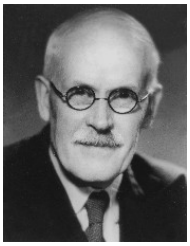
True and false discoveries with *e*-values

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 #3

First posted 31 December 2019 (on arXiv) and 14 February 2020
(as Working Paper). Last revised March 29, 2020.

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

Abstract

The topic of this paper is multiple hypothesis testing based on e-values, which are Bayes factors stripped of their Bayesian content. Using e-values instead of p-values, which are standard in this area, leads to simple and efficient procedures that control the number of false discoveries under arbitrary dependence of the base e-values. We prove an optimality result for our main procedure and demonstrate advantages of our methods over standard methods using simulated and real-world datasets.

Contents

1 Introduction

There are numerous ways of merging p-values, both under the assumption of independence and in general. One of the more exotic ways is merging by averaging [Vovk and Wang, 2019a]; however, the arithmetic average has to be scaled up by a factor of 2 [Rüschendorf, 1982; Meng, 1993] to get a valid merging function. The situation with e-values is radically different: no scaling is required for arithmetic averaging, and moreover, arithmetic averaging becomes essentially the only symmetric method of merging [Vovk and Wang, 2019b, Proposition 2.1].

In this paper we define an analogue for e-values of a known procedure of controlling the number of false discoveries [Genovese and Wasserman, 2004, 2006; Goeman and Solari, 2011a; Goeman et al., 2019b]. Whereas the original procedure is based on a merging function for p-values, our procedures are based, naturally, on arithmetic averaging of e-values. These procedures control the number of false discoveries in a stronger sense than the well-known procedure of Benjamini and Hochberg [1995] controlling the false discovery rate (FDR). Whereas FDR is the expected value of the false discovery proportion, in this paper we are interested in upper bounds on the number of false discoveries that can be asserted with various degrees of confidence.

This paper continues the programme of exploring the notion of e-values outside the framework of Bayesian statistics. See, e.g., Shafer [2019], Shafer and Vovk [2019, Section 11.5], Grünwald et al. [2019], and Vovk and Wang [2019b]. Different papers and books use different terms for our "e-value", such as "betting score", "Skeptic's capital", or "S-value".

E-values can be defined as values taken by e-variables, and an *e-variable* is a random variable taking values in $[0, \infty]$ whose expectation is at most 1 under the null hypothesis. If E is an e-variable and $c > 0$, Markov's inequality implies that, under the null hypothesis, $E \geq c$ with probability at most $1/c$. Therefore, observing a large value of E provides evidence against the null hypothesis.

In many areas of statistics e-variables appear naturally as likelihood ratios: if Q is a simple null hypothesis and Q' is an alternative probability measure, the Radon–Nikodym derivative $\mathrm{d}Q'/\mathrm{d}Q$ is an e-variable. In Bayesian statistics, Q or Q' or both may be defined as marginal probability measures for Bayesian models, in which case they are known as Bayes factors. The fundamental monograph treating Bayes factors is Jeffreys's [1961]; see, e.g., Ly et al. [2016] for a recent appreciation.

We start the main part of the paper by describing our procedures in Section 2 and establishing the validity of the main one (Theorem 2.1). In Sections 3 and 4 we demonstrate their advantages in simulation and empirical studies, respectively. In Section 5 we state an optimality result for our main procedure (Theorem 5.1). Section 6 concludes.

The main content of the paper is complemented by four appendixes, A, B, C, and D. Appendix A explores other procedures of multiple hypothesis testing, including the one based on a Bonferroni-type procedure of merging e-values. If the goal is family-wise validity, such procedures (such as the one in Holm

[1979]) usually work very well, but if the goal is to control the number of false discoveries, they work much worse than the optimal procedure. In this appendix we also discuss a Simes-type procedure based on e-values. In Appendix B we give a computationally efficient implementation of our method. Appendix C makes connections with Goeman and Solari's [2011a] work explicit. As Hemerik et al. [2019, Supplementary material] explain, the method of Goeman and Solari [2011a] is equivalent to a method in Genovese and Wasserman [2004, 2006]. Finally, Appendix D points out the importance of generalized Bayes factors.

2 Controlling true and false discoveries

Let us fix a measurable space $(\Omega, \mathcal{A})$ for now. This is our *sample space*; to complete it to a probability space we need a probability measure $Q \in \mathfrak{P}(\Omega)$, where $\mathfrak{P}(\Omega)$ is the set of all probability measures on $(\Omega, \mathcal{A})$. When given such a Q, we have the notion of expectation $\mathbb{E}^Q[E] := \int E \, \mathrm{d}Q \in [0, \infty]$ for each extended random variable E taking nonnegative values (we call it "extended" since it may take value ∞).

A *simple statistical hypothesis* is an element Q of $\mathfrak{P}(\Omega)$. An *e-test* is a family of extended nonnegative random variables E_Q, $Q \in \mathfrak{P}(\Omega)$, that satisfies

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

We will refer to an extended nonnegative E satisfying $\mathbb{E}^Q[E] \le 1$ as an *e-variable for testing* Q. Therefore, for each simple statistical hypothesis Q, an e-test fixes an e-variable E_Q for testing Q.

A *statistical hypothesis* (or *composite statistical hypothesis*, or simply *hypothesis*) is a set $H \subseteq \mathfrak{P}(\Omega)$ of probability measures on the sample space. An *e-variable for testing* H is an extended random variable $E : \Omega \to [0, \infty]$ satisfying $\mathbb{E}^Q[E] \le 1$ for all $Q \in H$. The values taken by e-variables are referred to as *e-values*. Intuitively, observing a large e-value provides evidence against H; in Section 3 we will discuss Jeffreys's rule of thumb for evaluating the strength of such evidence. We embed the simple statistical hypotheses into the composite statistical hypotheses by identifying $Q \in \mathfrak{P}(\Omega)$ with the corresponding singleton $\{Q\} \subseteq \mathfrak{P}(\Omega)$.

Suppose we are given $K \ge 2$ e-values $e_1, \dots, e_K$ for testing hypotheses $H_1, \dots, H_K$, which are our *base hypotheses*; we would like to reject some of them (in fact, as many of them as possible under a validity constraint).

If we do not know anything about the nature of the hypotheses $H_1, \dots, H_K$, it makes sense to reject a number of H_k with the largest e_k. But in general, we can consider an arbitrary *rejection set*

$$R \subseteq \{1, \dots, K\}; \tag{1}$$

this is the set of base hypotheses, represented by their indices, that the researcher chooses to reject. Goeman and Solari [2011a] argue convincingly that in some practically relevant cases R will not necessarily correspond to the largest

e_k; e.g., R may include hypotheses connected by a common theme, such as being related to the gastrointestinal tract [Goeman and Solari, 2011a, Subsection 4.1]. (For a much more elaborate picture, see Ramdas et al. [2019].)

If the researcher rejects H_k, we refer to this decision as a *discovery*. If Q is the true probability measure, which is not known to the researcher, then we say that the discovery is *true* if $Q \notin H_k$, and it is *false* if $Q \in H_k$. For the rejection set (1), the number of true discoveries is $|\{k \in R \mid Q \notin H_k\}|$ and the number of false discoveries is $|\{k \in R \mid Q \in H_k\}|$. The sum of these two numbers is K (the total number of discoveries).

Let us say that a family of measurable functions $D_R : \{1, \dots, |R|\} \times \Omega \to [0, \infty]$ indexed by non-empty rejection sets $R \subseteq \{1, \dots, K\}$ is a *discovery vector* if there exists an e-test $E = (E_Q)_{Q \in \mathfrak{P}(\Omega)}$ such that

$$\forall R \, \forall j \in \{1, \dots, |R|\} \, \forall Q \in \mathfrak{P}(\Omega) \, \forall \omega \in \Omega : \\ \left(|\{k \in R \mid Q \notin H_k\}| \geq j \right) \vee \left(E_Q(\omega) \geq D_R(j, \omega) \right), \quad (2)$$

where $\vee$ stands for "or". We will say that such an e-test E *witnesses* that D is a discovery vector. Since D_R is a random function of $j \in \{1, \dots, |R|\}$ (namely, a random vector), we sometimes write $D_R(j)$ instead of $D_R(j, \omega)$ suppressing the dependence on ω (as we do for other random functions).

The disjunction in (2) is of the kind discussed by Fisher [1973, Section III.1]: assuming $D_R(j, \omega)$ is large for some j, either there are at least j true discoveries (rejections of false null hypotheses) or a rare chance has occurred (namely, the observed e-value is at least $D_R(j, \omega)$). In this sense D controls the number of true discoveries. (Notice that, intuitively, controlling true discoveries and controlling false discoveries are the same thing, since the total number of discoveries $|R|$ is known.)

Let E_k be an e-variable for testing H_k, $k = 1, \dots, K$. The *arithmetic-mean discovery vector* is defined as

$$\mathrm{AV}_R(j) := \min_{I \subseteq \{1, \dots, K\} : |R \setminus I| < j} \frac{1}{|I|} \sum_{i \in I} E_i, \quad j \in \{1, \dots, |R|\} \quad (3)$$

(notice that $I = \emptyset$ is excluded for any j). The first main mathematical result of this paper (the very simple Theorem 2.1 below) says that the arithmetic-mean discovery vector controls the number of true discoveries. Later (Theorem 5.1) we will see that it is optimal in a natural sense.

Theorem 2.1. *The arithmetic-mean discovery vector* AV_R *is a discovery vector, for any e-variables* $E_1, \dots, E_K$.

Proof. The function witnessing that AV_R is a discovery vector will be the arithmetic mean

$$E_Q := \frac{1}{|I_Q|} \sum_{k \in I_Q} E_k, \quad (4)$$

Algorithm 1 Counting true discoveries for a given R

Require: An increasing sequence of e-values $e_1 \le \dots \le e_K$.
Require: The rejected hypotheses $R \subseteq \{1, \dots, K\}$.
for $j = 1, \dots, |R|$ **do**
  let R_j be R without its $j-1$ largest elements
  $\mathrm{AV}_R(j) := F_{\mathbf{e}}(R_j)$
  for $i = 1, \dots, K$ **do**
    $e := F_{\mathbf{e}}(R_j \cup \{1, \dots, i\})$
    if $e < \mathrm{AV}_R(j)$ **then**
      $\mathrm{AV}_R(j) := e$

where $I_Q := \{k \mid Q \in H_k\}$. If $I_Q = \emptyset$, the disjunction in (2) is obvious for any j; therefore, we can define the e-variable E_Q arbitrarily in this case (e.g., as 0 or 1).

To check the disjunction in (2) for $D := \mathrm{AV}$ and for given $Q \in \mathfrak{P}(\Omega)$, R, $j \in \{1, \dots, |R|\}$, and $\omega \in \Omega$, let us assume that the second term of the disjunction is false, namely $E_Q(\omega) < \mathrm{AV}_R(j, \omega)$. By the definitions (3) and (4), this means

$$\frac{1}{|I_Q|} \sum_{k \in I_Q} E_k(\omega) < \min_{I \subseteq \{1,\dots,K\} : |R \setminus I| < j} \frac{1}{|I|} \sum_{i \in I} E_i(\omega),$$

and we can see that $|R \setminus I_Q| \ge j$. We have at least j true discoveries, which establishes the disjunction. □

Our definitions so far and Theorem 2.1 are essentially translations of Goeman and Solari's [2011a, Section 2] definitions and their statement of validity into the language of e-values. We will explain the connection in detail in Appendix C. As the procedure for p-values was first proposed in Genovese and Wasserman [2004, 2006], we will refer to it as the *GWGS procedure*.

A polynomial-time algorithm for computing AV_R is given as Algorithm 1. It uses the notation

$$F_{\mathbf{e}}(I) := \frac{1}{|I|} \sum_{i \in I} e_i, \quad I \subseteq \{1, \dots, K\}, \ I \ne \emptyset, \tag{5}$$

where $\mathbf{e} := (e_1, \dots, e_K)$ (as we said earlier, arithmetic averaging is essentially the only symmetric e-merging function [Vovk and Wang, 2019b, Proposition 2.1]). Without loss of generality we assume that the e-values are sorted in the ascending order, $e_1 \le \dots \le e_K$.

Next we will discuss a less flexible method in which we consider a family of rejection sets R that are chosen in an optimal way, in some sense. For each $r \in \{1, \dots, K\}$, the set

$$R_r := \{K - r + 1, \dots, K\} \tag{6}$$

Algorithm 2 Arithmetic-mean discovery matrix

Require: An increasing sequence of e-values $e_1 \le \dots \le e_K$.

for $r = 1, \dots, K$ **do**
  for $j = 1, \dots, r$ **do**
    $S_{r,j} := \{K - r + 1, \dots, K - j + 1\}$
    $\mathrm{AM}_{r,j} := F_{\mathbf{e}}(S_{r,j})$
    for $i = 1, \dots, K - r$ **do**
      $e := F_{\mathbf{e}}(S_{r,j} \cup \{1, \dots, i\})$
      if $e < \mathrm{AM}_{r,j}$ **then**
        $\mathrm{AM}_{r,j} := e$

is the optimal rejection set of size r, meaning that, for any other set $R \subseteq \{1, \dots, K\}$ of size r, we have $\mathrm{AV}_{R_r} \ge \mathrm{AV}_R$. In the terminology of statistical decision theory [Wald, 1950, Section 1.3], R_r is the minimal complete class of rejection sets. Of course, this notion of optimality depends on the symmetric treatment of the base e-values (in which case we know AV to be optimal).

The output of Algorithm 2 is a matrix $\mathrm{AM}_{r,j}$, where $r \in \{1, \dots, K\}$ and $j \in \{1, \dots, r\}$. We regard AM as a $K \times K$ matrix whose elements above the main diagonal are undefined and refer to it as the *arithmetic-mean discovery matrix*. It is defined as

$$\mathrm{AM}_{r,j} := \mathrm{AV}_{R_r}(j), \quad r \in \{1, \dots, K\}, \ \ j \in \{1, \dots, r\},$$

where R_r is the optimal rejection set (6) of size r. By the definition of a discovery vector, a large value of $\mathrm{AM}_{r,j}$ is evidence for the statement "there are at least j true discoveries among the r hypotheses (with the largest e-values) that we choose to reject". Possible uses of such a matrix will be discussed in the next section.

3 Simulation studies

In our simulation studies we will visualize the arithmetic-mean discovery matrix in some simple cases and compare Algorithm 2 with a method based on p-values. Our setting will be similar to that of Vovk and Wang [2019b, Section 8], where we study family-wise validity.

The observations are generated from the Gaussian model $N(\mu, 1)$. The null hypotheses are $N(0, 1)$ and the alternatives are $N(\delta, 1)$, where we take $\delta := -3$ throughout the section. We generate $K/2$ observations from $N(\delta, 1)$ (the alternative distribution) and then $K/2$ observations from $N(0, 1)$ (the null distribution), where K (the overall number of hypotheses) is an even number.

Figure 1 shows the arithmetic-mean discovery matrix that Algorithm 2 gives for $K = 20$: we generate 10 observations from $N(\delta, 1)$ and then 10 from $N(0, 1)$. The base e-values are the likelihood ratios

$$E(x) := \frac{\exp(-(x - \delta)^2/2)}{\exp(-x^2/2)} = \exp(\delta x - \delta^2/2) \tag{7}$$

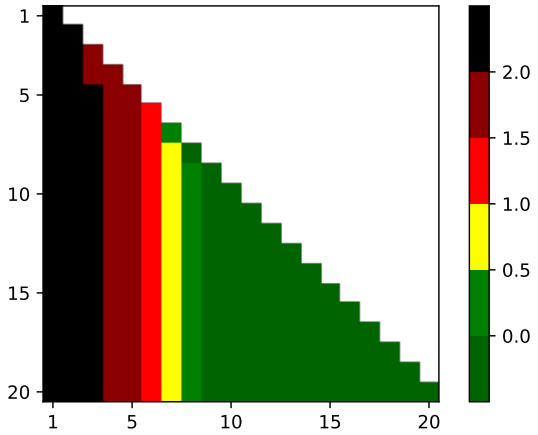


Figure 1: The arithmetic-mean discovery matrix for 10 false and 10 true null hypotheses, as described in text. The colour map on the right gives Jeffreys's thresholds, the boundaries between different colours in most of our plots, on the decimal log scale.

of the alternative to the null density, where $x \sim N(\mu, 1)$ is the corresponding observation.

In this paper, we colour-code the entries of discovery matrices containing e-values according to Jeffreys's [1961, Appendix B] rule of thumb:

- If an entry $\mathrm{AM}_{r,j}$ is below 1, the null hypothesis is supported. Such entries are shown in dark green.
- If $\mathrm{AM}_{r,j} \in (1, \sqrt{10}) \approx (1, 3.16)$, the evidence against the null hypothesis is not worth more than a bare mention. Such entries are shown in green.
- If $\mathrm{AM}_{r,j} \in (\sqrt{10}, 10) \approx (3.16, 10)$, the evidence against the null hypothesis is substantial. Such entries are shown in yellow.
- If $\mathrm{AM}_{r,j} \in (10, 10^{3/2}) \approx (10, 31.6)$, the evidence against the null hypothesis is strong. Such entries are shown in red.
- If $\mathrm{AM}_{r,j} \in (10^{3/2}, 100) \approx (31.6, 100)$, the evidence against the null hypothesis is very strong. Such entries are shown in dark red.
- If $\mathrm{AM}_{r,j} > 100$, the evidence against the null hypothesis is decisive. Such entries are shown in black.

The full colour map is shown on the right of Figure 1 with the thresholds between different colours given in terms of the decimal logarithm of $\mathrm{AM}_{r,j}$. The most interesting parts of our plots of discovery matrices are those in yellow and red; green and dark green parts carry little or no evidence and so are useless for us,

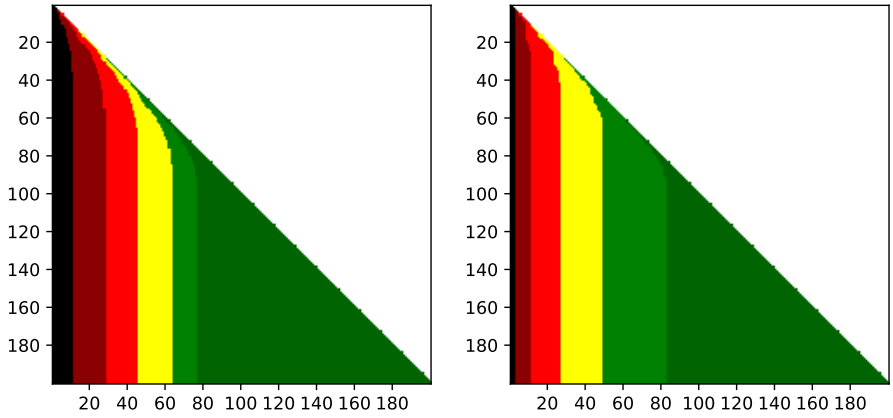


Figure 2: Left panel: the arithmetic-mean discovery matrix for 100 false and 100 true null hypotheses. Right panel: the discovery p-matrix in the same situation for the VS bound for the GWGS procedure and Jeffreys's thresholds, as described in text.

and dark red and black parts carry so much evidence that they are rare in a wide range of practical applications (cf. Section 4). In all our discussions below we will ignore the boundaries between the green and dark green parts.

The left panel of Figure 2 is the counterpart of Figure 1 for a larger number of hypotheses, $K = 200$: we generate 100 observations from $N(\delta, 1)$ and then 100 from $N(0, 1)$. We will refer to this set of observations as the *simulation data*.

In practice, a discovery matrix, such as that shown in the left panel of Figure 2, can be used in different ways, for example:

- The researcher may have budget for a limited number of follow-up studies of the hypotheses. For example, if in the situation of Figure 2 her budget is 50 hypotheses, she just concentrates on row 50 (studying the 50 hypotheses H_k with the largest e-values). For the first 11 entries in this row the e-value exceeds 100, and so she has decisive evidence that there are at least 11 true discoveries among those 50 hypotheses. Similarly,
  - she has at least very strong evidence that there are at least 27 true discoveries,
  - she has at least strong evidence that there are at least 40 true discoveries,
  - she has at least substantial evidence that there are at least 46 true discoveries.
- The researcher might have some idea of what proportion of false discoveries she is willing to tolerate (in the spirit of choosing the false discovery rate

a priori [Benjamini and Hochberg, 1995]). For example, if she is willing to tolerate 10% of false discoveries and willing to use Jeffreys's standard (e-value greater than 10) of strong evidence, she should concentrate on row 31 (i.e., study the 31 hypotheses with the largest e-values), which is the lowest row with at most 10% of entries below 10.

- Alternatively, the researcher might have some idea of how many false discoveries she is willing to tolerate (in the spirit of k-FWER [Romano and Wolf, 2007]). If she is willing to tolerate at most 10 false discoveries and still willing to use Jeffreys's standard of strong evidence, she should concentrate on row 51, which is the lowest row with at most 10 entries (in fact, exactly 10 entries) below 10.

Of course, the researcher may know her hypotheses and relations between them very well, and after looking at the discovery matrix she may come up with her own rejection set R, as discussed in Section 2. In this case she should also use Algorithm 1.

Comparisons

This paper concentrates on multiple hypothesis testing using e-values, but in scientific practice p-values are more popular, despite recent criticism. In this section we will report results of our simulation studies in terms of p-values, but first let us discuss ways to turn p-values into e-values (known as calibrating p-values) and vice versa. For further details, see Vovk and Wang [2019b, Section 5].

A *p-variable* w.r. to a statistical hypothesis H (which we will not mention explicitly) is a random variable $P : \Omega \to [0, \infty)$ satisfying

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

A decreasing function $f : [0, 1] \to [0, \infty]$ is a *calibrator* if, for any p-variable P, $f(P)$ is an e-variable. In other words, a calibrator transforms p-values to e-values.

A very natural family of calibrators is

$$f_\kappa(p) := \kappa p^{\kappa - 1}, \tag{8}$$

where $\kappa \in (0, 1]$. The best e-value

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

attainable by this family will be referred to as the *VS bound* (abbreviating "Vovk–Sellke bound" [Shafer and Vovk, 2019, Section 11.5]). Of course, $\mathrm{VS}(P)$ need not be an e-variable even if P is a p-variable.

In the opposite direction, a decreasing function $f : [0, \infty] \to [0, 1]$ is an *e-to-p calibrator* if, for any e-variable E, $f(E)$ is a p-variable. It is a function

transforming e-values to p-values. As explained and formalized in Vovk and Wang [2019b, Proposition 5.2],

$$t \in [0, \infty] \mapsto \min(1, 1/t) \tag{9}$$

is the only reasonable e-to-p calibrator.

In general, calibrating p-values and e-values are crude processes. A strong e-value of 20 barely attains statistical significance when transformed into a p-value (namely, 5%) using (9). The VS bound for the borderline significant p-value of 5% is approximately 2.456, and so "is not worth more than a bare mention", according to Jeffreys. The low "round-trip efficiency" in the domain of p-values can be illustrated by

$$1/\operatorname{VS}(0.005) \approx 0.072. \tag{10}$$

The round trip turns the highly significant p-value of 0.5% into the non-significant p-value of 7.2%. And this is despite the VS bound being achievable as e-value only in hindsight.

For comparison with methods based on p-values, we use the GWGS procedure applied to Simes's [1986] procedure for combining p-values and to the same nested rejection sets (initial subsets of $\{1, \dots, 200\}$ assuming the p-values are given in descending order). This procedure admits a computationally efficient shortcut [Goeman et al., 2019b, Theorem 1] and is implemented in the R package `hommel` [Goeman et al., 2019a]. As the base p-values we take $P(x) := N(x)$, where N is the standard Gaussian distribution function; these are the p-values found using the most powerful test given by the Neyman–Pearson lemma. The GWGS procedure can be interpreted as producing an analogue of a discovery matrix, which we call a *discovery p-matrix*, with e-values replaced by p-values. For details, see Appendix C. The right panel of Figure 2 shows, using Jeffreys's thresholds, the VS bounds for the discovery p-matrix found using `hommel` applied to the simulation data.

Our method does not require any assumptions about the dependence structure of the e-values, whereas such assumptions (such as independence or positive dependence [Sarkar, 2011]) about p-values are essential for Simes's method. It is true that our simulated data are independent, but this information is typically unavailable, and the performance of methods that do not depend on independence is still interesting. The package `hommel` includes functions for multiple hypothesis testing that do not rely on Simes's inequality. Figure 3 compares the GWGS discovery p-matrix that does not rely on Simes's inequality in `hommel` with the arithmetic-mean matrix e-to-p-calibrated using the trivial e-to-p calibrator $e \mapsto 1/e$. It uses Fisher's thresholds 1% and 5%; the values below 1% are shown in red, between 1% and 5% in yellow, and above 5% in green (so that red means "highly significant" and yellow means "significant but not highly significant"). It is interesting that even after the crude step of calibration (remember its woeful round-trip efficiency illustrated by (10)) we obtain competitive results. (But of course assuming independence makes direct treatment of p-values more efficient: see Figure 9 in Appendix C.)

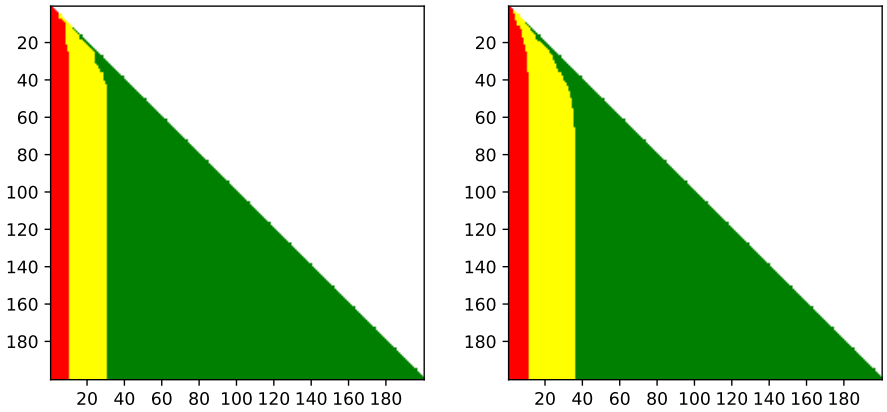


Figure 3: Left panel: the GWGS discovery p-matrix for the simulation data (as in Figure 2) for Fisher's thresholds 1% and 5% under arbitrary dependence, as described in text. Right panel: the e-to-p-calibrated arithmetic-mean discovery matrix, still using Fisher's thresholds (with values below 1% shown in red and between 1% and 5% in yellow).

Table 1: The Benjamini–Hochberg and Benjamini–Yekutieli procedures applied to the simulation data for FDR (false discovery rate) 5% and 1%.

assumption	5%	1%
independence	87	61
arbitrary dependence	55	28

Table 1 gives the numbers of null hypotheses rejected by the Benjamini–Hochberg procedure [Benjamini and Hochberg, 1995] and its version for arbitrary dependence [Benjamini and Yekutieli, 2001]. Here the results are more difficult to interpret, since the kind of guarantees provided by those procedures is so different from the guarantees provided by our methods and GWGS methods. Roughly, we get comparable results.

4 Empirical studies

In this section we will demonstrate how the methods of this paper can, in principle, be used in practice. It is important that we will make no independence or exchangeability assumptions.

We will use the classical dataset first described in Hedenfalk et al. [2001] and then carefully studied in Storey and Tibshirani [2003]. Essentially, we will adapt Storey and Tibshirani's analysis to using e-values in place of p-values (see the end of the section for a discussion of differences). In our experiments we use the version of the dataset made available as part of the R package `qvalue` [Storey et al., 2019].

The main content of the dataset is the expression levels of 3226 genes in 15 samples of tissues taken from 14 patients. Seven patients are carriers of mutations in the BRCA1 gene, and another seven are carriers of mutations in the BRCA2 gene. One of the carriers of a BRCA2 mutation had two tissue samples in the dataset. Therefore, there are 15 samples overall in the dataset, seven labelled BRCA1 and eight labelled BRCA2. The core of the dataset is the 3226×15 matrix of gene expressions in the samples; all entries are positive numbers. Following Storey and Tibshirani [2003, Appendix, Remark C], we remove all rows containing at least one entry exceeding 20, which leaves us with a 3170×15 data matrix.

Storey and Tibshirani's version of the dataset also contains some further information, such as the p-value for each gene corresponding to the statistic (14), to be discussed below. For further information about this dataset, which we will refer to as the *BRCA dataset*, see, e.g., Storey et al. [2007, Section 5] and Guindani et al. [2009, Section 5.2].

For this dataset methods ensuring family-wise validity do not work well. For example, the three smallest p-values in Storey and Tibshirani's list multiplied by the number of genes 3170 are

$$0.0000, 0.0400, 0.0501, \tag{11}$$

and so the Bonferroni correction leads to only two statistically significant p-values. Moreover, one of the p-values is exactly zero, and so cannot be a valid p-value (for details, see p. 14). Hedenfalk et al. conclude that 9–11 genes are differentially expressed. Storey and Tibshirani's informal analysis suggests that many more, at least 33%, of the examined genes are differentially expressed. However, their informal analysis assumes what they call "weak independence": they rely on the law of large numbers when inspecting histograms of p-values,

assuming that the probabilities of the p-values lying in various ranges will manifest themselves as empirical frequencies seen in the histograms. Their formal analysis is asymptotic and also assumes weak independence: see their Appendix, Remark D.

For computing base e-values, we use the formula

$$e_k := \frac{T_k}{\frac{1}{B+1}\left(\sum_{b=1}^{B} T_k^{0b} + T_k\right)}, \quad k = 1, \dots, 3170, \tag{12}$$

where T_k is a nonnegative number (*nonconformity score*, to be defined momentarily) computed from the kth row of the data matrix with each element labelled BRCA1 or BRCA2, T_k^{0b} is the nonconformity score computed from the same row with randomly permuted labels, and B is the number of permutations. In our experiments, the case $0/0$ of the right-hand side of (12) never occurs. We will call (12) the *conformal e-value*. We are justified in calling it an e-value since, under the null hypothesis that the label are uninformative, the expected value of the right-hand side of (12) is 1 if we set $0/0 := 1$.

We define the nonconformity score as $T_k := f(t_k)$ for some function f of the t-statistic t_k, which we will define shortly (see (14)). Our choice of f is motivated by the Bayesian two-sample t-test widely discussed in recent literature starting from Gönen et al. [2005] and briefly reviewed in Gönen et al. [2019, Section 3]. A standard expression for the Bayes factor produced by such a test via the t-statistic t is

$$f(t) := c\left(1 + at^2\right)^{d/2} \tag{13}$$

for positive constants a, b, and d involving the number of degrees of freedom and effective sample size; see, e.g., Wang and Liu [2016, (14)] and Rouder et al. [2009, (1)]; the form (13) goes back to Jeffreys [Ly et al., 2016, (12)]. However, different constants are used in different papers. We set, without loss of generality, $c := 1$, since c cancels out when using (12). We further simplify (13) by ignoring the "$1+$"; this makes a and any constant factors in the definition of the t-statistic t (there is a non-trivial factor under the assumption of equal variances for the two groups) irrelevant, as they also cancel out when applying (12). Of course, this step does not affect the validity of our methods.

Let x_{kj} be the base two logarithm of the value in row k and column j of the data matrix (although the base does not matter in our empirical studies). The two-sample t-statistic for the kth gene is

$$t_k := \frac{\bar{x}_{k2} - \bar{x}_{k1}}{\sqrt{s_{k1}^2/n_1 + s_{k2}^2/n_2}}, \tag{14}$$

where $n_1 = 7$ is the number of BRCA1 columns, $n_2 = 8$ is the number of BRCA2 columns, and

$$\bar{x}_{k1} := \frac{1}{n_1} \sum_{j \in \text{BRCA1}} x_{kj}, \quad s_{k1}^2 := \frac{1}{n_1 - 1} \sum_{j \in \text{BRCA1}} (x_{kj} - \bar{x}_{k1})^2$$

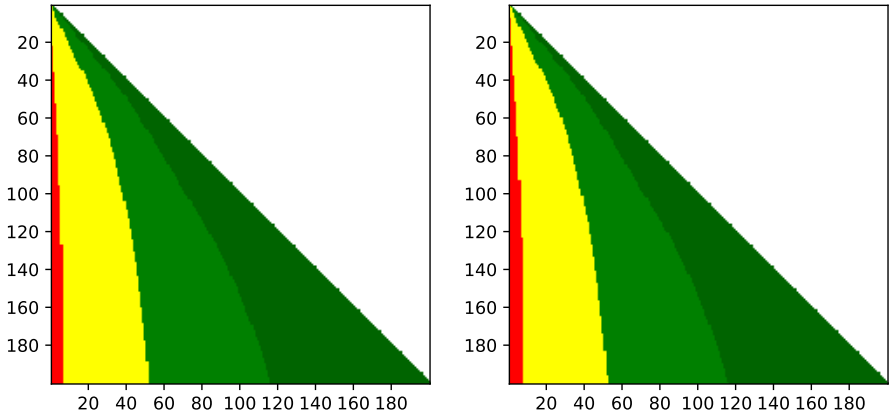


Figure 4: Left panel: the top-left 200×200 corner of the arithmetic-mean discovery matrix for the `BRCA` dataset for $B := 10000$, using Jeffreys's thresholds. Right panel: its version (based on (17)) that is only approximately valid.

are the sample mean and variance for the BRCA1 entries, with the analogous expressions for BRCA2. The variances of the two groups (BRCA1 and BRCA2) are not assumed to be equal (following Storey and Tibshirani [2003]), but using equal-variance two-sample t-statistics would lead to similar results.

The left panel of Figure 4 gives a key part of the arithmetic-mean discovery matrix for the BRCA dataset with $f(t) := |t|^d$ for $d := 10$, with $B := 10000$, and with base conformal e-values (12). We can see that there is strong evidence that the number of differentially expressed genes is at least as large as Hedenfalk at al.'s number. If we settle for substantial evidence, the number is much larger. Arguably, it is not as large as in Storey and Tibshirani's study, but we are not using any exchangeability or independence assumptions.

Dependence on the initial state of the random numbers generator (always set to 1 in our experiments) is fairly significant but does not affect our conclusions. Dependence on the value of d is also significant; the values below 10 tend to lead to higher numbers of true discoveries for Jeffrey's standard of substantial evidence, and the values above 10 to higher numbers of true discoveries for strong evidence (up to a limit; see Table 2). The literature on the Bayesian two-sample t-test quoted above seems to suggest that d should have the same order of magnitude as the number of degrees of freedom.

Comparisons

We start by comparing our methodology with that of Storey and Tibshirani [2003], which was our main source of data and ideas in this section. The main differences are:

Table 2: Summaries of the last row of the arithmetic-mean discovery matrix for different values of d. Column "strong" contains the number of entries that are greater than 10 (all of them are below $10^{3/2}$, and so they provide strong evidence, i.e., red in our pictures). Column "substantial" contains the number of entries that are greater than $10^{1/2}$ (providing at least substantial evidence).

d	strong	at least substantial
4	0	62
6	0	82
8	4	70
10	7	56
12	8	46
20	9	29
50	8	17
100	7	14

- Storey and Tibshirani use p-values whereas we use e-values.
- Storey and Tibshirani implicitly assume that the genes are exchangeable under the null hypothesis.
- Moreover, Storey and Tibshirani assume that the p-values are weakly independent.

Strictly speaking, Storey and Tibshirani's method does not produce valid p-values, even under their null hypothesis implicitly involving gene exchangeability. This can be seen from their formula for computing the p-values,

$$p_k := \frac{\sum_{b=1}^{B} \left|\{j : \left|t_j^{0b}\right| \geq |t_k|, j = 1, \dots, 3170\}\right|}{3170 \cdot B} \tag{15}$$

(the last displayed equation in their Appendix, Remark C), where t_k is the t-statistic for gene k and t_j^{0b} is the t-statistic for gene j with the labels BRCA1 and BRCA2 randomly permuted (for the bth random permutation, $b = 1, \dots, B$ and $B := 100$). The numerator of (15) can well be zero (and it is in one case: see (11)).

To turn the expression (15) into a valid p-value (under the null hypothesis of gene exchangeability and label uninformativeness), it suffices to add 1 to the numerator and denominator of (15), as in the method of conformal prediction [Vovk et al., 2005]. Let us call

$$\frac{\sum_{b=1}^{B} \left|\{j : \left|t_j^{0b}\right| \geq |t_k|, j = 1, \dots, 3170\}\right| + 1}{3170 \cdot B + 1} \tag{16}$$

the *conformal p-value*. The intuition behind the expression (16) is that, to see how well t_k conforms to the multiset of size $3170 \cdot B$ consisting of t_j^{0b}, we add

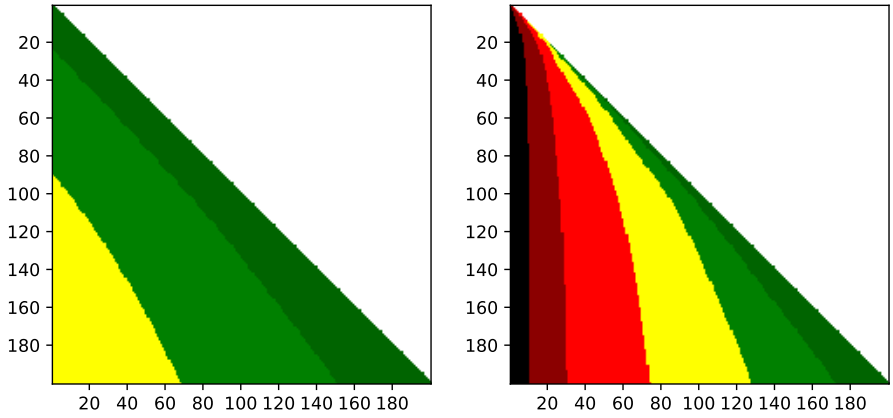


Figure 5: Left panel: the top-left 200×200 corner of the arithmetic-mean discovery matrix for the `BRCA` dataset for $B := 100$. Right panel: its simplified version whose lack of validity is visible.

t_k to the multiset before computing the rank p-value. We may regard (15) as a simplified version of the conformal p-value. Since we are comparing the t-statistic for gene k with t-statistics for other genes in (15) and (16), we are implicitly assuming gene exchangeability.

Under gene exchangeability, for computing base e-values, we can use the formula

$$e_k := \frac{T_k}{\frac{1}{3170 \cdot B + 1} \left(\sum_{b=1}^{B} \sum_{j=1}^{3170} T_j^{0b} + T_k \right)},$$

in analogy with (16). This gives an e-variable under the assumption that the labels are uninformative and the genes are exchangeable.

To avoid the assumption of gene exchangeability, we use the expression (12) thus avoiding comparing the statistic pertaining to gene k to statistics pertaining to other genes. Our value of B, $B = 10000$, is much larger than Storey and Tibshirani's $B = 100$.

We can also introduce a simplified version of (12):

$$e_k := \frac{T_k}{\frac{1}{B} \sum_{b=1}^{B} T_k^{0b}}, \tag{17}$$

in analogy with (15). This version may be more intuitive, but it is only approximately valid. When $B = 10000$, there is not much difference between using (12) and using (17): see the right panel of Figure 4, which uses (17).

A useful role of the version (17) may be to check whether the value of B in (12) is sufficiently large. In the case of Figure 4, the approximation is good, which suggests that B is sufficiently large. However, in the case of Figure 5,

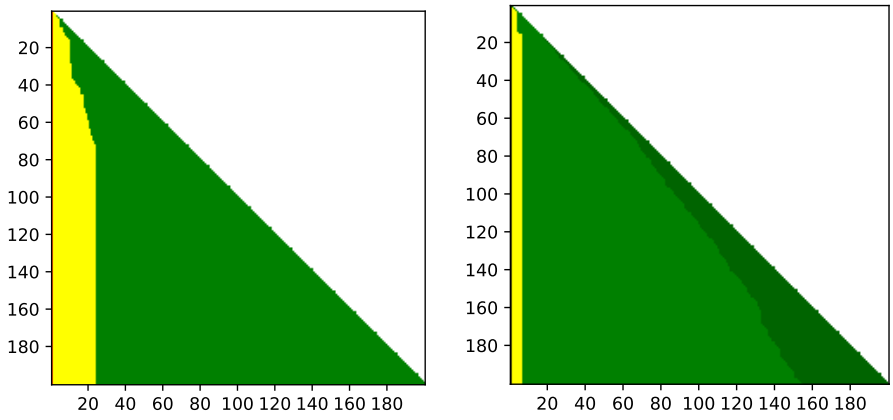


Figure 6: Left panel: the top-left corner of the GWGS discovery p-matrix for the `BRCA` dataset for Fisher's thresholds 1% and 5%, assuming independence. Right panel: analogous picture for Jeffreys's thresholds applied to the VS bounds of the entries of this matrix.

Table 3: The Benjamini–Hochberg and Benjamini–Yekutieli procedures applied to the `BRCA` dataset (the three entries of "1" are unreliable as they are based on a zero p-value).

assumption	5%	1%
independence	88	1
arbitrary dependence	1	1

where $B = 100$ (as in Storey and Tibshirani [2003]), the right-hand panel, which uses (17), looks far too good to be valid. On the other hand, the left-hand panel, which uses (12), is valid but extremely conservative.

Results given by `hommel` are either poor (when independence is assumed) or extremely poor (under arbitrary dependence). The former are given in Figure 6 and the latter are given in Appendix C (Figure 10).

The Benjamini–Hochberg procedure [Benjamini and Hochberg, 1995] rejects 88 null hypotheses at FDR $q := 0.05$ and 1 null hypothesis at FDR $q := 0.01$ for Storey and Tibshirani's list of p-values (of course, we will always reject at least 1 null hypothesis because of the zero p-value on their list). However, this procedure assumes independence. Under arbitrary dependence, we can control FDR by replacing q by $q/\sum_{k=1}^{K} 1/k$ [Benjamini and Yekutieli, 2001, Theorem 1.3]. This leads to rejecting 1 null hypothesis even at FDR 0.05, which is both poor and unwarranted. These results are summarized in Table 3.

5 Optimal symmetric discovery vectors

In our general definition (2) of discovery vectors, we allow them to depend on $\omega \in \Omega$ directly. In this section we specialize the definition to the practically interesting case where they depend on ω only via the e-values $E_1(\omega), \dots, E_K(\omega)$. This will allow us to define symmetric discovery vectors, and we will see that the arithmetic-mean discovery vector is optimal in this class, in a natural sense.

Fix the number $K \in \{2, 3, \dots\}$ of e-values. We say that a family of measurable functions $D_R : \{1, \dots, |R|\} \times [0, \infty]^K \to [0, \infty]$ indexed by the non-empty rejection sets $R \subseteq \{1, \dots, K\}$ is a *universal discovery vector* if, for all measurable spaces $(\Omega, \mathcal{A})$, all hypotheses $H_1, \dots, H_K \subseteq \mathfrak{P}(\Omega)$, and all e-variables $E_1, \dots, E_K$ for testing $H_1, \dots, H_K$, respectively, there exists an e-test $(E_Q)_{Q \in \mathfrak{P}(\Omega)}$ such that

$$\forall R \, \forall j \in \{1, \dots, |R|\} \, \forall Q \in \mathfrak{P}(\Omega) \, \forall \omega \in \Omega : \\ \left(|\{k \in R \mid Q \notin H_k\}| \ge j \right) \vee \left(E_Q(\omega) \ge D_R(j, E_1(\omega), \dots, E_K(\omega)) \right); \quad (18)$$

we also require that, for all $k \in \{1, \dots, K\}$, $D_R(j, e_1, \dots, e_K)$ is increasing in e_k. When $D_R(j, E_1, \dots, E_K)$ is large, we have the same interpretation as for (2): either there are at least j true discoveries or a rare chance has occurred (we have observed a large $E_Q \ge D_R(j, E_1, \dots, E_K)$). We say that D is "universal" to emphasize that it is required to work for all hypotheses and all e-variables for testing those hypotheses.

It is clear that the arithmetic-mean discovery vector

$$\mathrm{AV}_R(j, e_1, \dots, e_K) := \min_{I : |R \setminus I| < j} \frac{1}{|I|} \sum_{i \in I} e_i \quad (19)$$

(cf. (3)), if considered a function of R, j, and $e_1, \dots, e_K$, is a universal discovery vector. It is *symmetric*, in the sense of being invariant w.r. to permutations of $e_1, \dots, e_K$.

Let us say that a universal discovery vector D *essentially dominates* another universal discovery vector D' if

$$\forall R \, \forall j \in \{1, \dots, |R|\} \, \forall (e_1, \dots, e_K) \in [0, \infty]^K : \\ D'_R(j, e_1, \dots, e_K) > 1 \Longrightarrow D_R(j, e_1, \dots, e_K) \ge D'_R(j, e_1, \dots, e_K),$$

R ranging over the non-empty subsets of $\{1, \dots, K\}$. Intuitively, D dominates D' if D guarantees at least as many true discoveries as D' does at each reasonable confidence level (but we are not trying to compare the outputs of D and D' when both outputs are clearly useless; in terms of the colour code used in Figures 1–2 and 4–5, we are not comparing the values in the dark green cells).

Theorem 5.1. *The arithmetic-mean discovery vector* AV *essentially dominates any other symmetric universal discovery vector.*

Proof. We will prove a stronger version of Theorem 5.1 modifying the definition of a universal discovery vector; namely, we will allow the e-test (E_Q) to depend on R and j. In other words, we will only assume that, for all probability spaces $(\Omega, \mathcal{A}, Q)$, all hypotheses $H_1, \dots, H_K \subseteq \mathfrak{P}(\Omega)$, all e-variables $E_1, \dots, E_K$ for testing $H_1, \dots, H_K$, respectively, all non-empty rejection sets $R \subseteq \{1, \dots, K\}$, and all $j \in \{1, \dots, |R|\}$,

$$\left(|\{k \in R \mid Q \notin H_k\}| \geq j\right) \vee \left(\mathbb{E}^Q[D_R(j, E_1, \dots, E_K)] \leq 1\right). \tag{20}$$

For the purpose of contradiction, suppose there is a symmetric universal discovery vector D such that, for some R, j, and $e_1, \dots, e_K$,

$$D_R(j, e_1, \dots, e_K) > 1 \tag{21}$$

and

$$\mathrm{AV}_R(j, e_1, \dots, e_K) < D_R(j, e_1, \dots, e_K).$$

Fix such R, j, and $e_1, \dots, e_K$. By definition (see (19)), there exists $I \subseteq \{1, \dots, K\}$ such that

$$|R \setminus I| < j \quad \text{and} \quad \frac{1}{|I|} \sum_{i \in I} e_i < D_R(j, e_1, \dots, e_K). \tag{22}$$

Fix such an I. Choose a probability space $(\Omega, \mathcal{A}, Q)$ in such a way that Q is a continuous probability measure on the measurable space $(\Omega, \mathcal{A})$. We will choose hypotheses $H_1, \dots, H_K$ on $(\Omega, \mathcal{A})$ in such a way that

$$Q \in \left(\bigcap_{k \in I} H_k\right) \bigcap \left(\bigcap_{k \notin I} H_k^{\mathsf{c}}\right).$$

Since the first term of the disjunction (20) is false, the second term will be true for all e-variables $E_1, \dots, E_K$ for testing $H_1, \dots, H_K$, respectively. Specifically, we let H_k for $k \in I$ be the singleton $\{Q\}$ and H_k for $k \notin I$ be the singleton $\{Q'\}$ consisting of $Q' \in \mathfrak{P}(\Omega)$ such that Q and Q' are mutually singular (e.g., Q' is a Dirac measure). Further, let, for $k \in I$, E_k be an arbitrary e-variable for testing Q, and for $k \notin I$,

$$E_k = \begin{cases} e_k \ Q\text{-a.s.} \\ 1 \ Q'\text{-a.s.} \end{cases}$$

It is clear that E_k is an e-variable for testing H_k, $k = 1, \dots, K$, and thus $D_R(j, E_1, \dots, E_K)$ is an e-variable. By fixing E_k for $k \notin I$ above and treating E_k for $k \in I$ as inputs, we obtain a function $(E_k)_{k \in I} \mapsto D_R(j, E_1, \dots, E_K)$ which produces a valid e-variable for testing Q. By the symmetry assumption and the fact that $E_k = e_k$ a.s. under Q for $k \notin I$, Proposition 2.1 of Vovk and Wang [2019b] gives

$$D_R(j, e_1, \dots, e_K) \leq \frac{1}{|I|} \sum_{k \in I} e_k$$

for all $(e_k)_{k \in I}$ satisfying (21). This contradicts (22). □

Similarly, we can define formally the notion of essential domination for discovery matrices and prove an analogue of Theorem 5.1 for them. However, this does not appear an instructive exercise since the sets R_r defined by (6) are clearly optimal and we can rely on Theorem 5.1.

6 Conclusion

We have described methods for multiple hypothesis testing using e-values and demonstrated their use in simulation and empirical studies. We believe that these methods, being simpler and more powerful, are preferred to methods using p-values when the final result is to be stated in terms of e-values. However, our methods do not depend on the base e-values being independent, and under arbitrary dependence, they are sometimes competitive with results based on p-values even when the final result is to be stated in terms of p-values.

One of the obvious directions of further research is to extend our methods to non-symmetric problems of multiple hypothesis testing (cf. Genovese et al. [2006]), in which different e-values may be assigned different weights. Our procedure for multiple hypothesis testing is generic and does not have to rely on the arithmetic average.

Acknowledgments

We are grateful to Peter Westfall for his advice about the literature on Bayesian two-sample t-tests.

For most of our simulation and empirical studies in Sections 3–4 we used Python. We also used the R package `hommel` [Goeman et al., 2019a] and a dataset available in the R package `qvalue` [Storey et al., 2019].

V. Vovk's research has been partially supported by 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

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Hafner, New York, third edition, 1973.

Christopher R. Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, 32:1035–1061, 2004.

Christopher R. Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101: 1408–1417, 2006.

Christopher R. Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93:509–524, 2006.

Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26:584–597, 2011a. Correction: 28:464.

Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research: Rejoinder. *Statistical Science*, 26:608–612, 2011b.

Jelle J. Goeman, Rosa Meijer, and Thijmen Krebs. `hommel`*: Methods for closed testing with Simes inequality, in particular Hommel's method*, 2019a. R package version 1.5, available on CRAN.

Jelle J. Goeman, Rosa J. Meijer, Thijmen J. P. Krebs, and Aldo Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106:841–856, 2019b.

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, 2019c.

Mithat Gönen, Wesley O. Johnson, Yonggang Lu, and Peter H. Westfall. The Bayesian two-sample t test. *American Statistician*, 59:252–257, 2005.

Mithat Gönen, Wesley O. Johnson, Yonggang Lu, and Peter H. Westfall. Comparing objective and subjective Bayes factors for the two-sample comparison: the classification theorem in action. *American Statistician*, 73:22–31, 2019.

Peter Grünwald and Thijs van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12:1069–1103, 2017.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. Technical Report arXiv:1906.07801 [math.ST], arXiv.org e-Print archive, June 2019.

Michele Guindani, Peter Müller, and Song Zhang. A Bayesian discovery procedure. *Journal of the Royal Statistical Society B*, 71:905–925, 2009.

Ingrid Hedenfalk, David Duggan, Yidong Chen, Michael Radmacher, Michael Bittner, Richard Simon, Paul Meltzer, Barry Gusterson, Manel Esteller, Olli-P. Kallioniemi, Benjamin Wilfond, Åke Borg, and Jeffrey Trent. Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344:539–548, 2001.

Leonhard Held. On the Bayesian interpretation of the harmonic mean p-value. *Proceedings of the National Academy of Sciences*, 116:5855–5856, 2019.

Jesse Hemerik, Aldo Solari, and Jelle J. Goeman. Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106: 635–649, 2019.

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

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

Alexander Ly, Josine Verhagen, and Eric-Jan Wagenmakers. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32, 2016.

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

Aaditya K. Ramdas, Rina F. Barber, Martin J. Wainwright, and Michael I. Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *Annals of Statistics*, 47:2790–2821, 2019.

Joseph P. Romano and Michael Wolf. Control of generalized error rates in multiple testing. *Annals of Statistics*, 35:1378–1408, 2007.

Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, and Richard D. Morey. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, 16:225–237, 2009.

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

Sanat K. Sarkar. Simes' test in multiple testing. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1325–1327. Springer, Berlin, 2011.

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.

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

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 of the USA*, 100: 9440–9445, 2003.

John D. Storey, James Y. Dai, and Jeffrey T. Leek. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, 8:414–432, 2007.

John D. Storey, Andrew J. Bass, Alan Dabney, and David Robinson. `qvalue`*: Q-value estimation for false discovery rate control*, 2019. R package version 2.18.0, available on Bioconductor.

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

Vladimir Vovk and Ruodu Wang. Combining e-values and p-values. Technical Report arXiv:1912.06116 [math.ST], arXiv.org e-Print archive, December 2019b.

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

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

Min Wang and Guangying Liu. A simple two-sample Bayesian t-test for hypothesis testing. *American Statistician*, 70:195–201, 2016.

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

A Some general theory

An *e-merging function* is a function mapping any non-empty finite sequence of e-variables to an e-variable; in this paper we only consider e-merging functions that are symmetric, i.e., do not depend on the order of their arguments. An example is the arithmetic mean

$$F(E_1, \dots, E_n) := \frac{1}{n} \sum_{i=1}^{n} E_i \tag{23}$$

considered in the main part of the paper, but in this appendix we will also discuss two other e-merging functions. They are much less useful than the arithmetic mean, and we are interested in them only because they are analogous to standard merging functions for p-values.

Given an e-merging function F, we can generalize (3) to

$$D_R^F(j) := \min_{I : |R \setminus I| < j} F(E_i : i \in I), \quad j \in \{1, \dots, |R|\}. \tag{24}$$

Theorem 2.1 also generalizes, with the same proof.

Theorem A.1. *For any e-variables $E_1, \dots, E_K$, (24) is a discovery vector.*

The statement of Theorem A.1 is witnessed by the e-test

$$E_Q := F(E_i : i \in I_Q)$$

in the notation of (4).

The reader may have noticed that all discovery matrices in our figures satisfy properties of monotonicity. See, for concreteness, Figure 1. The entries $\mathrm{AM}_{r,j}$ are decreasing in j and increasing in r. They are also monotonic along the main diagonal and the lines parallel to it: for any $c = 0, 1, \dots$, $\mathrm{AM}_{r,r-c}$ is decreasing in r. These properties hold in general, as the following proposition shows.

Proposition A.2. *For any $R, R' \subseteq \{1, \dots, K\}$ and $j, j' \in \{1, \dots, |R|\}$:*

(1) $D^F_R \le D^F_{R'}$ if $R \subseteq R'$;

(2) $D^F_R(j) \le D^F_R(j')$ if $j' \le j$;

(3) $D^F_{R'}(j + |R' \setminus R|) \le D^F_R(j)$.

Proof. For item (1), we can rewrite $D^F_R(j) \le D^F_{R'}(j)$ as

$$\min_{I : |R \setminus I| < j} F(E_i : i \in I) \le \min_{I : |R' \setminus I| < j} F(E_i : i \in I),$$

and it suffices to notice that any I satisfying $|R' \setminus I| < j$ satisfies $|R \setminus I| < j$.

Item (2) can be rewritten as

$$\min_{I : |R \setminus I| < j} F(E_i : i \in I) \le \min_{I : |R \setminus I| < j'} F(E_i : i \in I),$$

which follows from $|R \setminus I| < j'$ implying $|R \setminus I| < j$.

Item (3) can be rewritten as

$$\min_{I : |R' \setminus I| < j + |R' \setminus R|} F(E_i : i \in I) \le \min_{I : |R \setminus I| < j} F(E_i : i \in I),$$

which follows from

$$|R \setminus I| < j \Longrightarrow |R' \setminus I| < j + |R' \setminus R|,$$

which in turn follows from the obvious

$$|R' \setminus I| \le |R \setminus I| + |R' \setminus R|. \qquad \square$$

The *Bonferroni e-merging function* is the following lower bound for (23):

$$B(E_1, \dots, E_n) := \frac{1}{n} \max_{i \in \{1, \dots, n\}} E_i. \tag{25}$$

A better lower bound for (5) is given by the *Simes e-merging function*

$$S(E_1, \dots, E_n)(\omega) := \max_{i \in \{1, \dots, n\}} \frac{i E_{[i]}(\omega)}{n}, \tag{26}$$

Algorithm 3 Matrix BM of true discoveries for the e-Bonferroni method

Require: An increasing sequence of e-values $e_1 \le \cdots \le e_K$.

1: $a := \infty$
2: **for** $r = 1, \dots, K$ **do**
3: $\quad B := e_{K-r+1}/(K-r+1)$
4: $\quad$ **if** $a > B$ **then**
5: $\quad\quad a := B$
6: $\quad$ **for** $i = r, \dots, K$ **do**
7: $\quad\quad \mathrm{BM}_{i,r} := a$

where $E_{[i]}(\omega)$ is the ith largest e-value among $E_i(\omega)$, $i \in \{1, \dots, n\}$ [Vovk and Wang, 2019b, end of Section 6]: $E_{[1]}(\omega), \dots, E_{[n]}(\omega)$ is the permutation of $E_1(\omega), \dots, E_n(\omega)$ satisfying $E_{[1]}(\omega) \ge \cdots \ge E_{[n]}(\omega)$.

Our discussion of e-values and p-values in Section 3 suggests that the function $t \mapsto 1/t$ transforms e-values into p-values (cf. (9)) and transforms p-values into approximate e-values (cf. (8) for a small $\kappa \in (0, 1)$); of course, the word "approximate" is used here in a crude sense (in the spirit of the algorithmic theory of randomness). Under this correspondence, the Bonferroni e-merging function (25) turns into the Bonferroni merging function for p-values, and the Simes e-merging function (26) turns into the Simes merging function for p-values. The dominating arithmetic-mean e-merging function (23) corresponds to using the harmonic mean for merging p-values, and indeed the harmonic mean has been discussed recently in this role [Wilson, 2019], sometimes with a similar justification based on the VS bound [Held, 2019]. However, the harmonic mean is not a p-merging function [Goeman et al., 2019c] unless multiplied by, say, $2.5 \ln K$ for $K \ge 3$ [Vovk and Wang, 2019a].

With $F_{\mathbf{e}}(I)$ replaced by the Bonferroni lower bound

$$B_{\mathbf{e}}(I) := \frac{1}{|I|} \max_{i \in I} e_i, \quad I \subseteq \{1, \dots, K\}, \ I \neq \emptyset,$$

for (5), Algorithms 1 and 2 have the same interpretation as before (although the results are not as good since they are based on more conservative e-values). However, they simplify, especially Algorithm 2, whose Bonferroni counterpart (*e-Bonferroni*) is given as Algorithm 3. In line 1 we initialize the adjusted Bonferroni e-value, in line 3 we compute the raw Bonferroni e-value, and in lines 4–5 we adjust it. The algorithm produces a matrix BM with constant columns.

Whereas Bonferroni-type procedures often perform well when the goal is family-wise validity (see, e.g., Vovk and Wang [2019b, Figures 4 and 5]), their performance tends to deteriorate for less demanding notions of validity. (In terms of p-values, this phenomenon is discussed in, e.g., Goeman and Solari [2011b, Section 1].) Comparing the left panel of Figure 7 with the left panel of Figure 2 we can see that the e-Bonferroni method is much worse than the arithmetic average when the goal is to control the number of false discoveries.

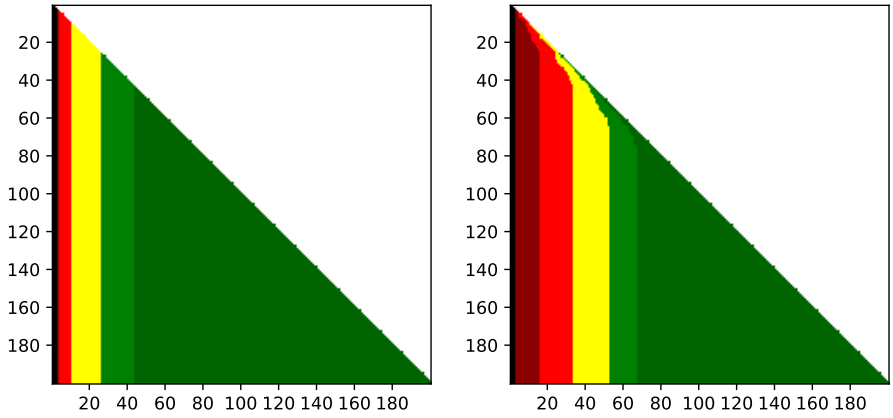


Figure 7: Left panel: the Bonferroni discovery matrix (given by Algorithm 3) in the situation of Figure 2, using Jeffreys's thresholds. Right panel: the corresponding Simes discovery matrix.

The poor performance of the e-Bonferroni method is clear already from the left panel of Figure 2: the areas of different colours are far from been vertical at the top, where they curve left. It is clear that every discovery matrix that is dominated by this one and has vertical boundaries between different colours (such as e-Bonferroni) is going to be much worse.

The right panel of Figure 7 shows the Simes discovery matrix, based on (26), in the situation of Figure 2. It is intermediate between AM and Bonferroni and, remarkably, it looks better than the GWGS discovery p-matrix transformed by applying the VS bound to its elements (the right panel of Figure 2).

Remark A.3. We can quantify the quality of the lower bounds (25) and (26) of the arithmetic mean (23) by the inequalities

$$\begin{aligned} &B(E_1,\dots,E_n) \le S(E_1,\dots,E_n) \le F(E_1,\dots,E_n), \\ &1 \le \frac{F(E_1,\dots,E_n)}{S(E_1,\dots,E_n)} \le \sum_{k=1}^{n} \frac{1}{k} \le \ln n + 1, \\ &1 \le \frac{F(E_1,\dots,E_n)}{B(E_1,\dots,E_n)} \le n, \quad 1 \le \frac{S(E_1,\dots,E_n)}{B(E_1,\dots,E_n)} \le n, \end{aligned} \tag{27}$$

all of which are tight (achievable as equality for any n), apart from the last one in (27) (which is tight only for $n = 1$). Since $n := |I| \le K$ in (24), this gives bounds for the ratios of the corresponding elements of the discovery matrices built on top of B, S, and F.

Algorithm 4 Arithmetic-mean discovery matrix (specialized)

Require: An increasing sequence of e-values $e_1 \le \dots \le e_K$.

1: $s_1 := e_1$
2: **for** $k = 2, \dots, K$ **do**
3: $\quad s_k := s_{k-1} + e_k$
4: **for** $r = 1, \dots, K$ **do**
5: $\quad \sigma_{r,r} := e_{K-r+1}$
6: $\quad$ **for** $j = r-1, \dots, 1$ **do**
7: $\quad\quad \sigma_{r,j} := \sigma_{r,j+1} + e_{K-j+1}$
8: **for** $r = 1, \dots, K$ **do**
9: $\quad$ **for** $j = 1, \dots, r$ **do**
10: $\quad\quad \mathrm{AM}_{r,j} := \sigma_{r,j}/(r-j+1)$
11: $\quad\quad$ **for** $i = 1, \dots, K-r$ **do**
12: $\quad\quad\quad e := (\sigma_{r,j} + s_i)/(r-j+1+i)$
13: $\quad\quad\quad$ **if** $e < \mathrm{AM}_{r,j}$ **then**
14: $\quad\quad\quad\quad \mathrm{AM}_{r,j} := e$

B Efficient implementation of Algorithm 2

Algorithm 2 is a generic algorithm that works for any e-merging function F, not necessarily the arithmetic mean. In general, computing one row of the discovery matrix takes time $O(K^3)$ if we assume that the base e-merging function F can be computed in time linear in the number of arguments. This assumption is correct for the arithmetic mean and, provided the arguments are sorted, the Simes e-merging function (and it is more than correct for the Bonferroni e-merging function, which takes constant time for sorted arguments). The overall computational complexity for the full discovery matrix is very high, $O(K^4)$.

For specific e-merging functions it is often possible to find more efficient implementations, so called shortcuts. In the case of the Bonferroni method, Algorithm 3 takes time $O(K)$ per column, and this time is spent simply by writing one value. The resulting computation complexity $O(K^2)$ is clearly the optimal one.

A more efficient implementation of Algorithm 2 is given as Algorithm 4, which uses arrays s_k (the sum of the first k base e-values) and $\sigma_{r,j}$ (the sum of the base e-values with indices in $S_{r,j}$ in the notation of Algorithm 2) There is a preprocessing stage (lines 1–3) taking time $O(K)$. After that computing each row of the discovery matrix takes time $O(K^2)$. The overall time, $O(K^3)$, is intermediate between Algorithms 2 and 3.

An even more efficient implementation of Algorithm 2 is given as Algorithm 5. This algorithm computes one row of the discovery matrix in time $O(K)$, which gives the overall time $O(K^2)$. Both $O(K)$ and $O(K^2)$ are clearly optimal in this context. The ability to compute efficiently individual rows is useful when the discovery matrix is big; e.g., it can be too big to fit in computer memory.

Algorithm 5 One row of the arithmetic-mean discovery matrix in time $O(K)$

Require: Increasing sequence of e-values $e_1 \le \cdots \le e_K$.
Require: Row number $r \in \{1, \dots, K\}$ of the discovery matrix.

1: $s_0 := 0$
2: **for** $k = 1, \dots, K - r$ **do**
3: $\quad s_k := s_{k-1} + e_k$
4: $\sigma_r := e_{K-r+1}$
5: **for** $j = r - 1, \dots, 1$ **do**
6: $\quad \sigma_j := \sigma_{j+1} + e_{K-j+1}$
7: $k := K - r$
8: **for** $j = 1, \dots, r$ **do**
9: $\quad$ `slope` $:= \frac{s_k + \sigma_j}{k+r-j+1}$
10: $\quad$ **for** $i = k - 1, \dots, 0$ **do**
11: $\quad\quad$ `new.slope` $:= \frac{s_i + \sigma_j}{i+r-j+1}$
12: $\quad\quad$ **if** `new.slope` $>$ `slope` **then break**
13: $\quad\quad k := i$
14: $\quad\quad$ `slope` $:=$ `new.slope`
15: $\quad \mathrm{AM}_{r,j} :=$ `slope`

We are using essentially the same array s as in Algorithm 4 (now we extend it by adding $s_0 := 0$), and the array σ in Algorithm 5 is one row of the array σ in Algorithm 4;

$$s_k = e_1 + \cdots + e_k, \qquad k = 1, \dots, K - r,$$
$$\sigma_j := e_{K-r+1} + \cdots + e_{K-j+1}, \qquad j = 1, \dots, r.$$

Of course, there is no need to recompute the array s for each row r of the discovery matrix.

The command **break** in line 12 means leaving the loop, as in Python or R; in this context, it is equivalent to "go to line 15". The variable k in line 13 is the index of the "current vertex" P_k; we start from the rightmost P_k in line 7 and then keep moving left.

The geometry behind Algorithm 5 is shown in Figure 8. The coordinates of each of the points P_k, $k = 0, \dots, K - r$, are (k, s_k), and the coordinates of the point Q_j, where $j \in \{1, \dots, r\}$, are $-(r - j + 1, \sigma_j)$. Since the sequence $e_1, \dots, e_{K-r}$ is increasing, connecting the points $P_0, P_1, \dots$ in this order (see the red line in Figure 8) gives us the graph of a convex function.

For each $j = 1, \dots, r$, the iteration of the loop in lines 10–15 of Algorithm 5 computes the slope of the line (shown in blue) passing through Q_j and touching the red line from below. The validity of the algorithm follows from the point Q_{j+1} lying at or above the blue line, for each $j = 1, \dots, r - 1$. If $k < K - r$, the slope of the blue line is at most e_{K-r+1}. On the other hand, the slope of the line going from Q_j to Q_{j+1} is $e_{K-j+1} \ge e_{K-r+1}$. It remain to consider the case $k = K - r$. In this case, it suffices to notice that the slope e_{K-j+1} of the line

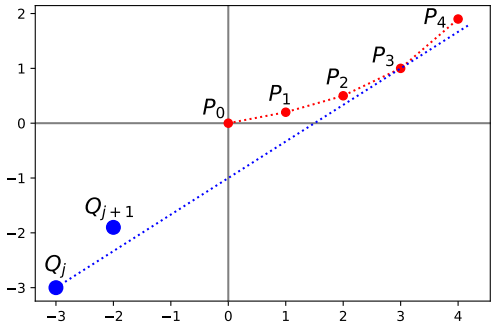


Figure 8: Geometry behind Algorithm 5.

going from Q_j to Q_{j+1} is greater than or equal to the average of $e_1, \dots, e_{K-j+1}$.

C Comparison with the GWGS procedure

In this appendix we will discuss the GWGS multiple testing procedure, in the form described in Goeman and Solari [2011a, Section 2] and in terms of our definitions. Let $F : \cup_{n=1}^{\infty}[0,1]^n \to [0,1]$ be a *p-merging function*, in the sense of transforming p-variables into a p-variable: whenever $P_1, \dots, P_n$ are p-variables for some $n \in \{1, 2, \dots\}$, $F(P_1, \dots, P_n)$ is a p-variable. Suppose that F is symmetric: for any $n \in \{1, 2, \dots\}$ and any $p_1, \dots, p_n \in [0,1]$, $F(p_1, \dots, p_n)$ does not depend on the order of $p_1, \dots, p_n$. We will write $F(p_i, i \in I)$ for the value of F on a sequence consisting of $|I|$ elements p_i, $i \in I$ (in any order), for any non-empty subset I of $\{1, \dots, K\}$. With such an F we can associate the following analogue of (3) in terms of p-values:

$$D_R(j) := \max_{I : |R \setminus I| < j} F(p_i, i \in I); \tag{28}$$

we leave the dependence on $p_1, \dots, p_K$ implicit, following Goeman and Solari.

Of course, the analogue of Theorem 2.1 holds for (28). Let $P_1, \dots, P_K$ be p-variables w.r. to statistical hypotheses $H_1, \dots, H_K$, respectively. Define $D_R(j, \omega)$ as the right-hand side of (28) with p_i replaced by $P_i(\omega)$.

Theorem C.1. *There exists a family $P = (P_Q)_{Q \in \mathfrak{P}(\Omega)}$ of p-variables such that*

$$\forall Q \in \mathfrak{P}(\Omega)\, \forall R\, \forall j \in \{1, \dots, |R|\}\, \forall \omega \in \Omega : \\ \Big(|\{k \in R \mid Q \notin H_k\}| \geq j\Big) \vee \Big(P_Q(\omega) \leq D_R(j, \omega)\Big). \tag{29}$$

Proof. A family satisfying (29) is

$$P_Q := F(P_k, k \in I_Q), \tag{30}$$

in the notation of (4); if $I_Q = \emptyset$ (in which case the disjunction in (29) is obvious), we set $P_Q := 1$. Let us check the disjunction in (29) for given $Q \in \mathfrak{P}(\Omega)$, R, $j \in \{1, \dots, |R|\}$, and $\omega \in \Omega$. If the second term of the disjunction is false, $P_Q(\omega) > D_R(j, \omega)$, the definition (30) of P_Q gives

$$F(P_k(\omega), k \in I_Q) > \max_{I : |R \setminus I| < j} F(P_i(\omega), i \in I),$$

and so $|R \setminus I_Q| \geq j$; therefore, there are at least j true discoveries, which establishes the first term of the disjunction. □

Goeman and Solari prefer the inverse to the function (28), which they denote $f_\alpha(R)$, suppressing the dependence on $p_1, \dots, p_K$; we consider it as function of $\alpha \in [0, 1]$, which is interpreted as significance level. We will see that this function satisfies

$$f_\alpha(R) \geq j \iff D_R(j) \leq \alpha \tag{31}$$

(and this equivalence can serve as definition of f). Therefore, it gives us the number of true discoveries that are warranted at significance level α.

For the reader familiar with Goeman and Solari [2011a], we will check that their definition indeed satisfies (31). They first define their bound

$$t_\alpha(R) := \max\{|I| \mid I \subseteq R, I \notin \mathcal{X}\}$$

on the number of false discoveries, where

$$\mathcal{X} := \{I \mid \forall J \supseteq I : J \in \mathcal{U}\}$$

are the subsets of $\{1, \dots, K\}$ rejected by the closed testing procedure, and

$$\mathcal{U} := \{I \mid F(p_i, i \in I) \leq \alpha\}$$

are the subsets of $\{1, \dots, K\}$ rejected by F; in general, I and J will run over the subsets of $\{1, \dots, K\}$. Then they define their bound on the number of true discoveries as

$$f_\alpha(R) := |R| - t_\alpha(R).$$

The equivalence (31) can be checked as follows:

$$\begin{aligned}
&f_\alpha(R) \geq j \iff t_\alpha(R) \leq |R| - j \\
&\iff \max\{|I| \mid I \subseteq R, I \notin \mathcal{X}\} \leq |R| - j \\
&\iff (\forall I \subseteq R : I \notin \mathcal{X} \Rightarrow |I| \leq |R| - j) \\
&\iff (\forall I \subseteq R : |I| > |R| - j \Rightarrow I \in \mathcal{X}) \\
&\iff (\forall I \subseteq R : |I| > |R| - j \Rightarrow (\forall J \supseteq I : J \in \mathcal{U})) \\
&\iff (\forall J : |J \cap R| > |R| - j \Rightarrow J \in \mathcal{U}) \\
&\iff (\forall J : |R \setminus J| < j \Rightarrow J \in \mathcal{U}) \\
&\iff \max_{J : |R \setminus J| < j} F(p_i, i \in J) \leq \alpha \iff D_R(j) \leq \alpha.
\end{aligned}$$

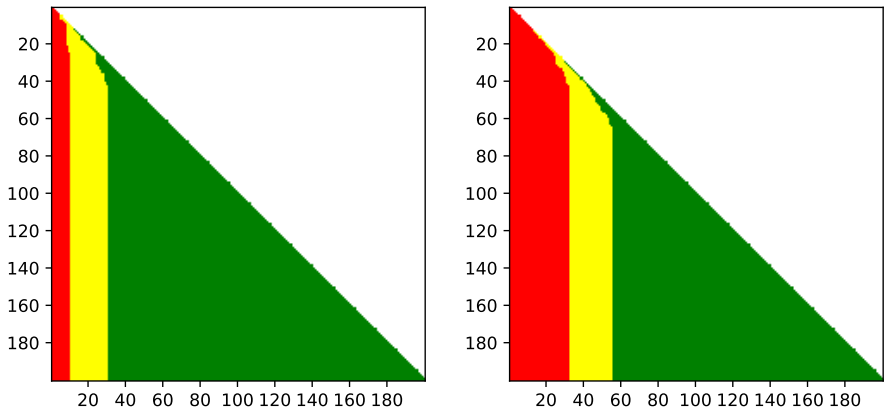


Figure 9: Left panel: the GWGS discovery p-matrix for the simulation data (as in Figure 2) for Fisher's thresholds 1% and 5%, under arbitrary dependence. Right panel: assuming independence (and so representing the same matrix as the right panel of Figure 2 with a different colour code).

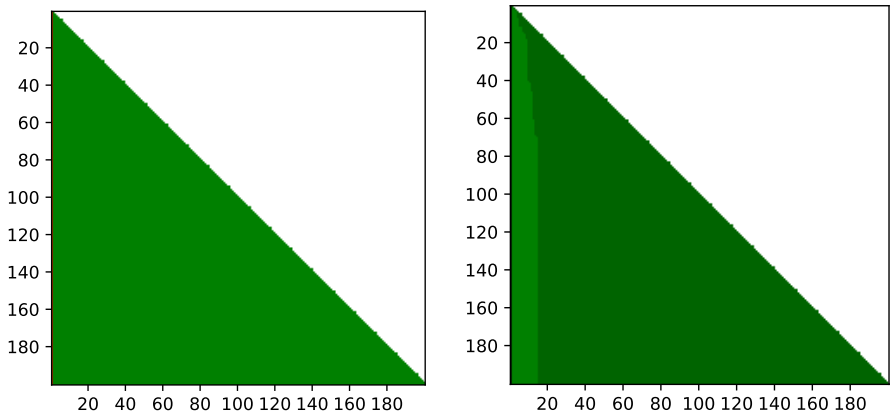


Figure 10: Left panel: the top-left corner of the GWGS discovery p-matrix for the `BRCA` dataset for Fisher's thresholds 1% and 5%, under arbitrary dependence. Right panel: analogous picture with each entry replaced by the corresponding VS bound and using Jeffreys's thresholds.

In conclusion, we give two more figures demonstrating the work of the `hommel` package. Figure 9 shows how much the assumption of independence helps the GWGS procedure in our simulation studies.

Figure 10 shows results (very poor) for the `BRCA` dataset without assuming independence. In particular, the right panel is much worse than the left panel of Figure 4, which also does not assume independence.

D Generalized Bayes and boosting a weak signal

When defining the base e-values for use in our empirical studies we just used the likelihood ratio $E(x)$ defined by (7). This is the simplest version of a Bayes factor. It usually works very well, but in some cases can be improved. Later in this appendix we will see an example where a weak signal needs to be boosted, but we start from developing tools that will allow us to do so.

Let us choose a constant $\eta > 0$ (the *learning rate*) and refer to

$$E_\eta(x) := \frac{1}{c} E(x)^\eta = \frac{1}{c} \exp(\eta\delta x - \eta\delta^2/2) \tag{32}$$

as the *generalized Bayes factor* (see, e.g., Grünwald and van Ommen [2017, Section 2.4] and references therein). Here $c > 0$ is the normalizing constant ensuring $\int E_\eta \, \mathrm{d}N(0,1) = 1$; a simple calculation gives

$$c = \exp\left(\eta(\eta-1)\delta^2/2\right).$$

Plugging this into (32) we obtain

$$E_\eta(x) = \exp(\eta\delta x - \eta^2\delta^2/2).$$

This gives a useful interpretation of the generalized Bayes factor: it is still the likelihood ratio, but we replace the true alternative $N(\delta, 1)$ by a false one, $N(\eta\delta, 1)$. For $\eta > 1$ we are boosting the difference between the null and alternative hypotheses.

One situation in which the likelihood ratio (7) does not work well is where we have a large number of false null hypotheses, but the true data-generating distributions are fairly close to the null hypotheses (as it were, we have a weak signal). Figures 11–12 illustrate the case of 10,000 null hypotheses $N(0,1)$ of which 1000 are false, the true alternatives being $N(-2,1)$ (which makes the signal much weaker than in Section 3). In the left panel of Figure 11 we use the Bayes factor (7), whereas in its right panel we use the generalized Bayes factor (32) for $\eta = 2$. The plot for the Bayes factor looks even worse than the plot for the GWGS procedure using Simes's inequality (right panel of Figure 12). Using the generalized Bayes factor greatly improves the discovery matrix.

Table 4 gives the numbers of rejections for the Benjamini–Hochberg and Benjamini–Yekutieli procedures. In view of Figure 11 (right panel), the results are particularly poor for arbitrary dependence.

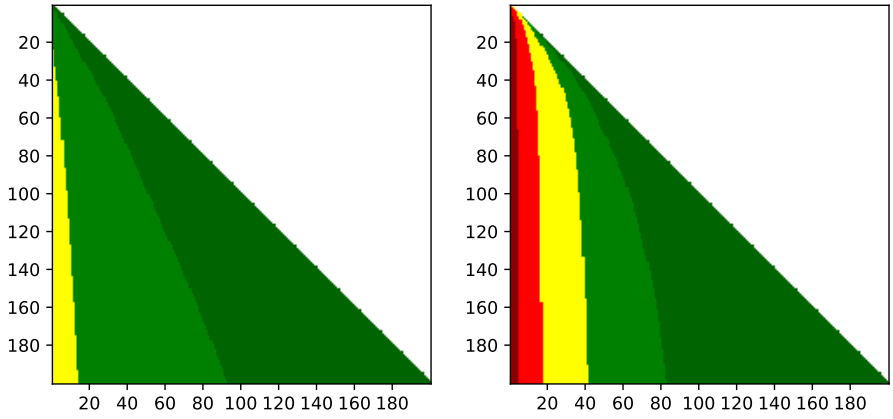


Figure 11: Left panel: the top-left 200×200 corner of the arithmetic-mean discovery matrix for the simulation data with 10,000 observations, 10% of false hypotheses, and weak signal, using Bayes factors as base e-values, as described in text. Right panel: using generalized Bayes factors with learning rate $\eta = 2$. Both panels use Jeffreys's thresholds.

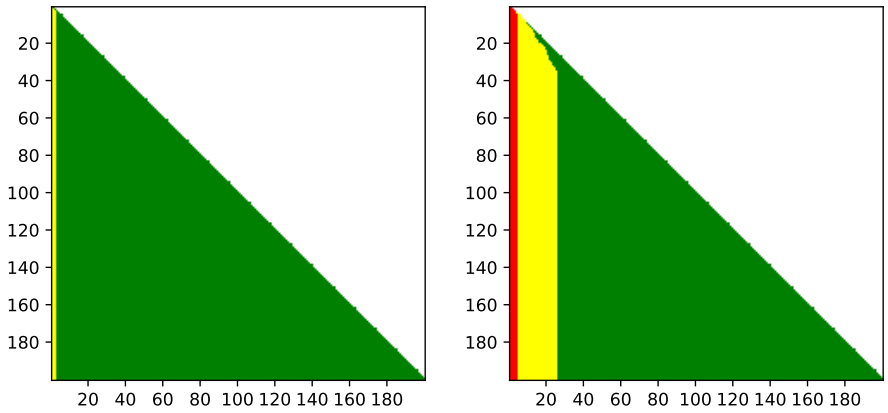


Figure 12: Left panel: the top-left 200×200 corner of the GWGS discovery p-matrix for the simulation data with 10,000 observations, 10% of false hypotheses, and weak signal, under general dependence. Right panel: assuming independence. The colour code is based on Fisher's thresholds.

Table 4: The Benjamini–Hochberg and Benjamini–Yekutieli procedures applied to the simulation data with 10,000 observations, 10% of false hypotheses, and weak signal for FDR 5% and 1%.

assumption	5%	1%
independence	84	18
arbitrary dependence	10	0