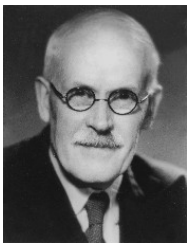
Multiple testing in game-theoretic probability: pictures and questions

Vladimir Vovk



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 #10

First posted March 18, 2024. Last revised March 19, 2024.

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

Abstract

The usual way of testing probability forecasts in game-theoretic probability is via construction of test martingales. The standard assumption is that all forecasts are output by the same forecaster. In this paper I will discuss possible extensions of this picture to testing probability forecasts output by several forecasters. This corresponds to multiple hypothesis testing in statistics. One interesting phenomenon is that even a slight relaxation of the requirement of family-wise validity leads to a very significant increase in the efficiency of testing procedures. The main goal of this paper is to report results of preliminary simulation studies and list some directions of further research.

Contents

1 Introduction

Game-theoretic probability, as presented in, e.g., my joint books [12] and [13] with Glenn Shafer, is based on the idea that a null hypothesis can be tested dynamically by gambling against it. More generally, we are testing a player called Forecaster, which can be a scientific theory, a computer program, a human forecaster, etc. The gambler starts from an initial capital of 1 and is required to keep his capital nonnegative. His current capital is interpreted as the degree to which the null hypothesis has been undermined.

The idea of testing via gambling goes back at least to Richard von Mises's principle of the impossibility of a gambling system (Unmöglichkeit eines Spielsystems [15, p. 58]; see also [16]), but von Mises's notion of gambling was too narrow, and it was only applicable to infinite sequences. The narrowness of von Mises's notion of gambling was demonstrated by Ville [14, Sect. II.4] (for an English translation, see [10]). Ville proposed extending von Mises's testing procedure to using nonnegative martingales [14, Chap. IV], but he is surprisingly terse when using his wider notion of testing to restate von Mises's principle of the impossibility of a gambling system, especially in the two philosophical chapters [14, preliminary chapter and Chap. 6] (even though he had been interested in the impossibility of gambling systems long before he started writing his book [14]: see [4, Sect. 5.3]). It appears that the idea of testing using nonnegative martingales emerged gradually in various fields, including the algorithmic theory of randomness.

In this paper we will be interested in testing several forecasters in one go, with different forecasters being tested at different steps. Testing by gambling can be studied in the usual setting of measure-theoretic probability, and this is what we will do in this paper, for simplicity and as a first step. Replacing measure-theoretic probability by game-theoretic probability as mathematical foundation for our definitions and results will be one of directions of future research. For now, each forecaster will be formalized as a composite null hypothesis, represented by a set of probability measures on the sample space.

In principle, we can consider two settings for testing multiple null hypotheses. In the *closed* setting, we have a fixed number K of null hypotheses. In the *open* setting, the number of null hypotheses is not known in advance and is potentially infinite. In this paper we will concentrate on the closed setting.

This paper has been prepared in support of my planned talk at the Oberwolfach workshop "Game-theoretic statistical inference: optional sampling, universal inference, and multiple testing based on e-values" organized by Peter Grünwald, Aaditya Ramdas, Ruodu Wang, and Johanna Ziegel (5–10 May 2024).

2 Dynamic necessity and possibility

Let us fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $\mathcal{F} = (\mathcal{F}_n)_{n=0}^{\infty}$, so that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$ is a nested sequence of σ-algebras. Apart from the

true probability measure $\mathbb{P}$ we will often consider other probability measures Q on the measurable space $(\Omega, \mathcal{F})$. Our notation for the expectation of a random variable $f : \Omega \to [0, \infty]$ w.r. to Q will be $\mathbb{E}_Q(f) := \int f \, \mathrm{d}Q$, abbreviated to $\mathbb{E}(f)$ when $Q = \mathbb{P}$ (in general, "w.r. to Q" or the indication of Q is usually omitted when $Q = \mathbb{P}$). Let $\mathcal{Q}$ be the family of all probability measures on $(\Omega, \mathcal{F})$.

A *test martingale* S w.r. to Q is a sequence $S_0, S_1, \ldots$ of random variables taking values in $[0, \infty]$ such that $S_0 = 1$ and $\mathbb{E}_Q(S_n \mid \mathcal{F}_{n-1}) = S_{n-1}$ for all $n = 1, 2, \ldots$. A *martingale test* is a family $(S^Q)_{Q \in \mathcal{Q}}$ of test martingales S^Q w.r. to Q. At each time n, we interpret $S^Q_n(\omega)$ as a measure of disagreement between the realized outcome ω and its putative explanation Q; we may say that ω is *α-strange* at time n w.r. to Q if $S^Q_n(\omega) \geq \alpha$.

Fix a martingale test (S^Q) and let $A \subseteq \mathcal{Q}$ be a property of a probability measure Q (with the property being satisfied if and only if $Q \in A$). The *necessity measure* of A at time n in view of the realized outcome $\omega \in \Omega$ is

$$\Box_n(A \mid \omega) := \inf_{Q : Q \notin A} S^Q_n(\omega),$$

and the *possibility measure* of A in view of ω is

$$\Diamond(A \mid \omega) := \inf_{Q : Q \in A} S^Q_n(\omega) = \Box_n(A^{\mathsf{c}} \mid \omega).$$

The interpretation is that A holds unless ω is $\Box_n(A \mid \omega)$-strange at time n, and similarly for $\Diamond$. It is important that this property of validity can be applied to all A at the same time; the martingale test, however, should be chosen in advance.

3 Multiple testing of a single null hypothesis

In this section, we fix a probability measure $Q \in \mathcal{Q}$ on the sample space; we are interested in testing whether Q is the true probability measure, $Q = \mathbb{P}$. For that, we would like to have one test martingale w.r. to Q.

Instead, we are given K test martingales $S^{(k)}$ for $k = 1, \ldots, K$. In the language of game-theoretic probability as presented in [13], we have K Sceptics testing Q as null hypothesis. Suppose the test martingales $S^{(1)}, \ldots, S^{(K)}$ are *uncorrelated*, meaning that there exists a predictable sequence k_n, $n = 1, 2, \ldots$, such that $S^{(k)}_n = S^{(k)}_{n-1}$ for all n and all $k \neq k_n$. (And the requirement of predictability means that each k_n is $\mathcal{F}_{n-1}$-measurable.) This concept and terminology goes back to Shafer [9, Sect. 12.3] (at least for the case $K = 2$). The interpretation is that Forecaster is being tested by Sceptic k_n on step n.

A convex combination of test martingales is always a test martingale. In this section we discuss how else we can combine test martingales. First we notice that the product $S^{(1)} \ldots S^{(K)}$ (as well as the product of a subset of $S^{(1)}, \ldots, S^{(K)}$) is a test martingale [9, Proposition 12.5(1)]. Indeed, dropping the lower index Q,

$$\mathbb{E}\left(S^{(1)}_n \ldots S^{(K)}_n \mid \mathcal{F}_{n-1}\right) = \mathbb{E}\left(S^{(1)}_{n-1} \ldots S^{(k_n-1)}_{n-1} S^{(k_n)}_n S^{(k_n+1)}_{n-1} \ldots S^{(K)}_{n-1} \mid \mathcal{F}_{n-1}\right)$$

$$= S_{n-1}^{(1)} \dots S_{n-1}^{(k_n-1)} S_{n-1}^{(k_n+1)} \dots S_{n-1}^{(K)} \mathbb{E}\left(S_n^{(k_n)} \mid \mathcal{F}_{n-1}\right)$$
$$= S_{n-1}^{(1)} \dots S_{n-1}^{(K)}.$$

A *martingale merging function* is a measurable function $F : [0,\infty)^K \to [0,\infty)$ such that $F(S_n^{(1)}, \dots, S_n^{(K)})$, $n = 0, 1, \dots$, is a test martingale whenever $S^{(1)}, \dots, S^{(K)}$ are test martingales (and we require this to hold for any probability space and any test martingales on it). We will apply such functions to base test martingales to get a new test martingale that can be used for testing. Our definition allows test martingales to take value ∞, and so we extend each martingale merging function in a canonical way (as in [18, Sect. 3]): namely, we set $F := \infty$ whenever one or more of its arguments are ∞.

An example of a martingale merging function is (cf. [18, 17])

$$U_n(s_1, \dots, s_K) := \frac{1}{\binom{K}{n}} \sum_{\{k_1,\dots,k_n\} \subseteq \{1,\dots,K\}} s_{k_1} \dots s_{k_n}$$
$$= \frac{1}{\binom{K}{n}} \sigma_n(s_1, \dots, s_K), \quad n \in \{1, \dots, K\},$$

where σ_n is the *nth elementary symmetric polynomial in K variables*. In other words, U_n is σ_n normalized by dividing by $\sigma(1, \dots, 1)$; normalization ensures that the initial value of the combination of test martingales starts from 1 as initial capital (and then it is a test martingale). We will be particularly interested in the cases $n = 1$ and $n = 2$.

In my previous joint papers with Ruodu Wang [18, 17], we referred to the functions U_n as U-statistics, but this is potentially confusing as we are omitting "with product as kernel" as far as the standard statistical notion of U-statistics is concerned. In this paper I will call U_n *normalized elementary symmetric polynomials* (NESPs).

A *multiaffine polynomial* is defined as a multivariate polynomial such that none of its monomials has any variable raised to power 2 or more. (The less formal version "multilinear polynomial" of this term is more popular in literature, but would have been awkward in this paper.) A multiaffine polynomial is *positive* if each of its (non-zero) coefficients is positive. It is *normalized* if its value is 1 when all its arguments are 1.

Proposition 1. *For a fixed number of arguments K, every martingale merging function is a multiaffine polynomial that is positive and, of course, normalized.*

Let us say that a function of several variables is *symmetric* if it is invariant w.r. to the permutations of its arguments. Specializing Proposition 1 to symmetric functions, we obtain the following statement.

Proposition 2. *For a fixed number of arguments K, every symmetric martingale merging function is a convex mixture of the NESPs U_n, $n = 0, \dots, K$.*

In Proposition 2, U_0 is understood to be 1. For proofs of Propositions 1 and 2, see Appendix A. From now on we will consider symmetric martingale merging functions.

4 Family-wise multiple testing

We are given K adapted sequences $S^{(k)} = (S_1^{(k)}, S_2^{(k)}, \dots)$, $k = 1, \dots, K$, of random variables taking values in $[0, \infty]$ and a predictable sequence $k_1, k_2, \dots$ of random variables taking values in $\{1, \dots, K\}$ such that $S_n^{(k)} = S_{n-1}^{(k)}$ whenever $k_n \neq k$. Let us say that $k \in \{1, \dots, K\}$ is an *anomalous index* for $Q \in \mathcal{Q}$ if $S^{(k)}$ is not a test martingale w.r. to Q (with $S_0^{(k)}$ understood to be 1).

The interpretation is that at each step n we are testing a null hypothesis $H_k \subseteq \mathcal{Q}$, which leads to a change in $S_n^{(k)}$. There are K null hypotheses $H_1, \dots, H_K$, and at step n we are testing H_{k_n}. The process of gambling is fair, in the sense of leading to a test martingale $S^{(k)}$, under each $Q \in H_k$. However, it does not have to be a test martingale under the true probability measure $\mathbb{P}$. (A more realistic picture arises when we replace "test martingale" by "e-process", i.e., a process dominated by a test martingale, but let us concentrate on the simpler case of test martingales in this paper.)

After observing the values of $S^{(k)}$ over steps $1, \dots, n$, we might come up with a *rejection set* $R \subseteq \{1, \dots, K\}$ containing the indices of the hypotheses that we decide to reject at step n. It is natural to include in R the indices k with the largest values of $S_n^{(k)}$. The elements of R are *discoveries*. A discovery $k \in R$ is a *true discovery* if $\mathbb{P} \notin H_k$, and it is a *false discovery* if $\mathbb{P} \in H_k$. Let us also say that $k \in R$ is a *justified discovery* if k is an anomalous index. Every justified discovery is a true discovery. (The notion of a justified discovery is simpler than that of a true discovery in that it does not involve the null hypotheses H_k.)

In this section we are interested in the necessity of all $k \in R$ being justified discoveries. This number is a lower bound on the necessity of all $k \in R$ being true discoveries. In other words, we are interested in conclusions that are family-wise valid.

For each $Q \in \mathcal{Q}$, let

$$J_Q := \left\{ k \in \{1, \dots, K\} \mid S^{(k)} \text{ is a test martingale under } Q \right\}.$$

We are interested in the necessity of the property

$$R \cap J_Q = \emptyset \tag{1}$$

that all discoveries in R are justified.

The most natural martingale test in our current context is obtained by applying a martingale merging function F to the test martingales among the $S^{(k)}$. In other words, for a given Q, F should be applied to $S^{(k)}$ for $k \in J_Q$. Let us fix F. Notice that we need F for any number of arguments from 1 to K, so formally we need a family $(F_k)_{k=1}^K$ of martingale merging functions. We abbreviate $F_k(\dots)$ to $F(\dots)$ since k is determined by the number of arguments and so redundant.

The optimal discovery sets at time n are $R_{r,n}$, $r = 1, \dots, K$, where each $R_{r,n} \subseteq \{1, \dots, K\}$ has size r and consists of the indices of the r largest values

Algorithm 1 Chronological discovery diagonal $d_{r,n}$

Input: symmetric martingale merging functions F_k, $k \in \{1, \dots, K\}$.
Input: decreasing sequence of martingale values $S^1 \geq \dots \geq S^K$.

for $n = 1, 2, \dots$
  for $r = 1, \dots, K$
    $d_{r,n} := F((S^r))$
    for $k = r+1, \dots, K$
      $S := F((S^i)_{i \in \{r\} \cup \{k, \dots, K\}})$
      if $S < d_{r,n}$
        $d_{r,n} := S$

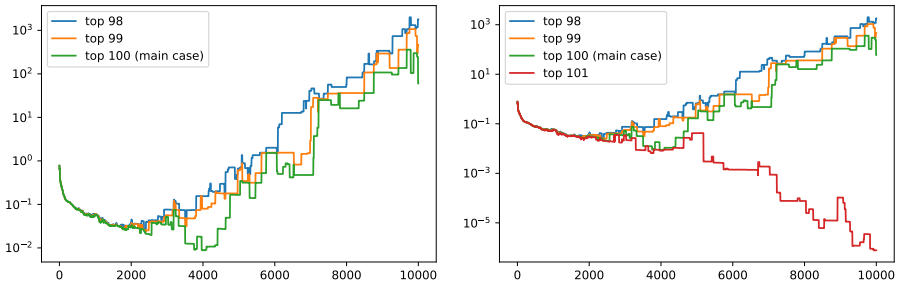


Figure 1: Discovery plots for 100, 99, and 98 hypotheses. The right panel also adds the case of 101 hypotheses (at least one of which is bound to be wrong).

in the set of $S_n^{(k)}$, $k = 1, \dots, K$; in the case of ties, let us give preference to smaller k. Define the *chronological discovery diagonal* by

$$\begin{aligned} \square_n(R_{r,n} \cap J_Q = \emptyset) &= \inf_{Q \in \mathcal{Q}: R_{r,n} \cap J_Q \neq \emptyset} F\left(\left(S_n^{(k)}\right)_{k \in J_Q}\right) \\ &\geq \inf_{I \subseteq \{1, \dots, K\}: R_{r,n} \cap I \neq \emptyset} F\left(\left(S_n^{(i)}\right)_{i \in I}\right) =: d_{r,n}. \end{aligned} \tag{2}$$

(I will explain the origin of the term "diagonal" in Sect. 7.)

Algorithm 2 spells out the computation of the infinite $K \times \infty$ matrix $d_{r,n}$, although in our simulation studies we will only plot paths $n \mapsto d_{r,n}$ for a few fixed r. The algorithm assumes that the final martingale values $S_n^{(k)}$ are sorted in the descending order, and the sorted values are denoted $S^1 \geq \dots \geq S^K$. One of its inputs is a family of martingale merging functions F_k, but as before, $F_k(\dots)$ is abbreviated to $F(\dots)$.

In our simulation studies we have 200 null hypotheses, all of them being $N(0,1)$, numbered from 1 to 200. The first 100 null hypotheses are false, and the true distribution is $N(-1,1)$; and the remaining 100 null hypotheses are true. At each step, from 1 to 10 000, we choose the hypothesis being tested

randomly with equal probabilities, so that each hypothesis is chosen with probability 0.5%. Figure 1 gives the plots $n \mapsto d_{r,n}$ for $r := 100$ (meaning that we aim to discover all 100 false null hypotheses), $r := 99$, and $r := 98$. Let us call such plots *discovery plots*. We generate the 10 000 observations randomly with the standard seed of 42 for the random number generator (in fact, the results are very sensitive to the chosen value for the seed). The final value $d_{100,10000}$ of the discovery plot for the top 100 martingale values ($r := 100$) is approximately 60.1. Using Jeffreys's [7, Appendix B] expression, there is very strong evidence that the top 100 martingale values exactly pinpoint the 100 false null hypotheses.

The martingale merging function used in Fig. 1 is U_1. It is clear that any symmetric martingale merging function, which is a convex mixture of U_n (Proposition 1 above), that does not have U_1 as its component, will produce very poor results for all discovery plots shown in Fig. 1: e.g., $U_2(S^{100}, S^{200})$ will be very small (approximately 7.80×10^{-25} in our case), and $U_2(S^{100}, S^k, \dots, S^{200})$, $k = 101, \dots, 199$, will be even smaller.

While Fig. 1 uses the U_1 martingale merging function, using, e.g., $(U_1+U_2)/2$ would give similar results.

5 Almost family-wise multiple testing

Let us now relax the requirement (1) to

$$|R \cap J_Q| \le 1.$$

This requirement can be interpreted as almost all discoveries in R being justified: we are allowing only one exception. The chronological discovery diagonal (2) now becomes the *chronological discovery subdiagonal*

$$\begin{aligned}\square_n(|R_{r,n} \cap J_Q| \le 1) &= \inf_{Q \in \mathcal{Q}: |R_{r,n} \cap J_Q| > 1} F\left(\left(S_n^{(k)}\right)_{k \in J_Q}\right) \\ &\ge \inf_{I \subseteq \{1,\dots,K\}: |R_{r,n} \cap I| > 1} F\left(\left(S_n^{(i)}\right)_{i \in I}\right) =: d'_{r,n}.\end{aligned}$$

The analogue of Algorithm 1 for the chronological discovery subdiagonal is given as Algorithm 2. In our simulation study we apply it to the martingale merging function U_2. One complication is that it sometimes has to be applied to sequences of length 1, in which case we understand it to be the same as U_1.

Figure 2 is analogous to Fig. 1 but allows one exception and uses U_2 as martingale merging function. In this case using U_2 works much better than U_1. The final value of the discovery plot for the top 100 martingale values ($r = 100$) is approximately 1.07×10^8.

6 Dynamic confidence regions

Necessity measures discussed in Sect. 2 are only one possible way to package the idea of necessity. A much more standard way is to use confidence regions,

Algorithm 2 Chronological discovery subdiagonal $d'_{r,n}$

Input: symmetric martingale merging functions F_k, $k \in \{1, \dots, K\}$.
Input: decreasing sequence of martingale values $S^1 \geq \dots \geq S^K$.

for $n = 1, 2, \dots$
  for $r = 1, \dots, K$
    if $r = 1$
      $I_r := \{r\}$
    else
      $I_r := \{r-1, r\}$
    $d'_{r,n} := F(I_r)$
    for $k = r+1, \dots, K$
      $S := F((S^i)_{i \in I_r \cup \{k, \dots, K\}})$
      if $S < d'_{r,n}$
        $d'_{r,n} := S$

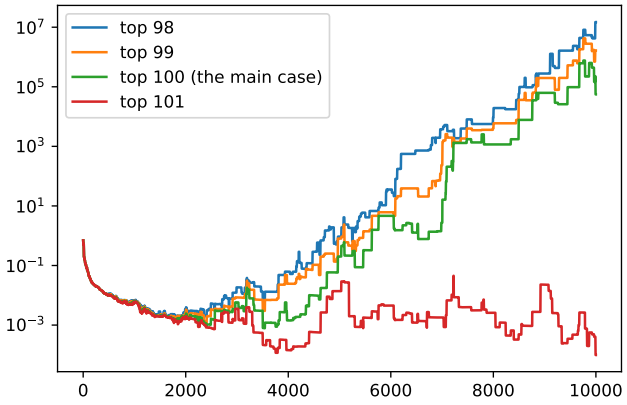


Figure 2: Discovery plots for 101, 100, 99, and 98 hypotheses when one exception is allowed (almost family-wise case).

which we will define in this section in our dynamic setting; our definitions will be natural modifications of the standard definitions (in the case of p-values) and the definitions given in [18, 17] (in the case of e-values). Let (S^Q) be a martingale test, fixed throughout this section.

We will be interested in confidence regions for the *parameter* $g(\mathbb{P})$, where $g : \mathcal{Q} \to \Theta$ is a mapping from the probability measures on the sample space to the parameter space Θ (which can be any set). The *exact confidence region* for $g(\mathbb{P})$ at time n corresponding to the realized outcome ω and significance level $\alpha > 0$ is defined as

$$\Gamma^g_{\alpha,n}(\omega) := \{g(Q) \mid S^Q_n(\omega) < \alpha\};$$

as usual, the dependence on ω is often suppressed. A *confidence region* is a set

of parameter values containing the exact confidence region.

Alternatively, we can define a confidence region as a set $A \subseteq \Theta$ such that

$$\square_n(\{Q \mid g(Q) \in A\}) \geq \alpha. \tag{3}$$

The exact confidence region $\Gamma^g_{\alpha,n}$ is the smallest such set. In other words, $A := \Gamma^g_{\alpha,n}$ satisfies (3), and any A satisfying (3) contains $\Gamma^g_{\alpha,n}$, $A \supseteq \Gamma^g_{\alpha,n}$.

Finally, we can define the exact confidence region $\Gamma^g_{\alpha,n}$ as the set of all $\theta \in \Theta$ satisfying $\Diamond_n(g^{-1}(\theta)) < \alpha$.

One disadvantage of the dynamic notion of exact confidence regions $\Gamma^g_{\alpha,n}$ is that, as a function of n, $\Gamma^g_{\alpha,n}$ is not decreasing: we are not guaranteed to have $\Gamma^g_{\alpha,n+1} \subseteq \Gamma^g_{\alpha,n}$. This phenomenon of "losing evidence" and ways of partially preventing it are discussed in [5, 11] and [13, Chap. 11].

It is essential to have the martingale test (S^Q) fixed in advance in order to have valid confidence regions; on the other hand, confidence regions corresponding to different g are valid simultaneously.

7 Multiple testing *en masse*

In this section we define, for each rejection set $R \subseteq \{1, \dots, K\}$, a confidence region for the number of justified discoveries in R (i.e., anomalous $k \in R$). Such confidence regions are often of the form $\{L, \dots, K\}$ for some lower bound L (we only have a lower confidence bound since $S^{(k)}$ can be arbitrarily close to being a test martingale without being one).

Given a rejection set R, we are interested in the parameter

$$g_R(Q) := |R \setminus J_Q|,$$

which is the number of justified discoveries. While in this paper we concentrate on the parameter function g_R, this function can be generalized in various directions; see, e.g., [19, Remark 6.1].

The confidence region for $g_R(\mathbb{P})$ at time n at significance level α consists of $j \in \{1, \dots, K\}$ satisfying $\Diamond_n(g_R^{-1}(j)) < \alpha$, where the possibility measure $\Diamond_n(g_R^{-1}(j))$ is

$$\begin{aligned} \Diamond_n(g_R^{-1}(j)) &= \min_{Q \in \mathcal{Q}: g_R(Q) = j} S_n^Q = \inf_{Q \in \mathcal{Q}: |R \setminus J_Q| = j} F\left(\left(S_n^{(k)}\right)_{k \in J_Q}\right) \\ &\geq \min_{I \subseteq \{1,\dots,K\}: |R \setminus I| = j} F\left(\left(S_n^{(i)}\right)_{i \in I}\right) =: D^R(j). \end{aligned} \tag{4}$$

Replacing $\Diamond_n(g_R^{-1}(j))$ by D^R_j we also obtain a valid (perhaps conservative) confidence region. In the case of the optimal $R := R_{r,n}$, we refer to

$$D_{r,j} := D^{R_{r,n}}(j)$$

as the *discovery matrix* at time n. It is a lower triangular matrix with $r \in \{1, \dots, K\}$ and $j \in \{0, \dots, r\}$. In the case $j = r$, the range of I includes the empty set $\emptyset$, and in this case we set F to 1.

Algorithm 3 Discovery matrix (lower triangular) $D_{r,j}$

Input: symmetric martingale merging functions F_k, $k \in \{1, \dots, K\}$.
Input: decreasing sequence of final martingale values $S^1 \geq \dots \geq S^K$.

```
for r = 1, ..., K
    for j = 0, ..., r
        I_{r,j} := {j + 1, ..., r}
        D_{r,j} := F((S^i)_{i ∈ I_{r,j}})
        for k = r + 1, ..., K
            e := F((S^i)_{i ∈ I_{r,j} ∪ {k,...,K}})
            if e < D_{r,j}
                D_{r,j} := e
```

In the computational experiments reported in this paper, the discovery matrix $D_{r,j}$ is always monotonically decreasing in j, and so

$$\Diamond_n(g_R^{-1}(j)) = \Diamond_n(g_R^{-1}(\{0, \dots, j\})).$$

This is essential for the interpretation of our results. However, in general, the discovery matrix $D_{r,j}$ is not guaranteed to be decreasing in j [19, 17], and so might need to be regularized by redefining $D_{r,j} := \min_{j' \leq j} D_{r,j'}$,

Algorithm 3 implements (4). In the case $j = r$ we have $I_{r,j} = \emptyset$, and as discussed earlier, we set $F(\emptyset) := 1$. This algorithm computes the discovery matrix in time $O(K^4)$ when F is a fixed U_n or a fixed convex mixture of the first few U_n; this follows from F being computable in time $O(K)$, which in turn follows from, e.g., Newton's identities (see Appendix B). It is interesting that for the simulation studies reported in this paper we do not need more efficient algorithms such as the $O(K^3)$ algorithm given in [17] and, in the case of U_1, the $O(K^2)$ algorithm given in [19]; computations take at most a couple of minutes on an ordinary laptop.

See Figs 3 and 4 for discovery matrices at time 10 000 for the same simulated data as before. The colour coding used in both figures involves much more extreme values of the possibility measures than the usual scheme used in [19, 17] (the scheme in [19, 17] uses the thresholds proposed by Jeffreys [7, Appendix B]):

- The final martingale values below 10 are shown in green. For such cells (in row r and column j) we have $D_{r,j} < 10$, and these are exactly the cells for which we do not have strong evidence for there being at least j justified discoveries.
- The final martingale values between 10 and 100 are shown in yellow. These are exactly the cells for which we have strong but not decisive evidence for there being at least j justified discoveries ($10 \leq D_{r,j} < 100$).
- The final martingale values between 100 and 10^8 are shown in orange. For these cells (and for the cells in the darker colours) we have decisive evidence for there being at least j justified discoveries.

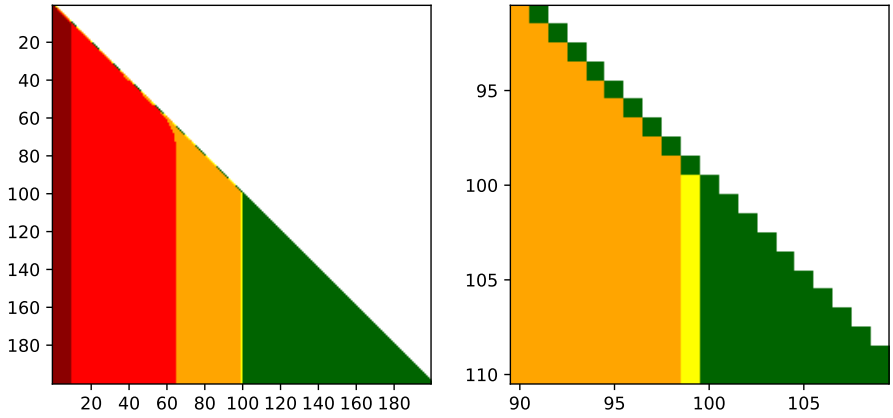


Figure 3: Left panel: Discovery matrix for the mean U_1 as martingale merging function. The right panel shows its middle portion.

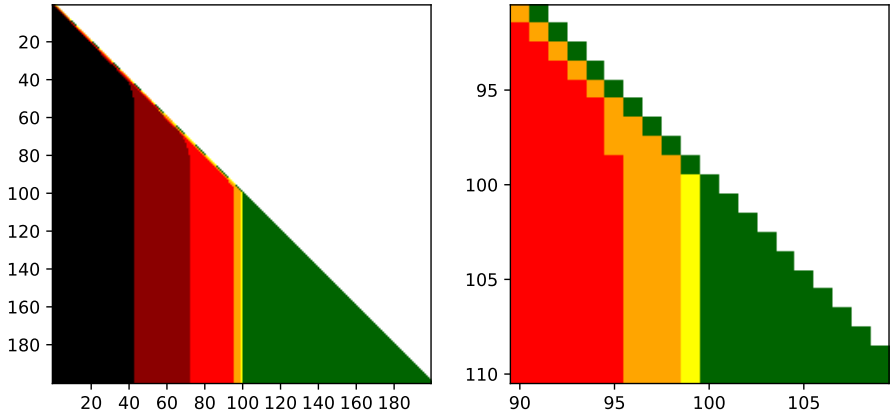


Figure 4: Left panel: Discovery matrix for the mixture $(U_1+U_2)/2$ as martingale merging function. The right panel shows its middle portion.

- The final martingale values between 10^8 and 10^{14} are shown in red.
- The final martingale values between 10^{14} and 10^{20} are shown in dark red.
- The final martingale values above 10^{20} are shown in black.

The *diagonal* of a discovery matrix consists of the cells $D_{r,r-1}$, the justification for the offset of 1 being that j in $D_{r,j}$ starts from 0 rather than 1. Correspondingly, the *subdiagonal* consists of $D_{r,r-2}$, and the *superdiagonal* consists of $D_{r,r}$.

The diagonal, subdiagonal, and superdiagonal can be clearly seen in the right panels of the two figures. The diagonal can be traced starting from the top left corner of either bounding box. In both figures the diagonal (when moving south-east) is first orange, then yellow (just one cell), and then green. In Fig. 3 the subdiagonal is also first orange, then yellow (just one cell), and then green, but in Fig. 4 the subdiagonal is first red and only later becomes orange. The superdiagonal is green in both figures.

The corresponding confidence intervals (i.e., confidence regions that happen to be intervals) can be read off the two figures. For example, for each row r, the non-green cells represent the confidence interval for the number of justified discoveries among the top r martingale values at significance level 10. We can see that for $r = 100$ the confidence interval is $\{100\}$; it is degenerate and contains only one value: we are predicting that all 100 null hypotheses with the largest final martingale values are justified (and *a fortiori* true) discoveries, and we have strong evidence for that. On the other hand, the green superdiagonal entry $D_{100,100}$ is very small (it is, approximately, 1.13×10^{-20}). If we raise the significance level to 100 (Jeffreys's threshold for decisive evidence), the confidence interval widens to $\{99, 100\}$. And when we raise it further to the huge value of 10^8, the confidence interval (given by the non-red entries in the right panel) widens to $\{96, 97, 98, 99, 100\}$.

Figures 3 and 4 suggest that discovery matrices are monotonically decreasing in the eastern and south-eastern directions and monotonically increasing in the southern direction. These properties of monotonicity (except for the monotonicity in j discussed earlier) are stated and proved in [19] and [17].

Figures 1 and 2 show the evolution of various entries of discovery matrices such as those in Figs 3 and 4 over time. The green lines in both panels of Fig. 1 show the evolution of the diagonal entry $D_{100,99}$ over the 10 000 observations. The orange and blue lines in Fig. 1 show the evolution of the entries $D_{99,98}$ and $D_{98,97}$, respectively. All these entries lie on the diagonal

$$d_{r,10000} := D_{r,r-1}$$

of the discovery matrix at time 10 000. We talked about family-wise validity in Sect. 4 since $D_{r,r-1}$ is the largest significance level at which the confidence interval is a one-element set, namely $\{r\}$.

The right panel of Fig. 1 also shows, as red line, the evolution of the diagonal entry $D_{101,100}$. This line shows that the green entry $D_{101,100}$ in Fig. 3 is very small; in numbers, the final value of the red line in Fig. 1 (i.e., $D_{101,100}$ in Fig. 3) is, approximately, 7.80×10^{-7}). The value of $D_{101,100}$ in Fig. 4 is even smaller.

The green line in Fig. 2 shows the evolution of the entry $D_{100,98}$, which determines the significance levels at which the confidence interval for the number of justified discoveries is $\{99, 100\}$ (allowing one unjustified discovery). Its final value corresponds to the entry $D_{100,98}$ in Fig. 4, and the two numbers have the same order of magnitude (they are, however, different because Figs 2 and 4 use different martingale merging functions, U_2 vs $(U_1 + U_2)/2$). The orange and blue lines in Fig. 2 are interpreted in the same way; they correspond to the

entries $D_{99,97}$ and $D_{98,96}$, respectively, of Fig. 4. The red line in Fig. 2, however, disagrees sharply with the entry $D_{101,99}$ of Fig. 4, because the martingale merging function $(U_1 + U_2)/2$ used in Fig. 4 has U_1 as its component.

8 Conclusion

These are some possible directions of further research:

- The motivation behind this paper is coming from game-theoretic probability and statistics, but its mathematical setting is that of measure-theoretic probability. Replacing measure-theoretic probability by purely game-theoretic probability (as developed in [13]) would simplify the exposition and lead to more natural and general definitions.
- This paper concentrates on simulation studies. It would be interesting to conduct empirical studies on benchmark or real-world datasets, for example ones collected in the course of statistical meta-analyses.
- The experimental results of Sect. 7 establish confidence regions for the numbers of true discoveries, which can be restated as results about the false discovery proportions, FDP. Are there any interesting theoretical results in this context about false discovery rates, FDR (as in [1] in the case of p-values and [21] in the case of e-values)?
- This paper concentrates on the closed setting (when the number of null hypotheses K is given in advance). The open setting, where new hypotheses may appear at any moment, may be even more interesting. In this case we need, of course, to break the symmetry between the null hypotheses: there is no uniform probability measure on $\{1, 2, \dots\}$.

Acknowledgments

As usual, I am grateful to Glenn Shafer for his thoughts and comments. Many thanks to Jean Gallier for his advice on literature and for correcting the statement of Lemma 4.1.3 in [6]. My research has been partially supported by Mitie.

References

[1] 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.

[2] Nicolas Bourbaki. *Elements of Mathematics. Algebra II. Chapters 4–7*. Springer, Berlin, 2003.

[3] Henri Cartan. *Calcul différentiel*. Hermann, Paris, 1967.

[4] Pierre Crépel. Jean Ville remembers martingales. In Laurent Mazliak and Glenn Shafer, editors, *The Splendors and Miseries of Martingales: Their History from the Casino to Mathematics*, pages 375–391. Birkhäuser, Cham, 2022.

[5] A. Philip Dawid, Steven de Rooij, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Insuring against loss of evidence in game-theoretic probability. *Statistics and Probability Letters*, 81:157–162, 2011.

[6] Jean Gallier. *Curves and Surfaces in Geometric Modeling: Theory and Algorithms*. Morgan Kaufmann, San Francisco, CA, 2000. An updated version (2024) is available from the author's website (accessed in March 2024).

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

[8] W. Keith Nicholson. *Introduction to Abstract Algebra*. Wiley, New York, second edition, 1999. Fourth edition: 2012.

[9] Glenn Shafer. *The Art of Causal Conjecture*. MIT Press, Cambridge, MA, 1996.

[10] Glenn Shafer. A counterexample to Richard von Mises's theory of collectives, by Jean Ville. Available on the web (accessed in March 2024), 2005.

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

[12] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

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

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

[15] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919.

[16] Richard von Mises. *Wahrscheinlichkeit, Statistik, und Wahrheit*. Springer, Berlin, 1928. English translation: *Probability, Statistics and Truth*. William Hodge, London, 1939.

[17] 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 2020.

[18] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49:1736–1754, 2021.

[19] Vladimir Vovk and Ruodu Wang. Confidence and discoveries with e-values. *Statistical Science*, 38:329–354, 2023. Revised arXiv version: arXiv:1912.13292 [math.ST] (November 2022).

[20] Vladimir Vovk and Ruodu Wang. Merging sequential e-values via martingales. *Electronic Journal of Statistics*, 18:1185–1205, 2024.

[21] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society B*, 84:822–852, 2022.

A Proofs

In this appendix I will prove Propositions 1 and 2 (and state and prove a new Proposition 4). A *multiaffine function* is a multivariate function that is affine in each of its arguments. (So that multiaffine polynomials are multiaffine functions, and in Sect. A.1 we will see that these two notions are equivalent.) The proofs will follow from the following lemma.

Lemma 3. *A martingale merging function must be multiaffine.*

Proof. Let F be a martingale merging function. We are required to prove

$$\begin{aligned} F(s_1,\dots,s_{k-1},\alpha s'_k+(1-\alpha)s''_k,s_{k+1},\dots,s_K) \\ = \alpha F(s_1,\dots,s_{k-1},s'_k,s_{k+1},\dots,s_K) \\ + (1-\alpha)F(s_1,\dots,s_{k-1},s''_k,s_{k+1},\dots,s_K), \end{aligned} \tag{5}$$

where $\alpha\in(0,1)$. Let us fix $s_1,\dots,s_{k-1},s'_k,s''_k,s_{k+1},\dots,s_K$, and α. Consider the sample space $\Omega:=\{0,1\}^N$ with the natural filtration and a positive probability measure $\mathbb{P}$ (i.e., $\mathbb{P}(E)>0$ for any $E\neq\emptyset$). Suppose that the set of sample points where

$$S^{(1)}_{N-1}=s_1,\dots,S^{(k-1)}_{N-1}=s_{k-1},S^{(k)}_{N-1}=\alpha s'_k+(1-\alpha)s''_k,$$
$$S^{(k+1)}_{N-1}=s_{k+1},\dots,S^{(K)}_{N-1}=s_K$$

for some uncorrelated test martingales $S^{(1)},\dots,S^{(K)}$ is non-empty, and let $(\omega_1,\dots,\omega_N)\in\Omega$ be such a sample point. Suppose that in our probability space we have the branching probability

$$\frac{\mathbb{P}(\{(\omega_1,\dots,\omega_{N-1},1)\})}{\mathbb{P}(\{(\omega_1,\dots,\omega_{N-1},0),(\omega_1,\dots,\omega_{N-1},1)\})}=\alpha$$

and that the martingale $S^{(k)}$ satisfies

$$S^{(k)}_N((\omega_1,\dots,\omega_{N-1},1))=s'_k$$

$$S_N^{(k)}((\omega_1,\dots,\omega_{N-1},0)) = s''_k.$$

The existence of such a probability space and uncorrelated test martingales $S^{(1)},\dots,S^{(K)}$ is obvious. Since

$$T_n := F(S_n^{(1)},\dots,S_n^{(K)})$$

is a test martingale, we have

$$T_{N-1}((\omega_1,\dots,\omega_N)) = \alpha T_N((\omega_1,\dots,\omega_{N-1},1)) + (1-\alpha)T_N((\omega_1,\dots,\omega_{N-1},0)),$$

which is equivalent to (5). □

A.1 Proof of Proposition 1

To show that a martingale merging function is a multiaffine polynomial, we combine Lemma 3 with Lemma 4.1.3 in [6] (whose proof relies on Cartan's method of successive differences [3, Sect. 6.3]). According to [6, Lemma 4.1.3], a multiaffine function f of K arguments has the form

$$f(s_1,\dots,s_K) = f(0,\dots,0) + \sum_{\substack{n\in\{1,\dots,K\}\\ \{1\le k_1\le\dots\le k_n\le K\}}} f_{k_1,\dots,k_n}(s_{k_1},\dots,s_{k_n}),$$

where $f_{k_1,\dots,k_n}$ are multilinear functions (i.e., functions linear in each argument). It remains to notice that

$$f_{k_1,\dots,k_n}(s_{k_1},\dots,s_{k_n}) = c s_{k_1}\dots s_{k_n}$$

for some constant c; indeed,

$$\begin{aligned} f_{k_1,\dots,k_n}(s_{k_1},\dots,s_{k_n}) &= s_{k_1} f_{k_1,\dots,k_n}(1,s_{k_2},\dots,s_{k_n}) \\ &= s_{k_1}s_{k_2} f_{k_1,\dots,k_n}(1,1,s_{k_3},\dots,s_{k_n}) = \dots \\ &= s_{k_1}\dots s_{k_n} f_{k_1,\dots,k_n}(1,\dots,1). \end{aligned}$$

Let us now check that a martingale merging function F is a positive multiaffine polynomial. Suppose there is a negative coefficient in front of one or more of its monomials. Choose and fix a monomial with a negative coefficient. Set all variables that do not occur in this monomial to zero. Set each of the variables that do occur in this monomial to C and let $C\to\infty$. For a large enough C, the value of the polynomial (the value being a univariate polynomial in C with a negative leading coefficient) will become negative, which is impossible.

There is a minor gap in our derivation of Proposition 1 from [6, Lemma 4.1.3]: the latter assumes that the multiaffine function f is defined on an affine space whereas in our context f is defined on $[0,\infty)^K$. Let us check that every affine $f:[0,\infty)^K\to\mathbb{R}$ can be extended to an affine $f':\mathbb{R}^K\to\mathbb{R}$. We proceed by induction and show that if $f(x_1,\dots,x_k,\dots,x_K)$ is an affine function with x_k ranging over $[0,\infty)$ we can extend it to an affine function with x_k ranging over

$\mathbb{R}$ (with the ranges of the other arguments of f unchanged). Without loss of generality, let $k := 1$. We extend f to f' by the affinity in x_1: for any $x_1 < 0$,

$$f'(x_1, x_2, \dots) := x_1 f'(1, x_2, \dots) + (1 - x_1) f'(0, x_2, \dots). \tag{6}$$

We only need to check that f' is multiaffine. The affinity in x_1 holds by construction, so we only need to check that f' is affine in x_k for $k \neq 1$. Without loss of generality, let $k := 2$. Since the arguments $x_3, \dots, x_K$ of f and f' are kept fixed, we will ignore them. Our goal is to show that

$$f'\left(x_1, \alpha x_2' + (1-\alpha) x_2''\right) = \alpha f'\left(x_1, x_2'\right) + (1-\alpha) f'\left(x_1, x_2''\right) \tag{7}$$

for $x_1 < 0$. By the definition (6), the equality (7) is equivalent to

$$\begin{aligned} x_1 f\left(1, \alpha x_2' + (1-\alpha) x_2''\right) + (1 - x_1) f\left(0, \alpha x_2' + (1-\alpha) x_2''\right) \\ = \alpha x_1 f\left(1, x_2'\right) + \alpha (1 - x_1) f\left(0, x_2'\right) \\ + (1-\alpha) x_1 f\left(1, x_2''\right) + (1-\alpha)(1 - x_1) f\left(0, x_2''\right). \end{aligned} \tag{8}$$

It remains to notice that (8) can be derived as linear combination of

$$f\left(1, \alpha x_2' + (1-\alpha) x_2''\right) = \alpha f\left(1, x_2'\right) + (1-\alpha) f\left(1, x_2''\right) \tag{9}$$

and

$$f\left(0, \alpha x_2' + (1-\alpha) x_2''\right) = \alpha f\left(0, x_2'\right) + (1-\alpha) f\left(0, x_2''\right) \tag{10}$$

(with the coefficients x_1 for (9) and $1 - x_1$ for (10)).

A.2 Proof of Proposition 2

We proceed as in Sect. A.1 replacing Lemma 4.1.3 in [6] by Lemma 4.1.4. Alternatively, we could have derived Proposition 2 from Proposition 1.

A.3 Comparisons with merging independent and sequential e-values

In this subsection we will discuss ie-merging and se-merging functions, to be defined momentarily; for a further discussion of these functions see, e.g., [20].

Suppose $E_1, \dots, E_K$ are admissible independent e-variables (i.e., nonnegative random variables that are independent and satisfy $\mathbb{E}(E_k) = 1$, $k = 1, \dots, K$) or admissible sequential e-variables (i.e., nonnegative, adapted, and satisfying $\mathbb{E}(E_k \mid \mathcal{F}_k - 1) = 1$, $k = 1, \dots, K$). Then

$$S_n^{(k)} := \begin{cases} 1 & \text{if } n < k \\ E_k & \text{if } n \geq k \end{cases}$$

are uncorrelated test martingales with final values $E_1, \dots, E_k$. Therefore, any normalized positive multiaffine polynomial (NPMAP) is an ie-merging function, in the sense of mapping any (admissible) independent e-variables to an

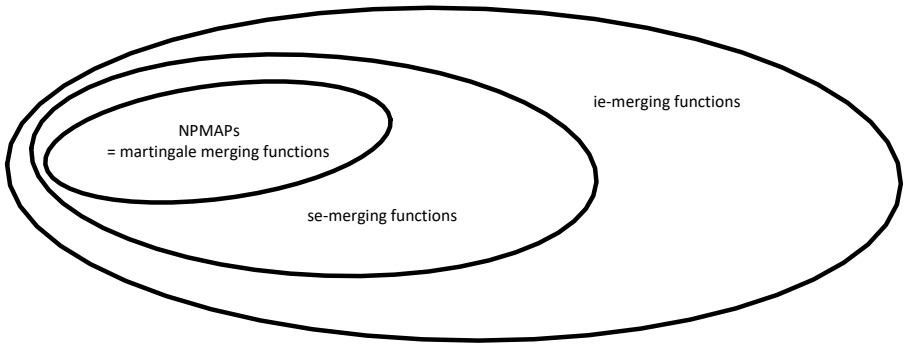


Figure 5: Three families of merging functions; all inclusions in this Euler diagram are strict.

e-variable (i.e., nonnegative random variable E satisfying $\mathbb{E}(E) \le 1$); moreover, any NPMAP is an se-merging function, in the sense of mapping any (admissible) sequential e-variables to an e-variable. This gives us the structure shown in Fig. 5: it is obvious that every se-merging function is an ie-merging function.

Let us check that both inclusions in the Euler diagram shown in Fig. 5 are strict. The outer inclusion is strict since the function

$$f(e_1, e_2) := \frac{1}{2}\left(\frac{e_1}{1+e_1} + \frac{e_2}{1+e_2}\right)(1 + e_1 e_2) \tag{11}$$

is an admissible ie-merging function [18, Remark 4.3] while it is not an se-merging function [20, Example 2]. To see that the inner inclusion is strict, notice that

$$(e_1, \dots, e_K) \mapsto 1 + g(e_1, \dots, e_{K-1})(e_K - 1)$$

is an se-merging function for any function g taking values in $[0, 1]$, even highly non-linear one, such as $g(e_1, \dots, e_{K-1}) := (\sin e_1 + 1)/2$.

Specializing Fig. 5 to symmetric merging functions we obtain Fig. 6. The function (11) is symmetric and so can also serve as an example demonstrating that the outer inclusion in Fig. 6 is strict. On the other hand, the following proposition shows that the inner inclusion in Fig. 6 is strict in an uninteresting way.

Proposition 4. *Every symmetric se-merging function is dominated by a convex mixture of NESPs.*

Proof. Theorem 1 in [20] says that any se-merging function is dominated by a martingale merging function (where "martingale merging function" is used in a sense different from this paper; this proof uses "martingale merging function" in the sense of [20]). By definition, a martingale merging function is affine in its last argument. By symmetry, it is affine in each argument. It remains to follow the reasoning of Sects A.1 and A.2. □

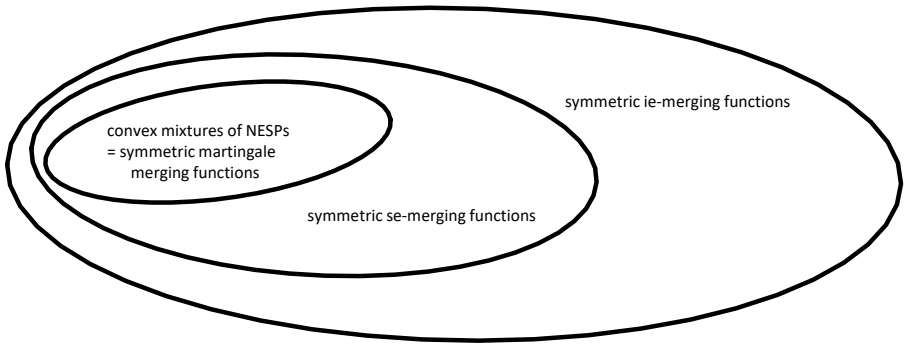


Figure 6: Three families of symmetric merging functions. All inclusions in this Euler diagram are strict, but each symmetric se-merging function is dominated by a convex mixture of NESPs.

Proposition 4 appears to be less interesting than Propositions 1 and 2: the family of symmetric se-merging functions is not as natural as the other two symmetric families in Fig. 6.

B Computing NESPs

In this appendix we will discuss how to compute the NESP $U_n = U_n(s_1, \dots, s_K)$ for a fixed n efficiently, namely, in time $O(K)$. We will use the fact that the polynomials $p_n := s_1^n + \cdots + s_K^n$ can be computed in time $O(K)$, and so one way to compute U_n efficiently is to express them via p_n.

These are the efficient representations for the first few NESPs:

$$
\begin{aligned}
U_1(s_1, \dots, s_K) &= \frac{1}{K}(s_1 + \cdots + s_K) \\
U_2(s_1, \dots, s_K) &= \frac{1}{K(K-1)}\left((s_1 + \cdots + s_K)^2 - (s_1^2 + \cdots + s_K^2)\right) \\
U_3(s_1, \dots, s_K) &= \frac{1}{K(K-1)(K-2)}\big((s_1 + \cdots + s_K)^3 \\
&\quad - 3(s_1^2 + \cdots + s_K^2)(s_1 + \cdots + s_K) + 2(s_1^3 + \cdots + s_K^3)\big) \\
U_4(s_1, \dots, s_K) &= \frac{1}{K(K-1)(K-2)(K-3)}\big((s_1 + \cdots + s_K)^4 \\
&\quad - 6(s_1^2 + \cdots + s_K^2)(s_1 + \cdots + s_K)^2 \\
&\quad + 8(s_1^3 + \cdots + s_K^3)(s_1 + \cdots + s_K) \\
&\quad + 3(s_1^2 + \cdots + s_K^2)^2 - 6(s_1^4 + \cdots + s_K^4)\big).
\end{aligned}
$$

It is clear that such a representation exists for any fixed n, and it allows us to

compute $U_n(s_1, \ldots, s_K)$ in time $O(K)$. In terms of *Bell polynomials*

$$B_n(x_1, \ldots, x_n) := n! \sum_{\substack{(j_1,\ldots,j_n)\in\mathbb{N}^n: \\ j_1+2j_2+\cdots+nj_n=n}} \prod_{i=1}^{n} \frac{x_i^{j_i}}{(i!)^{j_i} j_i!},$$

where $\mathbb{N} := \{0, 1, \ldots\}$ is the set of natural numbers, the general expression is

$$U_n(s_1, \ldots, s_K) = \frac{(K-n)!}{K!} B_n(p_1, -p_2, 2!p_3, -3!p_4, \ldots, (-1)^{n-1}(n-1)!p_n),$$

where $p_n := s_1^n + \cdots + s_K^n$.

A less straightforward way of computing the elementary symmetric polynomials (and therefore, U_n) via $p_1, p_2, \ldots$ in time $O(K)$ would be to use recursion and Newton's identities (see, e.g., [2, Lemma 4 of Chap. 4] or [8, Theorem 4.5.5]).