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 May 13, 2023.

Project web site:
http://alrw.net

Abstract

The topic of this paper is testing exchangeability using e-values in the batch mode, with 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. In the appendixes I explain connections with the algorithmic theory of randomness and with the traditional theory of testing statistical hypotheses. In the standard statistical terminology, this paper proposes a new permutation test. This test can also be interpreted as a poor man's version of Kolmogorov's deficiency of randomness.

Contents

1 Introduction

The usual approach to testing exchangeability in statistics is based on using p-values, as in [12, Sect. 7.2]. In this paper we will use e-values instead [36, 3], which facilitates computations. E-values have been used for testing exchangeability via conformal prediction [32, Part III] in the online protocol, while in this paper we will use the standard batch protocol: we are given the data sequence as one batch rather than getting its elements sequentially one by one.

The null hypothesis of exchangeability will be defined in Sect. 2 using the terminology of compression modelling [32, Chap. 11]. Compression modelling is an algorithm-free version of Kolmogorov's way of stochastic modelling: cf. [29], [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, our tools for obtaining 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.

In Appendix A I describe 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 *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. Our *null hypothesis* is the KCM, which we call the *Kolmogorov exchangeability model*, $t_E(z_1, \dots, z_N) := z_1 + \dots + z_N$.

Let us say that a probability measure P *agrees* with a summarising statistic t if the data sequences with the same summary have the same P-probability.

Lemma 2.1. *The exchangeable probability measures on Ω are exactly the probability measures that agree with the Kolmogorov exchangeability model (the mixtures of the uniform probability measures on $t_E^{-1}(k)$).*

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. Alternatively, it is an exchangeability e-variable w.r. to any 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 considered in this paper will be Kolmogorov compression models. In the main part of the paper we will concentrate on the exchangeability model, but in this and next section 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 any $\sigma \in t(\Omega)$. *E-values* are values taken by e-variables.

Disintegration of the alternative hypothesis

Let us fix an *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 $P(\sigma)$ are the uniform probability measures on $t^{-1}(\sigma)$.

Remark 2.3. Batch compression models are in fact standard and are often used without giving them any name, as in [11].

With an alternative hypothesis Q and a null hypothesis t we associate the *alternative Markov kernel*

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

As compared with Q, the alternative Markov kernel loses the information about $Q(\{\sigma\})$ for $\sigma \in \Sigma$.

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]

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

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,

$$\operatorname{ep}_Q(E) = \int \ln E \, \mathrm{d}Q. \tag{2}$$

The interpretation of (1) is that our capital $\prod_{i=1}^{I} E(z_1^i, \dots, z_N^i)$ grows exponentially fast (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 (2), which we will refer to as the *e-power* of E under the alternative Q.

Proof. It suffices to rewrite (1) as

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

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 (2) using frequentist considerations, we do not really need the IID picture, as emphasized by Neyman [18, 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 horizons $N = N_i$, against different alternatives $Q = Q_i$. 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 [24, Chap. 4], which allows non-stochastic choice of N_i, t_i, and Q_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.

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

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

$$E(\omega) := \left|t^{-1}(t(\omega))\right| Q_{t(\omega)}(\{\omega\}), \quad \omega \in \Omega. \tag{3}$$

In this case,

$$\mathrm{mep}(Q) := \mathrm{ep}_Q(E) = \int \ln \left|t^{-1}(\sigma)\right| (t_* Q)(\mathrm{d}\sigma) + H(t_* Q) - H(Q), \tag{4}$$

where $t_* Q$ *stands for the push-forward measure of* Q *by* t *(the summarising statistic of the null hypothesis), and* H *stands for the entropy.*

We will call $\mathrm{mep}(Q)$ defined by (4) the *maximum e-power* of the alternative Q. A sizeable $\mathrm{mep}(Q)$ 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 (3) not to be too volatile.

Proof. In this paper we let U_A stand for the uniform probability measure on a finite 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 $\sigma \in t(\Omega)$,

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

for each e-variable E' w.r. to t, which implies the first statement (about (3)) of the lemma. The second statement (4) follows from

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

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.

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:

$$\begin{aligned}
&\frac{1}{2}\int (1-\pi_{01})^{N_{00}}\pi_{01}^{N_{01}}\pi_{10}^{N_{10}}(1-\pi_{10})^{N_{11}}\,\mathrm{d}\pi_{01}\mathrm{d}\pi_{10} \\
&\quad = \frac{1}{2}\mathrm{B}(N_{00}+1, N_{01}+1)\mathrm{B}(N_{10}+1, N_{11}+1) \\
&\quad = \frac{1}{2}\frac{\Gamma(N_{00}+1)\Gamma(N_{01}+1)\Gamma(N_{10}+1)\Gamma(N_{11}+1)}{\Gamma(N_{0*}+2)\Gamma(N_{1*}+2)} \\
&\quad = \frac{1}{2}\frac{N_{00}!N_{01}!N_{10}!N_{11}!}{(N_{0*}+1)!(N_{1*}+1)!},
\end{aligned} \tag{5}$$

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 (5)).

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

The expression (5) gives us, analogously to [32, Chap. 9] (who follow [21]), the *lower benchmark*

$$\mathrm{LB} := \frac{1}{2}\frac{N_{00}!N_{01}!N_{10}!N_{11}!}{(N_{0*}+1)!(N_{1*}+1)!(N_0/N)^{N_0}(N_1/N)^{N_1}}. \tag{6}$$

The idea behind the lower benchmark is that, for any power probability measure Q^N (Q being a probability measure on $\{0, 1\}$), it is an e-variable w.r. to Q^N, i.e., satisfies $\int \mathrm{LB}\,\mathrm{d}Q^N \le 1$. To ensure this, (6) is defined as the ratio of the alternative probability measure to the maximum likelihood under the IID model.

However, the IID model is not our null hypothesis, and our null hypothesis of exchangeability is slightly more challenging. Replacing in (6) the maximum likelihood over the IID model by the maximum likelihood over the exchangeability model, we obtain the *exchangeability lower benchmark*

$$\mathrm{ELB} := \frac{1}{2}\binom{N}{N_1}\frac{N_{00}!N_{01}!N_{10}!N_{11}!}{(N_{0*}+1)!(N_{1*}+1)!}. \tag{7}$$

For the e-power of the exchangeability lower benchmark we have the formula (4) with the second term $H(t_*Q)$ omitted.

To compute efficiently the likelihood ratio of the alternative to null probability measures, we will 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). 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).

Lemma 4.1. *In any Markov graph σ with the set of vertices V the number of Eulerian paths from the source to the sink equals*

$$T(\sigma)\frac{\operatorname{out}(\text{sink})\prod_{v\in V}(\operatorname{out}(v)-1)!}{\prod_{u,v\in V}N_{u,v}!}, \tag{8}$$

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

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

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

If source = sink, the number of Eulerian paths is obtained by multiplying by out(source). Finally, we identify all darts from u to v for all pairs of vertices (u,v) by dividing by $N_{u,v}!$; the resulting expression agrees with (8).

Now suppose source $\neq$ sink. Create a new digraph by adding another dart leading from the source to the sink. 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.,

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

where out refers to the old digraph. It remains to identify all darts from u to v for all pairs of vertices (u,v) in the old digraph; the resulting expression again agrees with (8).

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

Lemma 4.2. *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 can compute the co-factor of any diagonal element $a_{i,i}$, but it 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 4.1 and 4.2 to the binary case $\mathbf{Z} := \{0, 1\}$.

Corollary 4.3. *Let σ be a Markov graph with vertices in $\{0, 1\}$ with F as its source. The number of Eulerian paths from the source to the sink equals*

$$N(\sigma) := \begin{cases} N_{F,1-F} \frac{(N_0-1)!(N_1-1)!}{N_{00}!N_{01}!N_{10}!N_{11}!} & \text{if } N_0 \wedge N_1 > 0 \\ 1 & \text{otherwise,} \end{cases} \tag{9}$$

where $N_i := \mathrm{in}(i) \vee \mathrm{out}(i)$ and $N_{i,j}$ (with the comma omitted) is the number of darts leading from i to j.

Proof. 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 4.2 and is obvious anyway. It remains to plug this in into Lemma 4.1: assuming $N_0 \wedge N_1 > 0$, if the source F and sink L coincide, $F = L$, we obtain

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

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 (9). The case $N_0 \wedge N_1 = 0$ is obvious. □

Combining (5) and (9), we obtain the total alternative weight of

$$W(\sigma) := \begin{cases} \frac{1}{2} N_{F,1-F} \frac{(N_0-1)!(N_1-1)!}{(N_{0*}+1)!(N_{1*}+1)!} & \text{if } N_{1-F} > 0 \\ \frac{1}{2N} & \text{otherwise} \end{cases} \tag{10}$$

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 null of exchangeability) is

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

(see (5) and (10)), where the σ in $\sum_\sigma$ ranges over the Markov types $(f, n_{00}, \ldots, n_{11}, l)$ compatible with the exchangeability type (N_0, N_1). The equality in (11) holds when $N_0 \wedge N_1 = 0$; in the case $N_0 \wedge N_1 = 0$ the likelihood value is 1 (and we will treat this case separately in Algorithm 1). We will refer

to (11) (interpreted as 1 when $N_0 \wedge N_1 = 0$) as the *uniformly mixed Markov* (*UMM*) e-variable; this is our main object of interest in this paper.

It remains to explain how to compute the sum $\sum_\sigma$ in (11). For the $\sigma = (f, n_{00}, \dots, n_{11}, l)$ with $f = l = 0$ (which is only possible when $N_0 \geq 2$), each 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

$$\sum_{n_{01}=1}^{(N_0-1)\wedge N_1} \frac{n_{01}}{N_0 N_1 (N_1 + 1)} = \frac{((N_0 - 1) \wedge N_1)((N_0 - 1) \wedge N_1 + 1)}{2 N_0 N_1 (N_1 + 1)}. \tag{12}$$

This component should only be used when $N_0 \geq 2$; otherwise, it is 0.

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

$$\sum_{n_{01}=1}^{N_0 \wedge N_1} \frac{n_{01}}{N_0 (N_0 + 1) N_1} = \frac{(N_0 \wedge N_1)(N_0 \wedge N_1 + 1)}{2 N_0 (N_0 + 1) N_1}. \tag{13}$$

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

$$\sum_{n_{10}=1}^{N_0 \wedge N_1} \frac{n_{10}}{N_0 N_1 (N_1 + 1)} = \frac{(N_0 \wedge N_1)(N_0 \wedge N_1 + 1)}{2 N_0 N_1 (N_1 + 1)}. \tag{14}$$

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

$$\sum_{n_{10}=1}^{N_0 \wedge (N_1 - 1)} \frac{n_{10}}{N_0 N_1 (N_1 + 1)} = \frac{(N_0 \wedge (N_1 - 1))(N_0 \wedge (N_1 - 1) + 1)}{2 N_0 (N_0 + 1) N_1}. \tag{15}$$

This component is used only when $N_1 \geq 2$; otherwise, we set it to 0.

The overall algorithm is presented as Algorithm 1. The value of the uniformly mixed Markov e-variable UMM is computed according to (11), and the value ELB of the exchangeability lower benchmark in line 5 is just (11) with the sum over the Markov types σ omitted. In line 6 we initialize the sum over the Markov types σ, and in lines 7, 8, 9, and 10 we compute it according to the right-hand sides of (12), (13), (14), and (15), respectively. The symbol += is

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 $\mathrm{UMM}(z_1, \dots, z_N)$.

Set N_0 and N_1 to the numbers of 0s and 1s in $(z_1, \dots, z_N)$, respectively.
if $N_0 \wedge N_1 = 0$: **return** 1
for $i, j \in \{0,1\}$:
    Set $N_{i,j}$ to the number of substrings (i,j) in $(z_1, \dots, z_N)$.
ELB $:= \frac{N_{00}!N_{01}!N_{10}!N_{11}!\binom{N}{N_1}}{(N_{0*+1})!(N_{1*+2})!}$.
Sum $:= 0$.
if $N_0 \geq 2$: Sum $+= \frac{((N_0-1)\wedge N_1)((N_0-1)\wedge N_1+1)}{2N_0N_1(N_1+1)}$.
Sum $+= \frac{(N_0\wedge N_1)(N_0\wedge N_1+1)}{2N_0(N_0+1)N_1}$.
Sum $+= \frac{(N_0\wedge N_1)(N_0\wedge N_1+1)}{2N_0N_1(N_1+1)}$.
if $N_1 \geq 2$: Sum $+= \frac{(N_0\wedge(N_1-1))(N_0\wedge(N_1-1)+1)}{2N_0(N_0+1)N_1}$.
return ELB/Sum.

used in the Python sense: $A \mathrel{+}= B$ is equivalent to $A := A + B$. 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 $O(N)$ time, which is clearly optimal.

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 we will see the weakness of the 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 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 in (4), 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 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. (5))

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

(we are ignoring the 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

$$\begin{aligned}
&\int_0^1 \int_0^1 \Bigg(\frac{\pi_{00}\pi_{10}}{\pi_{01}+\pi_{10}} \ln \pi_{00} + \frac{\pi_{01}\pi_{10}}{\pi_{01}+\pi_{10}} \ln \pi_{01} \\
&\qquad\qquad + \frac{\pi_{01}\pi_{10}}{\pi_{01}+\pi_{10}} \ln \pi_{10} + \frac{\pi_{01}\pi_{11}}{\pi_{01}+\pi_{10}} \ln \pi_{11} \Bigg) \mathrm{d}\pi_{01}\, \mathrm{d}\pi_{10} \\
&\qquad = \frac{2}{3} \ln 2 + \frac{2}{3} \ln^2 2 - \frac{1}{9}\pi^2 - \frac{1}{6} \approx -0.481. \qquad (16)
\end{aligned}$$

Now let us estimate the first term

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

in (4). 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

$$\begin{aligned}-\int_0^1\int_0^1 (\pi_0\ln\pi_0 + \pi_1\ln\pi_1)\,\mathrm{d}\pi_{01}\,\mathrm{d}\pi_{10} &= -2\int_0^1\int_0^1 (\pi_0\ln\pi_0)\,\mathrm{d}\pi_{01}\,\mathrm{d}\pi_{10} \\ &= -2\int_0^1\int_0^1 \left(\frac{\pi_{10}}{\pi_{01}+\pi_{10}}\ln\frac{\pi_{10}}{\pi_{01}+\pi_{10}}\right)\mathrm{d}\pi_{01}\,\mathrm{d}\pi_{10} \\ &= 2\ln 2 - \frac{1}{12}\pi^2 \approx 0.564. \qquad (17)\end{aligned}$$

The final term $H(t_*Q)$ in (4) can be ignored. Indeed, using the last expression in (5), we can bound the probability $(t_*Q)(\{K\})$, for any $K \in \{1,\dots,N-1\}$, by 1 from above and by $1/(2N)$ from below:

$$(t_*Q)(\{K\}) \geq \frac{1}{2}\frac{(N-K-1)!0!1!(K-1)!}{(N-K-1)!K!} = \frac{1}{2K} \geq \frac{1}{2N} \qquad (18)$$

(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 (16) and (17), we obtain the coefficient

$$\frac{8}{3}\ln 2 + \frac{2}{3}\ln^2 2 - \frac{7}{36}\pi^2 - \frac{1}{6} \approx 0.083 \qquad (19)$$

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. □

The proof of Proposition 5.1 contains the following relation between the UMM e-variable and the exchangeability lower benchmark; in particular, it shows once again that the exchangeability lower benchmark is also an e-variable.

Proposition 5.2. *It is always true that*

$$1 \le \frac{\text{UMM}}{\text{ELB}} \le 2N. \tag{20}$$

Moreover, it is true that

$$1 \le \frac{\text{UMM}}{\text{ELB}} \le N \tag{21}$$

unless $N_1 \in \{0, N\}$.

Proof. Unless $N_1 = 0$, we can improve (18) to

$$(t_* Q)(\{K\}) \ge \frac{1}{N}$$

by considering, alongside the sequence consisting of K 1s followed by $N-K$ 0s, the sequence consisting of K 0s followed by $N-K$ 1s. It remains to notice that

$$\text{UMM} = \frac{\text{ELB}}{(t_* Q)(\{N_1\})}. \qquad \square$$

6 Computational experiments

We will conduct two groups of experiments for the two lower benchmarks and the UMM exchangeability e-variable. In the first group, the true data distribution will be a specific Markov probability measure with initial probability of 1 equal to 1/2. In this case, we define the *upper benchmark* as

$$\text{UB} := \frac{1}{2} \frac{N_{00}! N_{01}! N_{10}! N_{11}!}{(N_{0*}+1)!(N_{1*}+1)! \pi_0^{N_0} \pi_1^{N_1}}, \tag{22}$$

where π_0 and π_1 are the stationary probabilities under the true data-generating distribution. Therefore, the upper benchmark is an e-variable 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 10^5 simulations of four values: the exchangeability lower benchmark ELB (given by (7)), the lower benchmark LB (given by (6)), the upper benchmark UB (given by (22)), and the UMM exchangeability e-variable UMM (given by Algorithm 1). Only two of these, ELB and UMM, are bona fide exchangeability e-values. It is interesting that UMM is often even higher than the upper benchmark, as in the right panel of Fig. 1. The horizon N and the transition probabilities for the two panels are given in the caption. In both cases, the alternative probability measure is Markov.

N	$\pi_{01} = \pi_{10}$	$\overline{\text{ELB}}$	$\overline{\text{UMM}}$	$\overline{\text{UMM}} - \overline{\text{ELB}}$	upper bound
20	0.1	-0.116	1.226	1.342	1.301
400	0.4	0.084	2.427	2.343	2.602

Table 1: Comparison between the decimal logarithms of the exchangeability lower benchmark and the UMM e-variable; the upper bound for the difference is $\log_{10} N$, as per (21).

According to Proposition 5.2, the UMM e-value cannot differ from the exchangeability lower benchmark by much. Table 1 gives the means of their decimal logarithms (over the same 10^5 simulations as in Fig. 1) and the upper bound (21) for the difference between them. The bars stand for the empirical averages (over all 10^5 replications). The upper bound (21) is violated in the first row because $\pi_{01} = \pi_{10}$ and N are so small, which often leads to $N_1 \in \{0, N\}$; of course, the upper bound (20) (whose value is approximately 1.602 in this case) still holds.

The second group of experiments involves generating the binary observations from the UMM alternative (which is not Markov any more). The explicit formula for this alternative is given in (5), 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. Figure 2 shows results for this case, with the same four values as in Fig. 1; in the expression (22) for the upper benchmark, we now set $\pi_{01} := \pi_{10} := 1/2$. It is striking how spread out the distributions for the three benchmarks and the UMM e-variable are. They are also skewed, with the mean very different from the median. Now the lack of validity for the upper benchmark is very obvious: it takes much larger values, and we will ignore it from now on.

Table 2 gives corresponding figures. Now the bars stand for the empirical averages over K replications (for three different values of K), N is the horizon, and "as." is the common theoretical asymptotic value for the UMM e-variable

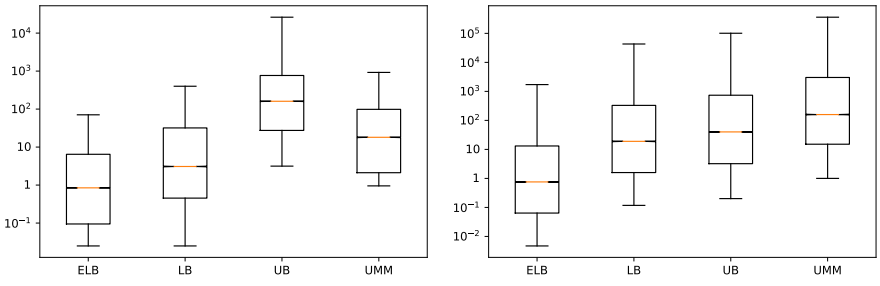

Figure 1: The four e-values and related quantities, as described in text. Left panel: $N = 20$ and $\pi_{01} = \pi_{10} = 0.1$. Right panel: $N = 400$ and $\pi_{01} = \pi_{10} = 0.4$.

N	K	$\overline{\mathrm{ELB}}$	$\overline{\mathrm{LB}}$	$\overline{\mathrm{UMM}}$	as.	UMM quantiles
10^3	10^3	31.05	32.56	34.02	36.02	$[-1.01, 2.56, 14.21, 49.04, 134.22]$
10^4	10^4	356.8	358.8	360.8	360.2	$[0.1, 35.7, 170.0, 509.8, 1378.3]$
10^5	10^5	3570	3573	3575	3602	$[12, 366, 1684, 5033, 13632]$

Table 2: Some figures for the decimal logarithms of the two lower benchmarks and the UMM e-variable; "as." stands for the asymptotic expression The UMM quantiles are for 5%, 25% (first quartile), 50% (median), 75% (third quartile), and 95%.

N	K	$\overline{\mathrm{ELB}}$	$\overline{\mathrm{UMM}}$	$\overline{\mathrm{UMM}} - \overline{\mathrm{ELB}}$	upper bound
10^3	10^3	31.05	34.02	2.968	3
10^4	10^4	356.8	360.8	3.964	4
10^5	10^5	3570	3575	4.966	5

Table 3: Analogue of Table 1 in the situation of Fig. 2.

and exchangeability lower benchmark obtained from (19) by dividing by $\ln 10$ (to convert natural logarithms to decimal ones) and multiplying by the sample size N.

Table 3, analogous to Table 1, 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 (21). The upper bound is surprisingly tight.

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],

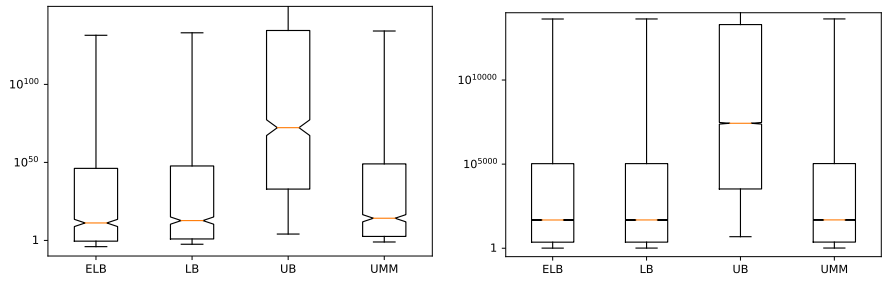

Figure 2: Two exchangeability e-values and two approximations for the UMM alternative. Left panel: $N = K = 10^3$. Right panel: $N = K = 10^5$.

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 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 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 increased 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.

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, June 2020.

[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] Bjørn Kjos-Hanssen, André Nies, Frank Stephan, and Liang Yue. Higher Kurtz randomness. *Annals of Pure and Applied Logic*, 161:1280–1290, 2010.

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

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

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

[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] Per Martin-Löf. The literature on von Mises' Kollektivs revisited. *Theoria*, 35:12–37, 1969.

[17] Per Martin-Löf. On the notion of randomness. In Akiko Kino, John Myhill, and Richard E. Vesley, editors, *Intuitionism and Proof Theory. Proceedings of the Summer Conference at Buffalo NY 1968*, pages 73–78. North-Holland, Amsterdam, 1970.

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

[19] 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.

[20] 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).

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

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

[23] Alexey Semenov, Alexander Shen, and Nikolay Vereshchagin. Kolmogorov's last discovery? (Kolmogorov and algorithmic statistics). Technical Report arXiv:2303.13185 [math.LO], arXiv.org e-Print archive, March 2023.

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

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

[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 Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986. Russian original: О понятии бернуллиевости. Another English translation with proofs: [30].

[29] 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.

[30] Vladimir Vovk. On the concept of Bernoulliness. Technical Report arXiv:1612.08859 [math.ST], arXiv.org e-Print archive, December 2016.

[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*, 2023. Submitted.

[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*, 2023. To appear.

[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. Die Widerspruchfreiheit des Kollectivbegriffes der Wahrscheinlichkeitsrechnung. *Ergebnisse eines Mathematischen Kolloquiums*, 8:38–72, 1937.

A Algorithmic theory of randomness

In the main part of the paper we avoided using computability theory that plays such an important role in Kolmogorov's original approach, which is the topic of this appendix. Kolmogorov's complexity models were introduced, in the most complete form, in what appears to be Kolmogorov's last talk, given on 14 October 1982 at what later became known as the Kolmogorov seminar; see [23, note 12], containing Shen's notes taken during the talk, and [29, 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 [23, note 12]; the two talks were conflated in my paper [29, Sect. 4].

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

Mathematical results

Let X be an infinitely countable set with a fixed bijection $f : X \to \{1, 2, \dots\}$ between X and the natural numbers. When talking about the computability of functions involving elements X, we mean the computability of those functions with $x \in X$ replaced by their "codes" $f(x)$. An example (of primary interest to us in this paper) is the set of all finite binary sequences with a computable bijection f. Alternatively, X can be any aggregate of constructive objects in the sense of [27, Sect. 1.0.6]; in general, we will regard elements of X as constructive objects.

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 constructive objects, such as elements of X or finite subsets of X.

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

$$d_A^C(x) := \log|A| - C(x \mid A) \tag{23}$$

[10, Sect. 2.3], where log stands for binary logarithm. 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 (23) can be stated in terms of p-values. Let A be a finite subset of X and U_A be 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 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 $A \subset X$, 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., [25, 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 logarithm, following [35].

Now we can state Martin-Löf's result expressing Kolmogorov's deficiency of randomness via the universal p-test. In this appendix log always stands for base 2 logarithm.

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

$$\left| d_A^C(x) + \log \tilde{P}_A(x) \right| \le c. \tag{24}$$

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

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

which implies the part

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

of (24).

To prove the other part of (24), i.e.,

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

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

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

The last inequality follows immediately from the definition of a p-test (with $c := 0$). □

Prefix complexity K has important technical advantages over C, and so a natural modification of (23) is

$$d_A^K(x) := \log|A| - K(x \mid A). \tag{25}$$

Analogously to expressing (23) in terms of p-values, we can express (25) in terms of e-values.

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

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

A family E of functions $E_A : A \to [0, 1]$, A ranging over the finite 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 $A \subset X$, 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 $\tilde{E} \ge E/c$.*

The proof of Lemma A.4 is again standard (but [25, 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 (25) 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,*

$$\left| d_A^K(x) - \log \tilde{E}_A(x) \right| \le c. \tag{26}$$

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 set $A \subset X$ is a *subprobability measure* (or *semimeasure* [25, 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 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 $A \subset X$, 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 $\tilde{m} \geq m/c$.*

For a proof of, essentially, Lemma A.6, see the proof of [25, 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,*

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

Proof. Follow [25, Sect. 4.5]. □

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

$$\frac{1}{c} \leq \frac{\tilde{m}_A(x)|A|}{\tilde{E}_A(x)} \leq c. \tag{27}$$

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 (27) 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 $1/|A|$.

Now we can easily prove Proposition A.5.

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

$$\begin{aligned} \left| d_A^K(x) - \log \tilde{E}_A(x) \right| &= \left| \log|A| - K(x \mid A) - \log \tilde{E}_A(x) \right| \\ &\leq |\log|A| + \log \tilde{m}_A(x) - \log(\tilde{m}_A(x)|A|)| + c = c, \end{aligned} \tag{28}$$

i.e., (26). The first equality in (28) just uses the definition of $d_A^K(x)$, and the inequality "$\leq$" 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 $A \subset X$ and all $x \in A$,*

$$C(x \mid A) - c \leq K(x \mid A) \leq C(x \mid A) + 2 \log C(x \mid A) + c \tag{29}$$

and

$$d_A^K(x) - c \leq d_A^C(x) \leq d_A^K(x) + 2 \log d_A^K(x) + c. \tag{30}$$

Proof. See [25, Theorem 65] for inequalities stronger than (29). For (30), follow the proof of [20, Proposition 1]. □

Discussion

Kolmogorov's original definition of randomness deficiency of an element of a finite set is (23). It can be interpreted as the universal p-value on the logarithmic scale (Proposition A.3). A natural modification of Kolmogorov's definition is (25), 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 [29, 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 (23) (or (25), which is close 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 [8, 9] 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.

- Another complexity model [9] 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 [29]), 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 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 Gaussian and Poisson (in his 1982 seminar talk [23, note 12]; see also [1, 2] and [29, Sect. 4]). A complexity model formalizing the property of being IID rather than exchangeability was introduced in work [28] 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 $A \subset X$ 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 Kolmogorov complexity model if $d_A^C(x) \le \Delta$, where A is the block of the model containing x. Data sequences that are modelled using Kolmogorov 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 $(\log N + c, \Delta + c)$-stochastic;
- if a data sequence of length N is Δ-Markov (i.e., Δ-random w.r. to the Markov model), it is $(2\log N + c, \Delta + c)$-stochastic;
- if a data sequence of length N is Δ-Markov of second order, it is $(4\log N + c, \Delta + c)$-stochastic;
- if a data sequence of length N is Δ-random w.r. to the IID model introduced in [28], it is $(\frac{1}{2}\log N + c, \Delta + c)$-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 $(2\log N + c, \Delta + c)$-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 alternative hypotheses, 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., [25, 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 this paper can, however, be regarded as a computable approximation to the ideal picture. The idea (which is an old one) 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 $\tilde{E}$.

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 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 $k \to \infty$.
- The first-order Markov model as the null, and 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 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., [22] 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

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

where the denominator includes all terms that exceed 1 and $c \approx 0.865$ is the normalizing constant [22, Appendix A]. The word "universal", however, is sometimes used in a more limited sense in information theory and statistics: it may be universality, in some sense, for a given statistical model, without attempting to make the statistical model wider.

Kolmogorov's ideal picture is based on computability, but when discussing practical approximations it may be useful to replace computability by expressibility in a given language. The idea of using expressibility in logic rather than computability is much older than the algorithmic theory of randomness (see, e.g., [16, Sect. 1]) and goes back to Wald [39]. This idea has led to higher-level algorithmic randomness, as in [17] and, e.g., [7].

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]. However, in our current context they do not have any obvious advantages. (Among their advantages in other contexts are their invariance w.r. to smooth reparametrizations and attaining minimax optimality in some cases [4, Sect. 8.2].) They do not always exist and many Bayesian statisticians find them objectionable (see, e.g., [34]). Using the uniform prior in this paper leads to simple analytical expressions and efficient calculations. Similar problems (using the Markov model as alternative when testing exchangeability) are considered in [21] and [32, Sect. 9.2.7], which use priors that are built on top of Jeffreys priors but are not Jeffreys priors themselves.

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

$$Q(\{(z_1, \ldots, z_N)\}) := \frac{1}{N-1} \sum_{n=1}^{N-1} \int_0^1 \int_0^1 \pi_0^{z_1+\cdots+z_n} (1-\pi_0)^{n-z_1-\cdots-z_n} \tag{31}$$

$$\pi_1^{z_{n+1}+\cdots+z_N}(1-\pi_1)^{N-n-z_{n+1}-\cdots-z_N}\,\mathrm{d}\pi_0\,\mathrm{d}\pi_1 \tag{32}$$

$$= \frac{1}{N-1}\sum_{n=1}^{N-1} \mathrm{B}(z_1+\cdots+z_n+1, n-z_1-\cdots-z_n+1)$$
$$\mathrm{B}(z_{n+1}+\cdots+z_N+1, N-n-z_{n+1}-\cdots-z_N+1)$$
$$= \frac{1}{N-1}\sum_{n=1}^{N-1} \frac{(z_1+\cdots+z_n)!(n-z_1-\cdots-z_n)!(N-n+1)!}{(z_{n+1}+\cdots+z_N)!(N-n-z_{n+1}-\cdots-z_N)!(n+1)!} \tag{33}$$

as quasi-universal alternative probability measure. The expression inside the double integral in (31)–(32) 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 (33) is

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

where $N_1 := z_1+\cdots+z_N$ 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

$$t_\tau(z_1,\dots,z_N) := \left(\sum_{n=1}^{\tau} z_n, \sum_{n=\tau+1}^{N} z_n\right) \tag{34}$$

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

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

for $\pi_0,\pi_1 \in [0,1]$. Of course, these are not all probability measures that agree with (34); 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 (33) 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

$$E_\tau(z_1,\dots,z_N) := \binom{\tau}{K_0}\binom{N-\tau}{K_1} Q_{\tau,K_0,K_1}(\{(z_1,\dots,z_N)\}). \tag{35}$$

Once we have the e-values (35), 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 (35).

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 [19, 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 $\int E \, \mathrm{d}P_\theta \leq 1$ for all $\theta \in \Theta$. It has *one-sided Neyman structure* w.r. to a sufficient statistic T if $\mathbb{E}_\theta(E \mid T) \leq 1$ 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,

$$\Big(\mathbb{E}_\theta(f(T)) \leq 0 \text{ for all } \theta \in \Theta\Big) \Longrightarrow \Big(f(T) \leq 0 \; P_\theta\text{-a.s. for all } \theta \in \Theta\Big). \tag{36}$$

(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 $\mathbb{E}_\theta(f(T) - 1) \leq 0$ for all θ, $f(T) \leq 1$ 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 $\mathbb{E}_\theta(f(T)) \leq 0$ 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 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 (36) can be written as

$$\sum_{k=0}^{N} f(k)\binom{N}{k}\rho^k \le 0, \quad \text{for all } \rho \in (0,\infty), \tag{37}$$

and under the supercompleteness we would have concluded that $f \le 0$. But on the left-hand side of (37) 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

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