Conformal testing: binary case with Markov alternatives

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman

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 #36

November 9, 2021

Project web site:
http://alrw.net

Abstract

We continue study of conformal testing in binary model situations. In this note we consider Markov alternatives to the null hypothesis of exchangeability. We propose two new classes of conformal test martingales; one class is statistically efficient in our experiments, and the other class partially sacrifices statistical efficiency to gain computational efficiency.

Contents

1 Introduction

This note treats a problem similar to the one considered in [3]: we would like to test online the null hypothesis of exchangeability of binary observations under Markov alternatives.

The simplest way of online hypothesis testing is to use *test martingales*, which are defined as nonnegative processes with initial value 1 that are martingales under the null hypothesis; see, e.g., [4]. Such processes, for the null hypothesis of exchangeability, can be constructed using the method of conformal prediction [7], and we will refer to them as conformal test martingales. A previous paper [6] constructs custom-made conformal test martingales for different alternative hypotheses, those of a changepoint.

The method of [3], which is specifically devoted to Markov alternatives, is more general: instead of a test martingale the authors construct a "safe e-process" (to be defined in the next section). Safe e-processes are closely related to test martingales and admit a similar interpretation as the capital of a gambler trying to discredit the null hypothesis. Our methods give similar results to the methods of [3] in the model situations that we consider (following [3]). The advantage of our methods is that they extend easily to the usual setting of machine learning, where the observations are pairs (x, y) consisting of a potentially complex object x and its label y.

In this note we only design conformal test martingales for a simple alternative hypothesis (a specific probability measure). This is different from [3], who are interested in testing against the composite alternative Markov hypothesis. As in [3], we could mix our conformal test martingales over the possible alternative hypotheses, but we leave this step for future research.

2 Model situations

This section introduces the model situations considered in this paper, following [3, Section 4.2]. Our data consist of binary observations generated from

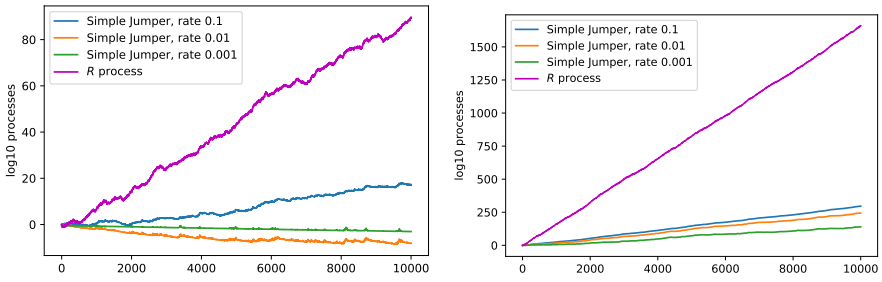

Figure 1: The process R of [3] and the Simple Jumper in the large scenario. Left panel: the hard case. Right panel: the easy case.

a Markov model. We will use the notation $\mathrm{Markov}(\pi_{1|0}, \pi_{1|1})$ for the probability distribution of a Markov chain with the transition probabilities $\pi_{1|0}$ for transitions $0 \to 1$ and $\pi_{1|1}$ for transitions $1 \to 1$; the probability that the first observation is 1 will always be assumed 0.5. In the *hard case*, the model is $\mathrm{Markov}(0.4, 0.6)$, and in the *easy case*, the model is $\mathrm{Markov}(0.1, 0.9)$. The number of observations is $N := 10^4$ (as in [3]) or $N := 10^3$ or $N := 10^2$; we will refer to these scenarios as *large*, *medium*, and *small*, respectively.

In all our experiments we use 2021 as the seed for the NumPy pseudorandom number generator. (This, however, does not make the trajectories in our plots comparable between different scenarios.) The dependence on the seed will be explored in boxplots reported in Section 5; the seed affects not only the data but also the values of conformal martingales, which are randomized processes, given the data.

Let B_π be the Bernoulli distribution on $\{0, 1\}$ with parameter $\pi \in [0, 1]$: $B_\pi(\{1\}) = \pi$. Set $\mathrm{Ber}(\pi) := B_\pi^\infty$. Our null hypothesis is the *IID model*, under which the observations are generated from $\mathrm{Ber}(\pi)$ with unknown parameter π.

Ramdas et al. construct a *safe e-process* $R = R_n$: namely, under any $\mathrm{Ber}(\pi)$, R is dominated by a test martingale $M_n^{(\pi)}$ w.r. to $\mathrm{Ber}(\pi)$, in the sense that $R_n \le M_n^{(\pi)}$ for all n and π. The trajectories of their process for the two cases, hard and easy, are shown in Figure 1 (they coincides with those in Figure 4 in [3] apart from using base 10 logarithms and a different randomly generated dataset). The figure also shows trajectories of the Simple Jumper martingale (see, e.g., [5]) for various values of the jumping rate; it performs poorly in this context.

3 Two benchmarks

In this section we will discuss possible benchmarks that we can use for evaluating the quality of our conformal test martingales. The *upper benchmark* is

$$\mathrm{UB}_n := \frac{\mathrm{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\mathrm{Ber}(0.5)([z_1, \dots, z_n])}, \tag{1}$$

where $[z_1, \dots, z_n]$ is the set of all infinite sequences of binary observations starting from $z_1, \dots, z_n$, and $z_1, z_2, \dots$ are the actual observations. The *lower benchmark* is

$$\mathrm{LB}_n := \frac{\mathrm{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\mathrm{Ber}(\hat{\pi})([z_1, \dots, z_n])}, \tag{2}$$

where $\hat{\pi} := k/n$ (the maximum likelihood estimate) and $k = k(n)$ is the number of 1s among $z_1, \dots, z_n$. By definition, $\mathrm{UB}_0 = \mathrm{LB}_0 := 1$.

The trajectories of the upper and lower benchmarks are shown in Figure 2 in red and green; the figure also shows the trajectory the R process discussed in the previous section, and the other two trajectories should be ignored for now. The two benchmarks coincide or almost coincide. Figure 3 should the same trajectories "under the lens", over the last 1000 observations. Notice that the upper benchmark can never be less than the lower benchmark.

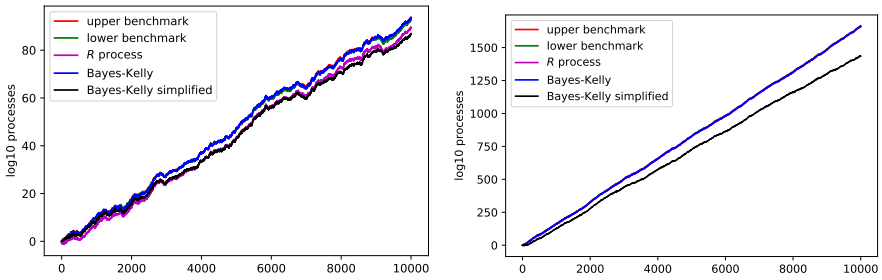

Figure 2: The two benchmarks, R process, Bayes–Kelly conformal test martingale, and its simplified version in the large scenario. Left panel: hard case (the trajectories for the two benchmarks and Bayes–Kelly almost coincide). Right panel: easy case (the trajectories for the two benchmarks, R process, and Bayes–Kelly virtually coincide).

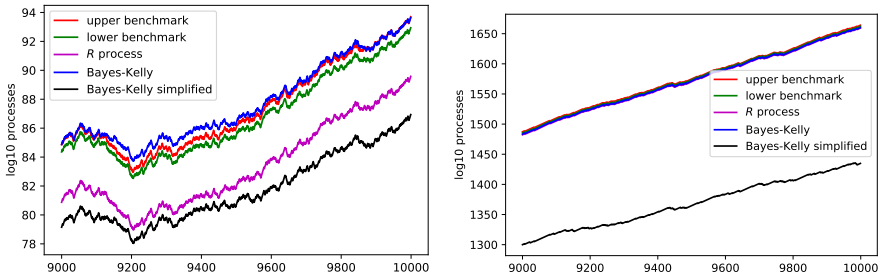

Figure 3: The analogue of Figure 2 for the last 1000 observations (as in that figure, in the right panel, the trajectories for the two benchmarks, R process, and Bayes–Kelly virtually coincide).

4 Bayesian conformal testing

In this section we will use a Bayesian method that is statistically efficient in our experiments but whose computational efficiency will be greatly improved in the next section. The p-values $p_1, p_2, \dots$ are generated as described in [6]; in particular, we are using the identity nonconformity measure (the nonconformity score of an observation z is z). Under the alternative hypothesis, the p-values are generated by a completely specified stochastic mechanism. According to [1, Theorem 2], the optimal (in the Kelly-type sense of that paper) betting functions f_n are given by the density of the predictive distribution of p_n conditional on knowing $p_1, \dots, p_{n-1}$. Let us find these predictive distributions. We will use the notation $U[a, b]$, where $a < b$, for the uniform probability distribution on the interval $[a, b]$ (so that its density is $1/(b - a)$).

Algorithm 1 Bayes–Kelly ($(p_1, p_2, \dots) \mapsto (S_1, S_2, \dots)$)

1: $S_0 := S_1 := 1$
2: Set the initial weights as per (3).
3: **for** $n = 2, 3, \dots$:
4: $\quad S_n := f_n(p_n) S_{n-1}$, with f_n defined by (7).
5: $\quad$ Update the weights as per (4)–(6).

We are in a typical situation of Bayesian statistics. The Bayesian parameter is the binary sequence $(z_1, z_2, \dots) \in \{0,1\}^\infty$ of observations, and the prior distribution on the parameter is Markov$(\pi_{1|0}, \pi_{1|1})$. The Bayesian observations are the conformal p-values $p_1, p_2, \dots$. Given the parameter, the distribution of p_n is

$$p_n \sim \begin{cases} U[0, k/n] & \text{if } z_n = 1 \\ U[k/n, 1] & \text{if } z_n = 0, \end{cases}$$

where $k := z_1 + \dots + z_n$ is the number of 1s among the first n observations.

Let $w^n_{k,j}$, where $n = 1, 2, \dots$, $k = 0, \dots, n$, and $j \in \{0, 1\}$, be the total posterior probability of the parameter values $z_1, z_2, \dots$ for which $z_1 + \dots + z_n = k$ and $z_n = j$; we will use them as the weights when computing the predictive distributions for the p-values. We can compute the weights $w^n_{k,j}$ recursively in n as follows. We start from

$$w^1_{0,0} := w^1_{1,1} := 0.5, \quad w^1_{0,1} := w^1_{1,0} := 0. \tag{3}$$

At each step $n \geq 2$, first we compute the unnormalized weights

$$\tilde{w}^n_{k,0} := \left(w^{n-1}_{k,0} \pi_{0|0} + w^{n-1}_{k,1} \pi_{1|0} \right) l^{n-1}_k(0, p_n), \tag{4}$$

$$\tilde{w}^n_{k,1} := \left(w^{n-1}_{k-1,0} \pi_{1|0} + w^{n-1}_{k-1,1} \pi_{1|1} \right) l^{n-1}_{k-1}(1, p_n), \tag{5}$$

where l is the likelihood defined by

$$l^n_k(1, p) := \begin{cases} \frac{n+1}{k+1} & \text{if } p \leq \frac{k+1}{n+1} \\ 0 & \text{otherwise,} \end{cases}$$

$$l^n_k(0, p) := \begin{cases} \frac{n+1}{n-k+1} & \text{if } p \geq \frac{k}{n+1} \\ 0 & \text{otherwise,} \end{cases}$$

and then we normalize them:

$$w^n_k := \tilde{w}^n_k \Big/ \sum_{k=0}^{m} \sum_{j=0}^{1} \tilde{w}^n_{k,j}. \tag{6}$$

Given the posterior weights for the previous step, we can find the predictive distribution for p_n as

$$p_n \sim \sum_{k=0}^{n-1} \sum_{j=0}^{1} w^{n-1}_{k,j} \left(\pi_{1|j} U \left[0, \frac{k+1}{n} \right] + \pi_{0|j} U \left[\frac{k}{n}, 1 \right] \right),$$

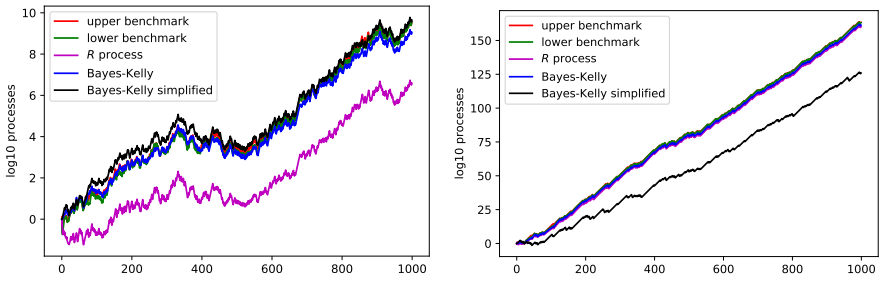

Figure 4: The Bayes–Kelly and Bayes–Kelly simplified conformal test martingales, the R-process, and the two benchmarks in the middle scenario. Left panel: hard case. Right panel: easy case (the trajectories for the two benchmarks, R process, and Bayes–Kelly virtually coincide).

where we use the shorthand $\pi_{0|j} := 1 - \pi_{1|j}$. Therefore, the betting functions for the resulting *Bayes–Kelly conformal test martingale* are

$$f_n(p) = \sum_{k=0}^{n-1} \sum_{j=0}^{1} w_{k,j}^{n-1} \left(\frac{n}{k+1} \pi_{1|j} 1_{p \le \frac{k+1}{n}} + \frac{n}{n-k} \pi_{0|j} 1_{p \ge \frac{k}{n}} \right). \tag{7}$$

The procedure is summarized as Algorithm 1.

For experimental results, see Figure 4, in addition to Figure 2. The Bayes–Kelly conformal test martingale appears to be very close to the two benchmarks. Its simplified version is described in the next section. The relatively poor performance of the R-process in the left panel of Figure 4 should not be interpreted as it being inferior to the Bayes–Kelly conformal test martingale: remember that R works against all Markov alternatives, whereas the other processes in Figures 2–8 are adapted to the specific alternative hypothesis (Markov$(0.4, 0.6)$ in the hard case and Markov$(0.1, 0.9)$ in the easy case).

5 Simplified Bayesian conformal testing

In this section we consider a radical simplification of the Bayes–Kelly conformal test martingale (7). We still assume that the Markov chain is symmetric, as in our model situations. If we assume that the weights $w_{k,j}^n$, $k = 0, \dots, n$, are concentrated at

$$k \approx k + 1 \approx n/2,$$

(7) will simplify to

$$f_n(p) = 2\pi_{1|j} 1_{p \le 0.5} + 2\pi_{0|j} 1_{p > 0.5}. \tag{8}$$

Figure 5 shows the weights (averaged over $j \in \{0, 1\}$) for the last step of the Bayes–Kelly conformal test martingale in the medium scenario (10^3 observa-

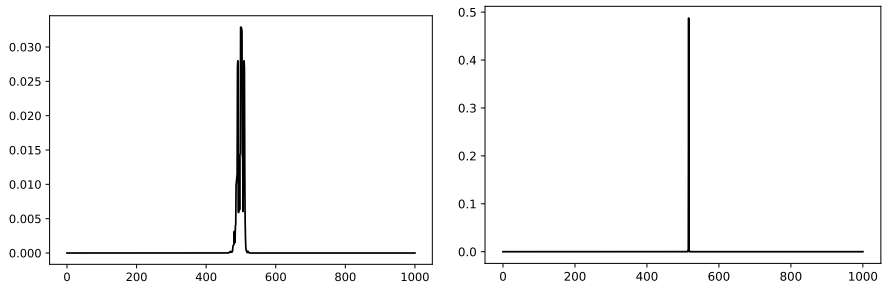

Figure 5: The weights w_k^{1000}, $k = 0, \dots, 1000$, at the last step for the Bayes–Kelly conformal test martingale in the medium scenario (the hard case on the left and easy on the right).

Algorithm 2 Simplified Bayes–Kelly $((p_1, p_2, \dots) \mapsto (S_1, S_2, \dots))$

$S_0 := S_1 := 1$
for $n = 2, 3, \dots$:
    if $p_{n-1} \le 0.5$:
        $j := 1$
    else:
        $j := 0$
    if $p_n \le 0.5$:
        $S_n := 2\pi_{1|j} S_{n-1}$
    else:
        $S_n := 2\pi_{0|j} S_{n-1}$

tions). They are indeed concentrated around values of k not so different from $0.5N = 500$. The procedure is summarized as Algorithm 2.

As a second step, we make (8) straightforward to compute by setting

$$j := \begin{cases} 1 & \text{if } p_{n-1} \le 0.5 \\ 0 & \text{if not.} \end{cases}$$

(If $k(n-1) := z_1 + \dots + z_{n-1} \approx (n-1)/2$, then $j = z_{n-1}$ with high probability.) The performance of the simplified version is shown in Figures 2–4 and 6. It is usually worse than that of the Bayes–Kelly conformal test martingale and the two benchmarks, but is comparable on the log scale apart from the right panel of Figure 6.

The right panel of Figure 6 and Figures 7 and 8 show that the statistical performance of the simplified Bayes–Kelly martingale particularly suffers in the easy case. The notches in the boxplots in Figures 7 and 8 indicate confidence intervals for the median.

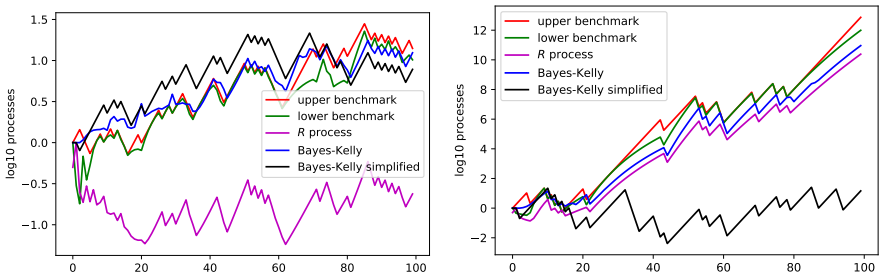

Figure 6: The analogue of Figures 2 and 4 for the small scenario.

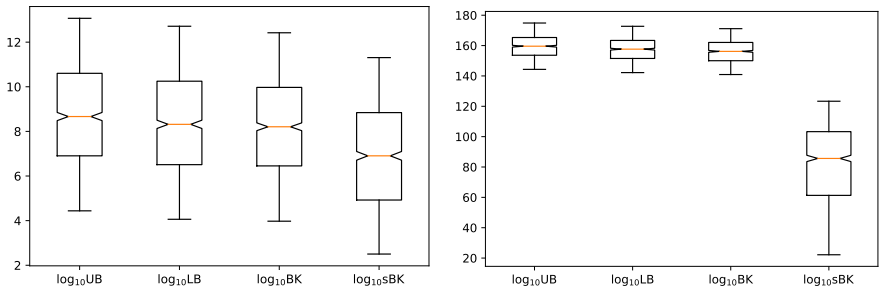

Figure 7: Boxplots based on 10^3 runs for the final values of the two benchmarks (upper UB and lower LB), the Bayes–Kelly conformal test martingale (BK), and its simplified version (sBK) in the medium scenario. Left panel: hard case. Right panel: easy case.

References

[1] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 4, April 2012. Conference version: ICML 2012.

[2] James R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1997.

[3] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. How can one test if a binary sequence is exchangeable? Fork-convex hulls, supermartingales, and Snell envelopes. Technical Report arXiv:2102.00630 [math.ST], arXiv.org e-Print archive, July 2021 (version 4).

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

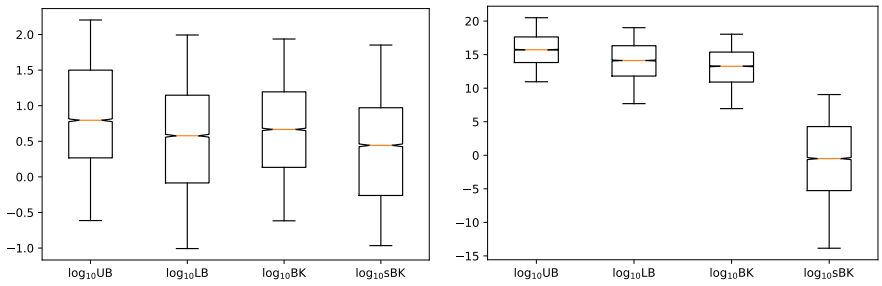

Figure 8: The analogue of Figure 7 for 10^4 runs in the small scenario.

[5] Vladimir Vovk. Testing randomness online, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 24, June 2019. Journal version: *Statistical Science* 36:595–611, 2021.

[6] Vladimir Vovk. Conformal testing in a binary model situation, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 33, April 2021. Conference version: COPA 2021.

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

A Asymmetric Markov alternatives

In the main part of this note we considered, following [3, Section 4.2],the case of symmetric Markov alternatives (i.e., the case $\forall i, j : \pi_{i|j} = \pi_{j|i}$). In this appendix we do not assume symmetry and only assume $\min_{i,j} \pi_{i|j} > 0$; in particular, the Markov chain is aperiodic and irreducible. We still assume that the initial distribution of the Markov chain is uniform (although Proposition 1 below only needs the initial distribution to be positive, i.e., both probabilities to be positive).

The definition of the lower benchmark (2) still works in the asymmetric case, but in the definition of the upper benchmark (1) we replace Ber(0.5) in the denominator by Ber(π_1), where π_1 is the probability of 1 under the stationary distribution for the Markov chain. By definition, the stationary distribution (π_0, π_1), where π_0 is the probability of 0, satisfies

$$\begin{cases} \pi_{0|0}\pi_0 + \pi_{0|1}\pi_1 = \pi_0 \\ \pi_{1|0}\pi_0 + \pi_{1|1}\pi_1 = \pi_1. \end{cases} \tag{9}$$

By the ergodic theorem [2, Theorem 1.10.2], this choice of the denominator for the likelihood ratio process makes the upper benchmark as close to the lower benchmark as possible asymptotically. The following proposition says that this choice of the denominator is asymptotically optimal.

Proposition 1. *For any* $x \in (0,1)$,

$$\frac{\mathrm{Markov}(\pi_{1|0},\pi_{1|1})([z_1,\dots,z_n])}{\mathrm{Ber}(x)([z_1,\dots,z_n])} > \frac{\mathrm{Markov}(\pi_{1|0},\pi_{1|1})([z_1,\dots,z_n])}{\mathrm{Ber}(\pi_1)([z_1,\dots,z_n])}$$

from some n *on almost surely under* $\mathrm{Markov}(\pi_{1|0},\pi_{1|1})$.

Proof. We have, almost surely as $n \to \infty$ (by the ergodic theorem and strong law of large numbers for martingales),

$$\begin{aligned}
&\frac{1}{n}\log\frac{\mathrm{Markov}(\pi_{1|0},\pi_{1|1})([z_1,\dots,z_n])}{\mathrm{Ber}(x)([z_1,\dots,z_n])} \\
&= \pi_{0|0}\pi_0 \log\frac{\pi_{0|0}}{1-x} + \pi_{1|0}\pi_0\log\frac{\pi_{1|0}}{x} + \pi_{0|1}\pi_1\log\frac{\pi_{0|1}}{1-x} + \pi_{1|1}\pi_1\log\frac{\pi_{1|1}}{x} + o(1) \\
&= \pi_{0|0}\pi_0 \log\frac{1}{1-x} + \pi_{1|0}\pi_0\log\frac{1}{x} + \pi_{0|1}\pi_1\log\frac{1}{1-x} + \pi_{1|1}\pi_1\log\frac{1}{x} + c + o(1) \\
&= \pi_0 \log\frac{1}{1-x} + \pi_1\log\frac{1}{x} + c + o(1) > \pi_0\log\frac{1}{\pi_0} + \pi_1\log\frac{1}{\pi_1} + c + o(1) \\
&= \frac{1}{n}\log\frac{\mathrm{Markov}(\pi_{1|0},\pi_{1|1})([z_1,\dots,z_n])}{\mathrm{Ber}(\pi_1)([z_1,\dots,z_n])} + o(1),
\end{aligned}$$

where c is a constant (depending only on the πs), the penultimate "=" follows from (9), and the last inequality, ">", disregards the $o(1)$ terms and follows from the positivity of the Kullback–Leibler distance in this context. □

The Bayes–Kelly conformal test martingale (Algorithm 1) also works for asymmetric Markov chains. Let us derive the simplified Bayes–Kelly conformal test martingale (Algorithm 2) in the non-symmetric case. The solution to (9) is

$$\pi_1 = \frac{\pi_{1|0}}{\pi_{1|0} + \pi_{0|1}}.$$

When

$$k \approx k+1 \approx n\pi_1,$$

(7) will lead to

$$f_n(p) = \frac{\pi_{1|j}}{\pi_1}1_{p\le\pi_1} + \frac{\pi_{0|j}}{\pi_0}1_{p>\pi_1}$$

in place of (8). It remains to set

$$j := \begin{cases} 1 & \text{if } p_{n-1} \le \pi_1 \\ 0 & \text{if not.} \end{cases}$$

Examples of the performance of various processes in simulation studies with asymmetric Markov alternatives are shown in Figures 9 and 10 (the poor performance of the R process in the left panel of Figure 9 should be ignored, since the comparison is not fair, as discussed earlier). In the *semi-hard case* the model is $\mathrm{Markov}(0.4, 0.5)$, and in the *semi-easy case*, the model is $\mathrm{Markov}(0.1, 0.5)$.

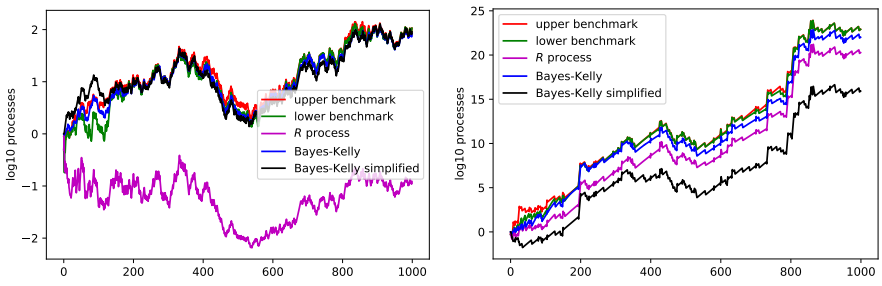

Figure 9: The analogue of Figures 2, 4, and 6 for the medium scenario and the semi-hard (left) and semi-easy (right) cases.

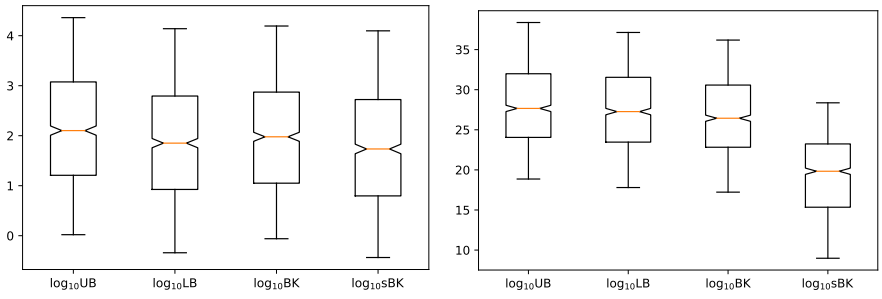

Figure 10: The analogue of Figures 7 and 8 for the medium scenario and the semi-hard (left) and semi-easy (right) cases.