

Conformal testing: binary case with Markov alternatives

Vladimir Vovk, Ilija Nouretdinov, and Alex Gammerman

Centre for Reliable Machine Learning
Department of Computer Science
Royal Holloway, University of London

COPA 2022
University of Brighton
25 August, 2022

Plan

- 1 Conformal testing
- 2 Useful benchmarks
- 3 Comparisons (pictures)

Conformal testing and this paper

- Conformal prediction depends on the IID (or exchangeability) assumption.
- Under the IID assumption, conformal p-values are IID and uniformly distributed on $[0, 1]$.
- This is at the basis of conformal prediction but can also be used for **conformal testing**: if we want to test the IID assumption, we can instead test the independence and uniformity of the p-values.
- This turns the composite (and massive) null hypothesis of exchangeability into a simple null hypothesis.
- A possible application: when do we retrain a prediction algorithm (traditional one, or conformal predictor)?

Batch vs online hypothesis testing

- The usual mode of testing in statistics is **batch**: we are given a batch of data, and the task is to decide whether to reject the null hypothesis.
- In the **online mode**, we start from a unit capital and keep gambling against the null hypothesis (making sure capital ≥ 0).
- Our current capital then measures the degree to which the null hypothesis has been falsified.
- By the Ville inequality: the probability our capital ever exceeds c is at most $1/c$.
- In many cases the online mode is more relevant: think, e.g., of retraining prediction algorithms.
- It's becoming a popular direction of research: “game-theoretic statistics” (Shafer, Grünwald, Ramdas, Wang, ...).

Efficiency of conformal testing

- Conformal testing is a valid mode of testing.
- But is it efficient?
- For a long time I doubted it was, but actually it might be.
- In simple cases where we have natural benchmarks, it is competitive with the benchmarks.
- COPA 2021: efficiency in the problem of changepoint detection.
- This paper and talk: efficiency against Markov alternatives to IID.

Conformal test martingales

- The process describing the evolution of our capital when gambling against conformal p-values: “conformal test martingale”.
- For gambling against the uniformity of the p-values we use **betting functions**, i.e., functions $f : [0, 1] \rightarrow [0, \infty]$ that integrate to 1.
- In conformal testing, at step n a betting function f_n is chosen (in a measurable manner) with the knowledge of the first $n - 1$ p-values p_1, \dots, p_{n-1} .
- The product $S_n := f_1(p_1) \dots f_n(p_n)$, $n = 0, 1, \dots$ (with $S_0 := 1$), is the corresponding **conformal test martingale**.

Gambling against a Markov alternative

- The details can be found in the paper; my description will be high-level.
- We are interested in the binary case: we observe z_1, z_2, \dots , and the observations are $z_i \in \{0, 1\}$.
- The null hypothesis: IID; we are flipping a coin with the same probability of 1 (= "heads").
- Suppose that in fact z_i are generated by a Markov distribution.
- It is sufficient to use the trivial nonconformity measure: the nonconformity score is $\alpha_i := z_i$.

Details of gambling (1)

- Under the alternative hypothesis, the distribution of the conformal p-values is not uniform.
- It is known that the optimal betting function is the true probability density (“Kelly gambling”); shown in, e.g., the paper by Fedorova et al. (ICML 2012).
- If we know the true data-generating distribution, we can compute the probability density for p_n (the n th conformal p-value) after observing p_1, \dots, p_{n-1} using Bayesian methods.
- The “parameter” is z_1, z_2, \dots ; the prior distribution is the Markov alternative; and the “observations” p_n are computed using the usual formulas of conformal prediction.

Details of gambling (2)

- In the paper we have two versions of our conformal test martingale: proper (“Bayes–Kelly”) and simplified (“simplified Bayes–Kelly”).
- The **Bayes–Kelly martingale** uses as the betting function the posterior distribution of p_n given p_1, \dots, p_{n-1} .
- It can be implemented efficiently as an algorithm that maintains weights for the parameters.
- Its sufficient to have the “aggregated” weights $w_{k,L}^n$ at time n , where k is the number of 1s so far and L is the last bit z_n .
- But $w_{k,L}^n \approx 0$ outside a very narrow interval of k ($k \approx n/2$ in the symmetric case).
- The **simplified Bayes–Kelly martingale**: we just ignore the ws outside the expected value of k .

Two ways of mixing

- I have explained how to gamble against a simple alternative.
- What to do if the alternative is composite?
- For each element of the composite alternative we construct a conformal test martingale, and then average all those martingales.
- Two natural ways of doing so:
 - We can do it on paper, evaluating all the integrals.
 - Or we can replace the composite alternative by a dense grid and add an external loop to our code.

Plan

- 1 Conformal testing
- 2 Useful benchmarks
- 3 Comparisons (pictures)

Ramdas et al.'s work



Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen (2022).

Testing exchangeability: fork-convexity, supermartingales, and e-processes.

International Journal of Approximate Reasoning
141:83–109 (Glenn Shafer Special Issue).

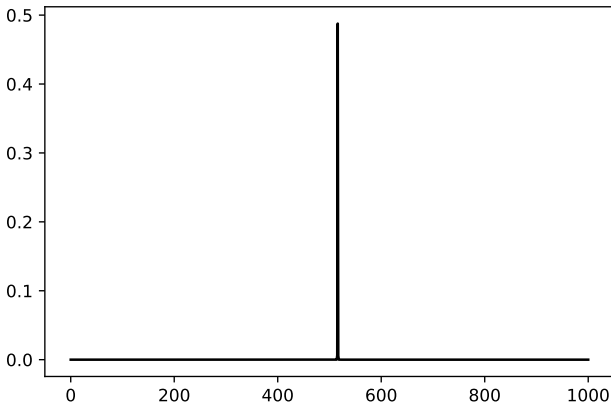
- They introduce a process R_n that is, essentially, a test martingale under the IID assumption (more later).
- The process is designed to be efficient for all Markov alternatives, but the construction only works in toy situations (such as binary with Markov or changepoint alternatives).

Model situation

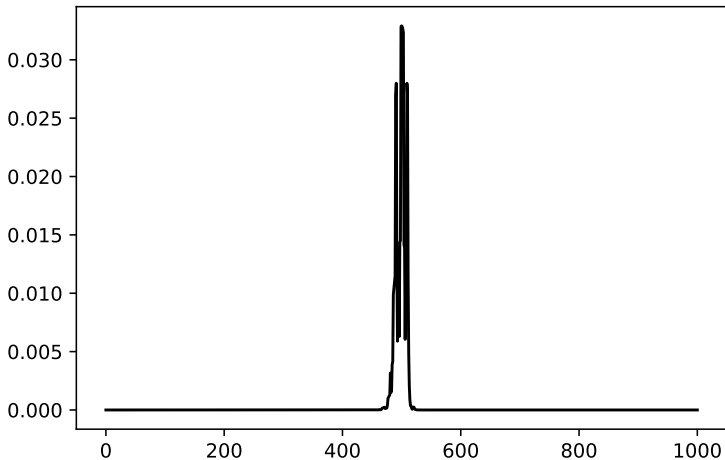
- In choosing the model situation, we essentially follow Ramdas et al.
- $\text{Markov}(\pi_{1|0}, \pi_{1|1})$ is the probability distribution of a Markov chain with the transition probabilities $\pi_{1|0}$ for transitions $0 \rightarrow 1$ and $\pi_{1|1}$ for transitions $1 \rightarrow 1$. The probability that the first observation is 1 will always be assumed 0.5 (plays a very minor role).
- Our null hypothesis is, essentially, that $\pi_{1|0} = \pi_{1|1}$.
- We consider two cases:
 - In the **hard case**, the model is $\text{Markov}(0.4, 0.6)$.
 - In the **easy case**, the model is $\text{Markov}(0.1, 0.9)$.
 - So that the hard case is harder to distinguish from the null hypothesis than the easy case.
- The number of observations is $N := 10^3$.

Simplifying Bayes–Kelly (easy case)

Here we plot $w_{k,0}^n + w_{k,1}^n$ vs k (at the last step, $n = 10^3$).



Simplifying Bayes–Kelly (hard case)



Benchmarks (1)

- The **upper benchmark** is

$$UB_n := \frac{\text{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\text{Ber}(0.5)([z_1, \dots, z_n])},$$

where $\text{Ber}(\pi)$ is the coin-tossing distribution with probability of success π .

- The **lower benchmark** is

$$LB_n := \frac{\text{Markov}(\pi_{1|0}, \pi_{1|1})([z_1, \dots, z_n])}{\text{Ber}(\hat{\pi})([z_1, \dots, z_n])},$$

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, $UB_0 = LB_0 := 1$.

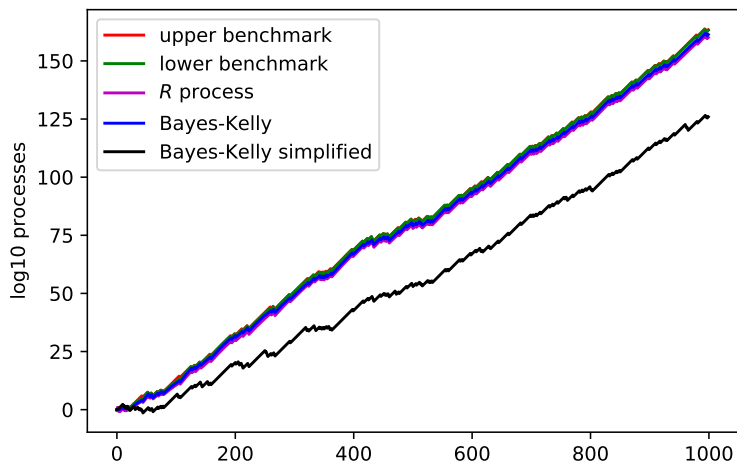
Benchmarks (2)

- The upper benchmark (a likelihood ratio) is a martingale only under $\text{Ber}(0.5)$ (and not under any other element of the null hypothesis), and so impossible to attain with “honest” methods such as conformal testing.
- The lower benchmark is valid under any element of the null hypothesis, but it does not generalize to complicated non-binary cases.
- Ramdas et al.’s process R : mixture of the lower benchmark over all Markov alternatives (over a Jeffreys-type prior).

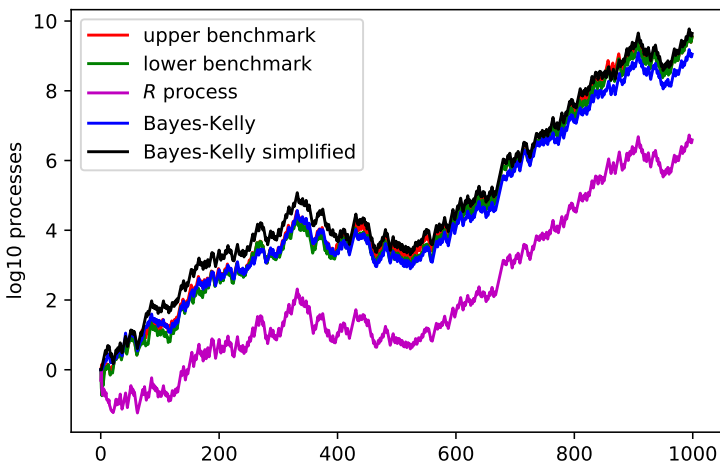
Plan

- 1 Conformal testing
- 2 Useful benchmarks
- 3 Comparisons (pictures)

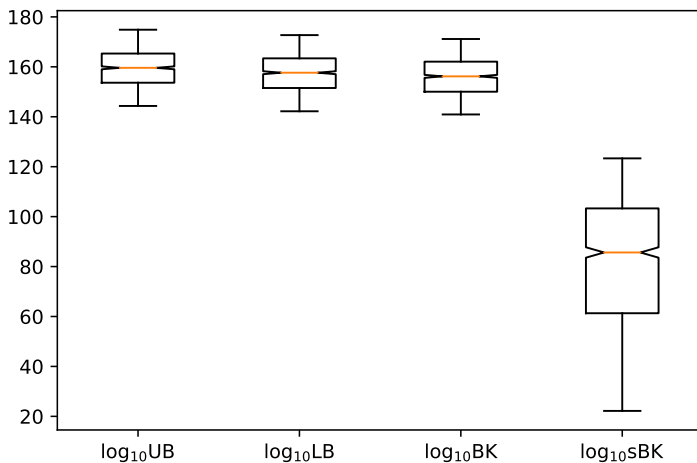
Paths in the easy case



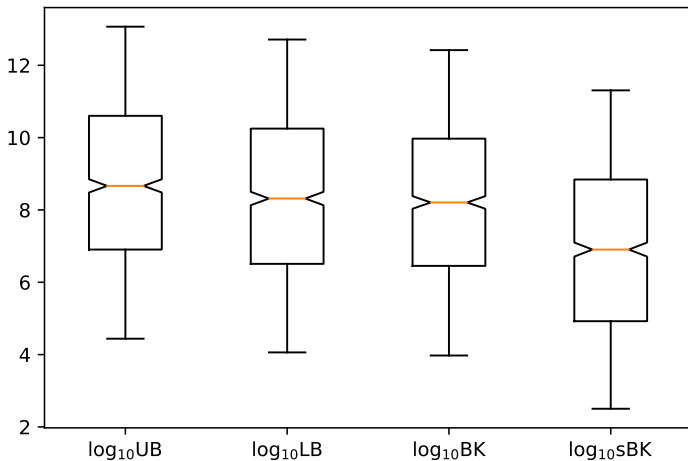
Paths in the hard case



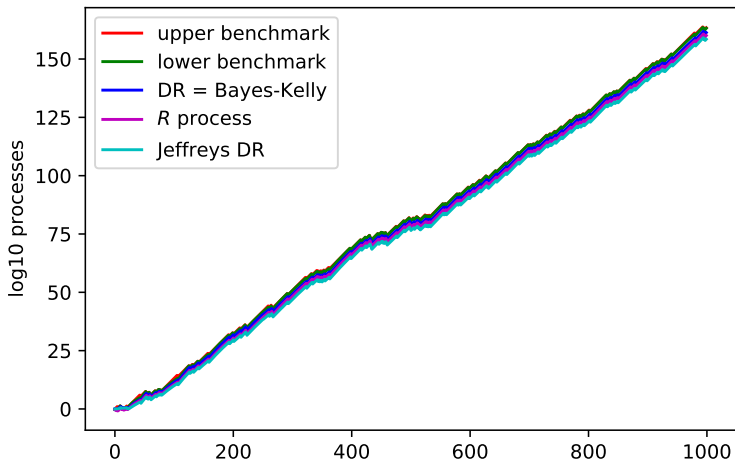
Boxplots in the easy case (10^3 simulations)



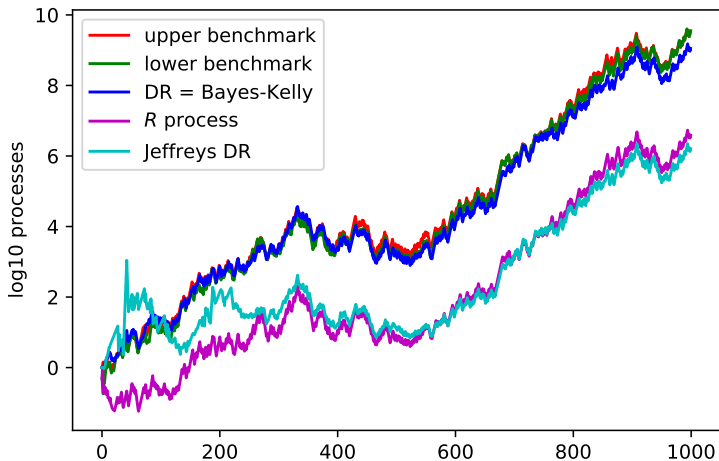
Boxplots in the hard case





Easy case (in the book)



Hard case (in the book)



References

-  Valentina Fedorova, Ilya Nourtdinov, Alex Gammerman, and Vladimir Vovk.
[Plug-in martingales for testing exchangeability on-line.](#)
ICML 2012, pp. 1639–1646; also arXiv.
Kelly gambling in conformal prediction.
-  Vladimir Vovk, Alex Gammerman, and Glenn Shafer.
[Algorithmic learning in a random world.](#)
New York: Springer, 2022.
There is a new chapter on the efficiency conformal testing (Chapter 9).

Thank you for your attention!