal{Q}$ is defined as an extended nonnegative random variable that integrates to at most 1 under any probability measure in $\mathcal{Q}$. The main results that we state for simple null hypotheses remain true for composite null hypotheses; see Appendix D.

Our emphasis in this paper is on e-values, but we start from discussing their connections with the familiar notion of p-values. A *p-variable* is a random

variable $P : \Omega \to [0,1]$ satisfying

$$\forall \epsilon \in (0,1) : Q(P \le \epsilon) \le \epsilon.$$

The set of all p-variables is denoted by $\mathcal{P}_Q$.

A calibrator is a function transforming p-values to e-values. Formally, a decreasing function $f : [0,1] \to [0,\infty]$ is a *calibrator* (or, more fully, *p-to-e calibrator*) if, for any probability space $(\Omega, \mathcal{A}, Q)$ and any p-variable $P \in \mathcal{P}_Q$, $f(P) \in \mathcal{E}_Q$. A calibrator f is said to *dominate* a calibrator g if $f \ge g$, and the domination is *strict* if $f \ne g$. A calibrator is *admissible* if it is not strictly dominated by any other calibrator.

The following proposition says that a calibrator is a nonnegative decreasing function integrating to at most 1 over the uniform probability measure.

Proposition 2.2. *A decreasing function $f : [0,1] \to [0,\infty]$ is a calibrator if and only if $\int_0^1 f \le 1$. It is admissible if and only if f is upper semicontinuous, $f(0) = \infty$, and $\int_0^1 f = 1$.*

Of course, in the context of this proposition, being upper semicontinuous is equivalent to being left-continuous.

Proof. Proofs of similar statements are given in, e.g., Vovk [1993, Theorem 7], Shafer et al. [2011, Theorem 3], and Shafer and Vovk [2019, Proposition 11.7], but we will give an independent short proof using our definitions. The first "only if" statement is obvious. To show the first "if" statement, suppose that $\int_0^1 f \le 1$, P is a p-variable, and P' is uniformly distributed on $[0,1]$. Since $Q(P < x) \le Q(P' < x)$ for all $x \ge 0$ and f is decreasing, we have

$$Q(f(P) > y) \le Q(f(P') > y)$$

for all $y \ge 0$, which implies

$$\mathbb{E}[f(P)] \le \mathbb{E}[f(P')] = \int_0^1 f(p)\,\mathrm{d}p \le 1.$$

The second statement in Proposition 2.2 is obvious. □

The following is a simple family of calibrators. Since $\int_0^1 \kappa p^{\kappa-1}\,\mathrm{d}p = 1$, the functions

$$f_\kappa(p) := \kappa p^{\kappa-1} \tag{1}$$

are calibrators, where $\kappa \in (0,1)$. To solve the problem of choosing the parameter κ, sometimes the maximum

$$\mathrm{VS}(p) := \max_{\kappa \in [0,1]} f_\kappa(p) = \begin{cases} -\exp(-1)/(p \ln p) & \text{if } p \le \exp(-1) \\ 1 & \text{otherwise} \end{cases}$$

is used (see, e.g., Benjamin and Berger [2019], Recommendations 2 and 3); we will refer to it as the *VS bound* (abbreviating "Vovk–Sellke bound", as used in,

e.g., the JASP package). It is important to remember that $\mathrm{VS}(p)$ is not a valid e-value, but just an overoptimistic upper bound on what is achievable with the class (1). Another way to get rid of κ is to integrate over it, which gives

$$F(p) := \int_0^1 \kappa p^{\kappa-1} \, \mathrm{d}\kappa = \frac{1 - p + p \ln p}{p(-\ln p)^2}. \tag{2}$$

(See Appendix B for more general results and references. We are grateful to Aaditya Ramdas for pointing out the calibrator (2).) An advantage of this method is that it produces a bona fide e-value, unlike the VS bound. As $p \to 0$, $F(p) \sim p^{-1}(-\ln p)^{-2}$, so that $F(p)$ is closer to the ideal (but unachievable) $1/p$ (cf. Remark 2.4 below) than any of (1).

In the opposite direction, an e-to-p calibrator is a function transforming e-values to p-values. Formally, a decreasing function $f : [0, \infty] \to [0, 1]$ is an *e-to-p calibrator* if, for any probability space $(\Omega, \mathcal{A}, Q)$ and any e-variable $E \in \mathcal{E}_Q$, $f(E) \in \mathcal{P}_Q$. The following proposition, which is the analogue of Proposition 2.2 for e-to-p calibrators, says that there is, essentially, only one e-to-p calibrator, $f(t) := \min(1, 1/t)$.

Proposition 2.3. *The function $f : [0, \infty] \to [0, 1]$ defined by $f(t) := \min(1, 1/t)$ is an e-to-p calibrator. It dominates every other e-to-p calibrator. In particular, it is the only admissible e-to-p calibrator.*

Proof. The fact that $f(t) := \min(1, 1/t)$ is an e-to-p calibrator follows from Markov's inequality: if $E \in \mathcal{E}_Q$ and $\epsilon \in (0, 1)$,

$$Q(f(E) \le \epsilon) = Q(E \ge 1/\epsilon) \le \frac{\mathbb{E}^Q[E]}{1/\epsilon} \le \epsilon.$$

On the other hand, suppose that f is another e-to-p calibrator. It suffices to check that f is dominated by $\min(1, 1/t)$. Suppose $f(t) < \min(1, 1/t)$ for some $t \in [0, \infty]$. Consider two cases:

- If $f(t) < \min(1, 1/t) = 1/t$ for some $t > 1$, fix such t and consider an e-variable E that is t with probability $1/t$ and 0 otherwise. Then $f(E)$ is $f(t) < 1/t$ with probability $1/t$, whereas it would have satisfied $P(f(E) \le f(t)) \le f(t) < 1/t$ had it been a p-variable.
- If $f(t) < \min(1, 1/t) = 1$ for some $t \in [0, 1]$, fix such t and consider an e-variable E that is 1 a.s. Then $f(E)$ is $f(t) < 1$ a.s., and so it is not a p-variable. □

Proposition 2.2 implies that the domination structure of calibrators is very rich, whereas Proposition 2.3 implies that the domination structure of e-to-p calibrators is trivial.

Remark 2.4. A possible interpretation of this section's results is that e-variables and p-variables are connected via a rough relation $1/e \sim p$. In one direction, the statement is precise: the reciprocal (truncated to 1 if needed) of an e-variable

is a p-variable by Proposition 2.3. On the other hand, using a calibrator (1) with a small $\kappa > 0$ and ignoring positive constant factors (as customary in the algorithmic theory of randomness, discussed in Section A.2), we can see that the reciprocal of a p-variable is approximately an e-variable. In fact, $f(p) \le 1/p$ for all p when f is a calibrator; this follows from Proposition 2.2. However, $f(p) = 1/p$ is only possible in the extreme case $f = 1_{[0,p]}/p$.

3 Merging e-values

An important advantage of e-values over p-values is that they are easy to combine. This is the topic of this section, in which we consider the general case, without any assumptions on the joint distribution of the input e-variables. The case of independent e-variables is considered in the next section.

Let $K \ge 2$ be a positive integer (fixed throughout the paper apart from Section 7). An *e-merging function* of K e-values is an increasing Borel function $F : [0,\infty]^K \to [0,\infty]$ such that, for any probability space $(\Omega, \mathcal{A}, Q)$ and random variables $E_1, \dots, E_K$ on it,

$$E_1, \dots, E_K \in \mathcal{E}_Q \implies F(E_1, \dots, E_K) \in \mathcal{E}_Q \tag{3}$$

(in other words, F transforms e-values into an e-value). In this paper we will also refer to increasing Borel functions $F : [0,\infty)^K \to [0,\infty)$ satisfying (3) for all probability spaces and all e-variables $E_1, \dots, E_K$ taking values in $[0,\infty)$ as e-merging functions; such functions are canonically extended to e-merging functions $F : [0,\infty]^K \to [0,\infty]$ by setting them to ∞ on $[0,\infty]^K \setminus [0,\infty)^K$ (see Proposition C.1 in Appendix C).

An e-merging function F *dominates* an e-merging function G if $F \ge G$ (i.e., $F(\mathbf{e}) \ge G(\mathbf{e})$ for all $\mathbf{e} \in [0,\infty)^K$). The domination is *strict* (and we say that F *strictly dominates* G) if $F \ge G$ and $F(\mathbf{e}) > G(\mathbf{e})$ for some $\mathbf{e} \in [0,\infty)^K$. We say that an e-merging function F is *admissible* if it is not strictly dominated by any e-merging function; in other words, admissibility means being maximal in the partial order of domination.

A fundamental fact about admissibility is proved in Appendix E (Proposition E.5): any e-merging function is dominated by an admissible e-merging function.

Merging e-values via averaging

In this paper we are mostly interested in symmetric merging functions (i.e., those invariant w.r. to permutations of their arguments). The main message of this section is that the most useful (and the only useful, in a natural sense) symmetric e-merging function is the *arithmetic mean*

$$M_K(e_1, \dots, e_K) := \frac{e_1 + \dots + e_K}{K}, \qquad e_1, \dots, e_K \in [0,\infty). \tag{4}$$

In Theorem 3.2 below we will see that M_K is admissible (this is also a consequence of Proposition 4.1). But first we state formally the vague claim that M_K is the only useful symmetric e-merging function.

An e-merging function F *essentially dominates* an e-merging function G if, for all $\mathbf{e} \in [0,\infty)^K$,

$$G(\mathbf{e}) > 1 \Longrightarrow F(\mathbf{e}) \geq G(\mathbf{e}).$$

This weakens the notion of domination in a natural way: now we require that F is not worse than G only in cases where G is not useless; we are not trying to compare degrees of uselessness. The following proposition can be interpreted as saying that M_K is at least as good as any other symmetric e-merging function.

Proposition 3.1. *The arithmetic mean M_K essentially dominates any symmetric e-merging function.*

In particular, if F is an e-merging function that is symmetric and positively homogeneous (i.e., $F(\lambda \mathbf{e}) = \lambda F(\mathbf{e})$ for all $\lambda > 0$), then F is dominated by M_K. This includes the e-merging functions discussed later in Section 6.

Proof of Proposition 3.1. Let F be a symmetric e-merging function. Suppose for the purpose of contradiction that there exists $(e_1, \dots, e_K) \in [0,\infty)^K$ such that

$$b := F(e_1, \dots, e_K) > \max\left(\frac{e_1 + \dots + e_K}{K}, 1\right) =: a. \tag{5}$$

Let Π_K be the set of all permutations of $\{1, \dots, K\}$, π be randomly and uniformly drawn from Π_K, and $(D_1, \dots, D_K) := (e_{\pi(1)}, \dots, e_{\pi(K)})$. Further, let $(D'_1, \dots, D'_K) := (D_1, \dots, D_K)1_A$, where A is an event independent of π and satisfying $P(A) = 1/a$ (the existence of such random π and A is guaranteed for any atomless probability space by Lemma D.1 in Appendix D).

For each k, since D_k takes the values $e_1, \dots, e_K$ with equal probability, we have $\mathbb{E}[D_k] = (e_1 + \dots + e_K)/K$, which implies $\mathbb{E}[D'_k] = (e_1 + \dots + e_K)/(Ka) \leq 1$. Together with the fact that D'_k is nonnegative, we know $D'_k \in \mathcal{E}_Q$. Moreover, by symmetry,

$$\mathbb{E}[F(D'_1, \dots, D'_K)] = Q(A)F(e_1, \dots, e_K) + (1 - Q(A))F(0, \dots, 0) \geq b/a > 1,$$

a contradiction. Therefore, we conclude that there is no $(e_1, \dots, e_K)$ such that (5) holds. □

It is clear that the arithmetic mean M_K does not dominate every symmetric e-merging function; for example, the convex mixtures

$$\lambda + (1 - \lambda)M_K, \quad \lambda \in [0, 1], \tag{6}$$

of the trivial e-merging function 1 and M_K are pairwise non-comparable (with respect to the relation of domination). In the theorem below, we show that each of these mixtures is admissible and that the class (6) is, in the terminology of statistical decision theory [Wald, 1950, Section 1.3], a complete class of symmetric e-merging functions: every symmetric e-merging function is dominated

by one of (6). In other words, (6) is the minimal complete class of symmetric e-merging functions.

Theorem 3.2. *Suppose that F is a symmetric e-merging function. Then F is dominated by the function $\lambda + (1-\lambda)M_K$ for some $\lambda \in [0,1]$. In particular, F is admissible if and only if $F = \lambda + (1-\lambda)M_K$, where $\lambda = F(\mathbf{0}) \in [0,1]$.*

The proof of Theorem 3.2 is put in Appendix E as it requires several other technical results in the appendix. Finally, we note that, for $\lambda \neq 1$, the functions in the class (6) carry the same statistical information.

4 Merging independent e-values

In this section we consider merging functions for independent e-values. An *ie-merging function* of K e-values is an increasing Borel function $F : [0,\infty)^K \to [0,\infty)$ such that $F(E_1,\dots,E_K) \in \mathcal{E}_Q$ for all independent $E_1,\dots,E_K \in \mathcal{E}_Q$ in any probability space $(\Omega, \mathcal{A}, Q)$. As for e-merging functions, this definition is essentially equivalent to the definition involving $[0,\infty]$ rather than $[0,\infty)$ (by Proposition C.1 in Appendix C, which is still applicable in the context of merging independent e-values). The definitions of domination, strict domination, and admissibility are obtained from the definitions of the previous section by replacing "e-merging" with "ie-merging".

Let $i\mathcal{E}_Q^K \subseteq \mathcal{E}_Q^K$ be the set of (component-wise) independent random vectors in $\mathcal{E}_Q^K$, and $\mathbf{1} := (1,\dots,1)$ be the all-1 vector in $\mathbb{R}^K$. The following proposition has already been used in Section 3 (in particular, it implies that the arithmetic mean M_K is an admissible e-merging function).

Proposition 4.1. *For an increasing Borel function $F : [0,\infty)^K \to [0,\infty)$, if $\mathbb{E}[F(\mathbf{E})] = 1$ for all $\mathbf{E} \in \mathcal{E}_Q^K$ with $\mathbb{E}[\mathbf{E}] = \mathbf{1}$ (resp., for all $\mathbf{E} \in i\mathcal{E}_Q^K$ with $\mathbb{E}[\mathbf{E}] = \mathbf{1}$), then F is an admissible e-merging function (resp., an admissible ie-merging function).*

Proof. It is obvious that F is an e-merging function (resp., ie-merging function). Next we show that F is admissible. Suppose for the purpose of contradiction that there exists an ie-merging function G such that $G \geq F$ and $G(e_1,\dots,e_K) > F(e_1,\dots,e_K)$ for some $(e_1,\dots,e_K) \in [0,\infty)^K$. Take $(E_1,\dots,E_K) \in i\mathcal{E}_Q^K$ with $\mathbb{E}[(E_1,\dots,E_K)] = \mathbf{1}$ such that $Q((E_1,\dots,E_K) = (e_1,\dots,e_K)) > 0$. Such a random vector is easy to construct by considering any distribution with a positive mass on each of $e_1,\dots,e_K$. Then we have

$$Q(G(E_1,\dots,E_K) > F(E_1,\dots,E_K)) > 0,$$

which implies

$$\mathbb{E}[G(E_1,\dots,E_K)] > \mathbb{E}[G(E_1,\dots,E_K)] = 1,$$

contradicting the assumption that G is an ie-merging function. Therefore, no ie-merging function strictly dominates F. Noting that an e-merging function is also an ie-merging function, admissibility of F is guaranteed under both settings. □

If $E_1, \dots, E_K$ are independent e-variables, their product $E_1 \dots E_K$ will also be an e-variable. This is the analogue of Fisher's [1932] method for p-values (according to the rough relation $e \sim 1/p$ mentioned in Remark 2.4; Fisher's method is discussed at the end of Section 6). The ie-merging function

$$(e_1, \dots, e_K) \mapsto e_1 \dots e_K \tag{7}$$

is admissible by Proposition 4.1. It will be referred to as the *product* (or *multiplication*) ie-merging function. The betting interpretation of (7) is obvious: it is the result of K successive bets using the e-variables $E_1, \dots, E_K$ (starting with initial capital 1 and betting the full current capital $E_1 \dots E_{k-1}$ on each E_k).

More generally, we can see that the U-statistics

$$U_n(e_1, \dots, e_K) := \frac{1}{\binom{K}{n}} \sum_{\{k_1, \dots, k_n\} \subseteq \{1, \dots, K\}} e_{k_1} \dots e_{k_n}, \quad n \in \{0, 1, \dots, K\}, \tag{8}$$

and their convex mixtures are ie-merging functions. Notice that this class includes product (for $n = K$), arithmetic average M_K (for $n = 1$), and constant 1 (for $n = 0$). Proposition 4.1 implies that the U-statistics (8) and their convex mixtures are admissible ie-merging functions.

The betting interpretation of a U-statistic (8) or a convex mixture of U-statistics is implied by the betting interpretation of each component $e_{k_1} \dots e_{k_n}$. Assuming that $k_1, \dots, k_n$ are sorted in the increasing order, $e_{k_1} \dots e_{k_n}$ is the result of n successive bets using the e-variables $E_{k_1}, \dots, E_{k_n}$; and a convex mixture of bets corresponds to investing the appropriate fractions of the initial capital into those bets.

Let us now establish a very weak counterpart of Proposition 3.1 for independent e-values (on the positive side it will not require the assumption of symmetry). An ie-merging function F *weakly dominates* an ie-merging function G if, for all $e_1, \dots, e_K$,

$$(e_1, \dots, e_K) \in [1, \infty)^K \implies F(e_1, \dots, e_K) \geq G(e_1, \dots, e_K).$$

In other words, we require that F is not worse than G if all input e-values are useful (and this requirement is weak because, especially for a large K, we are also interested in the case where some of the input e-values are useless).

Proposition 4.2. *The product $(e_1, \dots, e_K) \mapsto e_1 \dots e_K$ weakly dominates any ie-merging function.*

Proof. Indeed, suppose that there exists $(e_1, \dots, e_K) \in [1, \infty)^K$ such that

$$F(e_1, \dots, e_K) > e_1 \dots e_K.$$

Let $E_1, \dots, E_K$ be independent random variables such that each E_k for $k \in \{1, \dots, K\}$ takes values in the two-element set $\{0, e_k\}$ and $E_k = e_k$ with probability $1/e_k$. Then each E_k is an e-variable but

$$\begin{aligned} \mathbb{E}[F(E_1, \dots, E_K)] &\geq F(e_1, \dots, e_K) Q(E_1 = e_1, \dots, E_K = e_K) \\ &> e_1 \dots e_K (1/e_1) \dots (1/e_K) = 1, \end{aligned}$$

which contradicts F being an ie-merging function. □

Remark 4.3. A natural question is whether the convex mixtures of (8) form a complete class. They do not: Proposition 4.1 implies that

$$f(e_1, e_2) := \frac{1}{2}\left(\frac{e_1}{1+e_1} + \frac{e_2}{1+e_2}\right)(1 + e_1 e_2)$$

is an admissible ie-merging function, and it is easy to check that it is different from any convex mixture of (8).

Testing with martingales

The assumption of the independence of e-variables $E_1, \dots, E_K$ is not necessary for the product $E_1 \dots E_K$ to be an e-variable. Below, we say that the e-variables $E_1, \dots, E_K$ are *sequential* if $\mathbb{E}[E_k \mid E_1, \dots, E_{k-1}] \le 1$ almost surely for all $k \in \{1, \dots, K\}$. Equivalently, the sequence of the partial products $(E_1 \dots E_k)_{k=0,1,\dots,K}$ is a supermartingale in the filtration generated by $E_1, \dots, E_K$ (or a *test supermartingale*, in the terminology of Shafer et al. [2011], Howard et al. [2020b], and Grünwald et al. [2020], meaning a nonnegative supermartingale with initial value 1). A possible interpretation of this test supermartingale is that the e-values $e_1, e_2, \dots$ are obtained by laboratories $1, 2, \dots$ in this order, and laboratory k makes sure that its result e_k is a valid e-value given the previous results $e_1, \dots, e_{k-1}$. The test supermartingale is a *test martingale* if $\mathbb{E}[E_k \mid E_1, \dots, E_{k-1}] = 1$ almost surely for all k (intuitively, it is not wasteful).

It is straightforward to check that all convex mixtures of (8) (including the product function) produce a valid e-value from sequential e-values; we will say that they are *se-merging functions*. On the other hand, independent e-variables are sequential, and hence se-merging functions form a subset of ie-merging functions. In the class of se-merging functions, the convex mixtures of (8) are admissible, as they are admissible in the larger class of ie-merging functions (by Proposition 4.1). For the same reason (and by Proposition 4.2), the product function in (7) weakly dominates every other se-merging function. This gives a (weak) theoretical justification for us to use the product function as a canonical merging method in Sections 5 and 7 for e-values as long as they are sequential. Finally, we note that it suffices for $E_1, \dots, E_K$ to be sequential in any order for these merging methods (such as Algorithm 2 in Section 5) to be valid.

5 Application to testing multiple hypotheses

As in Vovk and Wang [2019a], we will apply results for multiple testing of a single hypothesis (combining e-values in the context of Sections 3 and 4) to testing multiple hypotheses. As we explain in Appendix A (Section A.3), our algorithms just spell out the application of the closure principle [Marcus et al., 1976; Goeman and Solari, 2011], but our exposition in this section will be self-contained.

Algorithm 1 Adjusting e-values for multiple hypothesis testing

Require: A sequence of e-values $e_1, \dots, e_K$.

Find a permutation $\pi : \{1, \dots, K\} \to \{1, \dots, K\}$ such that $e_{\pi(1)} \le \dots \le e_{\pi(K)}$.
Set $e_{(k)} := e_{\pi(k)}$, $k \in \{1, \dots, K\}$ (these are the *order statistics*).
$S_0 := 0$
for $i = 1, \dots, K$ **do**
    $S_i := S_{i-1} + e_{(i)}$
for $k = 1, \dots, K$ **do**
    $e^*_{\pi(k)} := e_{\pi(k)}$
    for $i = 1, \dots, k-1$ **do**
        $e := \frac{e_{\pi(k)} + S_i}{i+1}$
        if $e < e^*_{\pi(k)}$ **then**
            $e^*_{\pi(k)} := e$

Let $(\Omega, \mathcal{A})$ be our sample space (formally, a measurable space), and $\mathfrak{P}(\Omega)$ be the family of all probability measures on it. A *composite null hypothesis* is a set $H \subseteq \mathfrak{P}(\Omega)$ of probability measures on the sample space. We say that E is an *e-variable* w.r. to a composite null hypothesis H if $\mathbb{E}^Q[E] \le 1$ for any $Q \in H$.

In multiple hypothesis testing we are given a set of composite null hypotheses H_k, $k = 1, \dots, K$. Suppose that, for each k, we are also given an e-variable E_k w.r. to H_k. Our multiple testing procedure is presented as Algorithm 1. The procedure adjusts the e-values $e_1, \dots, e_K$, perhaps obtained in K experiments (not necessarily independent), to new e-values $e^*_1, \dots, e^*_K$; the adjustment is downward in that $e^*_k \le e_k$ for all k. Applying the procedure to the e-values $e_1, \dots, e_K$ produced by the e-variables $E_1, \dots, E_K$, we obtain extended random variables $E^*_1, \dots, E^*_K$ taking values $e^*_1, \dots, e^*_K$. The output $E^*_1, \dots, E^*_K$ of Algorithm 1 satisfies a property of validity which we will refer to as *family-wise validity* (FWV); in Section A.3 we will explain its analogy with the standard family-wise error rate (FWER).

A *conditional e-variable* is a family of extended nonnegative random variables E_Q, $Q \in \mathfrak{P}(\Omega)$, that satisfies

$$\forall Q \in \mathfrak{P}(\Omega) : \mathbb{E}^Q[E_Q] \le 1$$

(i.e., each E_Q is in $\mathcal{E}_Q$). We regard it as a system of bets against each potential data-generating distribution Q.

Extended random variables $E^*_1, \dots, E^*_K$ taking values in $[0, \infty]$ are *family-wise valid* (*FWV*) for testing $H_1, \dots, H_K$ if there exists a conditional e-variable $(E_Q)_{Q \in \mathfrak{P}(\Omega)}$ such that

$$\forall k \in \{1, \dots, K\} \, \forall Q \in H_k : E_Q \ge E^*_k \tag{9}$$

(where $E_Q \ge E^*_k$ means, as usual, that $E_Q(\omega) \ge E^*_k(\omega)$ for all $\omega \in \Omega$). We can say that such $(E_Q)_{Q \in \mathfrak{P}(\Omega)}$ *witnesses* the FWV property of $E^*_1, \dots, E^*_K$.

The interpretation of family-wise validity is based on our interpretation of e-values. Suppose we observe an outcome $\omega \in \Omega$. If $E_Q(\omega)$ is very large, we may reject Q as the data-generating distribution. Therefore, if $E_k^*(\omega)$ is very large, we may reject the whole of H_k (i.e., each $Q \in H_k$). In betting terms, we have made at least $\$E_k^*(\omega)$ risking at most $\$1$ when gambling against any $Q \in H_k$.

Notice that we can rewrite (9) as

$$\forall Q \in \mathfrak{P}(\Omega) : \mathbb{E}^Q \left[\max_{k : Q \in H_k} E_k^* \right] \le 1.$$

In other words, we require joint validity of the e-variables E_k^*.

We first state the validity of Algorithm 1 (as well as Algorithm 2 given below), and our justification follows.

Theorem 5.1. *Algorithms 1 and 2 are family-wise valid.*

Let us check that the output $E_1^*, \dots, E_K^*$ of Algorithm 1 is FWV. For $I \subseteq \{1, \dots, K\}$, the composite hypothesis H_I is defined by

$$H_I := \left(\bigcap_{k \in I} H_k \right) \bigcap \left(\bigcap_{k \in \{1,\dots,K\} \setminus I} H_k^c \right), \tag{10}$$

where H_k^c is the complement of H_k. The conditional e-variable witnessing that $E_1^*, \dots, E_K^*$ are FWV is the arithmetic mean

$$E_Q := \frac{1}{|I_Q|} \sum_{k \in I_Q} E_k, \tag{11}$$

where $I_Q := \{k \mid Q \in H_k\}$ and E_Q is defined arbitrarily (say, as 1) when $I_Q = \emptyset$. The optimal adjusted e-variables E_k' can be defined as

$$E_k' := \min_{Q \in H_k} E_Q \ge \min_{I \subseteq \{1,\dots,K\} : k \in I} \frac{1}{|I|} \sum_{i \in I} E_i, \tag{12}$$

but for computational efficiency we use the conservative definition

$$E_k^* := \min_{I \subseteq \{1,\dots,K\} : k \in I} \frac{1}{|I|} \sum_{i \in I} E_i. \tag{13}$$

Remark 5.2. The inequality "$\ge$" in (12) holds as the equality "$=$" if all the intersections (10) are non-empty. If some of these intersections are empty, we can have a strict inequality. Algorithm 1 implements the definition (13). Therefore, it is valid regardless of whether some of the intersections (10) are empty; however, if they are, it may be possible to improve the adjusted e-values. According to Holm's [1979] terminology, we allow "free combinations". Shaffer [1986] pioneered methods that take account of the logical relations between the base hypotheses H_k.

Algorithm 2 Adjusting sequential e-values for multiple hypothesis testing

Require: A sequence of e-values $e_1, \dots, e_K$.

1: Let a be the product of all $e_k < 1$, $k = 1, \dots, K$ (and $a := 1$ if there are no such k).
2: **for** $k = 1, \dots, K$ **do**
3: $\quad e_k^* := a e_k$

To obtain Algorithm 1, we rewrite the definitions (13) as

$$E^*_{\pi(k)} = \min_{i \in \{0,\dots,k-1\}} \frac{E_{\pi(k)} + E_{(1)} + \dots + E_{(i)}}{i+1} = \min_{i \in \{1,\dots,k-1\}} \frac{E_{\pi(k)} + E_{(1)} + \dots + E_{(i)}}{i+1}$$

for $k \in \{1, \dots, K\}$, where π is the ordering permutation and $E_{(j)} = E_{\pi(j)}$ is the jth order statistic among $E_1, \dots, E_K$, as in Algorithm 1. In lines 3–5 of Algorithm 1 we precompute the sums

$$S_i := e_{(1)} + \dots + e_{(i)}, \qquad i = 1, \dots, K,$$

in lines 8–9 we compute

$$e_{k,i} := \frac{e_{\pi(k)} + e_{(1)} + \dots + e_{(i)}}{i+1}$$

for $i = 1, \dots, k-1$, and as result of executing lines 6–11 we will have

$$e^*_{\pi(k)} = \min_{i \in \{1,\dots,k-1\}} e_{k,i} = \min_{i \in \{1,\dots,k-1\}} \frac{e_{\pi(k)} + e_{(1)} + \dots + e_{(i)}}{i+1},$$

which shows that Algorithm 1 is an implementation of (13).

The computational complexity of Algorithm 1 is $O(K^2)$.

In the case of sequential e-variables, we have Algorithm 2. This algorithm assumes that, under any $Q \in \mathfrak{P}(\Omega)$, the base e-variables E_k, $k \in I_Q$, are sequential (remember that I_Q is defined by (11) and that independence implies being sequential). The conditional e-variable witnessing that the output of Algorithm 2 is FWV is the one given by the product ie-merging function,

$$E_Q := \prod_{k \in I_Q} E_k,$$

where the adjusted e-variables are defined by

$$E^*_k := \min_{I \subseteq \{1,\dots,K\} : k \in I} \prod_{i \in I} E_i. \tag{14}$$

A remark similar to Remark 5.2 can also be made about Algorithm 2. The computational complexity of Algorithm 2 is $O(K)$ (unusually, the algorithm does not require sorting the base e-values).

6 Merging p-values and comparisons

Merging p-values is a much more difficult topic than merging e-values, but it is very well explored. First we review merging p-values without any assumptions, and then we move on to merging independent p-values.

A *p-merging function* of K p-values is an increasing Borel function $F : [0,1]^K \to [0,1]$ such that $F(P_1, \dots, P_K) \in \mathcal{P}_Q$ whenever $P_1, \dots, P_K \in \mathcal{P}_Q$.

For merging p-values without the assumption of independence, we will concentrate on two natural families of p-merging functions. The older family is the one introduced by Rüger [1978], and the newer one was introduced in our paper Vovk and Wang [2019a]. Rüger's family is parameterized by $k \in \{1, \dots, K\}$, and its kth element is the function (shown by Rüger [1978] to be a p-merging function)

$$(p_1, \dots, p_K) \mapsto \frac{K}{k} p_{(k)} \wedge 1, \tag{15}$$

where $p_{(k)} := p_{\pi(k)}$ and π is a permutation of $\{1, \dots, K\}$ ordering the p-values in the ascending order: $p_{\pi(1)} \le \dots \le p_{\pi(K)}$. The other family [Vovk and Wang, 2019a], which we will refer to as the *M-family*, is parameterized by $r \in [-\infty, \infty]$, and its element with index r has the form $a_{r,K} M_{r,K} \wedge 1$, where

$$M_{r,K}(p_1, \dots, p_K) := \left(\frac{p_1^r + \dots + p_K^r}{K} \right)^{1/r} \tag{16}$$

and $a_{r,K} \ge 1$ is a suitable constant. We also define $M_{r,K}$ for $r \in \{0, \infty, -\infty\}$ as the limiting cases of (16), which correspond to the geometric average, the maximum, and the minimum, respectively.

The initial and final elements of both families coincide: the initial element is the Bonferroni p-merging function

$$(p_1, \dots, p_K) \mapsto K \min(p_1, \dots, p_K) \wedge 1, \tag{17}$$

and the final element is the maximum p-merging function

$$(p_1, \dots, p_K) \mapsto \max(p_1, \dots, p_K).$$

Similarly to the case of e-merging functions, we say that a p-merging function F *dominates* a p-merging function G if $F \le G$. The domination is *strict* if, in addition, $F(\mathbf{p}) < G(\mathbf{p})$ for at least one $\mathbf{p} \in [0,1]^K$. We say that a p-merging function F is *admissible* if it is not strictly dominated by any p-merging function G.

The domination structure of p-merging functions is much richer than that of e-merging functions. The maximum p-merging function is clearly inadmissible (e.g., $(p_1, \dots, p_K) \mapsto \max(p_1, \dots, p_K)$ is strictly dominated by $(p_1, \dots, p_K) \mapsto p_1$) while the Bonferroni p-merging function is admissible, as the following proposition shows.

Proposition 6.1. *The Bonferroni p-merging function* (17) *is admissible.*

Proof. Denote by M_B the Bonferroni p-merging function (17). Suppose the statement of the proposition is false and fix a p-merging function F that strictly dominates M_B. If $F = M_B$ whenever $M_B < 1$, then $F = M_B$ also when $M_B = 1$, since F is increasing. Hence for some point $(p_1, \dots, p_K) \in [0,1]^K$,

$$F(p_1, \dots, p_K) < M_B(p_1, \dots, p_K) < 1.$$

Fix such $(p_1, \dots, p_K)$ and set $p := \min(p_1, \dots, p_K)$; we know that $Kp < 1$. Since

$$F(p, \dots, p) \le F(p_1, \dots, p_K) < M_B(p_1, \dots, p_K) = Kp,$$

we can take $\epsilon \in (0, p)$ such that $F(p, \dots, p) < K(p - \epsilon)$. Let $A_1, \dots, A_K, B$ be disjoint events such that $Q(A_k) = p - \epsilon$ for all k and $Q(B) = \epsilon$ (their existence is guaranteed by the inequality $Kp < 1$). Define random variables

$$U_k := \begin{cases} p - \epsilon & \text{if } A_k \text{ happens} \\ p & \text{if } B \text{ happens} \\ 1 & \text{otherwise,} \end{cases}$$

$k = 1, \dots, K$. It is straightforward to check that $U_1, \dots, U_K \in \mathcal{P}_Q$. By writing $F := F(U_1, \dots, U_K)$ and $M_B := M_B(U_1, \dots, U_K)$, we have

$$\begin{aligned} Q(F \le K(p-\epsilon)) &= Q(M_B \le K(p-\epsilon)) + Q(F \le K(p-\epsilon) < M_B) \\ &\ge Q(\min(U_1, \dots, U_K) \le p - \epsilon) + Q(U_1 = \dots = U_k = p) \\ &= Q\left(\bigcup_{k=1}^{K} A_k\right) + Q(B) = \sum_{k=1}^{K} Q(A_k) + \epsilon \\ &= K(p-\epsilon) + \epsilon > K(p-\epsilon). \end{aligned}$$

Therefore, F is not a p-merging function, which gives us the desired contradiction. □

The general domination structure of p-merging functions appears to be very complicated, and is the subject of Vovk et al. [2020].

Connections to e-merging functions

The domination structure of the class of e-merging functions is very simple, according to Theorem 3.2. It makes it very easy to understand what the e-merging analogues of Rüger's family and the M-family are; when stating the analogues we will use the rough relation $1/e \sim p$ between e-values and p-values (see Remark 2.4). Let us say that an e-merging function F is *precise* if cF is not an e-merging function for any $c > 1$.

For a sequence $e_1, \dots, e_K$, let $e_{[k]} := e_{\pi(k)}$ be the order statistics numbered from the largest to the smallest; here π is a permutation of $\{1, \dots, K\}$ ordering e_k in the descending order: $e_{\pi(1)} \ge \dots \ge e_{\pi(K)}$. Let us check that the Rüger-type function $(e_1, \dots, e_K) \mapsto (k/K)e_{[k]}$ is a precise e-merging function. It is an

e-merging function since it is dominated by the arithmetic mean: indeed, the condition of domination

$$\frac{k}{K} e_{[k]} \le \frac{e_1 + \cdots + e_K}{K}, \tag{18}$$

can be rewritten as

$$k e_{[k]} \le e_1 + \cdots + e_K$$

and so is obvious. As sometimes we have a strict inequality, the e-merging function is inadmissible (remember that we assume $K \ge 2$). The e-merging function is precise because (18) holds as equality when the k largest e_i, $i \in \{1, \dots, K\}$, are all equal and greater than 1 and all the other e_i are 0.

In the case of the M-family, let us check that the function

$$F := (K^{1/r-1} \wedge 1) M_{r,K} \tag{19}$$

is a precise e-merging function, for any $r \in [-\infty, \infty]$. For $r \le 1$, $M_{r,K}$ is increasing in r [Hardy et al., 1952, Theorem 16], and so $F = M_{r,K}$ is dominated by the arithmetic mean M_K; therefore, it is an e-merging function. For $r > 1$ we can rewrite the function $F = K^{1/r-1} M_{r,K}$ as

$$F(e_1, \dots, e_K) = K^{1/r-1} M_{r,K}(e_1, \dots, e_K) = K^{-1} \left(e_1^r + \cdots + e_K^r\right)^{1/r},$$

and we know that the last expression is a decreasing function of r [Hardy et al., 1952, Theorem 19]; therefore, F is also dominated by M_K and so is a merging function. The e-merging function F is precise (for any r) since

$$\begin{aligned} r \le 1 &\implies F(e, \dots, e) = M_K(e, \dots, e) = e \\ r > 1 &\implies F(0, \dots, 0, e) = M_K(0, \dots, 0, e) = e/K, \end{aligned}$$

and so by Proposition 3.1 (applied to a sufficiently large e) cF is not an e-merging function for any $c > 1$. But F is admissible if and only if $r = 1$ as shown by Theorem 3.2.

Remark 6.2. The rough relation $1/e \sim p$ also sheds light on the coefficient, $K^{1/r-1} \wedge 1 = K^{1/r-1}$ for $r > 1$, given in (19) in front of $M_{r,K}$. The coefficient $K^{1/r-1}$, $r > 1$, in front of $M_{r,K}$ for averaging e-values corresponds to a coefficient of $K^{1+1/r}$, $r < -1$, in front of $M_{r,K}$ for averaging p-values. And indeed, by Proposition 5 of Vovk and Wang [2019a], the asymptotically precise coefficient in front of $M_{r,K}$, $r < -1$, for averaging p-values is $\frac{r}{r+1} K^{1+1/r}$. The extra factor $\frac{r}{r+1}$ appears because the reciprocal of a p-variable is only approximately, but not exactly, an e-variable.

Remark 6.3. Our formulas for merging e-values are explicit and much simpler than the formulas for merging p-values given in Vovk and Wang [2019a], where the coefficient $a_{r,K}$ is often not analytically available. Merging e-values does not involve asymptotic approximations via the theory of robust risk aggregation (e.g., Embrechts et al. [2015]), as used in that paper. This suggests that in some important respects e-values are easier objects to deal with than p-values.

Merging independent p-values

In this section we will discuss ways of combining p-values $p_1, \dots, p_K$ under the assumption that the p-values are independent.

One of the oldest and most popular methods for combining p-values is Fisher's [1932, Section 21.1], which we already mentioned in Section 4. Fisher's method is based on the product statistic $p_1 \dots p_K$ (with its low values significant) and uses the fact that $-2\ln(p_1 \dots p_K)$ has the χ^2 distribution with $2K$ degrees of freedom when p_k are all independent and distributed uniformly on the interval $[0, 1]$; the p-values are the tails of the χ^2 distribution.

Simes [1986] proves a remarkable result for Rüger's family (15) under the assumption that the p-values are independent: the minimum

$$(p_1, \dots, p_K) \mapsto \min_{k \in \{1,\dots,K\}} \frac{K}{k} p_{(k)} \tag{20}$$

of Rüger's family over all k turns out to be a p-merging function. The counterpart of Simes's result still holds for e-merging functions; moreover, now the input e-values do not have to be independent. Namely,

$$(e_1, \dots, e_K) \mapsto \max_{k \in \{1,\dots,K\}} \frac{k}{K} e_{[k]}$$

is an e-merging function. This follows immediately from (18), the left-hand side of which can be replaced by its maximum over k. And it also follows from (18) that there is no sense in using this counterpart; it is better to use the arithmetic mean.

7 Experimental results

In this section we will explore the performance of various methods of combining e-values and p-values and multiple hypothesis testing, both standard and introduced in this paper. For our code, see Vovk and Wang [2020c].

In order to be able to judge how significant results of testing using e-values are, Jeffreys's [1961, Appendix B] rule of thumb may be useful:

- If the resulting e-value e is below 1, the null hypothesis is supported.
- If $e \in (1, \sqrt{10}) \approx (1, 3.16)$, the evidence against the null hypothesis is not worth more than a bare mention.
- If $e \in (\sqrt{10}, 10) \approx (3.16, 10)$, the evidence against the null hypothesis is substantial.
- If $e \in (10, 10^{3/2}) \approx (10, 31.6)$, the evidence against the null hypothesis is strong.
- If $e \in (10^{3/2}, 100) \approx (31.6, 100)$, the evidence against the null hypothesis is very strong.

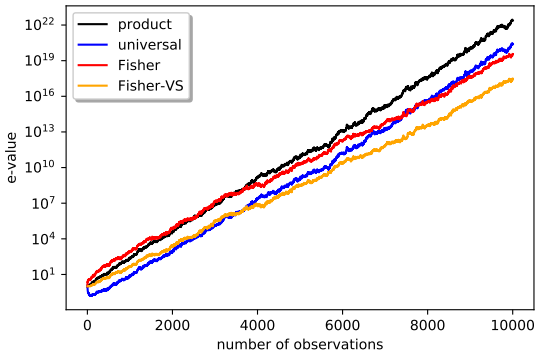


Figure 1: Combining p-values using Fisher's method vs combining e-values by multiplication (details in text).

- If $e > 100$, the evidence against the null hypothesis is decisive.

Our discussions in this section assume that our main interest is in e-values, and p-values are just a possible tool for obtaining good e-values (which is, e.g., the case for Bayesian statisticians in their attitude towards Bayes factors and p-values; cf. Section A.1 and Appendix B). Our conclusions would have been different had our goal been to obtain good p-values.

Combining independent e-values and p-values

First we explore combining independent e-values and independent p-values; see Figure 1. The observations are generated from the Gaussian model $N(\mu, 1)$ with standard deviation 1 and unknown mean μ. The null hypothesis is $\mu = 0$ and the alternative hypothesis is $\mu = \delta$; for Figures 1 and 2 we set $\delta := -0.1$. The observations are IID. Therefore, one observation does not carry much information about which hypothesis is true, but repeated observations quickly reveal the truth (with a high probability).

For Figures 1 and 2, all data (10,000 or 1000 observations, respectively) are generated from the alternative distribution (there will be an example where some of the data is coming from the null distribution in Appendix H). For each observation, the e-value used for testing is the likelihood ratio

$$E(x) := e^{-(x-\delta)^2/2} / e^{-x^2/2} = e^{x\delta - \delta^2/2} \tag{21}$$

of the alternative probability density to the null probability density, where x is the observation. It is clear that (21) is indeed an e-variable under the null hypothesis: its expected value is 1. As the p-value we take

$$P(x) := N(x), \tag{22}$$

where N is the standard Gaussian distribution function; in other words, the p-value is found using the most powerful test, namely the likelihood ratio test given by the Neyman–Pearson lemma.

In Figure 1 we give the results for the product e-merging function (7) and Fisher's method described in the last subsection of Section 6. (The other methods that we consider are vastly less efficient, and we show them in the following figure, Figure 2.) Three of the values plotted in Figure 1 against each $K = 1, \dots, 10{,}000$ are:

- the product e-value $E(x_1) \dots E(x_K)$; it is shown as the black line;
- the reciprocal $1/p$ of Fisher's p-value p obtained by merging the first K p-values $P(x_1), \dots, P(x_K)$; it is shown as the red line;
- the VS bound applied to Fisher's p-value; it is shown as the orange line.

The plot depends very much on the seed for the random number generator, and so we report the median of all values over 100 seeds.

The line for the product method is below that for Fisher's over the first 2000 observations but then it catches up. If our goal is to have an overall e-value summarizing the results of testing based on the first K observations (as we always assume in this section), the comparison is unfair, since Fisher's p-values need to be calibrated. A fairer (albeit still unfair) comparison is with the VS bound, and the curve for the product method can be seen to be above the curve for the VS bound. *A fortiori*, the curve for the product method would be above the curve for any of the calibrators in the family (1).

It is important to emphasize that the natures of plots for e-values and p-values are very different. For the red and orange lines in Figure 1, the values shown for different K relate to different batches of data and cannot be regarded as a trajectory of a natural stochastic process. In contrast, the values shown by the black line for different K are updated sequentially, the value at K being equal to the value at $K-1$ multiplied by $E(x_K)$, and form a trajectory of a test martingale. Moreover, for the black line we do not need the full force of the assumption of independence of the p-values. As we discuss at the end of Section 4, it is sufficient to assume that $E(x_K)$ is a valid e-value given $x_1, \dots, x_{K-1}$; the black line in Figure 1 is then still a trajectory of a test supermartingale.

What we said in the previous paragraph can be regarded as an advantage of using e-values. On the negative side, computing good (or even optimal in some sense) e-values often requires more detailed knowledge. For example, whereas computing the e-value (21) requires the knowledge of the alternative hypothesis, for computing the p-value (22) it is sufficient to know that the alternative hypothesis corresponds to $\mu < 0$. Getting μ very wrong will hurt the performance of methods based on e-values. To get rid of the dependence on μ, we can, e.g., integrate the product e-value over $\delta \sim N(0,1)$ (taking the standard deviation of 1 is somewhat wasteful in this situation, but we take the most standard probability measure). This gives the "universal" test martingale

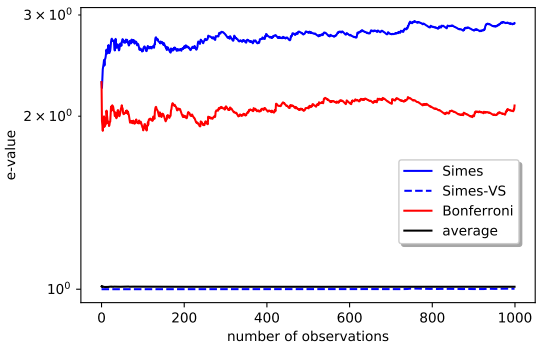


Figure 2: Combining p-values using Simes's and Bonferroni's methods and combining e-values using averaging (details in text).

(see e.g., Howard et al. [2020b])

$$S_K := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\delta^2/2) \prod_{k=1}^{K} \exp(x_k \delta - \delta^2/2) \, \mathrm{d}\delta$$
$$= \frac{1}{\sqrt{K+1}} \exp\left(\frac{1}{2(K+1)} \left(\sum_{k=1}^{K} x_k \right)^2 \right). \tag{23}$$

This test supermartingale is shown in blue in Figure 1. It is below the black line but at the end of the period it catches up even with the line for Fisher's method (and beyond that period it overtakes Fisher's method more and more convincingly).

Arithmetic average (4) and Simes's method (20) have very little power in the situation of Figure 1: see Figure 2, which plots the e-values produced by the averaging method, the reciprocals $1/p$ of Simes's p-values p, the VS bound for Simes's p-values, and the reciprocals of the Bonferroni p-values over 1000 observations, all averaged (in the sense of median) over 1000 seeds. They are very far from attaining statistical significance (a p-value of 5% or less) or collecting substantial evidence against the null hypothesis (an e-value of $\sqrt{10}$ or more according to Jeffreys).

Multiple hypothesis testing

Next we discuss multiple hypothesis testing. Figure 3 shows plots of adjusted e-values and adjusted p-values resulting from various methods for small numbers of hypotheses, including Algorithms 1 and 2. The observations are again generated from the statistical model $N(\mu, 1)$.

We are testing 20 null hypotheses. All of them are $\mu = 0$, and their alternatives are $\mu = -4$. Each null hypothesis is tested given an observation drawn

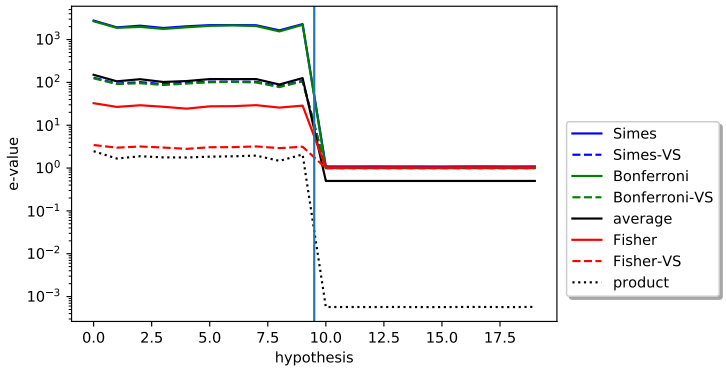


Figure 3: Multiple hypothesis testing for 20 hypotheses using p-values and e-values, with some graphs indistinguishable (details in text).

either from the null or from the alternative. The first 10 null hypotheses are false, and in fact the corresponding observations are drawn from the alternative distribution. The remaining 10 null hypotheses are true, and the corresponding observations are drawn from them rather than the alternatives. The vertical blue line at the centre of Figure 3 separates the false null hypotheses from the true ones: null hypotheses 0 to 9 are false and 10 to 19 are true. We can see that at least some of the methods can detect that the first 10 null hypotheses are false.

Since some of the lines are difficult to tell apart, we will describe the plot in words. The top two horizontal lines to the left of the vertical blue line are indistinguishable but are those labeled as Simes and Bonferroni in the legend; they correspond to e-values around 2×10^3. The following cluster of horizontal lines to the left of the vertical blue line (with e-values around 10^2) are those labeled as average, Simes-VS, and Bonferroni-VS, with average slightly higher. To the right of the vertical blue line, the upper horizontal lines (with e-values 10^0) include all methods except for average and product; the last two are visible.

Most of the methods (all except for Bonferroni and Algorithm 1) require the observations to be independent. The base p-values are (22), and the base e-values are the likelihood ratios

$$E(x) := \frac{1}{2} e^{x\delta - \delta^2/2} + \frac{1}{2} \tag{24}$$

(cf. (21)) of the "true" probability density to the null probability density, where the former assumes that the null or alternative distribution for each observation is decided by coin tossing. Therefore, the knowledge encoded in the "true" distribution is that half of the observations are generated from the alternative distribution, but it is not known that these observations are in the first half. We set $\delta := -4$ in (24), keeping in mind that accurate prior knowledge is essential for the efficiency of methods based on e-values.

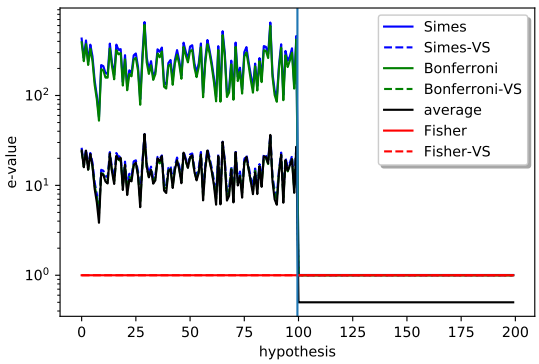


Figure 4: The analogue of Figure 3 without the product method, with 200 observations, and with some graphs indistinguishable (details in text).

A standard way of producing multiple testing procedures is applying the closure principle described in Appendix A and already implicitly applied in Section 5 to methods of merging e-values. In Figure 3 we report the results for the closures of five methods, three of them producing p-values (Simes's, Bonferroni's, and Fisher's) and two producing e-values (average and product); see Section 5 for self-contained descriptions of the last two methods (Algorithms 1 and 2). For the methods producing p-values we show the reciprocals $1/p$ of the resulting p-values p (as solid lines) and the corresponding VS bounds (as dashed lines). For the closure of Simes's method we follow the appendix of Wright [1992], the closure of Bonferroni's method is described in Holm [1979] (albeit not in terms of adjusted p-values), and for the closure of Fisher's method we use Dobriban's [2020] FACT (FAst Closed Testing) procedure. To make the plot more regular, all values are averaged (in the sense of median) over 1000 seeds of the Numpy random number generator.

According to Figure 3, the performance of Simes's and Bonferroni's methods is very similar, despite Bonferroni's method not depending on the assumption of independence of the p-values. The e-merging method of averaging (i.e., Algorithm 1) produces better e-values than those obtained by calibrating the closures of Simes's and Bonferroni's methods; remember that the line corresponding to Algorithm 1 should be compared with the VS versions (blue and green dashed, which almost coincide) of the lines corresponding to the closures of Simes's and Bonferroni's methods, and even that comparison is unfair and works in favour of those two methods (since the VS bound is not a valid calibrator). The other algorithms perform poorly.

Figure 4 is an analogue of Figure 3 that does not show results for merging by multiplication (for large numbers of hypotheses its results are so poor that, when shown, differences between the other methods become difficult to see). To get more regular and comparable graphs, we use averaging (in the sense of

median) over 100 seeds.

Since some of the graphs coincide, or almost coincide, we will again describe the plot in words (referring to graphs that are straight or almost straight as lines). To the left of the vertical blue line (separating the false null hypotheses 0–99 from the true null hypotheses 100–199) we have three groups of graphs: the top graphs (with e-values around 2×10^2) are those labeled as Simes and Bonferroni in the legend, the middle graphs (with e-values around 10^1) are those labeled as average, Simes-VS, and Bonferroni-VS, and the bottom lines (with e-values around 10^0) are those labeled as Fisher and Fisher-VS. To the right of the vertical blue line, we have two groups of lines: the upper lines (with e-values 10^0) include all methods except for average, which is visible.

Now the graph for the averaging method (Algorithm 1) is very close to (barely distinguishable from) the graphs for the VS versions of the closures of Simes's and Bonferroni's methods, which is a very good result (in terms of the quality of e-values that we achieve): the VS bound is a bound on what can be achieved whereas the averaging method produces a bona fide e-value. The two lines (solid and dotted) for Fisher's method are indistinguishable from the horizontal axis; the method does not scale up in our experiments (which is a known phenomenon in the context of p-values: see, e.g., Westfall [2011, Section 1]). And the four blue and green lines (solid and dotted) for Simes's and Bonferroni's methods are not visible to the right of 100 since they are covered by the lines for Fisher's method. The behaviour of the lines for Simes's, Bonferroni's, and Fisher's methods to the right of 100 demonstrates that they do not produce valid e-values: for validity, we have to pay by getting e-values below 1 when the null hypothesis is true in order to be able to get large e-values when the null hypothesis is false (which is the case for the averaging method, represented by the black line). Most of these remarks are also applicable to Figure 3.

A key advantage of the averaging and Bonferroni's methods over Simes's and Fisher's is that they are valid regardless of whether the base e-values or p-values are independent.

8 Conclusion

This paper systematically explores the notion of an e-value, which can be regarded as a betting counterpart of p-values that is much more closely related to Bayes factors and likelihood ratios. We argue that e-values often are more mathematically convenient than p-values and lead to simpler results. In particular, they are easier to combine: the average of e-values is an e-value, and the product of independent e-values is an e-value. We apply e-values in two areas, multiple testing of a single hypothesis and testing multiple hypotheses, and obtain promising experimental results. One of our experimental findings is that, for testing multiple hypotheses, the performance of the most natural method based on e-values almost attains the Vovk–Sellke bound for the closure of Simes's method, despite that bound being overoptimistic and not producing

bona fide e-values.

Acknowledgments

The authors thank Aaditya Ramdas, Alexander Schied, and Glenn Shafer for helpful suggestions. Thoughtful comments by the associate editor and four reviewers of the journal version have led to numerous improvements in presentation and substance.

V. Vovk's research has been partially supported by Amazon, Astra Zeneca, and Stena Line. R. Wang is supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-03823, RGPAS-2018-522590).

References

Daniel J. Benjamin and James O. Berger. Three recommendations for improving the use of p-values. *American Statistician*, 73(Supplement 1):186–191, 2019.

James O. Berger and Mohan Delampady. Testing precise hypotheses (with discussion). *Statistical Science*, 2:317–352, 1987.

Edgar Dobriban. FACT: Fast closed testing for exchangeable local tests. *Biometrika*, 107:761–768, 2020.

Joseph L. Doob. *Stochastic Processes*. Wiley, New York, 1953.

Boyan Duan, Aaditya Ramdas, Sivaraman Balakrishnan, and Larry Wasserman. Interactive martingale tests for the global null. Technical Report arXiv:1909.07339 [stat.ME], arXiv.org e-Print archive, December 2019.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1 of *IMS Monographs*. Cambridge University Press, Cambridge, 2010.

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

Alexander Etz and Eric-Jan Wagenmakers. J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, 32:313–329, 2017.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, fourth edition, 1932. Section 21.1 on combining independent p-values first appears in this edition and is present in all subsequent editions.

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

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

Jelle J. Goeman, Jonathan D. Rosenblatt, and Thomas E. Nichols. The harmonic mean p-value: Strong versus weak control, and the assumption of independence. *Proceedings of the National Academy of Sciences*, 116:23382–23383, 2019.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. Technical Report arXiv:1906.07801 [math.ST], arXiv.org e-Print archive, June 2020.

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

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

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020a.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. Technical Report arXiv:1810.08240 [math.ST], arXiv.org e-Print archive, May 2020b. To appear in the *Annals of Statistics*.

Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, third edition, 1961.

Kaniav Kamary, Kerrie Mengersen, Christian P. Robert, and Judith Rousseau. Testing hypotheses via a mixture estimation model. Technical Report arXiv:1412.2044 [stat.ME], arXiv.org e-Print archive, December 2014.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

Alexander S. Kechris. *Classical Descriptive Set Theory*. Springer, New York, 1995.

Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.

Andrei N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, IT-14:662–664, 1968.

Leonid A. Levin. Uniform tests of randomness. *Soviet Mathematics Doklady*, 17:337–340, 1976.

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.

Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

Laurent Mazliak and Glenn Shafer. The splendors and miseries of martingales, June 2009. Special Issue of the *Electronic Journal for History of Probability and Statistics* **5**(1) (ed. by Laurent Mazliak and Glenn Shafer), `www.jehps.net`.

Frank W. J. Olver, A. B. Olde Daalhuis, Daniel W. Lozier, Barry I. Schneider, Ronald F. Boisvert, Charles W. Clark, Bruce R. Miller, Bonita V. Saunders, Howard S. Cohl, and Marjorie A. McClain, editors. *NIST Digital Library of Mathematical Functions*. 2020. URL `http://dlmf.nist.gov/`. Release 1.0.28 of 2020-09-15.

Ulrich Rieder. Measurable selection theorems for optimization problems. *Manuscripta Mathematica*, 24:115–131, 1978.

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.

Thomas Sellke, M. J. Bayarri, and James Berger. Calibration of p-values for testing precise null hypotheses. *American Statistician*, 55:62–71, 2001.

Glenn Shafer. The language of betting as a strategy for statistical and scientific communication. Technical Report arXiv:1903.06991 [math.ST], arXiv.org e-Print archive, March 2019. To appear as discussion paper in the *Journal of the Royal Statistical Society A* (read in September 2020).

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ, 2019.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors, and p-values. *Statistical Science*, 26:84–101, 2011.

Juliet P. Shaffer. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81:826–831, 1986.

Alexander Shen, Vladimir A. Uspensky, and Nikolai Vereshchagin. *Kolmogorov Complexity and Algorithmic Randomness*. American Mathematical Society, Providence, RI, 2017.

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

John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 2003.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk. A logic of probability, with application to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society B*, 55: 317–351, 1993.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. Technical Report arXiv:1212.4966 [math.ST], arXiv.org e-Print archive, October 2019a. To appear in *Biometrika*.

Vladimir Vovk and Ruodu Wang. True and false discoveries with e-values. Technical Report arXiv:1912.13292 [math.ST], arXiv.org e-Print archive, December 2019b.

Vladimir Vovk and Ruodu Wang. True and false discoveries with independent e-values. Technical Report arXiv:2003.00593 [stat.ME], arXiv.org e-Print archive, March 2020a.

Vladimir Vovk and Ruodu Wang. A class of ie-merging functions. Technical Report arXiv:2003.06382 [math.ST], arXiv.org e-Print archive, July 2020b.

Vladimir Vovk and Ruodu Wang. Python code generating the figures in this paper. Jupyter notebook. Go to the abstract of this paper on arXiv, click on "v3", click on "Other formats", click on "Download source", and extract the Jupyter notebook from the downloaded archive (perhaps adding `.gz` as its extension), May 2020c.

Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. Technical Report arXiv:2007.14208 [math.ST], arXiv.org e-Print archive, July 2020.

Abraham Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences of the USA*, 117: 16880–16890, 2020.

Peter H. Westfall. Discussion of "Multiple testing for exploratory research" by J. J. Goeman and A. Solari. *Statistical Science*, 26:604–607, 2011.

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

S. Paul Wright. Adjusted p-values for simultaneous inference. *Biometrics*, 48: 1005–1013, 1992.

A Comparisons with existing literature

A.1 Bayes factors

Historically, the use of p-values versus e-values reflects the conventional division of statistics into frequentist and Bayesian (although a sizable fraction of people interested in the foundations of statistics, including the authors of this paper, are neither frequentists nor Bayesians). P-values are a hallmark of frequentist

statistics, but Bayesians often regard p-values as misleading, preferring the use of Bayes factors (which can be combined with prior probabilities to obtain posterior probabilities). In the case of simple statistical hypotheses, a Bayes factor is the likelihood ratio of an alternative hypothesis to the null hypothesis (or vice versa, as in Shafer et al. [2011]). From the betting point of view of this paper, the key property of the Bayes factor is that it is an e-variable.

For composite hypotheses, Bayes factors and e-values diverge. For example, a possible general definition of a Bayes factor is as follows [Kamary et al., 2014, Section 1.2]. Let $(f^0_\theta \mid \theta \in \Theta_0)$ and $(f^1_\theta \mid \theta \in \Theta_1)$ be two statistical models on the same sample space Ω, which is a measurable space, $(\Omega, \mathcal{A})$, with a fixed measure P, and Θ_1 and Θ_2 are measurable spaces. Each $f^n_\theta(\omega)$, $n \in \{0,1\}$, is a probability density as function of ω and a measurable function of $\theta \in \Theta_n$. The corresponding families of probability measures are $(f^0_\theta P)_{\theta\in\Theta_0}$ and $(f^1_\theta P)_{\theta\in\Theta_1}$, where fP is defined as the probability measure $(fP)(A) := \int_A f \,\mathrm{d}P$, $A \in \mathcal{A}$. Make them Bayesian models by fixing prior probability distributions μ_0 and μ_1 on Θ_0 and Θ_1, respectively. This way we obtain Bayesian analogues of the null and alternative hypotheses, respectively. The corresponding *Bayes factor* is

$$B(\omega) := \frac{\int_{\Theta_1} f^1_\theta(\omega)\mu_1(\mathrm{d}\theta)}{\int_{\Theta_0} f^0_\theta(\omega)\mu_0(\mathrm{d}\theta)}, \quad \omega \in \Omega. \tag{A.1}$$

If Θ_0 is a singleton, then B is an e-variable for the probability measure $Q := f^0 P$. In general, however, this is no longer true. Remember that, according to our definition in Section 5, for B to be an e-variable w.r. to the null hypothesis Θ_0 it needs to satisfy $\int B f^0_\theta \,\mathrm{d}P \le 1$ for all $\theta \in \Theta_0$. However, (A.1) only guarantees this property "on average", $\int B f^0_\theta \,\mathrm{d}P\mu_0(\mathrm{d}\theta) \le 1$. Therefore, for a composite null hypothesis a Bayes factor does not need to be an e-value w.r. to that null hypothesis (it is an e-value w.r. to its average).

The literature on Bayes factors is vast; we only mention the fundamental book by Jeffreys [1961], the influential review by Kass and Raftery [1995], and the historical investigation by Etz and Wagenmakers [2017]. Jeffreys's scale that we used in Section 7 was introduced in the context of Bayes factors, but of course it is also applicable to e-values in view of the significant overlap between the two notions. Kass and Raftery [1995, Section 3.2] simplify Jeffreys's scale by merging the "strong" and "very strong" categories into one, which they call "strong".

A.2 Algorithmic theory of randomness

One area where both p-values and e-values have been used for a long time is the algorithmic theory of randomness (see, e.g., Shen et al. [2017]), which originated in Kolmogorov's work on the algorithmic foundations of probability and information [Kolmogorov, 1965, 1968]. Martin-Löf [1966] introduced an algorithmic version of p-values, and then Levin [1976] introduced an algorithmic version of e-values. In the algorithmic theory of randomness people are often interested in low-accuracy results, and then p-values and e-values can be regarded

as slight variations of each other: if e is an e-value, $1/e$ will be a p-value; and vice versa, if p is a p-value, $1/p$ will be an approximate e-value. We discussed this approximation in detail in the main paper; see, e.g., Remark 2.4.

A.3 Standard methods of multiple hypothesis testing

Let us check what the notion of family-wise validity becomes when p-variables are used instead of e-variables. Now we have a procedure that, given p-variables P_k for testing H_k, $k \in \{1, \dots, K\}$, produces random variables $P_1^*, \dots, P_K^*$ taking values in $[0, 1]$. A *conditional p-variable* is a family of p-variables P_Q, $Q \in \mathfrak{P}(\Omega)$. The procedure's output $P_1^*, \dots, P_K^*$ is *family-wise valid* (*FWV*) if there exists a conditional p-variable $(P_Q)_{Q \in \mathfrak{P}(\Omega)}$ such that

$$\forall k \in \{1, \dots, K\} \, \forall Q \in H_k : P_Q \le P_k^*. \tag{A.2}$$

In this case we can see that, for any $Q \in \mathfrak{P}(\Omega)$ and any $\epsilon \in (0, 1)$,

$$Q(\exists k \in \{1, \dots, K\} : Q \in H_k \text{ and } P_k^* \le \epsilon) \le Q(P_Q \le \epsilon) \le \epsilon. \tag{A.3}$$

The left-most expression in (A.3) is known as the *family-wise error rate* (the standard abbreviation is FWER) of the procedure that rejects H_k when $P_k^* \le \epsilon$. The inequality between the extreme terms of (A.3) can be expressed as P_k^* being *family-wise adjusted p-values*. (See, e.g., Efron [2010, Section 3.2].)

On the other hand, we can check that any procedure satisfying (A.3) will satisfy (A.2) for some conditional p-variable (P_Q): indeed, we can set

$$P_Q := \min_{Q \in H_k} P_k^*.$$

Remark A.1. Notice that calibrators maintain the FWV property. Namely, if p-variables $P_1^*, \dots, P_K^*$ are FWV and f is a calibrator, the e-variables $f(P_1^*), \dots, f(P_K^*)$ are FWV. This follows immediately from the definitions (9) and (A.2). And in the opposite direction, if e-variables $E_1^*, \dots, E_K^*$ are FWV and g is an e-to-p-calibrator, the p-variables $g(E_1^*) \wedge 1, \dots, g(E_K^*) \wedge 1$ are FWV.

As we mentioned in Section 5, Algorithms 1 and 2 can be obtained from the e-merging function (4) by applying the closure principle. In our description of this principle we will follow Efron [2010, Section 3.3]. Suppose, for some $\epsilon > 0$ and all $I \subseteq \{1, \dots, K\}$, we have a level-$\epsilon$ test function $\phi_I : \Omega \to \{0, 1\}$:

$$\forall Q \in \cap_{i \in I} H_i : \mathbb{E}^Q[\phi_I] \le \epsilon;$$

$\phi_I = 1$ means that the combined null hypothesis $\cap_{i \in I} H_i$ is rejected. (Such a collection of "local tests", for all I and ϵ, is just a different representation of p-merging functions.) The principle then recommends the simultaneous test function

$$\Phi_J := \min_{I \supseteq J} \phi_I, \qquad J \subseteq \{1, \dots, K\};$$

this simultaneous test function rejects J if ϕ rejects all I such that $J \subseteq I \subseteq \{1, \dots, K\}$. If $P_1, \dots, P_K$ are p-variables, f is a symmetric p-merging function, and ϕ is defined by

$$\phi_I = 1 \iff f(P_i, i \in I) \le \epsilon$$

(which is clearly a level-ϵ test function), we have

$$\Phi_J = 1 \iff \max_{I \supseteq J} f(P_i, i \in I) \le \epsilon$$

(omitting the dependence of ϕ and Φ on ϵ). This corresponds to the simultaneous p-variable

$$P_J := \max_{I \supseteq J} f(P_i, i \in I). \tag{A.4}$$

In this paper we are only interested in the case where J is a singleton (analogues for general J are considered in Vovk and Wang [2019b, 2020a], to be discussed later). This gives us the adjusted p-values

$$P_k^* = P_{\{k\}} := \max_{I \ni k} f(P_i, i \in I).$$

The corresponding formula for the adjusted e-values is

$$E_k^* := \min_{I \ni k} f(E_i, i \in I).$$

This coincides with

- (13) when f is taken to be arithmetic average (which is implemented in Algorithm 1),
- and (14) when f is taken to be product (which is implemented in Algorithm 2).

When the J in (A.4) is allowed not to be a singleton and the p-values are replaced by e-values, we obtain the possibility of controlling false discovery proportion. This appears to us an interesting program of research; the ease of merging e-functions open up new possibilities. First steps in this directions are done in Vovk and Wang [2019b] and (under the assumption of independence) in Vovk and Wang [2020a].

Empirical Bayes methods

Several simple but informative models for multiple hypothesis testing have been proposed in the framework of empirical Bayes methods. Perhaps the simplest model [Efron, 2010, Chapter 2], known as the *two-groups model*, is where we are given a sequence of real values $z_1, \dots, z_N$, each of which is generated either from the null probability density function f_0 or from the alternative probability density function f_1, w.r. to Lebesgue measure. Each value is generated from f_0 with probability π_0 and from f_1 with probability π_1, where $\pi_0 + \pi_1 = 1$. This gives the overall probability density function $f := \pi_0 f_0 + \pi_1 f_1$.

From the Bayesian point of view, the most relevant value for multiple hypothesis testing is the conditional probability $\mathrm{fdr}(z) := \pi_0 f_0(z)/f(z)$ that an observed value z has been generated from the null probability density function f_0; it is knows as the *local false discovery rate*. The most natural e-value in this context is the likelihood ratio $e := f_1(z)/f_0(z)$, and the local false discovery rate can be written in the form $\mathrm{fdr}(z) = \pi_0/(\pi_0 + \pi_1 e)$. Efron [2010, Section 5.1] refers to the ratio $f_1(z)/f_0(z)$ as "Bayes factor"; as discussed in Section A.1, in this case the notions of e-values and Bayes factors happen to coincide.

A conventional threshold for reporting "interesting" cases z_i is $\mathrm{fdr}(z_i) \le 0.2$, where in practice the true $\mathrm{fdr}(z_i)$ is replaced by its empirical estimate [Efron, 2010, Section 5.1]. In terms of the likelihood ratio e, the criterion $\mathrm{fdr}(z_i) \le 0.2$ can be rewritten as $e \ge 4\pi_0/\pi_1$ [Efron, 2010, Exercise 5.1]; of course, the Bayesian decision depends on the ratio of the prior probabilities of the two hypotheses. When $\pi_0 \ge 0.1$ (which is a common case), we have $e \ge 4\pi_0/\pi_1 \ge 36$ [Efron, 2010, (5.9)], and so in large-scale hypothesis testing we need at least very strong evidence on Jeffreys's scale (Section 7) to declare a case interesting.

The two-groups model is highly idealized; e.g., all non-null z are assumed to be coming from the same distribution, f_1. In the empirical Bayesian approach the values $z_1, \dots, z_N$ are assumed to satisfy some independence-type conditions (e.g., Storey and Tibshirani [2003] assume what they call weak dependence), in order to be able to estimate relevant quantities and functions, such as f, from the data. In general, this approach makes different assumptions and arrives at different conclusions as compared with our approach.

A.4 Test martingales in statistics

This paper only scratches the surface of the huge topic of test martingales and their use in statistics. Martingales were introduced by Ville [1939] and popularized by Doob [1953]; see Mazliak and Shafer [2009] for their fascinating history, including their applications in statistics. Recent research includes exponential line-crossing inequalities [Howard et al., 2020a], nonparametric confidence sequences [Howard et al., 2020b], and universal inference [Wasserman et al., 2020].

B History and other classes of calibrators

The question of calibration of p-values into Bayes factors has a long history in Bayesian statistics. The idea was first raised by Berger and Delampady [1987, Section 4.2] (who, however, referred to the idea as "ridiculous"; since then the idea has been embraced by the Bayesian community). The class of calibrators (1) was proposed in Vovk [1993] and rediscovered in Sellke et al. [2001]. A simple characterization of the class of all calibrators was first obtained in Shafer et al. [2011]. A popular Bayesian point of view is that p-values tend to be misleading and need to be transformed into e-values (in the form of Bayes factors) in order to make sense of them.

Recall that the calibrator (2) is a mixture of (1), and it is closer to $1/p$ than any of (1) as $p \to 0$. Of course, (2) is not the only calibrator that is close to $1/p$. Since

$$\int_0^{e^{-1-\kappa}} v^{-1}(-\ln v)^{-1-\kappa} \, \mathrm{d}v = \frac{1}{\kappa(1+\kappa)^\kappa},$$

where $\kappa \in (0, \infty)$, each function

$$H_\kappa(p) := \begin{cases} \infty & \text{if } p = 0 \\ \kappa(1+\kappa)^\kappa p^{-1}(-\ln p)^{-1-\kappa} & \text{if } p \in (0, \exp(-1-\kappa)] \\ 0 & \text{if } p \in (\exp(-1-\kappa), 1] \end{cases} \tag{B.1}$$

is a calibrator [Shafer et al., 2011]. It is instructive to compare (B.1) with $\kappa := 1$ and (2); whereas the former benefits from the extra factor of 2, it kicks in only for $p \le \exp(-2) \approx 0.135$.

We can generalize the calibrator (2) by replacing the uniform distribution on the interval $[0, 1]$ by the distribution with density (1). Replacing κ in (2) by x to avoid clash of notation, we obtain the calibrator

$$F_\kappa(p) := \int_0^1 x p^{x-1} \kappa x^{\kappa-1} \, \mathrm{d}x = \frac{\kappa \gamma(1+\kappa, -\ln p)}{p(-\ln p)^{1+\kappa}},$$

where

$$\gamma(a, z) := \int_0^z t^{a-1} \exp(-t) \, \mathrm{d}t$$

is one of the incomplete gamma functions [Olver et al., 2020, 8.2.1]. For $\kappa := 1$ we have

$$\gamma(2, -\ln p) = 1 - p + p \ln p$$

[Olver et al., 2020, 8.4.7], which recovers (2). For other positive values of κ, we can see that

$$F_\kappa(p) \sim \frac{\kappa \Gamma(1+\kappa)}{p(-\ln p)^{1+\kappa}}$$

as $p \to 0$. The coefficient in (B.1) is better, $\Gamma(1+\kappa) < (1+\kappa)^\kappa$ for all $\kappa > 0$, but F_κ gives an informative e-value for all p, not just for $p \le \exp(-1-\kappa)$.

C Merging infinite e-values

Let us check that, despite the conceptual importance of infinite e-values, we can dispose of them when discussing e-merging functions.

Proposition C.1. *For any e-merging function F, the function $F' : [0, \infty]^K \to [0, \infty]$ defined by*

$$F'(\mathbf{e}) := \begin{cases} F(\mathbf{e}) & \text{if } \mathbf{e} \in [0, \infty)^K \\ \infty & \text{otherwise} \end{cases}$$

is also an e-merging function. Moreover, F' dominates F. Neither e-merging function takes value ∞ on $[0, \infty)^K$.

Proof. If $E_1, \dots, E_K$ are e-variables, each of them is finite a.s.; therefore,

$$F(E_1, \dots, E_K) = F'(E_1, \dots, E_K) \quad \text{a.s.},$$

and F' is an e-merging function whenever F is.

For the last statement, we will argue indirectly. Suppose $F(e_1, \dots, e_K) = \infty$ for some $e_1, \dots, e_K \in [0, \infty)$. Fix such $e_1, \dots, e_K \in [0, \infty)$ and let E_k, $k \in \{1, \dots, K\}$, be independent random variables such that E_k takes values in the set $\{0, e_k\}$ (of cardinality 2 or 1), takes value e_k with a positive probability, and has expected value at most 1. (For the existence of such random variables, see Lemma D.1 below.) Since $\mathbb{E}[F(E_1, \dots, E_K)] = \infty$, F is not an e-merging function. □

As we mentioned in Section 4, Proposition C.1 continues to hold for ie-merging functions.

D Atomless probability spaces

In several of our definitions, such as those of a calibrator or a merging function, we have a universal quantifier over probability spaces. Fixing a probability space in those definitions, we may obtain wider notions. More generally, in this appendix we will be interested in dependence of our notions on a chosen statistical model. We start our discussion from a well-known lemma that we have already used on a few occasions. (Despite being well-known, the full lemma is rarely stated explicitly; we could not find a convenient reference in literature.) Remember that a probability space $(\Omega, \mathcal{A}, Q)$ is *atomless* if it has no *atoms*, i.e., sets $A \in \mathcal{A}$ such that $P(A) > 0$ and $P(B) \in \{0, P(A)\}$ for any $B \in \mathcal{A}$ such that $B \subseteq A$.

Lemma D.1. *The following three statements are equivalent for any probability space $(\Omega, \mathcal{A}, Q)$:*

(i) *$(\Omega, \mathcal{A}, Q)$ is atomless;*

(ii) *there is a random variable on $(\Omega, \mathcal{A}, Q)$ that is uniformly distributed on $[0, 1]$;*

(iii) *for any Polish space S and any probability measure R on S, there is a random element on $(\Omega, \mathcal{A}, Q)$ with values in S that is distributed as R.*

Typical examples of a Polish space in item (iii) that are useful for us in this paper are $\mathbb{R}^K$ and finite sets.

Proof. The equivalence between (i) and (ii) is stated in Föllmer and Schied [2011, Proposition A.27]. It remains to prove that (ii) implies (iii). According to Kuratowski's isomorphism theorem [Kechris, 1995, Theorem 15.6], S is Borel isomorphic to $\mathbb{R}$, $\mathbb{N}$, or a finite set (the last two equipped with the discrete topology). The only nontrivial case is where S is Borel isomorphic to $\mathbb{R}$, in which case we can assume $S = \mathbb{R}$. It remains to apply Föllmer and Schied [2011, Proposition A.27] again. □

If $(\Omega, \mathcal{A})$ is a measurable space and $\mathcal{Q}$ is a collection of probability measures on $(\Omega, \mathcal{A})$, we refer to $(\Omega, \mathcal{A}, \mathcal{Q})$ as a *statistical model*. We say that it is *rich* if there exists a random variable on $(\Omega, \mathcal{A})$ that is uniformly distributed on $[0, 1]$ under any $Q \in \mathcal{Q}$.

Remark D.2. Intuitively, any statistical model $(\Omega, \mathcal{A}, \mathcal{Q})$ can be made rich by complementing it with a random number generator producing a uniform random value in $[0, 1]$: we replace Ω by $\Omega \times [0, 1]$, $\mathcal{A}$ by $\mathcal{A} \times \mathcal{U}$, and each $Q \in \mathcal{Q}$ by $Q \times U$, where $([0, 1], \mathcal{U}, U)$ is the standard measurable space $[0, 1]$ equipped with the uniform probability measure U. If $\mathcal{Q} = \{Q\}$ contains a single probability measure Q, being rich is equivalent to being atomless (by Lemma D.1).

For a statistical model $(\Omega, \mathcal{A}, \mathcal{Q})$, an *e-variable* is a random variable $E : \Omega \to [0, \infty]$ satisfying

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}^Q[E] \le 1.$$

(as in Section 5). As before, the values taken by e-variables are *e-values*, and the set of e-variables is denoted by $\mathcal{E}_{\mathcal{Q}}$.

An *e-merging function for* $(\Omega, \mathcal{A}, \mathcal{Q})$ is an increasing Borel function $F : [0, \infty]^K \to [0, \infty]$ such that, for all $E_1, \dots, E_K$,

$$(E_1, \dots, E_K) \in \mathcal{E}_{\mathcal{Q}}^K \Longrightarrow F(E_1, \dots, E_K) \in \mathcal{E}_{\mathcal{Q}}.$$

This definition requires that K e-values for $(\Omega, \mathcal{A}, \mathcal{Q})$ be transformed into an e-value for $(\Omega, \mathcal{A}, \mathcal{Q})$. Without loss of generality (as in Appendix C), we replace $[0, \infty]$ by $[0, \infty)$.

Proposition D.3. *Let $F : [0, \infty)^K \to [0, \infty)$ be an increasing Borel function. The following statements are equivalent:*

(i) F is an e-merging function for some rich statistical model;

(ii) F is an e-merging function for all statistical models;

(iii) F is an e-merging function.

Proof. Let us first check that, for any two rich statistical models $(\Omega, \mathcal{A}, \mathcal{Q})$ and $(\Omega', \mathcal{A}', \mathcal{Q}')$, we always have

$$\sup \left\{ \mathbb{E}^Q[F(\mathbf{E})] \mid Q \in \mathcal{Q}, \ \mathbf{E} \in \mathcal{E}_{\mathcal{Q}}^K \right\} = \sup \left\{ \mathbb{E}^{Q'}[F(\mathbf{E}')] \mid Q' \in \mathcal{Q}', \ \mathbf{E}' \in \mathcal{E}_{\mathcal{Q}'}^K \right\}. \tag{D.1}$$

Suppose

$$\sup \left\{ \mathbb{E}^Q[F(\mathbf{E})] \mid Q \in \mathcal{Q}, \ \mathbf{E} \in \mathcal{E}_{\mathcal{Q}}^K \right\} > c$$

for some constant c. Then there exist $\mathbf{E} \in \mathcal{E}_{\mathcal{Q}}^K$ and $Q \in \mathcal{Q}$ such that $\mathbb{E}^Q[F(\mathbf{E})] > c$. Take a random vector $\mathbf{E}' = (E'_1, \dots, E'_K)$ on $(\Omega', \mathcal{A}')$ such that $\mathbf{E}'$ is distributed under each $Q' \in \mathcal{Q}'$ identically to the distribution of $\mathbf{E}$ under Q. This is possible as $\mathcal{Q}'$ is rich (by Lemma D.1 applied to the probability space

$([0,1], \mathcal{U}, U)$, U being the uniform probability measure). By construction, $\mathbf{E}' \in \mathcal{E}^K_{\mathcal{Q}'}$ and $E^{Q'}[F(\mathbf{E}')] > c$ for all $Q' \in \mathcal{Q}'$. This shows

$$\sup\left\{\mathbb{E}^Q[F(\mathbf{E})] \mid Q \in \mathcal{Q},\ \mathbf{E} \in \mathcal{E}^K_{\mathcal{Q}}\right\} \le \sup\left\{\mathbb{E}^{Q'}[F(\mathbf{E}')] \mid Q' \in \mathcal{Q}',\ \mathbf{E}' \in \mathcal{E}^K_{\mathcal{Q}'}\right\},$$

and we obtain equality by symmetry.

The implications (ii) $\Rightarrow$ (iii) and (iii) $\Rightarrow$ (i) are obvious (remember that, by definition an e-merging function is an e-merging function for all singleton statistical models). To check (i) $\Rightarrow$ (ii), suppose F is an e-merging function for some rich statistical model. Consider any statistical model. Its product with the uniform probability measure on $[0,1]$ will be a rich statistical model (cf. Remark D.2). It follows from (D.1) that F will be an e-merging function for the product. Therefore, it will be an e-merging function for the original statistical model. □

Remark D.4. The assumption of being rich is essential in item (i) of Proposition D.3. For instance, if we take $\mathcal{Q} := \{\delta_\omega \mid \omega \in \Omega\}$, where δ_ω is the point-mass at ω, then $\mathcal{E}_{\mathcal{Q}}$ is the set of all random variables taking values in $[0,1]$. In this case, the maximum of e-variables is still an e-variable, but the maximum function is not a valid e-merging function as seen from Theorem 3.2.

An *ie-merging function for* $(\Omega, \mathcal{A}, \mathcal{Q})$ is an increasing Borel function $F : [0,\infty]^K \to [0,\infty]$ such that, for all $E_1, \dots, E_K \in \mathcal{E}_{\mathcal{Q}}$ that are independent under any $Q \in \mathcal{Q}$, we have $F(E_1, \dots, E_K) \in \mathcal{E}_{\mathcal{Q}}$. The proof of Proposition D.3 also works for ie-merging functions.

Proposition D.5. *Proposition D.3 remains true if all entries of "e-merging function" are replaced by "ie-merging function".*

Proof. The changes to the proof of Proposition D.3 are minimal. In (D.1), the components of $\mathbf{E}$ and $\mathbf{E}'$ should be assumed to be independent under any probability measure in $\mathcal{Q}$ and $\mathcal{Q}'$, respectively. The components of the vector $\mathbf{E}'$ constructed from $\mathbf{E}$ and Q will be independent under any $Q' \in \mathcal{Q}'$. □

Proposition D.3 shows that in the definition of an e-merging function it suffices to require that (3) hold for a fixed atomless probability space $(\Omega, \mathcal{A}, Q)$. Proposition D.5 extends this observation to the definition of an ie-merging function.

We can state similar propositions in the case of calibrators. A *p-variable for a statistical model* $(\Omega, \mathcal{A}, \mathcal{Q})$ is a random variable $P : \Omega \to [0,\infty)$ satisfying

$$\forall \epsilon \in (0,1)\, \forall Q \in \mathcal{Q} : Q(P \le \epsilon) \le \epsilon.$$

The set of p-variables for $(\Omega, \mathcal{A}, \mathcal{Q})$ is denoted by $\mathcal{P}_{\mathcal{Q}}$. A decreasing function $f : [0,1] \to [0,\infty]$ is a *calibrator for* $(\Omega, \mathcal{A}, \mathcal{Q})$ if, for any p-variable $P \in \mathcal{P}_{\mathcal{Q}}$, $f(P) \in \mathcal{E}_{\mathcal{Q}}$.

Proposition D.6. *Let $f : [0,1] \to [0,\infty]$ be a decreasing Borel function. The following statements are equivalent:*

(i) f is a calibrator for some rich statistical model;

(ii) f is a calibrator for all statistical models;

(iii) f is a calibrator.

We refrain from stating the obvious analogue of Proposition D.6 for e-to-p calibrators.

E Domination structure of the class of e-merging functions

In this appendix we completely describe the domination structure of the symmetric e-merging functions, showing that (6) is the minimal complete class of symmetric e-merging functions. We start, however, with establishing some fundamental facts about e-merging functions.

First, we note that for an increasing Borel function $F : [0,\infty)^K \to [0,\infty]$, its upper semicontinuous version F^* is given by

$$F^*(\mathbf{e}) = \lim_{\epsilon \downarrow 0} F(\mathbf{e} + \epsilon \mathbf{1}), \quad \mathbf{e} \in [0,\infty)^K; \tag{E.1}$$

remember that $\mathbf{1} := (1,\dots,1)$. Clearly, F^* is increasing, is upper semicontinuous (by a simple compactness argument), and satisfies $F^* \geq F$.

On the other hand, for an upper semicontinuous (and so automatically Borel) function $F : [0,\infty)^K \to [0,\infty]$, its increasing version $\widetilde{F}$ is given by

$$\widetilde{F}(\mathbf{e}) = \sup_{\mathbf{e}' \leq \mathbf{e}} F(\mathbf{e}'), \quad \mathbf{e} \in [0,\infty)^K, \tag{E.2}$$

where $\leq$ is component-wise inequality. Clearly, $\widetilde{F}$ is increasing, upper semicontinuous, and $\widetilde{F} \geq F$. Notice that the supremum in (E.2) is attained (as the supremum of an upper semicontinuous function on a compact set), and so we can replace sup by max.

Proposition E.1. *If F is an e-merging function, then its upper semicontinuous version F^* in (E.1) is also an e-merging function.*

Proof. Take $\mathbf{E} \in \mathcal{E}_Q^K$. For every rational $\epsilon \in (0,1)$, let A_ϵ be an event independent of $\mathbf{E}$ with $Q(A_\epsilon) = 1-\epsilon$, and $\mathbf{E}_\epsilon = (\mathbf{E} + \epsilon\mathbf{1})1_{A_\epsilon}$ (of course, here we use the convention that $\mathbf{E}_\epsilon = \mathbf{0} := (0,\dots,0)$ if the event A_ϵ does not occur). For each ϵ, $\mathbb{E}[\mathbf{E}_\epsilon] \leq (1-\epsilon)(\mathbf{1} + \epsilon\mathbf{1}) \leq \mathbf{1}$. Therefore, $\mathbf{E}_\epsilon \in \mathcal{E}_Q^K$ and hence

$$1 \geq \mathbb{E}[F(\mathbf{E}_\epsilon)] = (1-\epsilon)\mathbb{E}\left[F(\mathbf{E} + \epsilon\mathbf{1})\right] + \epsilon F(\mathbf{0}),$$

which implies

$$\mathbb{E}\left[F(\mathbf{E} + \epsilon\mathbf{1})\right] \leq \frac{1 - \epsilon F(\mathbf{0})}{1-\epsilon}.$$

Fatou's lemma yields

$$\mathbb{E}[F^*(\mathbf{E})] = \mathbb{E}\left[\lim_{\epsilon \downarrow 0} F(\mathbf{E} + \epsilon \mathbf{1})\right] \leq \lim_{\epsilon \downarrow 0} \mathbb{E}\left[F(\mathbf{E} + \epsilon \mathbf{1})\right] \leq \lim_{\epsilon \downarrow 0} \frac{1 - \epsilon F(\mathbf{0})}{1 - \epsilon} = 1.$$

Therefore, F^* is an e-merging function. □

Corollary E.2. *An admissible e-merging function is always upper semicontinuous.*

Proof. Let F be an admissible e-merging function. Using Proposition E.1, we obtain that $F^* \geq F$ is an e-merging function. Admissibility of F forces $F = F^*$, implying that F is upper semicontinuous. □

Proposition E.3. *If $F : [0, \infty)^K \to [0, \infty]$ is an upper semicontinuous function satisfying $\mathbb{E}[F(\mathbf{E})] \leq 1$ for all $\mathbf{E} \in \mathcal{E}_Q^K$, then its increasing version $\widetilde{F}$ in (E.2) is an e-merging function.*

Proof. Take any $\mathbf{E} \in \mathcal{E}_Q^K$ supported in $[0, M]^K$ for some $M > 0$. Define

$$u(\mathbf{x}, \mathbf{y}) := F(\mathbf{y}) \text{ and } D := \left\{(\mathbf{x}, \mathbf{y}) \in [0, M]^K \times [0, M]^K \mid \mathbf{y} \leq \mathbf{x}\right\};$$

as a closed subset of a compact set, D is compact. Since F is upper semicontinuous, the sets

$$U_c := \{(\mathbf{x}, \mathbf{y}) \in D \mid F(\mathbf{y}) \geq c\} \text{ and } U_c(\mathbf{x}) := \{\mathbf{y} \mid (\mathbf{x}, \mathbf{y}) \in U_c\}$$

are all compact (and therefore, Borel). Moreover, for each compact subset $\mathcal{K}$ of $[0, M]^K$, the set

$$\left\{\mathbf{x} \in [0, M]^K \mid \exists \mathbf{y} : (\mathbf{x}, \mathbf{y}) \in U_c \;\&\; \mathbf{y} \in \mathcal{K}\right\}$$

is compact (and therefore, Borel). These conditions justify the use of Theorem 4.1 of Rieder [1978], which gives the existence of a Borel function $g : [0, M]^K \to [0, M]^K$ such that $F(g(\mathbf{e})) = \widetilde{F}(\mathbf{e})$ and $g(\mathbf{e}) \leq \mathbf{e}$ for each $\mathbf{e} \in [0, M]^K$. Hence, $g(\mathbf{E}) \in \mathcal{E}_Q^K$, and we have

$$\mathbb{E}[\widetilde{F}(\mathbf{E})] = \mathbb{E}[F(g(\mathbf{E}))] \leq 1.$$

An unbounded $\mathbf{E} \in \mathcal{E}_Q^K$ can be approximated by an increasing sequence of bounded random vectors in $\mathcal{E}_Q^K$, and the monotone convergence theorem implies $\mathbb{E}[\widetilde{F}(\mathbf{E})] \leq 1$. □

Proposition E.4. *An admissible e-merging function is not strictly dominated by any Borel function G satisfying $\mathbb{E}[G(\mathbf{E})] \leq 1$ for all $\mathbf{E} \in \mathcal{E}_Q^K$.*

Proof. Suppose that an admissible e-merging function F is strictly dominated by a Borel function G satisfying $\mathbb{E}[G(\mathbf{E})] \leq 1$ for all $\mathbf{E} \in \mathcal{E}_Q^K$. Take a point $\mathbf{e} \in [0, \infty)^K$ such that $G(\mathbf{e}) > F(\mathbf{e})$. Define a function H by $H(\mathbf{e}) := G(\mathbf{e})$ and

$H := F$ elsewhere. By Corollary E.2, we know that F is upper semicontinuous, and so is H by construction. Clearly, $\mathbb{E}[H(\mathbf{E})] \le \mathbb{E}[G(\mathbf{E})] \le 1$ for all $\mathbf{E} \in \mathcal{E}_Q^K$. Using Proposition E.3, we obtain that $\widetilde{H}$ is an e-merging function. It remains to notice that $\widetilde{H}$ strictly dominates F. □

Proposition E.5. *Any e-merging function is dominated by an admissible e-merging function.*

Proof. Let R be any probability measure with positive density on $[0,\infty)$ with mean 1. Fix an e-merging function F. By definition, $\int F \,\mathrm{d}R^K \le 1$, and such an inequality holds for any e-merging function. Set $F_0 := F$ and let

$$c_i := \sup_{G : G \ge F_{i-1}} \int G \,\mathrm{d}R^K \le 1, \tag{E.3}$$

where $i := 1$ and G ranges over all e-merging functions dominating F_{i-1}. Let F_i be an e-merging function satisfying

$$F_i \ge F_{i-1} \quad \text{and} \quad \int F_i \,\mathrm{d}R^K \ge c_i - 2^{-i}, \tag{E.4}$$

where $i := 1$. Continue setting (E.3) and choosing F_i to satisfy (E.4) for $i = 2, 3, \ldots$. Set $G := \lim_{i\to\infty} F_i$. It is clear that G is an e-merging (by the monotone convergence theorem) function dominating F and that $\int G \,\mathrm{d}R = \int H \,\mathrm{d}R$ for any e-merging function H dominating G.

By Proposition E.1, the upper semicontinuous version G^* of G is also an e-merging function. Let us check that G^* is admissible. Suppose that there exists an e-merging function H such that $H \ge G^*$ and $H \ne G^*$. Fix such an H and an $\mathbf{e} \in [0,\infty)^K$ satisfying $H(\mathbf{e}) > G^*(\mathbf{e})$. Since G^* is upper semicontinuous and H is increasing, there exists $\epsilon > 0$ such that $H > G^*$ on the hypercube $[\mathbf{e}, \mathbf{e} + \epsilon\mathbf{1}] \subseteq [0,\infty)^K$, which has a positive R^K-measure. This gives

$$\int G \,\mathrm{d}R^K \le \int G^* \,\mathrm{d}R^K < \int H \,\mathrm{d}R^K,$$

a contradiction. □

The key component of the statement of completeness of (6) is the following proposition.

Proposition E.6. *Suppose that F is a symmetric e-merging function satisfying $F(\mathbf{0}) = 0$. Then F is admissible if and only if it is the arithmetic mean.*

Proof. For the "if" statement, see Proposition 4.1. Next we show the "only if" statement. Let F be an admissible symmetric e-merging function with $F(\mathbf{0}) = 0$. As always, all expectations $\mathbb{E}$ below are with respect to Q.

Suppose for the purpose of contradiction that there exists $(e_1, \ldots, e_K) \in [0,\infty)^K$ such that

$$F(e_1, \ldots, e_K) > \frac{e_1 + \cdots + e_K}{K} \in [0, 1)$$

(the case " $\in [1,\infty)$" is excluded by Proposition 3.1). We use the same notation as in the proof of Proposition 3.1. Since $F(\mathbf{0}) = 0$, we know that $b > a > 0$. Let $\delta := (b-a)/(1-a) > 0$, and define $G : [0,\infty)^K \to [0,\infty]$ by $G(\mathbf{0}) := \delta$ and $G := F$ otherwise. It suffices to show that $\mathbb{E}[G(\mathbf{E})] \le 1$ for all $\mathbf{E} \in \mathcal{E}_Q^K$; by Proposition E.4 this will contradict the admissibility of F.

Since F is an e-merging function, for any random vector $(E_1,\dots,E_K)$ taking values in $[0,\infty)^K$ and any non-null event B independent of $(E_1,\dots,E_K)$ and $(D_1,\dots,D_K)$ we have the implication: if

$$(E_1,\dots,E_K)1_B + (D_1,\dots,D_K)1_{B^c} \in \mathcal{E}_Q^K,$$

then

$$\mathbb{E}[F((E_1,\dots,E_K)1_B + (D_1,\dots,D_K)1_{B^c})] \le 1.$$

Write $\beta := Q(B)$. The above statement shows that if

$$\beta \bigvee_{k=1}^K \mathbb{E}[E_k] + (1-\beta)a \le 1,$$

or equivalently,

$$\bigvee_{k=1}^K \mathbb{E}[E_k] \le \frac{1-(1-\beta)a}{\beta}, \tag{E.5}$$

then

$$\beta\mathbb{E}[F(E_1,\dots,E_K)] + (1-\beta)b \le 1,$$

or equivalently,

$$\mathbb{E}[F(E_1,\dots,E_K)] \le \frac{1-(1-\beta)b}{\beta}. \tag{E.6}$$

Next, take an arbitrary random vector $(E_1,\dots,E_K)$ such that

$$Q((E_1,\dots,E_K) \in [0,\infty)^K \setminus \{\mathbf{0}\}) = 1. \tag{E.7}$$

Further, take an arbitrary non-null event C independent of $(E_1,\dots,E_K)$ such that

$$\bigvee_{k=1}^K \mathbb{E}[E_k] \le \frac{1}{Q(C)}, \tag{E.8}$$

which implies $(E_1,\dots,E_K)1_C \in \mathcal{E}_Q^K$. We will show that

$$\mathbb{E}[G((E_1,\dots,E_K)1_C)] \le 1.$$

Write $\lambda := Q(C)$ and choose $\beta \in (0,1]$ such that $\beta/(1-(1-\beta)a) = \lambda$. From $\bigvee_{k=1}^K \mathbb{E}[E_k] \le 1/\lambda$ we obtain (E.5), which implies (E.6). Using (E.6), we have

$$\begin{aligned}
&\mathbb{E}[G((E_1,\dots,E_K)1_C)] \\
&\quad = \lambda\mathbb{E}[F(E_1,\dots,E_K)] + (1-\lambda)\delta \\
&\quad = \lambda\mathbb{E}[F(E_1,\dots,E_K)] + (1-\lambda)\frac{b-a}{1-a} \\
&\quad \le \frac{\beta}{1-(1-\beta)a}\frac{1-(1-\beta)b}{\beta} + \left(1 - \frac{\beta}{1-(1-\beta)a}\right)\frac{b-a}{1-a} = 1.
\end{aligned}$$

Finally, we note that for any $\mathbf{E} \in \mathcal{E}_Q^K$, if $Q(\mathbf{E} = \mathbf{0}) = 0$, then $\mathbb{E}[G(\mathbf{E})] = \mathbb{E}[F(\mathbf{E})] \le 1$. If $Q(\mathbf{E} = \mathbf{0}) > 0$, then $\mathbf{E}$ is distributed as $(E_1, \dots, E_K)1_C$ for some event C and $(E_1, \dots, E_K)$ satisfying (E.7)–(E.8). In either case, we have $\mathbb{E}[G(\mathbf{E})] \le 1$. □

Finally, we are able to prove Theorem 3.2 based on Proposition E.6.

Proof of Theorem 3.2. In view of Proposition E.5, it suffices to check the characterization of admissibility. The "if" statement follows from Proposition 4.1. We next show the "only if" statement; let F be admissible. If $F(\mathbf{0}) \ge 1$, then $F \ge 1$. The fact that F is an e-merging function further forces $F = 1$. Next, assume $F(\mathbf{0}) \in [0, 1)$ and let $\lambda := F(\mathbf{0})$. Define another function $G : [0, \infty)^K \to [0, \infty)$ by

$$G(\mathbf{e}) := \frac{F(\mathbf{e}) - \lambda}{1 - \lambda}.$$

It is easy to see that G is a symmetric and admissible e-merging function satisfying $G(\mathbf{0}) = 0$. Therefore, using Proposition E.6, we have $G = M_K$. The statement of the theorem follows. □

F A maximin view of merging

This appendix is inspired by the comments by the Associate Editor of the journal version of this paper.

F.1 Informal maximin view

This section is high-level and informal; in it (and in this appendix in general) we will only discuss the case of e-merging.

In this paper, two particularly important sources of e-merging functions are:

- $\mathcal{F}_0$, the class of all increasing Borel functions $F : [0, \infty)^K \to [0, \infty)$;
- $\mathcal{F}_S$, the class of all symmetric functions in $\mathcal{F}_0$.

In the rest of this appendix, $\mathcal{F}$ will stand for either $\mathcal{F}_0$ or $\mathcal{F}_S$.

Let $(\Omega, \mathcal{A}, Q)$ be an atomless probability space (cf. Appendix D). For $\mathbf{E} \in \mathcal{E}_Q^K$, let $\mathcal{F}_{\mathbf{E}}$ be the set of all functions $F \in \mathcal{F}$ such that

$$\mathbb{E}[F(\mathbf{E})] \le 1;$$

intuitively, these are $\mathbf{E}$-specific e-merging functions.

We are looking for suitable e-merging functions to use, which can be interpreted as the problem of finding the "best elements", in some sense, of $\cap(\mathcal{F}_{\mathbf{E}} : \mathbf{E} \in \mathcal{E}_Q^K)$. More generally, we could specify a class $\mathcal{M}$ of joint models of K e-variables, and be interested in

$$\max \left(\bigcap_{\mathbf{E} \in \mathcal{M}} \mathcal{F}_{\mathbf{E}} \right), \tag{F.1}$$

where $\max(\cdot)$ gives the best element(s) of a set, in some sense. In the case of admissibility, it will be literally the set of maximal elements, but it can also be the element essentially or weakly dominating all other elements if it exists.

The problem (F.1) can be said to be a maximin problem, since the natural interpretation of $\cap$ (preceded by max) is minimum. For the e-merging and ie-merging functions, the informal problems are

$$\max\left(\bigcap_{\mathbf{E}\in\mathcal{E}_Q^K}\mathcal{F}_{\mathbf{E}}\right)\qquad\text{and}\qquad\max\left(\bigcap_{\mathbf{E}\in i\mathcal{E}_Q^K}\mathcal{F}_{\mathbf{E}}\right).$$

In the rest of this appendix, $\mathcal{M}$ will stand for either $\mathcal{E}_Q^K$ or $i\mathcal{E}_Q^K$.

F.2 Formal maximin

Our results about essential and weak domination, namely Propositions 3.1 and 4.2, have interesting connections with the maximin problem

$$\sup_{F\in\mathcal{F}}\min_{\mathbf{E}\in\mathcal{M}}F(\mathbf{e})1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}} \tag{F.2}$$

for a fixed $\mathbf{e}\in[0,\infty)^K$. The value (F.2) is the supremum of $F(\mathbf{e})$ over all e-merging functions (if $\mathcal{M}=\mathcal{E}_Q^K$) or over all ie-merging functions (if $\mathcal{M}=i\mathcal{E}_Q^K$). Intuitively, this corresponds to an overoptimistic way of merging e-values choosing the best merging function in hindsight (which makes (F.2) somewhat similar to the VS bound). Notice that the minimum in (F.2) (either $F(\mathbf{e})$ or 0) is indeed attained.

Fix $\mathbf{e}\in[0,\infty)^K$. Propositions 3.1 and 4.2 show that

$$\begin{aligned}\max_{F\in\mathcal{F}_S}\min_{\mathbf{E}\in\mathcal{E}_Q^K}F(\mathbf{e})1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}}&=M_K(\mathbf{e})\vee 1,\\ \max_{F\in\mathcal{F}_S}\min_{\mathbf{E}\in i\mathcal{E}_Q^K}F(\mathbf{e})1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}}&=P_K(\mathbf{e}\vee\mathbf{1})\\ &=\max_{F\in\mathcal{F}_0}\min_{\mathbf{E}\in i\mathcal{E}_Q^K}F(\mathbf{e})1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}},\end{aligned}$$

where M_K is the arithmetic mean and P_K is the product function. This follows from the maximin problem having a universal optimizer for $\mathbf{e}$ large enough: $M_K(\mathbf{e})\ge 1$ in the e-merging case and $\mathbf{e}\ge\mathbf{1}$ in the ie-merging case.

Let us check that

$$\max_{F\in\mathcal{F}_0}\min_{\mathbf{E}\in\mathcal{E}_Q^K}F(\mathbf{e})1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}}=\max(\mathbf{e})\vee 1.$$

To show this, first notice that, for each $k=1,\dots,K$, $(e_1,\dots,e_K)\mapsto e_k$ is an e-merging function, and so is $(e_1,\dots,e_K)\mapsto 1$. This shows the $\ge$ part of the equality. On the other hand, for any function $F\in\mathcal{F}$, if $F(\mathbf{e})>a:=\max(\mathbf{e})\vee 1$ for some $\mathbf{e}$, then by designing a vector $\mathbf{E}$ of e-variables with $Q(\mathbf{E}=\mathbf{e})=1/a$, we have $\mathbb{E}[F(\mathbf{E})]\ge F(\mathbf{e})/a>1$, and hence F is not an e-merging function. This gives the $\le$ part of the equality.

F.3 The minimax formulation

For a fixed $\mathbf{e} \in [0,\infty)^K$, we can also talk about the minimax problem corresponding to the maximin problem (F.2):

$$\inf_{\mathbf{E}\in\mathcal{M}} \max_{F\in\mathcal{F}} F(\mathbf{e}) 1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}}. \tag{F.3}$$

As usual, we have

$$\inf_{\mathbf{E}\in\mathcal{M}} \max_{F\in\mathcal{F}} F(\mathbf{e}) 1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}} \ge \sup_{F\in\mathcal{F}} \min_{\mathbf{E}\in\mathcal{M}} F(\mathbf{e}) 1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}}, \tag{F.4}$$

but the two sides are not always equal.

The minimax problem (F.3) is usually easy to solve. We first look at the case $\mathcal{F} = \mathcal{F}_0$. Note that for fixed $\mathbf{E} \in \mathcal{E}_Q^K$, using the Neyman–Pearson argument, we have

$$\max_{F\in\mathcal{F}_0} F(\mathbf{e}) 1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}} = \frac{1}{Q(\mathbf{E} \ge \mathbf{e})} \in [1,\infty],$$

for which a maximizer (typically the unique maximizer) is

$$\mathbf{e}' \mapsto \frac{1_{\{\mathbf{e}'\ge\mathbf{e}\}}}{Q(\mathbf{E} \ge \mathbf{e})}.$$

Therefore, the minimax problem (F.3) for $\mathcal{F} = \mathcal{F}_0$ has value

$$\inf_{\mathbf{E}\in\mathcal{M}} \max_{F\in\mathcal{F}_0} F(\mathbf{e}) 1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}} = \inf_{\mathbf{E}\in\mathcal{M}} \frac{1}{Q(\mathbf{E} \ge \mathbf{e})}.$$

Since $\mathcal{M}$ is $\mathcal{E}_Q^K$ or $i\mathcal{E}_Q^K$, we can compute this as

$$\begin{aligned}
\min_{\mathbf{E}\in\mathcal{E}_Q^K} \frac{1}{Q(\mathbf{E} \ge \mathbf{e})} &= \min_{E\in\mathcal{E}_Q} \frac{1}{Q(E \ge \max(\mathbf{e}))} = \max(\mathbf{e}) \vee 1, \\
\min_{\mathbf{E}\in i\mathcal{E}_Q^K} \frac{1}{Q(\mathbf{E} \ge \mathbf{e})} &= \prod_{k=1}^K \min_{E_k\in\mathcal{E}_Q} \frac{1}{Q(E_k \ge e_k)} = P_K(\mathbf{e} \vee \mathbf{1}).
\end{aligned} \tag{F.5}$$

In combination with results of Section F.2, this shows that (F.4) holds as an equality in this case.

Next, let us look at the case $\mathcal{F} = \mathcal{F}_S$. For $\mathcal{M} = i\mathcal{E}_Q^K$, the coincidence of the minimax and maximin follows from the previous results, so we assume $\mathcal{M} = \mathcal{E}_Q^K$. Let

$$A_{\mathbf{e}} := \bigcup_{\pi\in\Pi_K} \{\mathbf{e}' \in [0,\infty)^K : \mathbf{e}' \ge \mathbf{e}_\pi\},$$

where Π_K is the set of all K-permutations and $\mathbf{e}_\pi := (e_{\pi(1)}, \dots, e_{\pi(K)})$. Using a Neyman–Pearson argument again, we have, for a fixed $\mathbf{E} \in \mathcal{E}_Q^K$,

$$\max_{F\in\mathcal{F}_S} F(\mathbf{e}) 1_{\{\mathbb{E}[F(\mathbf{E})]\le 1\}} = \frac{1}{Q(\mathbf{E} \in A_{\mathbf{e}})},$$

for which a maximizer is

$$\mathbf{e}' \mapsto \frac{1_{\{\mathbf{e}' \in A_{\mathbf{e}}\}}}{Q(\mathbf{E} \in A_{\mathbf{e}})}.$$

Let $a := (1/M_K(\mathbf{e})) \wedge 1$, and the distribution of $\mathbf{E}'$ be given by

$$\frac{a}{K!} \sum_{\pi \in \Pi_K} \delta_{\mathbf{e}_\pi} + (1-a)\delta_{\mathbf{0}},$$

δ_ω being the point-mass at ω. It is clear that $\mathbf{E}' \in \mathcal{E}_Q^K$ and $Q(\mathbf{E}' \in A_{\mathbf{e}}) = a$. It follows that

$$\inf_{\mathbf{E} \in \mathcal{E}_Q^K} \frac{1}{Q(\mathbf{E} \in A_{\mathbf{e}})} \le \frac{1}{Q(\mathbf{E}' \in A_{\mathbf{e}})} = M_K(\mathbf{e}) \vee 1. \tag{F.6}$$

Hence, by (F.4) and the results of Section F.2, the values of the maximin and the minimax again coincide (and the inf in (F.6) is actually attained and so can be replaced by min).

The inequalities (F.5) and (F.6) give alternative proofs to the domination statements in Propositions 3.1 and (in the independent case) 4.2, since our overoptimistic upper bound coincides with a valid e-merging (ie-merging in the independent case) function when $M_K(\mathbf{e}) \ge 1$ (when $\mathbf{e} \ge \mathbf{1}$ in the independent case). For this, we do not need anything derived in the main paper or in Section F.2 above. However, the minimax approach (at least in the form presented here) does not produce the full domination structure of symmetric e-merging functions as given in Theorem 3.2.

G Cross-merging between e-values and p-values

In this appendix we will briefly discuss functions performing "cross-merging": either merging several e-values into a p-value or several p-values into an e-value. Formally, an *e-to-p merging function* is a decreasing Borel function $F : [0,\infty]^K \to [0,1]$ such that $F(E_1,\dots,E_K)$ is a p-variable whenever $E_1,\dots,E_K$ are e-variables, and a *p-to-e merging function* is a decreasing Borel function $F : [0,1]^K \to [0,\infty]$ such that $F(P_1,\dots,P_K)$ is an e-variable whenever $P_1,\dots,P_K$ are p-variables. The message of this appendix is that cross-merging can be performed as composition of pure merging (applying an e-merging function or a p-merging function) and calibration (either e-to-p calibration or p-to-e calibration); however, in some important cases (we feel in the vast majority of cases) pure merging is more efficient, and should be done, in the domain of e-values.

Let us start from e-to-p merging. Given e-values $e_1,\dots,e_K$, we can merge them into one e-value by applying the arithmetic mean, the only essentially admissible symmetric e-merging function (Proposition 3.1), and then by applying inversion $e \mapsto e^{-1} \wedge 1$, the only admissible e-to-p calibrator (Proposition 2.3). This gives us the e-to-p merging function

$$F(e_1,\dots,e_K) := \frac{K}{e_1 + \dots + e_K} \wedge 1. \tag{G.1}$$

The following proposition shows that in this way we obtain the optimal symmetric e-to-p merging function.

Proposition G.1. *The e-to-p merging function* (G.1) *dominates all symmetric e-to-p merging functions.*

Proof. Suppose that a symmetric e-to-p merging function G satisfies $G(\mathbf{e}) < F(\mathbf{e})$ for some $\mathbf{e} = (e_1, \dots, e_K) \in [0, \infty)^K$. The following arguments are similar to the proof of Proposition 3.1. As before, Π_K is the set of all permutations on $\{1, \dots, K\}$, π is randomly and uniformly drawn from Π_K, and $(D_1, \dots, D_K) := (e_{\pi(1)}, \dots, e_{\pi(K)})$. Further, let $(D'_1, \dots, D'_K) := (D_1, \dots, D_K)1_A$, where A is an event independent of π and satisfying $Q(A) = F(\mathbf{e})$. For each k, we have $\mathbb{E}[D'_k] = F(\mathbf{e})M_K(e_1, \dots, e_K) \le 1$, and hence $D'_k \in \mathcal{E}_Q$. By the symmetry of G, we have $Q(G(D'_1, \dots, D'_K) = G(\mathbf{e})) \ge Q(A) = F(\mathbf{e})$, and hence

$$Q\left(G(D'_1, \dots, D'_K) \le G(\mathbf{e})\right) \ge F(\mathbf{e}) > G(\mathbf{e}).$$

This contradicts G being an e-to-p merging function. □

It is interesting that (G.1) can also be obtained by composing e-to-p calibration and improper pure p-merging. Given e-values $e_1, \dots, e_K$ we first transform them into p-values $1/e_1, \dots, 1/e_K$ (in this paragraph we allow p-values greater than 1, as in Vovk and Wang [2019a]). Wilson [2019] proposed the harmonic mean as a p-merging function. The composition of these two transformations again gives us the e-to-p merging function (G.1). The problem with this argument is that, as Goeman et al. [2019, Wilson's second claim] point out, Wilson's method is in general not valid (one obtains a valid method if the harmonic mean is multiplied by $c \ln K$ for $K > 2$ and for some constant $c < \exp(1)$, according to Vovk and Wang [2019a]). Despite the illegitimate application of the harmonic mean, the resulting function (G.1) is still a valid e-to-p merging function. At least in this context, we can see that e-to-p merging should be done by first pure merging and then e-to-p calibration, not vice versa (which would result in an extra coefficient of $c \ln K$).

Now suppose we are given p-values $p_1, \dots, p_K$, and we would like to merge them into one e-value. Let $\kappa \in (0, 1)$. Applying the calibrator (1), we obtain e-values $\kappa p_1^{\kappa-1}, \dots, \kappa p_K^{\kappa-1}$, and since the average of e-values is an e-value,

$$F(p_1, \dots, p_K) := \frac{\kappa}{K} \sum_{k=1}^{K} p_k^{\kappa-1} \tag{G.2}$$

is a p-to-e merging function.

The following proposition will imply that all p-to-e merging functions (G.2) are admissible; moreover, it will show, in conjunction with Proposition 2.2, that for any admissible (p-to-e) calibrator g, the function

$$M_g(p_1, \dots, p_K) := \frac{1}{K} \sum_{k=1}^{K} g(p_k)$$

is an admissible p-to-e merging function.

Proposition G.2. *If $F : [0,1]^K \to [0,\infty]$ is an upper semicontinuous and decreasing Borel function, $\mathbb{E}[F(\mathbf{P})] = 1$ for all $\mathbf{P} \in \mathcal{P}_Q^K$ with margins uniform on $[0,1]$, and $F = \infty$ on $[0,1]^K \setminus (0,1]^K$, then F is an admissible p-to-e merging function.*

Proof. It is obvious (cf. the proof of Proposition 2.2) that F is a p-to-e merging function. To show that F is admissible, consider another p-to-e merging function G such that $G \geq F$. For independent $P_1, \dots, P_K$ distributed uniformly on $[0,1]$,

$$1 \geq \mathbb{E}[G(P_1, \dots, P_K)] \geq \mathbb{E}[F(P_1, \dots, P_K)] = 1,$$

forcing $G = F$ almost everywhere on $[0,1]^K$. The upper semicontinuity of F and G being decreasing further guarantee that $G = F$ on $(0,1]^K$; indeed, if $G(\mathbf{e}) > F(\mathbf{e})$ for $\mathbf{e} \in (0,1]^K$, there exists $\epsilon > 0$ such that $G > H$ on the hypercube $[\mathbf{e} - \epsilon\mathbf{1}, \mathbf{e}] \subseteq (0,1]^K$, which has a positive Lebesgue measure. Therefore, F is admissible. □

Let us see how we can obtain (G.2) reversing the order in which we do calibration and pure merging. If we first merge the p-values $p_1, \dots, p_K$ by naively (improperly) assuming that their generalized mean

$$\left(\frac{1}{K} \sum_{k=1}^{K} p_k^{\kappa-1} \right)^{\frac{1}{\kappa-1}} \tag{G.3}$$

is a p-value and then apply the calibrator (1), we will obtain exactly the p-to-e merging function (G.2). As shown in Vovk and Wang [2019a, Table 1], (G.3) is not a valid p-value in general (and has to be multiplied by at least $\kappa^{1/(\kappa-1)}$ to get a valid p-value). This lack of validity, however, does not matter in this context: the final result (G.2) is still a valid p-to-e merging function. This shows that, in the context of p-to-e merging, one should first perform p-to-e calibration and then pure merging, not vice versa.

Remark G.3. We can generalize (G.2) to

$$F(p_1, \dots, p_K) := \sum_{k=1}^{K} \lambda_k f_k(p_k),$$

where $\lambda_k \geq 0$ sum to 1 and f_k are calibrators. It is interesting that any p-to-e merging function is dominated by a p-to-e merging function of this form [Vovk et al., 2020, Theorem 4.1]. We can see that the classes of e-merging functions, e-to-p merging functions, and p-to-e merging functions admit simple explicit representations. But the class of p-merging functions is much more complex.

Cross-merging under dependency assumptions

In the main part of this appendix we have discussed cross-merging for arbitrarily dependent e-values and p-values. Cross-merging of independent p-values might

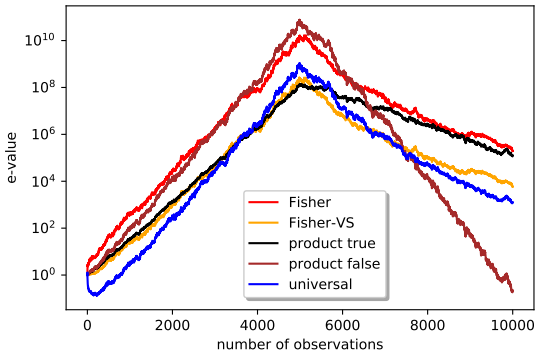


Figure 5: The analogue of Figure 1 where the alternative hypothesis is true half of the time (details in text).

be trivial (we can take any decreasing Borel $F : [0,1]^K \to [0,\infty]$ satisfying $\int F \, \mathrm{d}U \le 1$, U being the uniform probability measure on $[0,1]^K$), but cross-merging of independent e-values raises interesting problems.

Combining the only admissible e-to-p calibrator of Proposition 2.3 and the product ie-merging function (7), we obtain the "ie-to-p" merging function

$$(e_1, \dots, e_K) \mapsto \frac{1}{e_1 \dots e_K} \wedge 1;$$

this is even an *se-to-p merging function*, in the sense of mapping any sequential e-values to a p-value. However, we can do better: by Ville's theorem [Ville, 1939, p. 100],

$$(e_1, \dots, e_K) \mapsto \min_{k=0,\dots,K} \frac{1}{e_1 \dots e_k} \wedge 1$$

is also an se-to-p merging function. For further information, see Shafer et al. [2011].

Strengthening the assumption of $e_1, \dots, e_K$ being sequential to their independence opens up new possibilities for their combination; cf. Vovk and Wang [2020b].

H Additional experimental results

In our experiments in the subsection "Combining independent e-values and p-values" in Section 7 we considered the case where the alternative hypothesis was always true. In this appendix we will report results of experiments in the situation where it is true only part of the time.

Figure 5 uses a similar setting to Figure 1; in particular, the observations are generated from the Gaussian model $N(\mu, 1)$, the null hypothesis is $\mu = 0$ and the alternative hypothesis is $\mu = -0.1$. But now we generate only half (namely, the

first half) of the data (10,000 observations overall) from the alternative distribution, and the rest from the null distribution. The e-variable is the likelihood ratio (24) of the "true" probability density to the null probability density, so that we assume it known that half of the observations are generated from the alternative distribution. The results for (24) are shown in Figure 5 as the black line (all graphs in that figure use the medians over 100 seeds). Comparing the black line with the red line (representing Fisher's method), we can see that their final values are approximately the same. For the comparison to be fairer, we should compare the black line with the orange one (representing the VS bound for Fisher's method); the final value for the black line is significantly higher. Despite the method of multiplication lagging behind Fisher's and the VS bound for it over the first half of the data, it then catches up with them.

As we said in Section A.1, p-values are usually associated with frequentist statistics while e-values are closely connected to Bayesian statistics. As discussed in Section 7, the latter often require stronger assumptions, which is typical of Bayesian statistics. This can be illustrated using the two ways of generating data that we consider in Section 7 and in this appendix so far: always using $N(-0.1, 1)$ or first using $N(-0.1, 1)$ and then $N(0, 1)$. Whereas the p-value is always computed using the same formula (namely, (22)), the e-value is computed as the likelihood ratio (21) or the likelihood ratio (24). The fact that more knowledge is assumed in the case of e-values is further illustrated by the brown line in Figure 5, which is the graph for the product rule that uses the "wrong" likelihood ratio (21) in the case where the alternative hypothesis is true half of the time (as for the other graphs in that figure). Over the first half of the data the product rule performs very well (as in Figure 1), but then it loses all evidence gathered against the null hypothesis. Its final value is approximately 1, despite the null hypothesis being false. The blue line corresponds to the universal test martingale (23) and does not have this deficiency.

I FACT algorithm

Algorithm 3 is a generic procedure that turns any p-merging function F into a function performing multiple hypothesis testing. It is equivalent to the closed testing procedure provided the p-merging function F is symmetric and monotonically increasing in each (equivalently, any) of its arguments. It is a version of Dobriban's [2020] Algorithm 2.

When specialized to Fisher's combination method, Algorithm 3 becomes Algorithm 4, where $F_n^{\chi^2}$ stands for the χ^2 distribution function with n degrees of freedom and line 8 uses the easy-to-check identity

$$1 - F_2^{\chi^2}(-2 \ln p) = p.$$

Algorithm 4 is used in our code [Vovk and Wang, 2020c] for producing Figures 3 and 4.

Algorithm 3 FACT (FAst Closed Testing)

Require: A sequence of p-values $p_1, \dots, p_K$.

Find a permutation π of $\{1, \dots, K\}$ such that $p_{\pi(1)} \le \dots \le p_{\pi(K)}$.
Define the order statistics $p_{(k)} := p_{\pi(k)}$, $k \in \{1, \dots, K\}$.
for $i = 1, \dots, K$ **do**
    $P_i := F(p_{(i)}, \dots, p_{(K)})$
for $k = 1, \dots, K$ **do**
    $p^*_{\pi(k)} := F(p_{\pi(k)})$
    for $i = k+1, \dots, K$ **do**
        $p := F(p_{\pi(k)}, p_{(i)}, \dots, p_{(K)})$
        if $p > p^*_{\pi(k)}$ **then**
            $p^*_{\pi(k)} := p$
    for $i = 1, \dots, k$ **do**
        if $P_i > p^*_{\pi(k)}$ **then**
            $p^*_{\pi(k)} := P_i$

Algorithm 4 FACT on top of Fisher's method

Require: A sequence of p-values $p_1, \dots, p_K$.

Find a permutation π of $\{1, \dots, K\}$ such that $p_{\pi(1)} \le \dots \le p_{\pi(K)}$.
Define the order statistics $p_{(k)} := p_{\pi(k)}$, $k \in \{1, \dots, K\}$.
$S_{K+1} := 0$
for $i = K, \dots, 1$ **do**
    $S_i := S_{i+1} - 2 \ln p_{(i)}$
    $P_i := 1 - F^{\chi^2}_{2(K+1-i)}(S_i)$
for $k = 1, \dots, K$ **do**
    $p^*_{\pi(k)} := p_{\pi(k)}$
    for $i = K, \dots, k+1$ **do**
        $p := 1 - F^{\chi^2}_{2(K+2-i)}(-2 \ln p_{\pi(k)} + S_i)$
        if $p > p^*_{\pi(k)}$ **then**
            $p^*_{\pi(k)} := p$
    for $i = 1, \dots, k$ **do**
        if $P_i > p^*_{\pi(k)}$ **then**
            $p^*_{\pi(k)} := P_i$