Testing exchangeability in the batch mode with e-values and Markov alternatives

Vladimir Vovk

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

On-line Compression Modelling Project (New Series)

Working Paper #38

First posted May 9, 2023. Last revised February 6, 2025.

Project web site:
http://alrw.net

Abstract

The topic of this paper is testing the assumption of exchangeability, which is the standard assumption in mainstream machine learning. The common approaches are online testing by betting (such as conformal testing) and the older batch testing using p-values (as in classical hypothesis testing). The approach of this paper is intermediate in that we are interested in batch testing by betting; as a result, p-values are replaced by e-values. As a first step in this direction, this paper concentrates on the Markov model as alternative. The null hypothesis of exchangeability is formalized as a Kolmogorov-type compression model, and the Bayes mixture of the Markov model w.r. to the uniform prior is taken as simple alternative hypothesis. Using e-values instead of p-values leads to a computationally efficient testing procedure. Two appendixes discuss connections with the algorithmic theory of randomness; in particular, the test proposed in this paper can be interpreted as a poor man's version of Kolmogorov's deficiency of randomness.

Contents

1 Introduction

Exchangeability is the fundamental assumption in machine learning. Traditional machine learning studies prediction under exchangeability (see, e.g., [28]), while newer methods consider deviations from exchangeability (see, e.g., [19]). The role of exchangeability in conformal prediction, as subarea of machine learning, is briefly reviewed in [32, Sect. 13.5.1].

Testing the assumption of exchangeability is a traditional topic in conformal prediction (see, e.g., [32, Part III]). It is done in the online mode and is based on conformal test martingales. This area is often referred to as conformal testing.

The classical approach to testing exchangeability, which developed in statistics starting from at least 1943 [39], proceeds in the batch mode: we are given the data sequence as one batch rather than getting its elements sequentially one by one; see [12, Sect. 7.2] for a review. As always in classical hypothesis testing, testing exchangeability in the batch mode is based on p-values.

In this paper we will adapt standard methods of conformal testing to testing exchangeability in the batch mode. In particular, p-values will be replaced by e-values [36, 3], which are widely used in conformal testing: namely, conformal test martingales are obtained by compounding e-values. An important advantage of e-values is that their use facilitates efficient computations.

The null hypothesis of exchangeability will be defined in Sect. 2 using the terminology of compression modelling, widely used in conformal prediction [32, Chap. 11]. Compression modelling is an algorithm-free version of Kolmogorov's way of stochastic modelling: cf. [30], [33], [38, Sect. 2], and [32, Sect. 11.6.1]. Kolmogorov's original version will be discussed in Appendix A.

In Sect. 2 we also define e-variables, which are functions for producing e-values in testing exchangeability (or another null hypothesis). We will derive our main e-variable as likelihood ratio for a Markovian alternative hypothesis, which we will introduce in Sect. 4. A simple optimality property of the likelihood ratios is derived in Sect. 3.

After defining our main alternative hypothesis in Sect. 4, we derive an efficient algorithm for computing the corresponding e-variable. The power of this e-variable is the topic of Sect. 5. The algorithm's performance in view of the results of Sect. 5 is studied in Sect. 6 using simulated data. Section 7 concludes.

Appendix A describes Kolmogorov's original ideal picture of algorithmic randomness. In the following Appendix B we will discuss possible ways of making this picture more practical, and in Appendix C will go deeper into another class of alternatives for testing exchangeability (namely, into the changepoint alternatives).

In traditional statistics, the p-value version of the procedure of this paper is often presented in terms of the Neyman structure; see, e.g., [13, Sect. 4.3]. We discuss its counterpart for e-values in Appendix D.

2 Testing exchangeability

We consider the simplest binary case, and our *observation space* is $\mathbf{Z} := \{0, 1\}$. Fix an integer $N > 1$, which we will refer to as the *time horizon*. We are interested in binary *data sequences* $(z_1, \dots, z_N) \in \Omega := \mathbf{Z}^N$. A *Kolmogorov compression model* (KCM) is a *summarising statistic* $t : \Omega \to \Sigma$, where Σ is a finite set (the *summary space*), together with the implicit statement that given the *summary* $t(z_1, \dots, z_N)$ (for which we do not make any stochastic assumptions) the actual data sequence $(z_1, \dots, z_N)$ is generated from the uniform probability measure on the set $t^{-1}(t(z_1, \dots, z_N))$ of all data sequences compatible with the summary. Our *null hypothesis* is the KCM, which we call the *exchangeability compression model* (ECM), $t_E(z_1, \dots, z_N) := z_1 + \dots + z_N$. (In the current binary case this is equivalent to the more standard definition $t_E(z_1, \dots, z_N) := \wr z_1, \dots, z_N \int$ used in [32, Sect. 11.3.1], where $\wr \cdots \int$ stands for a multiset.)

KCM and ECM are two of the three main classes of models used in this paper. The third, largest, class will be introduced later in this section and called BCM. Therefore, the inclusions between the classes will be

$$\text{ECM} \subseteq \text{KCM} \subseteq \text{BCM}. \tag{1}$$

Let us say that a probability measure P on Ω *agrees* with a summarising statistic t if the data sequences with the same summary have the same P-probability. A probability measure P on Ω is *exchangeable* if $P(\{(z_1, \dots, z_N)\})$ depends on $z_1, \dots, z_N$ only via $z_1 + \dots + z_N$ (equivalently, via $\wr z_1, \dots, z_N \int$).

Lemma 2.1. *The exchangeable probability measures on Ω are exactly the probability measures that agree with the ECM (the mixtures of the uniform probability measures on $t_E^{-1}(k)$, $k \in \{0, \dots, N\}$).*

The easy proof of Lemma 2.1 is omitted. It shows that, in terms of standard statistical modelling, we can define our null hypothesis as the set of all exchangeable probability measures on Ω.

An *e-variable* w.r. to a probability measure is a nonnegative function on Ω with expectation at most 1. An *exchangeability e-variable* is a function $E : \Omega \to [0, \infty)$ whose average over each $t_E^{-1}(k)$ is at most 1. Such a function E can be used for testing the assumption of exchangeability: if E is chosen in advance, observing a very large $E(\omega)$ for the realized outcome $\omega \in \Omega$ casts doubt on the exchangeability assumption.

Alternatively (and equivalently), an exchangeability e-variable may be defined as an e-variable w.r. to every exchangeable probability measure.

Proposition 2.2. *The two meanings of an exchangeability e-variable coincide.*

Proof. If the average of E over each $t_E^{-1}(k)$ is at most 1, it will be an e-variable w.r. to each exchangeable probability measure by Lemma 2.1.

Now suppose E is an e-variable w.r. to each exchangeable probability measure. Since the uniform probability measure on $t_E^{-1}(k)$ is exchangeable, the average of E over $t_E^{-1}(k)$ will be at most 1. □

All null hypotheses discussed in this paper will be KCMs. In the main part of the paper we will concentrate on the ECM, but in this and next sections we will also give more general definitions. An *e-variable* w.r. to a KCM t is a function $E : \Omega \to [0, \infty)$ such that the arithmetic mean of E over $t^{-1}(\sigma)$ is at most 1 for every $\sigma \in t(\Omega)$. *E-values* are values taken by e-variables.

Disintegration of the alternative hypothesis

Let us fix a simple *alternative hypothesis* Q, which is a probability measure on Ω. Our statistical procedures will depend on Q only via the corresponding *batch compression model* (BCM). A *BCM* is a pair (t, P) such that $t : \Omega \to \Sigma$ is a summarising statistic and $P : \Sigma \hookrightarrow \Omega$ (to use the notation of [32, Sect. A.4]) is a Markov kernel such that $P(\sigma)$ is concentrated on $t^{-1}(\sigma)$ for each $\sigma \in \Sigma$. As before, we refer to $t(\omega)$ as the *summary* of ω. Kolmogorov compression models are a special case in which each $P(\sigma)$ is the uniform probability measure on $t^{-1}(\sigma)$.

Remark 2.3. Batch compression models are standard and are often used without giving them any name, as in [11]. They are the batch counterpart of online compression models used in conformal prediction [32, Chap. 11]. The three classes shown in (1) are used in different contexts in this paper: general BCMs serve as alternative hypotheses, the null hypothesis of interest in the main part of the paper is the ECM, and in the appendix we will discuss more general KCMs as null hypotheses.

With an alternative hypothesis Q and a summarising statistic $t : \Omega \to \Sigma$ (serving as null hypothesis) we associate the *alternative Markov kernel* $\sigma \in \Sigma \mapsto Q_\sigma$ defined by

$$Q_\sigma(\{\omega\}) := \frac{Q(\{\omega\})}{Q(t^{-1}(\sigma))}, \quad \sigma \in \Sigma, \ \ \omega \in t^{-1}(\sigma). \tag{2}$$

(We are mainly interested in alternative hypotheses Q for which the denominator of (2) is always positive, but in general we could set, e.g., $0/0 := 1/2$ in our binary context.) As compared with Q, the alternative Markov kernel loses the information about $Q(t^{-1}(\sigma))$ for $\sigma \in \Sigma$. (And of course, the reader should keep in mind that alternative Markov kernels and Markov alternative hypotheses are completely different objects, despite both being named after Andrei Andreevich Markov Sr.)

3 Frequentist performance of e-variables

Suppose Q (the alternative probability measure) is the true data-generating distribution, and we keep generating data sequences $(z_1, \dots, z_N) \in \Omega$ from Q in the IID fashion. The following lemma allows us to define the efficiency of an e-variable via its frequentist performance when we keep applying it repeatedly to accumulate capital. This is a special case of Kelly's criterion [6].

Lemma 3.1. *Consider an e-variable E w.r. to a Kolmogorov compression model $t : \Omega \to \Sigma$. For any alternative probability measure Q on Ω, the limit*[1]

$$\mathrm{ep}_Q(E) := \lim_{I\to\infty} \frac{1}{I} \ln \prod_{i=1}^{I} E(z_1^i, \dots, z_N^i) \tag{3}$$

where $(z_1^i, \dots, z_N^i)$ is the ith data sequence generated from Q independently, exists Q^∞-almost surely. Moreover, for all E and Q,

$$\mathrm{ep}_Q(E) = \int \ln E \, \mathrm{d}Q. \tag{4}$$

The interpretation of (3) is that our capital $\prod_{i=1}^{I} E(z_1^i, \dots, z_N^i)$ grows exponentially fast when betting repeatedly using E (we will see later, in Lemma 3.2, that we can indeed expect it to grow rather than shrink if we can guess a good Q), and its rate of growth is given by the expression (4), which we will refer to as the *e-power* of E under the alternative Q.

Proof. It suffices to rewrite (3) as

$$\mathrm{ep}_Q(E) = \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} \ln E(z_1^i, \dots, z_N^i)$$

and apply Kolmogorov's law of large numbers to the IID random variables $\ln E(z_1^i, \dots, z_N^i)$ with expectation $\int \ln E \, \mathrm{d}Q$ (which exists and is finite since the sample space is assumed to be finite). □

To justify the expression (4) using frequentist considerations, we do not really need the IID picture, as emphasized by Neyman [16, Sect. 10]. When generating $z_1^i, \dots, z_N^i$ for different i, we may test different Kolmogorov compression models $t = t_i$, perhaps with different time horizons $N = N_i$, against different alternatives $Q = Q_i$ and using different E_i. The corresponding generalization of Lemma 3.1 states that the long-term rate of growth of our capital will be asymptotically close to the arithmetic average of $\int \ln E_i \, \mathrm{d}Q_i$. It will involve certain regularity conditions needed for the applicability of the martingale strong law of large numbers (e.g., in the form of [23, Chap. 4], which allows non-stochastic choice of N_i, t_i, Q_i, and E_i). If the alternative hypothesis does not hold in all trials, Lemma 3.1 is still applicable to the trials where it does hold.

Now it is easy to find the optimal, in the sense of ep_Q, e-variable; it will be the ratio of the alternative Markov kernel to the null hypothesis.

Lemma 3.2. *The maximum of ep_Q is attained at*

$$E(\omega) := \left|t^{-1}(t(\omega))\right| Q_{t(\omega)}(\{\omega\}), \quad \omega \in \Omega. \tag{5}$$

[1] In this paper, our notation for logarithms is ln (natural) and log (binary, used only in Appendix A).

In this case,

$$\operatorname{mep}(Q) := \operatorname{ep}_Q(E) = \int \ln \left|t^{-1}(\sigma)\right| (t_* Q)(\mathrm{d}\sigma) + H(t_* Q) - H(Q), \tag{6}$$

where $t_* Q$ (a probability measure on the summary space Σ) is the push-forward measure

$$(t_* Q)(\{\sigma\}) := Q(t^{-1}(\sigma))$$

of Q by t (the summarising statistic of the null hypothesis), and $H(\cdot)$ stands for the entropy.

We will call $\operatorname{mep}(Q)$ defined by (6) the *maximum e-power* of the alternative Q. A sizeable $\operatorname{mep}(Q)$ for a plausible alternative Q means that the testing problem is not hopeless and has some potential. The guarantee given by Lemma 3.1, however, is frequentist and not applicable if testing is done only once, in which case we also want the optimal e-variable (5) not to be too volatile.

Proof. In this paper we let U_A stand for the uniform probability measure on a finite non-empty set A. The optimization $\int E \, \mathrm{d}Q \to \max$ can be performed inside each block $t^{-1}(\sigma)$ separately. Using the nonnegativity of the Kullback–Leibler divergence, we have, for each $\sigma \in t(\Omega)$,

$$\operatorname{ep}_{Q_\sigma}\left(\frac{Q_\sigma}{U_{t^{-1}(\sigma)}}\right) \geq \operatorname{ep}_{Q_\sigma}(E')$$

for each e-variable E' w.r. to t, which implies the first statement (about (5)) of the lemma. The second statement (6) follows from

$$\begin{aligned}
\operatorname{ep}_Q(E) &= \int \operatorname{KL}(Q_\sigma \parallel U_{t^{-1}(\sigma)})(t_* Q)(\mathrm{d}\sigma) \\
&= \int \left(\ln \left|t^{-1}(\sigma)\right| - H(Q_\sigma)\right) (t_* Q)(\mathrm{d}\sigma) \\
&= \int \ln \left|t^{-1}(\sigma)\right| (t_* Q)(\mathrm{d}\sigma) + H(t_* Q) - H(Q),
\end{aligned}$$

where KL stands for the Kullback–Leibler divergence. □

4 An explicit algorithm for Markov alternatives

Starting from this section we will consider a specific alternative hypothesis obtained by mixing Markov probability measures. The corresponding exchangeability e-variable will be computable in linear time, $O(N)$.

First let us fix some terminology. The *exchangeability summary*, or *exchangeability type*, of a data sequence $z_1, \dots, z_N$ is the numbers (N_0, N_1) of 0s and 1s in it. (It carries the same information as just the number of 1s, but we prefer a symmetric definition despite some redundancy.) By a "substring" we always mean a contiguous substring. The *Markov type* of $z_1, \dots, z_N$ is the sextuple

$(F, N_{00}, N_{01}, N_{10}, N_{11}, L)$, where $N_{i,j}$ is the number of times (i,j) occurs as substring in the sequence $z_1, \dots, z_N$ (with the comma often omitted), and F and L are the first and last bits of the sequence.

As our alternative hypothesis, we will take the uniform mixture of the Markov probability measures, defined as follows: π_{01} and π_{10} are generated independently from the uniform distribution $U_{[0,1]}$ on $[0,1]$; the first bit is chosen as 1 with probability 1/2, and after that each 0 is followed by 1 with probability π_{01}, and each 1 is followed by 0 with probability π_{10}. Let us compute the probability of a sequence of a Markov type $(F, N_{00}, \dots, N_{11}, L)$ under this probability measure:

$$\begin{aligned}
&\frac{1}{2}\int (1-\pi_{01})^{N_{00}} \pi_{01}^{N_{01}} \pi_{10}^{N_{10}} (1-\pi_{10})^{N_{11}} \,\mathrm{d}\pi_{01}\mathrm{d}\pi_{10} \\
&\quad = \frac{1}{2}\mathrm{B}(N_{00}+1, N_{01}+1)\mathrm{B}(N_{10}+1, N_{11}+1) \\
&\quad = \frac{1}{2}\frac{\Gamma(N_{00}+1)\Gamma(N_{01}+1)\Gamma(N_{10}+1)\Gamma(N_{11}+1)}{\Gamma(N_{0*}+2)\Gamma(N_{1*}+2)} \\
&\quad = \frac{1}{2}\frac{N_{00}!N_{01}!N_{10}!N_{11}!}{(N_{0*}+1)!(N_{1*}+1)!},
\end{aligned} \tag{7}$$

where $N_{i*} := N_{i,0} + N_{i,1}$. If $N_{1-F} = 0$, this probability is $\frac{1}{2N}$ (which in fact agrees with the general expression (7)). We will refer to (7) as the *UMM probability measure*, or *UMM alternative*, where "UMM" stands for "uniformly mixed Markov". The uniform prior in (7) is used for mathematical convenience and computational efficiency, and it is discussed in greater detail at the end of Appendix B.

For future use, set $\pi_{00} := 1 - \pi_{01}$ and $\pi_{11} := 1 - \pi_{10}$.

Following [32, Chap. 9], which in turn follows [20], let us define the *lower benchmark*

$$\mathrm{LB} := \frac{1}{2}\frac{N_{00}!N_{01}!N_{10}!N_{11}!}{(N_{0*}+1)!(N_{1*}+1)!(N_0/N)^{N_0}(N_1/N)^{N_1}} \tag{8}$$

as the ratio of the UMM alternative (7) to the maximum likelihood under the *IID model* (which consists of the *IID probability measures* B^N, B being a probability measure on $\{0,1\}$). The idea behind the lower benchmark is that, for any IID probability measure B^N, it is an e-variable w.r. to B^N, i.e., satisfies $\int \mathrm{LB}\,\mathrm{d}B^N \le 1$.

However, the IID model is not our null hypothesis, and our null hypothesis of exchangeability is slightly more challenging. Replacing in (8) the maximum likelihood over the IID model by the maximum likelihood over the exchangeable probability measures, we obtain the *exchangeability lower benchmark*

$$\mathrm{ELB} := \frac{1}{2}\binom{N}{N_1}\frac{N_{00}!N_{01}!N_{10}!N_{11}!}{(N_{0*}+1)!(N_{1*}+1)!}. \tag{9}$$

The exchangeability lower benchmark (9) is a bona fide exchangeability e-variable. However, our main object of interest in this paper is the more efficient

(in the sense of its e-power) e-variable given by Lemma 3.2 with t being the exchangeability model and Q being the UMM alternative (7). We will refer to this optimal e-variable as the *uniformly mixed Markov* (*UMM*) e-variable. A more explicit expression for it and a way of computing it are given below as (14) and Algorithm 1, respectively.

Remark 4.1. In the spirit of [10, Theorem 2], the value of the UMM e-variable on a data sequence $z_1, \dots, z_N$ can be written as

$$\frac{Q(\{(z_1, \dots, z_N)\})}{\frac{1}{N!}\sum_\sigma Q(\{(z_{\sigma(1)}, \dots, z_{\sigma(N)})\})}, \tag{10}$$

where Q is given by (7) and σ ranges over the permutations of $\{1, \dots, N\}$. Indeed, the denominator of (10) equals the average of $Q(\{\omega\})$ over $\omega \in t^{-1}(t(z_1, \dots, z_N))$, and so the whole expression (10) equals (5) for $\omega = (z_1, \dots, z_N)$ (and t the exchangeability model).

In fact the UMM e-variable dominates the exchangeability lower benchmark. Indeed, the exchangeability lower benchmark replaces the right-hand side of (5) by $\left|t^{-1}(t(\omega))\right| Q(\{\omega\})$, and so ignores the denominator in (2). Namely, we have

$$\mathrm{UMM}(\omega) = \frac{\mathrm{ELB}(\omega)}{Q(t^{-1}(t(\omega)))}.$$

For the e-power of the exchangeability lower benchmark we have the formula (6) with the second term $H(t_* Q)$ omitted. Indeed, according to the proof of Lemma 3.2, that term corresponds to the denominator in (2), which the lower benchmark ignores.

The UMM e-variable and the lower benchmark are not comparable. On the one hand, the lower benchmark is not an exchangeability e-variable in general; it is only an e-variable w.r. to the narrower IID model. This tends to make the lower benchmark larger. On the other hand, the lower benchmark is not admissible under any IID probability measure B^N, in the sense of $\int \mathrm{LB}\, \mathrm{d}B^N < 1$, while the UMM e-variable is admissible under any exchangeable probability measure Q, meaning $\int \mathrm{UMM}\, \mathrm{d}Q = 1$. This tends to make the UMM e-variable larger.

Remark 4.2. Notice that the difference between the assumptions of IID and exchangeability, while non-existent in the case of infinite data sequences (by de Finetti's theorem, every exchangeable probability measure on $\{0,1\}^\infty$ is a mixture of IID probability measures), becomes important for finite data sequences. The difference is quantified in [29].

In the rest of this section we will see how to compute efficiently the UMM e-variable, i.e., the likelihood ratio of the UMM alternative Markov kernel (2) to the null Markov kernel. In our derivation we will use the terminology of [31, Section 8.6] (such as "Markov graph") and consider an arbitrary finite observation space $\mathbf{Z}$ (instead of $\{0,1\}$, as in the rest of this paper); to avoid trivialities, let us assume $|\mathbf{Z}| > 1$. We will also use the following facts [31, Lemmas 8.5 and 8.6], which are versions of standard results in graph theory (the BEST theorem and the Matrix-Tree theorem).

Lemma 4.3. *In any Markov graph σ with the set of vertices V the number of Eulerian paths from the source to the sink equals*

$$T(\sigma)\frac{\mathrm{out}(\mathrm{sink})\prod_{v\in V}(\mathrm{out}(v)-1)!}{\prod_{u,v\in V}N_{u,v}!}, \tag{11}$$

where $T(\sigma)$ is the number of spanning out-trees in the underlying digraph rooted at the source, $N_{u,v}$ is the number of darts leading from u to v, and $\mathrm{out}(\cdot)$ *is the number of darts leaving a given vertex.*

Proof. According to Theorem VI.28 in [26] (and using the terminology of [26, Chap. VI]), the number of Eulerian tours in the underlying digraph is

$$T(\sigma)\prod_{v\in V}(\mathrm{out}(v)-1)!.$$

If the source and sink coincide, the number of Eulerian paths is obtained by multiplying this expression by out(source). Finally, we erase the identities of different darts going from u to v for each pair of vertices (u,v) by dividing by $N_{u,v}!$; the resulting expression agrees with (11).

Now suppose the source and sink are different vertices. Create a new digraph by adding another dart leading from the sink to the source. The number of Eulerian paths from the source to the sink in the old digraph will be equal to the number of Eulerian tours in the new graph, i.e.,

$$T(\sigma)\,\mathrm{out}(\mathrm{sink})\prod_{v\in V}(\mathrm{out}(v)-1)!,$$

where out refers to the old digraph. It remains to erase the identities of different darts going from u to v for each pair of vertices (u,v) in the old digraph; the resulting expression again agrees with (11).

Alternatively, we can combine the two cases by always adding another dart leading from the sink to the source. □

Lemma 4.4. *To find the number $T(\sigma)$ of spanning out-trees rooted at the source in the underlying digraph of a Markov graph σ with vertices $z_1,\dots,z_n$ (z_1 being the source),*

- *create the $n\times n$ matrix with the elements $a_{i,j}=-N_{z_i,z_j}$;*
- *change the diagonal elements so that each column sums to 0;*
- *compute the co-factor of $a_{1,1}$.*

Proof. This lemma can be derived from Theorem VI.28 in [26]. In that theorem we obtain $T(\sigma)$ by computing the co-factor of any diagonal element $a_{i,i}$, but that theorem is about Eulerian digraphs. We can make the underlying digraph of our Markov graph Eulerian by connecting the sink to the source. This operation does not affect the number of out-trees rooted at the source and does not change the co-factor of $a_{1,1}$. □

Let us specialize Lemmas 4.3 and 4.4 to the binary case $\mathbf{Z} := \{0, 1\}$.

Corollary 4.5. *Let σ be a Markov graph with vertices in $\{0, 1\}$ and with $F \in \{0, 1\}$ as its source. The number of Eulerian paths from the source to the sink equals*

$$N(\sigma) := \begin{cases} N_{F,1-F} \frac{(N_0-1)!(N_1-1)!}{N_{00}!N_{01}!N_{10}!N_{11}!} & \text{if } N_0 \wedge N_1 > 0 \\ 1 & \text{otherwise,} \end{cases} \tag{12}$$

where $N_i := \mathrm{in}(i) + 1_{\{F=i\}}$ ($\mathrm{in}(i)$ being the number of darts entering i, so that N_i is the number of i on any Eulerian path) and $N_{i,j}$ (with the comma often omitted) is the number of darts leading from i to j.

Proof. The case $N_0 \wedge N_1 = 0$ is obvious, so we will assume $N_0 \wedge N_1 > 0$. The number of spanning out-trees rooted at the source in the underlying digraph is

$$T(\sigma) = N_{F,1-F};$$

this follows from Lemma 4.4 and is obvious anyway. It remains to plug this in into Lemma 4.3: if the source F and sink L coincide, $F = L$, we obtain

$$N_{F,1-F} \frac{(N_F - 1)(N_F - 2)!(N_{1-F} - 1)!}{N_{00}!N_{01}!N_{10}!N_{11}!}$$

for the number of Eulerian paths from the source to the sink, and if $F \neq L$, we obtain

$$N_{F,1-F} \frac{(N_L - 1)(N_F - 1)!(N_L - 2)!}{N_{00}!N_{01}!N_{10}!N_{11}!};$$

both expression agree with (12). □

Combining (7) and (12), we obtain the total alternative weight (i.e., probability under the alternative hypothesis) of

$$W(\sigma) := \begin{cases} \frac{1}{2} N_{F,1-F} \frac{(N_0-1)!(N_1-1)!}{(N_{0*}+1)!(N_{1*}+1)!} & \text{if } N_{1-F} > 0 \\ \frac{1}{2N} & \text{otherwise} \end{cases} \tag{13}$$

for all data sequences of a given Markov type σ.

Under the null hypothesis the probability of a data sequence of exchangeability type (N_0, N_1) is

$$1/\binom{N}{N_1},$$

and so the likelihood ratio (the alternative over the ECM as the null hypothesis) is

$$\begin{aligned} &\frac{1}{2} \frac{N_{00}!N_{01}!N_{10}!N_{11}!\binom{N}{N_1}}{(N_{0*}+1)!(N_{1*}+1)! \sum_\sigma W(\sigma)} \\ &\qquad = \frac{N_{00}!N_{01}!N_{10}!N_{11}!\binom{N}{N_1}}{(N_{0*}+1)!(N_{1*}+1)! \sum_\sigma n_{f,1-f} \frac{(N_0-1)!(N_1-1)!}{(n_{0*}+1)!(n_{1*}+1)!}} \end{aligned} \tag{14}$$

(see (7) and (13)), where the σ in $\sum_\sigma$ ranges over the Markov types $(f, n_{00}, \dots, n_{11}, l)$ compatible with the exchangeability type (N_0, N_1). The equality in (14) holds when $N_0 \wedge N_1 > 0$; in the case $N_0 \wedge N_1 = 0$ the likelihood ratio is 1 (and we will treat this case separately in Algorithm 1).

The expression (14) (interpreted as 1 when $N_0 \wedge N_1 = 0$) is our main object of interest in this paper; remember that we refer to it as the *UMM e-variable*.

It remains to explain how to compute the second sum $\sum_\sigma$ in (14) (which is twice as large as $\sum_\sigma W(\sigma)$; in particular, it sums to 2 over all exchangeability types). Assume $N_0 \wedge N_1 > 0$ and remember that $N \geq 2$. For $\sigma = (f, n_{00}, \dots, n_{11}, l)$ with $f = l = 0$ (which is only possible when $N_0 \geq 2$), each such addend in the sum is

$$n_{f,1-f} \frac{(N_0-1)!(N_1-1)!}{(n_{0*}+1)!(n_{1*}+1)!} = n_{01} \frac{(N_0-1)!(N_1-1)!}{N_0!(N_1+1)!} = \frac{n_{01}}{N_0 N_1 (N_1+1)}.$$

A specific Markov type $(f, n_{00}, \dots, n_{11}, l)$ is determined (once we know that $f = l = 0$) by n_{01}, and its other components can be found from the equalities

$$\begin{cases} n_{01} = n_{10} \\ N_0 = n_{00} + n_{01} + 1 \\ N_1 = n_{01} + n_{11}. \end{cases}$$

The valid values for n_{01} are between 1 and $(N_0 - 1) \wedge N_1$, and so the part of the sum $\sum_\sigma$ corresponding to such σ is

$$\sum_{n_{01}=1}^{(N_0-1)\wedge N_1} \frac{n_{01}}{N_0 N_1 (N_1+1)} = \frac{((N_0-1)\wedge N_1)((N_0-1)\wedge N_1 + 1)}{2 N_0 N_1 (N_1+1)}. \tag{15}$$

Both sides are well defined since $N_0 \geq 2$.

For σ with $f = 0$ and $l = 1$, the part of the sum $\sum_\sigma$ corresponding to such σ is

$$\sum_{n_{01}=1}^{N_0\wedge N_1} \frac{n_{01}}{N_0 (N_0+1) N_1} = \frac{(N_0 \wedge N_1)(N_0 \wedge N_1 + 1)}{2 N_0 (N_0+1) N_1}. \tag{16}$$

For σ with $f = 1$ and $l = 0$, the part of the sum $\sum_\sigma$ corresponding to such σ is

$$\sum_{n_{10}=1}^{N_0\wedge N_1} \frac{n_{10}}{N_0 N_1 (N_1+1)} = \frac{(N_0 \wedge N_1)(N_0 \wedge N_1 + 1)}{2 N_0 N_1 (N_1+1)}. \tag{17}$$

Finally, for σ with $f = l = 1$, the part of the sum $\sum_\sigma$ corresponding to such σ is

$$\sum_{n_{10}=1}^{N_0\wedge(N_1-1)} \frac{n_{10}}{N_0 (N_0+1) N_1} = \frac{(N_0 \wedge (N_1-1))(N_0 \wedge (N_1-1) + 1)}{2 N_0 (N_0+1) N_1}. \tag{18}$$

Both sides of (18) are well defined since $N_1 \geq 2$.

Algorithm 1 Computing the UMM exchangeability e-variable

Input: $(z_1, \dots, z_N) \in \{0,1\}^N$.
Output: the value of the UMM e-variable $\mathrm{UMM}(z_1, \dots, z_N)$.
1: Set N_0 and N_1 to the numbers of 0s and 1s in $(z_1, \dots, z_N)$, respectively.
2: **if** $N_0 \wedge N_1 = 0$: **return** 1.
3: **for** $i, j \in \{0,1\}$:
4: Set $N_{i,j}$ to the number of substrings (i,j) in $(z_1, \dots, z_N)$.
5: ELB $:= \frac{1}{2} \frac{N_{00}!N_{01}!N_{10}!N_{11}!\binom{N}{N_1}}{(N_{0*}+1)!(N_{1*}+1)!}$.
6: **if** $N_0 = N_1$:
7: Sum $:= \frac{2}{N_0+1}$
8: **else**
9: Sum $:= \frac{N_0+N_1+1}{(N_0 \vee N_1)(N_0 \vee N_1+1)}$.
10: **return** ELB/Sum.

We can simplify the sum of (15), (16), (17), and (18) as follows. If $N_0 < N_1$, the sum simplifies to

$$\frac{N_0 + N_1 + 1}{N_1(N_1+1)},$$

and if $N_0 = N_1$, the sum simplifies to $2/(N_0+1)$. (There is no need to consider the case $N_1 < N_0$ because of the symmetry between N_0 and N_1.) Therefore, the sum over σ on the right-hand side of (14) is

$$2\sum_\sigma W(\sigma) = \begin{cases} \frac{N_0+N_1+1}{(N_0 \vee N_1)(N_0 \vee N_1+1)} & \text{if } N_0 \neq N_1 \\ \frac{2}{N_0+1} & \text{otherwise.} \end{cases} \tag{19}$$

The overall algorithm is presented as Algorithm 1. The value of the uniformly mixed Markov e-variable UMM is computed according to (14), and the value ELB of the exchangeability lower benchmark in line 5 is just (14) with the sum over the Markov types σ omitted. The variable Sum is set in lines 6–9 to $\sum_\sigma W(\sigma)$ and computed according to (19). The output is returned by the **return** command, and the algorithm stops as soon as the first such command is issued.

The computational complexity of Algorithm 1 is clearly optimal (to within a constant factor) both time-wise and memory-wise. Namely, the algorithm requires $O(N)$ steps and $O(1)$ memory.

5 Maximum e-power of the UMM alternative

In this section we will compute the asymptotic efficiency of the UMM e-variable under the UMM alternative. (In the next section, however, we will see the weakness of our notion of efficiency: it has a long-run frequency interpretation, but the logarithm of the UMM e-variable can be extremely volatile, and so its

mathematical expectation can be very different from what we actually expect to observe.)

Proposition 5.1. *Under the UMM alternative Q, the asymptotic e-power of the UMM e-variable* UMM *(for time horizon N) satisfies*

$$\lim_{N\to\infty} \mathrm{mep}(Q)/N = \lim_{N\to\infty} \mathrm{ep}_Q(\mathrm{UMM})/N = \frac{8}{3}\ln 2 + \frac{2}{3}\ln^2 2 - \frac{7}{36}\pi^2 - \frac{1}{6} \approx 0.083.$$

The same expression gives the asymptotic e-power of the exchangeability lower benchmark (and of the lower benchmark).

Proof. Let us compute separately the three components after the "=" in (6), starting from the last one.

When estimating $-H(Q)$, we need to estimate the frequencies N_{00}, N_{01}, N_{10}, N_{11} for a Markov chain with transition probabilities $\pi_{i,j}$. To this end, we define a new Markov chain whose states are the pairs $z_i z_{i+1}$, $i = 1, \dots, N-1$, of adjacent states of the old Markov chain with the matrix of transition probabilities

$$P := \begin{pmatrix} \pi_{00} & \pi_{01} & 0 & 0 \\ 0 & 0 & \pi_{10} & \pi_{11} \\ \pi_{00} & \pi_{01} & 0 & 0 \\ 0 & 0 & \pi_{10} & \pi_{11} \end{pmatrix};$$

the rows and columns of this matrix are labelled by the states 00, 01, 10, and 11 of the new Markov chain, in this order. The stationary probabilities for this 4×4 matrix are

$$\left(\frac{\pi_{00}\pi_{10}}{\pi_{01}+\pi_{10}}, \frac{\pi_{01}\pi_{10}}{\pi_{01}+\pi_{10}}, \frac{\pi_{01}\pi_{10}}{\pi_{01}+\pi_{10}}, \frac{\pi_{01}\pi_{11}}{\pi_{01}+\pi_{10}} \right).$$

Now, assuming that the observations are generated from a Markov chain with transition probabilities $\pi_{i,j}$, we obtain (cf. (7))

$$\begin{aligned}
&\mathbb{E} \ln \left(\frac{1}{2} \frac{N_{00}!N_{01}!N_{10}!N_{11}!}{(N_{0*}+1)!(N_{1*}+1)!} \right) \\
&\quad = \mathbb{E}\big(N_{00}\ln N_{00} - N_{00} + N_{01}\ln N_{01} - N_{01} \\
&\qquad + N_{10}\ln N_{10} - N_{10} + N_{11}\ln N_{11} - N_{11} \\
&\qquad - (N_{00}+N_{01}+1)\ln(N_{00}+N_{01}+1) + (N_{00}+N_{01}+1) \\
&\qquad - (N_{10}+N_{11}+1)\ln(N_{10}+N_{11}+1) + (N_{10}+N_{11}+1)\big) + O(N^{1/2}) \\
&\quad = \mathbb{E}\bigg(N_{00}\ln\frac{N_{00}}{N_{00}+N_{01}} + N_{01}\ln\frac{N_{01}}{N_{00}+N_{01}} \\
&\qquad + N_{10}\ln\frac{N_{10}}{N_{10}+N_{11}} + N_{11}\ln\frac{N_{11}}{N_{10}+N_{11}} \bigg) + O(N^{1/2}) \\
&\quad = N\frac{\pi_{00}\pi_{10}}{\pi_{01}+\pi_{10}}\ln\pi_{00} + N\frac{\pi_{01}\pi_{10}}{\pi_{01}+\pi_{10}}\ln\pi_{01} \\
&\qquad + N\frac{\pi_{01}\pi_{10}}{\pi_{01}+\pi_{10}}\ln\pi_{10} + N\frac{\pi_{01}\pi_{11}}{\pi_{01}+\pi_{10}}\ln\pi_{11} + O(N^{1/2})
\end{aligned}$$

(we are ignoring special cases such as $N_{00} = 0$, which should be considered separately). To find the expectation under the Bayes mixture of the Markov model with the uniform prior on (π_{01}, π_{10}), we integrate

$$\begin{aligned}\int_0^1 \int_0^1 \Bigg(& \frac{\pi_{00}\pi_{10}}{\pi_{01}+\pi_{10}} \ln \pi_{00} + \frac{\pi_{01}\pi_{10}}{\pi_{01}+\pi_{10}} \ln \pi_{01} \\ & + \frac{\pi_{01}\pi_{10}}{\pi_{01}+\pi_{10}} \ln \pi_{10} + \frac{\pi_{01}\pi_{11}}{\pi_{01}+\pi_{10}} \ln \pi_{11} \Bigg) \, \mathrm{d}\pi_{01} \, \mathrm{d}\pi_{10} \\ & = \frac{2}{3} \ln 2 + \frac{2}{3} \ln^2 2 - \frac{1}{9}\pi^2 - \frac{1}{6} \approx -0.481. \end{aligned} \tag{20}$$

Now let us estimate the first term

$$\int \ln \left| t^{-1}(\sigma) \right| (t_* Q)(\mathrm{d}\sigma)$$

after the "=" in (6). Set $K := \sigma$ (this is the number of 1s), and suppose the observations are generated from a Markov chain with given transition probabilities π_{01} and π_{10}. We then have

$$\begin{aligned} \mathbb{E}\left(\ln \binom{N}{K} \right) &= \mathbb{E}\left(\ln \frac{N!}{K!(N-K)!} \right) = \mathbb{E}\left(\ln \frac{(N/e)^N}{\left(\frac{K}{e}\right)^K \left(\frac{N-K}{e}\right)^{N-K}} \right) + O(N^{1/2}) \\ &= \mathbb{E}\left(-K \ln \frac{K}{N} - (N-K) \ln \left(1 - \frac{K}{N}\right) \right) + O(N^{1/2}) \\ &= -N\pi_1 \ln \pi_1 - N\pi_0 \ln \pi_0 + O(N^{1/2}), \end{aligned}$$

where π_0 and π_1 are the stationary probabilities

$$\pi_0 := \frac{\pi_{10}}{\pi_{01}+\pi_{10}} \text{ and } \pi_1 := \frac{\pi_{01}}{\pi_{01}+\pi_{10}}$$

of the Markov chain. It remains to take the integral

$$\begin{aligned} -\int_0^1 \int_0^1 (\pi_0 \ln \pi_0 + \pi_1 \ln \pi_1) \, \mathrm{d}\pi_{01} \, \mathrm{d}\pi_{10} &= -2 \int_0^1 \int_0^1 (\pi_0 \ln \pi_0) \, \mathrm{d}\pi_{01} \, \mathrm{d}\pi_{10} \\ &= -2 \int_0^1 \int_0^1 \left(\frac{\pi_{10}}{\pi_{01}+\pi_{10}} \ln \frac{\pi_{10}}{\pi_{01}+\pi_{10}} \right) \mathrm{d}\pi_{01} \, \mathrm{d}\pi_{10} \\ &= 2 \ln 2 - \frac{1}{12}\pi^2 \approx 0.564. \end{aligned} \tag{21}$$

The final term $H(t_* Q)$ in (6) can be ignored. Indeed, using the last expression in (7), we can bound the probability $(t_* Q)(\{K\})$, for any $K \in \{1, \dots, N-1\}$, by 1 from above and by $1/(2N^3)$ from below:

$$(t_* Q)(\{K\}) \geq \frac{1}{2} \frac{(N-K-1)!0!1!(K-1)!}{(N-K)!(K+1)!} = \frac{1}{2(N-K)K(K+1)} \geq \frac{1}{2N^3} \tag{22}$$

(the expression after the first "$\geq$" being the probability of the sequence consisting of K 1s followed by $N-K$ 0s). Therefore, $H(t_*Q) = O(\ln N)$. (As always, the extreme cases $K \in \{0, N\}$ should be considered separately.)

Combining (20) and (21), we obtain the coefficient

$$\frac{8}{3}\ln 2 + \frac{2}{3}\ln^2 2 - \frac{7}{36}\pi^2 - \frac{1}{6} \approx 0.083 \tag{23}$$

in front of N in the asymptotic expression for $\mathrm{ep}_Q(\mathrm{UMM})$.

The proof shows that the asymptotic e-power is the same for the exchangeability lower benchmark, and a simple calculation using Stirling's formula (see, e.g., [32, Proposition 9.2]) shows that we also have the same asymptotic e-power for the lower benchmark. □

Proposition 5.1 states that the e-powers of the UMM e-variable and of the exchangeability lower benchmark are close asymptotically, and its proof gives a crude argument that is still sufficient to demonstrate this. The following corollary of the previous section's results establishes much more precise relations between the UMM e-variable and the exchangeability lower benchmark.

Corollary 5.2. *It is always true that*

$$1 \leq \frac{\mathrm{UMM}}{\mathrm{ELB}} \leq 2N. \tag{24}$$

Moreover,

$$\frac{\mathrm{UMM}}{\mathrm{ELB}} = \begin{cases} \frac{2(N_0 \vee N_1)(N_0 \vee N_1 + 1)}{N_0 + N_1 + 1} & \text{if } N_0 \neq N_1 \\ N_0 + 1 & \text{otherwise.} \end{cases} \tag{25}$$

Proof. In the case $N_0 \wedge N_1 > 0$, the relation (25) follows from (19). If $N_0 = 0$ or $N_1 = 0$, the expression on the right-hand side of (25) becomes $2N$, which agrees with the last expression (which simplifies to $1/(2N)$) on the right-hand side of the chain (7).

For a fixed sum $N_0 + N_1$, the maximum of the right-hand side of (25) is attained for $N_0 = 0$ or $N_1 = 0$, and the maximum is $2N$. This proves (24). □

6 Computational experiments

In this section we will conduct three groups of experiments involving the two lower benchmarks and the UMM exchangeability e-variable. The first group is the main one, and in it the true data distribution is a specific Markov probability measure with the initial probability of 1 equal to 1/2. In this case, we define another benchmark (as in [32, Sect. 9.2.5]), the *upper benchmark*, as

$$\mathrm{UB} := \frac{1}{2}\frac{N_{00}!N_{01}!N_{10}!N_{11}!}{(N_{0*}+1)!(N_{1*}+1)!\pi_0^{N_0}\pi_1^{N_1}} \tag{26}$$

(cf. (7)), where π_0 and π_1 are the stationary probabilities under the true data-generating distribution. We can see that the upper benchmark is an e-variable

(likelihood ratio) w.r. to a specific IID probability measure, and so it is not even an IID e-variable. Therefore, we should not be surprised if the upper benchmark exceeds a bona fide exchangeability e-variable; there are two elements of cheating in interpreting the upper benchmark as measure of evidence against the null hypothesis of exchangeability: first, it tests IID rather than exchangeability, and second, it tests only one individual IID measure.

Our results for specific Markov alternatives are given in Fig. 1. This figure contains boxplots for $K := 10^5$ simulations of four values: the exchangeability lower benchmark ELB (given by (9)), the lower benchmark LB (given by (8)), the upper benchmark UB (given by (26)), and the UMM exchangeability e-variable UMM (given by Algorithm 1). Only two of these, ELB and UMM, are bona fide exchangeability e-variables. The time horizon N and the transition probabilities for the two panels are given in the caption.

In both panel of Fig. 1 we consider symmetric Markov chains, $\pi_{01} = \pi_{10}$, as alternatives to exchangeability. The observations are generated from those alternative probability measures. In the left panel we consider an "easy" case, $\pi_{01} = 0.1$, in the sense of being easily distinguishable from the case of exchangeability, $\pi_{01} = 0.5$. The case in the right panel, $\pi_{01} = 0.4$, is closer to exchangeability and thus more difficult. To decide which e-values are most interesting in practice I used Jeffrey's [5, Appendix B] rule of thumb involving thresholds for e-values between $10^{1/2}$ and 100. In the easy case, $N = 20$ observations are sufficient for the UMM e-variable to produce typical e-values that are of the same order of magnitude as Jeffreys's thresholds. In the difficult case, we need more observations for that, and we set $N := 400$.

UMM performs better than LB in both panels and, of course, better than ELB (we know that UMM dominates ELB). ELB and LB often fail to achieve Jeffreys's low threshold of $10^{1/2}$ for substantial evidence against the null hypothesis. It is interesting that UMM is often even higher than the upper benchmark,

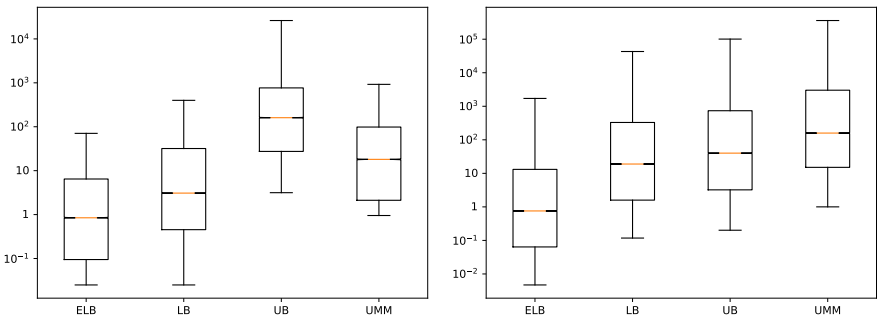

Figure 1: The four e-values and related quantities, as described in text. Left panel: $N = 20$ and $\pi_{01} = \pi_{10} = 0.1$. Right panel: $N = 400$ and $\pi_{01} = \pi_{10} = 0.4$. Only ELB and UMM are bona fide exchangeability e-values. The number of simulations is $K = 10^5$ in both panels.

N	$\pi_{01} = \pi_{10}$	$\overline{\text{ELB}}$	$\overline{\text{LB}}$	$\overline{\text{UMM}}$	UMM quantiles	$\overline{\text{UMM}} - \overline{\text{ELB}}$	upper bound
20	0.1	-0.116	0.471	1.226	$[-0.021, 0.325, 1.258, 1.993, 2.965]$	1.342	1.602
400	0.4	0.084	1.482	2.427	$[0.001, 1.179, 2.201, 3.479, 5.557]$	2.343	2.903

Table 1: Numerical values for the decimal logarithms of the two lower benchmarks and the UMM e-variable shown in Fig. 1. The bars stand for the averages of the decimal logarithms. The UMM quantiles (i.e., quantiles of the UMM e-variable) are for 5%, 25% (first quartile), 50% (median), 75% (third quartile), and 95%. The upper bound for the difference between the decimal logarithms of the UMM e-variable and the exchangeability lower benchmark is $\log_{10}(2N)$, as per (24).

as in the right panel of Fig. 1.

Table 1 gives more precise numerical values that can be read off Fig. 1 only very approximately. The bars stand for the empirical averages of the decimal logarithms of ELB, LB, and UMM over the same $K := 10^5$ simulations as in Fig. 1. The table also gives the difference between the empirical averages of the UMM and ELB and the upper bound for the difference given by (24). According to Corollary 5.2, the UMM e-value cannot differ from the exchangeability lower benchmark by much. The upper bound (24) holds and is not excessively loose.

Figure 2 describes the second group of experiments and explores the behaviour of ELB, LB, UB, and UMM under the null hypothesis (as suggested by a referee). In the left panel the probability of 1 is 0.5, and all four are valid e-variables; while UB is not valid under exchangeability in general, it is valid under this particular exchangeable probability measure. The number of observations is $N = 20$. The UMM e-variable performs best in this case. The right panel has 0.1 as the probability of 1, which makes UB (still based on

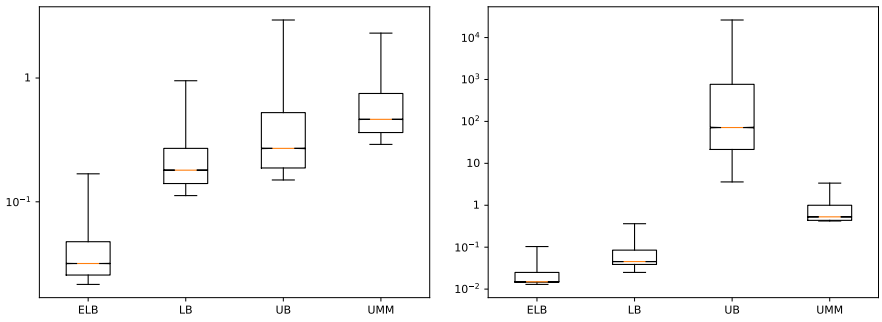

Figure 2: Two exchangeability e-values (ELB and UMM) and two approximations (LB and UB) under the null hypothesis. Left panel: the probability of 1 is 0.5. Right panel: the probability of 1 is 0.1. The number of observations is $N = 20$, and the number of simulations is $K = 10^5$.

$\pi_0 = \pi_1 = 0.5$) very invalid. Among the valid e-variables UMM still performs best.

The third group of experiments involves generating the binary observations from the UMM alternative (which is not Markov any longer). The explicit formula for this alternative is given in (7), but it is easier to generate π_{01} and π_{10} from the uniform distribution on $[0,1]^2$ and then generate the observations from the Markov chain with these parameters. This interpretation of the UMM alternative shows that our algorithm for testing exchangeability is now in a hostile environment: with a sizeable probability we will get $\pi_{01} \approx \pi_{10}$, i.e., difficult data sequences that look almost exchangeable.

Figure 3 shows results for this case; in the expression (26) for the upper benchmark, we still set $\pi_0 := \pi_1 := 0.5$. It is striking how spread out the distributions for the three benchmarks and the UMM e-variable are, demonstrating the hostile nature of the testing environment. They are also skewed, with the mean very different from the median. To obtain UMM e-values that are consistently in Jeffreys's range, now we need much larger values of N, such as 10^3, shown in the left panel of Figure 3. The lack of validity for the upper benchmark is very obvious in Figure 3: it takes much larger values, and I did not even bother to include the whole boxplots for it.

Table 2, which is analogous to Table 1, gives more precise numbers related to Fig. 3. As before, the bars stand for the empirical averages of the decimal logarithms over $K = 10^5$ replications, and N is the time horizon. Now we also have "as.", the common theoretical asymptotic value for the UMM e-variable and exchangeability lower benchmark obtained from (23) by dividing by $\ln 10$ (to convert natural logarithms to decimal ones) and multiplying by the sample size N. As expected, the approximation is least accurate for $N = 10^3$. The table also gives the average differences between the UMM e-variable and exchangeability lower benchmark on the $\log_{10}$ scale, together with the upper bound given by (24). The upper bound still holds.

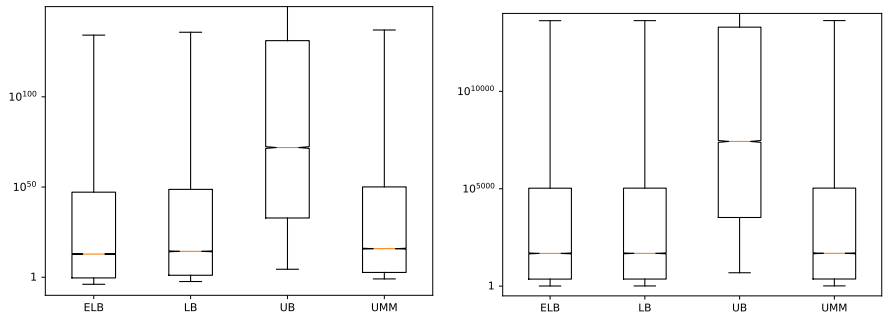

Figure 3: Two exchangeability e-values (ELB and UMM) and two approximations (LB and UB) under the UMM alternative. Left panel: $N = 10^3$. Right panel: $N = 10^5$. The number of simulations is still $K = 10^5$.

N	$\overline{\text{ELB}}$	$\overline{\text{LB}}$	$\overline{\text{UMM}}$	as.	UMM quantiles	$\overline{\text{UMM}} - \overline{\text{ELB}}$	upper bound
10^3	32.08	33.59	35.04	36.02	[−0.91, 2.67, 15.84, 50.08, 137.04]	2.965	3.301
10^4	354.9	356.9	358.8	360.2	[0.0, 34.6, 167.9, 505.6, 1379.6]	3.966	4.301
10^5	3571	3573	3575	3602	[12, 366, 1684, 5033, 13632]	4.966	5.301

Table 2: Some figures for the decimal logarithms of the two lower benchmarks and the UMM e-variable. The bars stand for the averages of the decimal logarithms, and "as." stands for the asymptotic expression, as described in text. The UMM quantiles are for 5%, 25%, 50%, 75%, and 95%. The upper bound for the difference between UMM and ELB is given by (24). The number of simulations is always $K = 10^5$.

7 Conclusion

In this paper the algorithm for computing the UMM e-variable was fully developed only in the binary case. A natural next step would be to extend it to any finite observation space $\mathbf{Z}$. (A big chunk of Sect. 4, following [31, Sect. 8.6], presented the combinatorics for an arbitrary finite observation space $\mathbf{Z}$.) It is interesting what the computational complexity of such an extension of Algorithm 1 will be in general as function of N and $|\mathbf{Z}|$.

The topic of this paper has been testing the exchangeability compression model in the batch mode using Markov alternatives. There are many other interesting null hypotheses among Kolmogorov compression models, and there are many interesting alternatives. For example, in [32, Chap. 9] we discussed, alongside Markov alternatives, detecting changepoints. Our discussion there was in the online mode, but for changepoint detection the batch mode is not less important [32, Remark 8.19]; e.g., its role has been increasing in bioinformatics (including DNA analysis). Using e-values in changepoint detection is particularly convenient when multiple hypothesis testing is involved (as it often is in batch changepoint detection). Some extensions will be discussed in Appendixes B–C, including changepoint detection in Appendix C.

Acknowledgements

Many thanks to Wouter Koolen and the reviewers of the journal version of this paper for useful comments and corrections. Research on this paper has been partially supported by Mitie.

References

[1] Eugene A. Asarin. Some properties of Kolmogorov Δ-random finite sequences. *Theory of Probability and its Applications*, 32:507–508, 1987.

[2] Eugene A. Asarin. On some properties of finite objects random in the algorithmic sense. *Soviet Mathematics Doklady*, 36:109–112, 1988.

[3] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. Technical Report arXiv:1906.07801 [math.ST], arXiv.org e-Print archive, March 2023. Journal version is to appear in the *Journal of the Royal Statistical Society B* (with discussion).

[4] Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

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

[6] John L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926, 1956.

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

[8] Andrei N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40, 1983.

[9] Andrei N. Kolmogorov and Vladimir A. Uspensky. Algorithms and randomness. *Theory of Probability and Its Applications*, 32:389–412, 1987.

[10] Nick W. Koning. Post-hoc and anytime valid inference for exchangeability and group invariance. Technical Report arXiv:2310.01153 [math.ST], arXiv.org e-Print archive, April 2024.

[11] Steffen L. Lauritzen. *Extremal Families and Systems of Sufficient Statistics*. Springer, New York, 1988.

[12] Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Springer, New York, revised first edition, 2006.

[13] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, Cham, fourth edition, 2022.

[14] Dennis V. Lindley. *Understanding Uncertainty*. Wiley, Hoboken, NJ, 2006.

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

[16] Jerzy Neyman. Frequentist probability and frequentist statistics. *Synthese*, 36:97–131, 1977.

[17] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, 231:289–337, 1933.

[18] Gleb Novikov. Relations between randomness deficiencies. Technical Report arXiv:1608.08246 [math.LO], arXiv.org e-Print archive, August 2016. Published in *Lecture Notes in Computer Science* 10307:338–350 (2017).

[19] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009.

[20] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M. Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022.

[21] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[22] Alexey Semenov, Alexander Shen, and Nikolay Vereshchagin. Kolmogorov's last discovery? (Kolmogorov and algorithmic statistics). *Theory of Probability and Its Applications*, 68:582–606, 2024.

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

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

[25] Jun'ichi Takeuchi, Tsutomu Kawabata, and Andrew R. Barron. Properties of Jeffreys mixture for Markov sources. *IEEE Transactions on Information Theory*, 59:438–457, 2013.

[26] W. T. Tutte. *Graph Theory*. Addison-Wesley, Reading, MA, 1984.

[27] Vladimir A. Uspensky and Alexei L. Semenov. *Algorithms: Main Ideas and Applications*. Kluwer, Dordrecht, 1993.

[28] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[29] Vladimir Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986. Another English translation with proofs: arXiv:1612.08859 (2016).

[30] Vladimir Vovk. Kolmogorov's complexity conception of probability. In Vincent F. Hendricks, Stig Andur Pedersen, and Klaus Frovin Jørgensen, editors, *Probability Theory: Philosophy, Recent History and Relations to Science*, pages 51–69. Kluwer, Dordrecht, 2001.

[31] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, first edition, 2005. Section 8.6 of the first edition is not part of the second edition [32].

[32] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, second edition, 2022.

[33] Vladimir Vovk and Glenn Shafer. Kolmogorov's contributions to the foundations of probability. *Problems of Information Transmission*, 39:21–31, 2003.

[34] Vladimir Vovk and Glenn Shafer. A conversation with A. Philip Dawid. *Statistical Science*, 40:148–166, 2025.

[35] Vladimir Vovk and Vladimir V. V'yugin. On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society B*, 55:253–266, 1993.

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

[37] Vladimir Vovk and Ruodu Wang. Confidence and discoveries with e-values. *Statistical Science*, 38:329–354, 2023.

[38] Vladimir V. V'yugin. Kolmogorov complexity in the USSR (1975–1982): Isolation and its end. Technical Report arXiv:1907.05056 [cs.GL], arXiv.org e-Print archive, July 2019.

[39] Abraham Wald and Jacob Wolfowitz. An exact test for randomness in the non-parametric case based on serial correlation. *Annals of Mathematical Statistics*, 14:378–388, 1943.

A Algorithmic theory of randomness

The topic of this appendix is Kolmogorov's original approach to compression modelling. While in the main part of the paper we avoided using computability theory, here it will play an important role.

Kolmogorov's complexity models were introduced, in their most complete form, in what appears to be Kolmogorov's last talk. It was given on 14 October 1982 at what later became known as the Kolmogorov seminar; see [22, note 12], containing Shen's notes taken during the talk, and [30, Sect. 4]. The Kolmogorov seminar at Moscow State University was opened by Kolmogorov on 28 October 1981, and Kolmogorov gave two talks in it, on 26 November 1981 and 14 October 1982 [22, note 12]; the two talks were conflated in my paper [30, Sect. 4].

All results listed in this appendix are either well known or immediately follow from well-known results.

Mathematical results

In this appendix, we assume a fixed sufficiently large aggregate X of constructive objects in the sense of [27, Sect. 1.0.6]. In particular, X contains the integers, the finite binary sequences, and the finite sets of those. Let us use the notation $C(x)$ for the Kolmogorov complexity of x, $C(x \mid y)$ for the conditional Kolmogorov complexity of x given y, $K(x)$ for the prefix complexity of x, and $K(x \mid y)$ for

the conditional prefix complexity of x given y. Here x and y are any constructive objects from X. See, e.g., [24, Chaps 1, 2, and 4] for definitions.

Kolmogorov's definition of randomness deficiency of an element x of a finite set $A \subset X$ is

$$d_A^C(x) := \log|A| - C(x \mid A) \tag{27}$$

[9, Sect. 2.3]. Informally, x is random in A if $d_A^C(x)$ is small. (And Kolmogorov called x Δ-random in A if $d_A^C(x) \le \Delta$.)

Martin-Löf [15] showed that Kolmogorov's definition (27) can be stated in terms of p-values. Let A be a finite non-empty subset of X; remember that U_A is the uniform probability measure on A. A function $f : A \to [0,1]$ is a *p-variable* if

$$\forall \epsilon > 0 : U_A(f \le \epsilon) \le \epsilon.$$

A family P of functions $P_A : A \to [0,1]$, A ranging over the finite non-empty subsets of X, is a *p-test* if

- the function $(A, x) \mapsto P_A(x)$ is upper semicomputable, i.e., there is an algorithm that eventually stops on input (A, x, ϵ), where ϵ is a rational number, if and only if $P_A(x) < \epsilon$, and
- for each finite non-empty $A \subset X$, P_A is a p-variable.

The values taken by p-variables are *p-values*.

Lemma A.1. *There exists a* universal *p-test $\tilde{P}$, in the sense that for any p-test P there exists a positive constant c such that $\tilde{P} \le cP$.*

The proof of Lemma A.1 is standard (cf., e.g., [24, Theorem 39]). Fix a universal p-test $\tilde{P}$. The universal p-test is unique to within a constant factor, and it is customary in the algorithmic theory of randomness to disregard such differences, which we will also do in this appendix.

Remark A.2. The usual definitions in the algorithmic theory of randomness are given in terms of $-\log P$, but for simplicity let us discard the minus logarithm, following [35].

Now we can state Martin-Löf's result expressing Kolmogorov's deficiency of randomness via the universal p-test.

Proposition A.3. *There exists a constant $c > 0$ such that, for all A and $x \in A$,*

$$\left| d_A^C(x) + \log \tilde{P}_A(x) \right| \le c. \tag{28}$$

Proof. Martin-Löf states and proves a slightly less general result in [15, Sect. II, Theorem on p. 607] (see also [15, Sect. V, Theorem on p. 616]), but his argument is general. Since, for each finite set $A \subset X$ and each $n \in \{0, 1, \dots\}$, we have

$$|\{x \in A \mid C(x \mid A) \le n\}| \le 2^{n+1},$$

we will also have

$$U_A\left(\{x \in A \mid \log|A| - C(x \mid A) \geq n\}\right) \leq 2^{-n+2},$$

which implies the part

$$d_A^C(x) + \log \tilde{P}_A(x) \leq c$$

of (28).

To prove the other part of (28), i.e.,

$$C(x \mid A) \leq \log|A| + \log \tilde{P}_A(x) + c,$$

it suffices to establish that, for some c (perhaps a different one),

$$\forall A : \left|\left\{x \in A \mid \log|A| + \log \tilde{P}_A(x) \leq n\right\}\right| \leq 2^{n+c},$$

A ranging over the finite non-empty subsects of X. The last inequality (with $c := 0$) follows immediately from the definition of a p-test. □

Prefix complexity K has important technical advantages over C (see, e.g., [24, Chap. 4]), and so a natural modification of (27) is

$$d_A^K(x) := \log|A| - K(x \mid A). \tag{29}$$

Analogously to expressing (27) in terms of p-values, we can express (29) in terms of e-values.

A function $f : A \to [0, \infty)$ on a finite non-empty set $A \subset X$ is an *e-variable* if

$$\int f \, \mathrm{d}U_A \leq 1.$$

A family E of functions $E_A : A \to [0, 1]$, A ranging over the finite non-empty subsets of X, is an *e-test* if

- the function $(A, x) \mapsto E_A(x)$ is *lower semicomputable*, i.e., there is an algorithm that eventually stops on input (A, x, t), where t is a rational number, if and only if $E_A(x) > t$, and
- for each finite non-empty $A \subset X$, E_A is an e-variable.

Lemma A.4. *There exists a* universal *e-test $\tilde{E}$, in the sense that for any e-test E there exists a positive constant c such that $\tilde{E} \geq E/c$.*

The proof of Lemma A.4 is again standard (but [24, Theorem 47] is now more relevant). Fix a universal e-test $\tilde{E}$. It is clear that the universal e-test is unique to within a constant factor.

Notice the difference between the universal tests in Lemma A.1 and Lemma A.4: whereas in the former "universal" means "smallest" (to within a constant factor), in the latter "universal" means "largest". The following result expresses the prefix version (29) of deficiency of randomness via the universal e-test.

Proposition A.5. *There exists a constant $c > 0$ such that, for all A and x,*

$$\left| d_A^K(x) - \log \tilde{E}_A(x) \right| \le c. \tag{30}$$

Proposition A.5 will follow from two other propositions (Propositions A.7 and A.8 below), which, despite their simplicity (especially Proposition A.8), are of great independent interest.

A function $f : A \to [0, 1]$ on a finite non-empty set $A \subset X$ is a *subprobability measure* (or *semimeasure* [24, Sect. 4.1]) if

$$\sum_{x \in A} f(x) \le 1.$$

A family m of functions $m_A : A \to [0, 1]$, A ranging over the finite non-empty subsets of X, is a *lower semicomputable subprobability measure* if

- the function $(A, x) \mapsto m_A(x)$ is lower semicomputable, and
- for each finite non-empty $A \subset X$, m_A is a subprobability measure.

Lemma A.6. *There exists a* universal *lower semicomputable subprobability measure $\tilde{m}$, in the sense that for any lower semicomputable subprobability measure m there exists a positive constant c such that $\tilde{m} \ge m/c$.*

For a proof of, essentially, Lemma A.6, see the proof of [24, Theorem 47]. Let us abbreviate "universal lower semicomputable subprobability measure" to *universal measure*.

Proposition A.7. *There exists a constant $c > 0$ such that, for all A and x,*

$$|K(x \mid A) + \log \tilde{m}_A(x)| \le c.$$

Proof. Follow [24, Sect. 4.5]. □

Proposition A.8. *There exists a constant $c > 0$ such that, for all A and x,*

$$\frac{1}{c} \le \frac{\tilde{m}_A(x)|A|}{\tilde{E}_A(x)} \le c. \tag{31}$$

Proof. It suffices to notice that $\tilde{m}_A(x)|A|$ is an e-test and that $\tilde{E}_A(x)/|A|$ is a lower semicomputable subprobability measure. □

The interpretation of (31) is that the universal e-test $\tilde{E}$ is a likelihood ratio: we divide the universal measure $\tilde{m}$ ("universal alternative hypothesis") by the null uniform probability measure, assigning weight $1/|A|$ to each $x \in A$.

Now we can easily prove Proposition A.5.

Proof of Proposition A.5. Combining the previous propositions, we obtain

$$\begin{aligned}\left|d_A^K(x) - \log \tilde{E}_A(x)\right| &= \left|\log|A| - K(x \mid A) - \log \tilde{E}_A(x)\right| \\ &\le |\log|A| + \log \tilde{m}_A(x) - \log(\tilde{m}_A(x)|A|)| + c = c,\end{aligned} \tag{32}$$

i.e., (30). The first equality in (32) just uses the definition of $d_A^K(x)$, and the inequality "$\le$" in (32) is obtained by applying Proposition A.7 to $K(x \mid A)$ and applying Proposition A.8 to $\tilde{E}_A(x)$. □

Both complexities C and K and randomness deficiencies d^C and d^K are close to each other.

Proposition A.9. *There is a constant $c > 0$ such that, for all finite non-empty $A \subset X$ and all $x \in A$,*

$$C(x \mid A) - c \le K(x \mid A) \le C(x \mid A) + 2\log C(x \mid A) + c \tag{33}$$

and

$$d_A^K(x) - c \le d_A^C(x) \le d_A^K(x) + 2\log d_A^K(x) + c. \tag{34}$$

Proof. See [24, Theorem 65] for inequalities stronger than (33). For (34), follow the proof of [18, Proposition 1]. □

Discussion

Kolmogorov's original definition of randomness deficiency of an element of a finite set is (27). It can be interpreted as the universal p-value on the logarithmic scale (Proposition A.3). A natural modification of Kolmogorov's definition is (29), given in terms of prefix complexity, and it can be interpreted as the universal e-value on the logarithmic scale (Proposition A.5).

The simplest context in which these definitions can be used is that of *complexity models*, in the terminology of [30, 33]. A complexity model is a computable partition of the sample space, and the implicit statement about the observed data sequence x is that it is random in the sense of (27) (or (29), which is close to (27) by Proposition A.9) in the block $A \ni x$ of the partition. Let me give several examples of such models, those that are most relevant in the context of this paper. The sample space in all these examples will be $\{0,1\}^*$.

- The main complexity model of interest to Kolmogorov [7, 8] was that of *exchangeability*, where the binary sequences $\{0,1\}^*$ are divided into the blocks of sequences of the same length and with the same number of 1s. Stripping this complexity model of the algorithmic theory of randomness, we obtain the exchangeability compression model introduced in the main part of the paper (Sect. 2).
- Another complexity model [8] is the Markov model, in which the blocks consist of the binary sequences with the identical first element and the same number of substrings 00, 01, 10, and 11. In the terminology of [32, Sect. 11.3.4], the exchangeability model is more specific than the Markov model.

- A further generalization of the exchangeability complexity model is the second order Markov model (suggested in Kolmogorov's 1982 seminar talk [30]), in which the blocks consist of the binary sequences with the identical first and second elements and the same number of substrings 000, 001, 010, 011, 100, 101, 110, and 111.
- A model not considered by Kolmogorov is the *changepoint model* (exchangeability with a changepoint), in which the blocks are indexed by (N, τ, K_0, K_1), where $N \in \{2, 3, \dots\}$ (the time horizon), $\tau \in \{1, \dots, N-1\}$ (the changepoint), $K_0 \in \{0, \dots, \tau\}$, and $K_1 \in \{0, \dots, N-\tau\}$, and the block (N, τ, K_0, K_1) consists of all binary sequences of length N with K_0 1s among their first τ elements and K_1 1s among their last $N-\tau$ elements.

Other complexity models introduced by Kolmogorov were the Gaussian and Poisson models (in his 1982 seminar talk [22, note 12]; see also [1, 2] and [30, Sect. 4]). A complexity model formalizing the property of being IID rather than exchangeability was introduced in work [29] done under Kolmogorov's supervision.

Stochastic sequences

Kolmogorov's 1981 seminar talk was devoted to what he called stochastic sequences, which can be interpreted as an overarching structure over complexity models. Let us say that a binary data sequence $x \in X$ is (α, β)-*stochastic* if there is a finite set $A \subset X$ containing x such that $C(A) \le \alpha$ and $d_A^C(x) \le \beta$. And let us say that $x \in X$ is Δ-*random* w.r. to a complexity model if $d_A^C(x) \le \Delta$, where A is the block of the complexity model containing x. Data sequences that are modelled using complexity models are stochastic; e.g., for some constant c:

- if a data sequence of length N is Δ-exchangeable (i.e., Δ-random w.r. to the exchangeability model), it is $(2\log N + c, \Delta + c)$-stochastic; ;
- if a data sequence of length N is Δ-Markov (i.e., Δ-random w.r. to the Markov model), it is $(4\log N + c, \Delta + c)$-stochastic;
- if a data sequence of length N is Δ-Markov of second order, it is $(8\log N + c, \Delta + c)$-stochastic;
- if a data sequence of length N is Δ-random w.r. to the IID model introduced in [29], it is $(\frac{3}{2}\log N + c, \Delta + c)$-stochastic;
- if a data sequence of length N is Δ-exchangeable with one change point (i.e., Δ-random w.r. to the changepoint model), it is $(4\log N + c, \Delta + c)$-stochastic.

B Quasi-universal e-variables

In this paper we are interested, at least implicitly, in the universal e-test $\tilde{E}$ introduced in Lemma A.4. It is a fundamental object in that its components $\tilde{E}_A$

are the largest e-variables; in this sense they are the most powerful e-variables. By Proposition A.8, $\tilde{E}_A$ is the likelihood ratio of the universal measure to the null hypothesis U_A. In the main part of the paper we discussed a specific alternative hypothesis (namely, UMM), and the universal measure can be regarded as the universal alternative.

The way the universal measure is constructed in the algorithmic theory of randomness is by averaging over all subprobability measures that are computable in a generalized sense (see, e.g., [24, Theorem 47], the alternative proof).

The algorithmic theory of randomness, however, provides only an ideal picture. It can serve as a model for more practical approaches, but it is not practical itself. The two most conspicuous reasons are that:

- the basic quantities used in the algorithmic theory of randomness, such as complexity or randomness deficiency, are not computable (they are only computable in a generalized sense, let alone efficiently computable); in particular, the universal alternative is not computable;
- these basic quantities are only defined to within a constant (additive or multiplicative).

What we did in the main part of this paper can, however, be regarded as a computable approximation to the ideal picture. The idea (which is an old one; see the references below) is to replace the universal alternative by a Bayesian average of a statistical model that is significantly richer than the null hypothesis. In particular, the UMM exchangeability e-variable discussed in the main part of this paper can be regarded as a practical approximation to the universal e-test $\tilde{E}$.

The justification that we had for the UMM e-variable is less convincing than the justification for its ideal counterpart $\tilde{E}$: it is the frequentist one given by Lemma 3.1 and assuming that the observed data sequence is generated by the UMM alternative. Its advantage, however, is that this justification does not involve an arbitrary constant factor.

It would be more in the spirit of the algorithmic theory of randomness to use a different principle for choosing the alternative hypothesis: instead of choosing an alternative probability measure likely to generate the data, we could choose an alternative probability measure likely to lead to a high likelihood ratio of the alternative to the null.

The general scheme of testing exemplified by this paper is that we test a Kolmogorov compression model as null hypothesis, and have a batch compression model with a more detailed summarising statistic as alternative. This paper has the exchangeability compression model as the null and a mixture of the first-order Markov model as the alternative. We can imagine lots of other testing problems of this kind:

- The exchangeability model as the null, and the uniform mixture of the second-order Markov model as the alternative.

- The exchangeability model as the null, and a mixture of the uniform mixtures of the kth order Markov models as the alternative; the weights w_k for those should sum to 1, $\sum_k w_k = 1$, and tend to 0 as slowly as possible as $k \to \infty$ (see below).
- The first-order Markov model as the null, and a mixture of the second-order Markov model as the alternative.
- The exchangeability model as the null and the changepoint model as alternative.
- A changepoint at a postulated time τ as the null, and a mixture of probability measures corresponding to a changepoint at a different time as alternative. (In order to obtain confidence regions for the changepoint.)

We can call them instances of quasi-universal testing.

In information theory and statistics, quasi-universal prediction and coding (similar to quasi-universal testing discussed here) was promoted by Rissanen; see, e.g., [21] and Grünwald's review [4]. Rissanen's suggestion for the weights w_k, $k = 1, 2, \ldots$, that sum to 1 and tend to 0 slowly was

$$w_k := \frac{1}{ck \log k \log\log k \log\log\log k \ldots}, \tag{35}$$

where the denominator includes all terms that exceed 1 and $c \approx 0.865$ is the normalizing constant [21, Appendix A].

In this paper we used the uniform prior on the Markov statistical model to obtain our alternative hypothesis. Another natural choice is Jeffreys priors [5]. They have some advantages, to be discussed in the next paragraph, but their advantages in our current context are much less pronounced than in other contexts, where they, e.g., are invariant w.r. to smooth reparametrizations [5] and attain minimax optimality in some cases [4, Sect. 8.2] (perhaps after modifications). They do not always exist, and many Bayesian statisticians find them objectionable (see, e.g., [34, Sect. 6]). Using the uniform prior in this paper leads to simple analytical expressions and efficient calculations.

A typical advantage of Jeffreys priors over uniform priors is that they assign larger weights to extreme values of parameters. Let us discuss, for simplicity, the priors considered in [20]: π_{01} and π_{10} are generated independently from Jeffreys's probability density

$$f(\theta) := \frac{1}{\pi\sqrt{\theta(1-\theta)}}, \quad \theta \in [0,1] \tag{36}$$

(where $\pi \approx 3.14$ is the standard mathematical constant, not to be confused with $\pi_{i,j}$ and not used outside of this and next paragraphs). These priors are built on top of Jeffreys priors but are not Jeffreys priors themselves [25, Sect. 1]. They are used in [20] for tackling problems that are similar to ours (using the Markov model as alternative when testing exchangeability).

The density f in (36) dominates the uniform density (if we ignore the constant factor π), and it can be much larger than the uniform density at the ends $\theta \approx 0$ and $\theta \approx 1$ of the interval $[0,1]$. This is a step towards quasi-universality, but the step is small: we can easily go further and consider, e.g., the beta distribution with density

$$f(\theta) := \frac{1}{\mathrm{B}(\alpha,\alpha)}\theta^{\alpha-1}(1-\theta)^{\alpha-1}, \quad \theta \in [0,1],$$

for a small $\alpha > 0$; this would not even complicate calculations. An even better choice would be in the direction of (35), which was an improvement on $w_k \propto k^{\alpha-1}$, but this would complicate calculations enormously. A natural next step would be to assign small but positive point masses to $\theta = 0$ and $\theta = 1$.

Using the uniform prior reflects an implicit assumption that we are making in this paper: all four probabilities $\pi_{i,j}$, $i,j \in \{0,1\}$, are middling ones (not too close to 0 or 1).

The idea of quasi-universal testing is closely related to Lindley's "Cromwell's rule" (see, e.g., [14, Sect. 6.8]). A possible interpretation of Cromwell's rule in our context is that, when designing a suitable e-variable, we should think of all kinds of alternative models (say, Markov models of all orders), and then mix all of them. Cromwell's rule as stated by Lindley is very general and encompasses two aspects: our statistical models should be as wide as possible, and our priors should be diffuse (at least non-zero).

C Changepoint models

In this appendix we will discuss in greater detail the changepoint compression models mentioned in the previous appendixes. But first we discuss a changepoint alternative hypothesis when testing exchangeability.

In the ideal picture, we just use $\tilde{E}$ of Lemma A.4 as e-test, but in practice we could use

$$Q(\{(z_1,\dots,z_N)\}) := \frac{1}{N-1}\sum_{n=1}^{N-1}\int_0^1\int_0^1 \pi_0^{z_1+\dots+z_n}(1-\pi_0)^{n-z_1-\dots-z_n} \tag{37}$$

$$\pi_1^{z_{n+1}+\dots+z_N}(1-\pi_1)^{N-n-z_{n+1}-\dots-z_N}\,\mathrm{d}\pi_0\,\mathrm{d}\pi_1 \tag{38}$$

$$= \frac{1}{N-1}\sum_{n=1}^{N-1}\mathrm{B}(z_1+\dots+z_n+1, n-z_1-\dots-z_n+1)$$

$$\mathrm{B}(z_{n+1}+\dots+z_N+1, N-n-z_{n+1}-\dots-z_N+1)$$

$$= \frac{1}{N-1}\sum_{n=1}^{N-1}\frac{(z_1+\dots+z_n)!(n-z_1-\dots-z_n)!(N-n+1)!}{(z_{n+1}+\dots+z_N)!(N-n-z_{n+1}-\dots-z_N)!(n+1)!} \tag{39}$$

as quasi-universal alternative probability measure. The expression inside the double integral in (37)–(38) is the likelihood of the observed data sequence

when the probability of 1 is π_0 before and including time $n \in \{1, \dots, N\}$ (the changepoint) and is π_1 strictly after time n. We average this likelihood over the uniform distribution for (π_0, π_1) and then over the uniform distribution for the changepoint n.

The alternative Markov kernel corresponding to (39) is

$$Q_{N_1}(\{(z_1, \dots, z_N)\}) = \frac{Q(\{(z_1, \dots, z_N)\})}{\sum_{z'_1, \dots, z'_N : z'_1 + \dots + z'_N = N_1} Q(\{(z'_1, \dots, z'_N)\})},$$

where $N_1 := z_1 + \dots + z_N$ is interpreted as the value of the summarising statistic. Finally, we can compute the quasi-universal e-value as

$$E(z_1, \dots, z_N) := \binom{N}{N_1} Q_{N_1}(\{(z_1, \dots, z_N)\}).$$

We do not discuss efficient ways of computing this e-value in this version of the paper.

Confidence regions

Now suppose we believe that there is at most one changepoint in a binary data sequence $z_1, \dots, z_N$ and would like to pinpoint its location. To obtain a confidence region, we need different null hypotheses.

The Kolmogorov compression model with the changepoint $\tau \in \{1, \dots, N-1\}$ has

$$t_\tau(z_1, \dots, z_N) := \left(\sum_{n=1}^{\tau} z_n, \sum_{n=\tau+1}^{N} z_n \right) \tag{40}$$

as its summarising statistic. Examples of probability measures that agree with this KCM are

$$P(\{(z_1, \dots, z_N)\}) := \\ \pi_0^{z_1 + \dots + z_\tau} (1 - \pi_0)^{\tau - z_1 - \dots - z_\tau} \pi_1^{z_{\tau+1} + \dots + z_N} (1 - \pi_1)^{N - \tau - z_{\tau+1} - \dots - z_N}$$

for $\pi_0, \pi_1 \in [0, 1]$. Of course, these are not all probability measures that agree with (40); those consist of all convex mixtures of the uniform probability measures on $t_\tau^{-1}(k_0, k_1)$, where $(k_0, k_1) \in \{0, \dots, \tau\} \times \{0, \dots, N - \tau\}$.

As alternative probability measure we can take (39) or, which is slightly more natural, its modification

$$Q_\tau(\{(z_1, \dots, z_N)\}) := \frac{1}{N-2} \\ \sum_{n \in \{1, \dots, N-1\} \setminus \{\tau\}} \frac{(z_1 + \dots + z_n)!(n - z_1 - \dots - z_n)!(N - n + 1)!}{(z_{n+1} + \dots + z_N)!(N - n - z_{n+1} - \dots - z_N)!(n+1)!}$$

that only considers changepoint locations different from τ, the one we are testing. The alternative Markov kernel becomes

$$Q_{\tau,K_0,K_1}(\{(z_1,\dots,z_N)\}) = \frac{Q_\tau(\{(z_1,\dots,z_N)\})}{\sum_{z'_1,\dots,z'_N : z'_1+\dots+z'_\tau=K_0, z'_{\tau+1}+\dots+z'_N=K_1} Q_\tau(\{(z'_1,\dots,z'_N)\})},$$

where $(K_0, K_1) := (z_1+\dots+z_\tau, z_{\tau+1}+\dots+z_N)$ is the value of the summarising statistic. Finally, we can compute the quasi-universal e-value as

$$E_\tau(z_1,\dots,z_N) := \binom{\tau}{K_0}\binom{N-\tau}{K_1} Q_{\tau,K_0,K_1}(\{(z_1,\dots,z_N)\}). \tag{41}$$

Once we have the e-values (41), we have the e-confidence regions for the changepoint τ: at a significance level α, the e-confidence region is $\{\tau \mid E_\tau \le 1/\alpha\}$ (see [37]). A natural direction of further research is to find a computationally efficient version of the e-confidence regions based on (41).

D Neyman structure

In this appendix we assume, as usual in this paper, that the sample space is finite. (In this case every function on the sample space is bounded, and we do not have to discuss completeness and bounded completeness separately; in fact, the most relevant notion of completeness for e-testing without this restriction would have been "semi-bounded completeness" only involving functions that are bounded below.)

Let us say that a statistic (i.e., function on the sample space) E is a *similar* (or *precise*) *e-variable* for a statistical model $\{P_\theta \mid \theta \in \Theta\}$ if $\int E \, \mathrm{d}P_\theta = 1$ for all $\theta \in \Theta$; this is an analogue for e-testing of Neyman and Pearson's [17, Sects IV(a) and V(a)] notion of a similar test. And we say that a statistic E has *Neyman structure* w.r. to a sufficient statistic T if $\mathbb{E}_\theta(E \mid T) = 1$ P_θ-a.s. for all $\theta \in \Theta$. This is analogous to the standard notion of Neyman structure (see, e.g., [13, Sect. 4.3]).

A statistic T is *complete* if, for any function f on its range,

$$\Big(\mathbb{E}_\theta(f(T)) = 0 \text{ for all } \theta \in \Theta\Big) \Longrightarrow \Big(f(T) = 0 \ P_\theta\text{-a.s. for all } \theta \in \Theta\Big).$$

The following is an analogue of Theorem 4.3.2 in [13].

Proposition D.1. *Let T be a sufficient statistic for a statistical model $\{P_\theta \mid \theta \in \Theta\}$. If T is complete, a statistic is a similar e-variable if and only if it has Neyman structure w.r. to T. The condition that T be complete is both sufficient and necessary.*

Proof. Suppose T is complete. It is clear that a statistic that has Neyman structure is a similar e-variable. Now suppose E is a similar e-variable. Set $f(T) := \mathbb{E}_\theta(E \mid T)$; f can be chosen independent of θ since T is sufficient. Since $\mathbb{E}_\theta(f(T) - 1) = 0$ for all θ, $f(T) = 1$ P_θ-a.s. for all θ, and so E has Neyman structure.

Now suppose that T is not complete. Choose a $[-1,\infty)$-valued function f such that $\mathbb{E}_\theta(f(T)) = 0$ for all $\theta \in \Theta$ but $f(T) \neq 0$ with a positive P_θ-probability for some $\theta \in \Theta$. Then $1+f(T)$ is a similar e-variable that does not have Neyman structure w.r. to T. □

For our purposes the following one-sided variation of having Neyman structure is more useful (although it is much less widely applicable). An *e-variable* w.r. to a statistical model $\{P_\theta \mid \theta \in \Theta\}$ is a nonnegative random variable E such that $\int E \, \mathrm{d}P_\theta \le 1$ for all $\theta \in \Theta$. It has *one-sided Neyman structure* w.r. to a sufficient statistic T if $\mathbb{E}_\theta(E \mid T) \le 1$ P_θ-a.s. for all $\theta \in \Theta$.

Let us say that a statistic T is *supercomplete* if, for any function f on its range,

$$\Big(\mathbb{E}_\theta(f(T)) \le 0 \text{ for all } \theta \in \Theta\Big) \implies \Big(f(T) \le 0 \; P_\theta\text{-a.s. for all } \theta \in \Theta\Big). \tag{42}$$

(It is clear that this property is stronger than completeness.) Now we have the following analogue of Proposition D.1.

Proposition D.2. *Let T be a sufficient statistic for a statistical model $\{P_\theta \mid \theta \in \Theta\}$. If T is supercomplete, a nonnegative random variable is an e-variable if and only if it has one-sided Neyman structure w.r. to T. The condition that T be supercomplete is both sufficient and necessary.*

Proof. Suppose T is supercomplete. It is clear that a nonnegative variable that has one-sided Neyman structure is an e-variable. Now suppose E is an e-variable. Set $f(T) := \mathbb{E}_\theta(E \mid T)$. Since $\mathbb{E}_\theta(f(T) - 1) \le 0$ for all θ, $f(T) \le 1$ P_θ-a.s. for all θ, and so E has one-sided Neyman structure.

Now suppose that T is not supercomplete. Choose a $[-1,\infty)$-valued function f such that $\mathbb{E}_\theta(f(T)) \le 0$ for all $\theta \in \Theta$ but $f(T) > 0$ with a positive P_θ-probability for some $\theta \in \Theta$. Then $1 + f(T)$ is an e-variable that does not have Neyman structure w.r. to T. □

The following two examples show that the notion of supercompleteness is limited albeit not vacuous.

Example D.3 (exchangeability)**.** The summarising statistic t_E of the exchangeability compression model (we can set t_E to the number of 1s in the data sequence) is supercomplete w.r. to the exchangeability statistical model (consisting of all exchangeable probability measures). This is because for each summary k there exists an exchangeable probability measure concentrated on $t_E^{-1}(k)$. (And it is clear that this argument is applicable to any batch compression model and the family of all probability measures that agree with it.)

Example D.4 (IID)**.** On the other hand, t_E is not supercomplete w.r. to the Bernoulli statistical model $(B_\theta \mid \theta \in (0,1))$ (where B_θ is the probability measure on $\{0,1\}$ satisfying $B_\theta(\{1\}) = \theta$). The standard argument for completeness as

given in [13, Example 4.3.1] now fails. A function f satisfying the first inequality in (42) can be written as

$$\sum_{k=0}^{N} f(k)\binom{N}{k}\rho^k \leq 0, \quad \text{for all } \rho \in (0,\infty), \tag{43}$$

and under the supercompleteness we would have concluded that $f \leq 0$. But on the left-hand side of (43) we can have any polynomial of degree N, and a polynomial can be nonpositive without all its coefficients being nonpositive. An example is $-(\rho-1)^2$, which corresponds to the function

$$f(k) := \begin{cases} -1 & \text{if } k = 0 \\ \frac{2}{N} & \text{if } k = 1 \\ -\frac{2}{N(N-1)} & \text{if } k = 2 \\ 0 & \text{otherwise.} \end{cases}$$