The power of forgetting in statistical hypothesis testing

Vladimir Vovk

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

COPA 2023
Limassol, Cyprus
14 September, 2023

Plan

1. Different ways of testing
2. Need for forgetting
3. Positive results

Testing statistical models

- Mainstream machine learning is based on the IID model. (Assumes it, modifies it, etc.)
- How do we test a composite statistical hypothesis? (Null hypothesis, such as IID, which is massive.)
- The view that I had accepted since I was a PhD student working in the algorithmic theory of randomness: to test a composite null,
  - test each of its elements;
  - reject it if each element is rejected.

  That is, test element-wise.
- This talk: this does not work in dynamic hypothesis testing (in particular, in conformal testing).

Dynamic testing with sequential observations (1)

- The usual way of testing in statistics: batch testing (Neyman–Pearson, Fisher).
- Dynamic testing of a simple null hypothesis: by a test martingale.
- The notion of a martingale depends on the available information; formally, on a filtration $\mathcal{F}_0, \mathcal{F}_1, \ldots$, whose elements $\mathcal{F}_n$ are called σ-algebras.
- The most basic filtrations:
  - the natural filtration is where $\mathcal{F}_n$ is generated by (i.e., carries the same information as) the first n observations.
  - the conformal filtration associated with a given conformal predictor: $\mathcal{F}_n$ is generated by the first n p-values.

Dynamic testing with sequential observations (2)

- A process (more fully, an adapted process) is a sequence of random variables $S_0, S_1, \dots$ such that S_n is $\mathcal{F}_n$-measurable ($\mathcal{F}_n$ determines S_n) for all n.
- A process S_n is a test martingale if $S_0 = 1$, it is nonnegative, and $\mathbb{E}(S_{n+1} \mid \mathcal{F}_n) = S_n$ [fairness] for all n.
- The fairness condition on the relative increment

$$\mathbb{E}\left(\frac{S_{n+1}}{S_n} \mid \mathcal{F}_n\right) = 1$$

might be easier to interpret.

Element-wise testing

- We are interested in testing a statistical model $(P_\theta \mid \theta \in \Theta)$.
- For each $\theta \in \Theta$, we fix a test martingale S^θ and define

$$S_n := \inf_{\theta \in \Theta} S_n^\theta.$$

- Any process S that can be obtained in this way: element-wise test. (And this is element-wise testing.)

Pivotal testing (example)

- Origin: Fisher's fiducial statistics.
- Only works for tiny statistical models. No machine learner would be interested in it.
- Suppose $Z_1, Z_2, \ldots$ are IID Gaussian $\mathcal{N}_{\mu,1}$ random variables. Then $Z'_1 := 0$ and

$$Z'_n := Z_n - Z_1, \quad n \geq 2,$$

have a known distribution and we can gamble against them (getting a martingale in the filtration $\mathcal{F}'_n := \sigma(Z'_1, \ldots, Z'_n)$).

Conformal testing

- The main property of validity of conformal prediction: the conformal p-values are distributed uniformly in $[0,1]^\infty$.
- Gambling against them gives us a conformal test martingale (in the conformal filtration).
- More interesting than pivotal testing; we can, e.g., decide when to retrain a machine-learning algorithm.

Suggested general testing scheme

- Way of constructing test martingales suggested in the paper.
- Conglomerate of all the useful tricks I know:
  - reducing the filtration
  - minimization over $\theta \in \Theta$
  - using a random number generator
- It looks awkward, and it would be good to have a provably general method.

Plan

Testing the IID assumption

- Suppose our null hypothesis is that the observations are IID.
- If we want to test with one martingale, no testing is possible without forgetting: there are no restrictions whatsoever on the distribution of the next observation, and so the capital can only go down.
- Conformal prediction works only because of forgetting (conformal martingales are martingales in a reduced filtration, namely the conformal filtration).

Example for pivotal testing

- The process

$$S_n := \begin{cases} 1 & \text{if } n \leq 1 \\ 1/\mathcal{N}_{0,2}([-1,1]) & \text{if } n \geq 2 \text{ and } Z'_2 \in [-1,1] \\ 0 & \text{if } n \geq 2 \text{ and } Z'_2 \notin [-1,1]. \end{cases}$$

  is a martingale in the reduced filtration $\mathcal{F}'_n$.
- In the paper I show that it cannot be lower bounded by the minimum of test martingales in the natural filtration.

Example for conformal testing

- The example for conformal testing (again given in the paper) is a simple modification of the example in the pivotal case.
- However, this example is based on an online compression model different from the standard IID (or exchangeability) model (namely, it uses the “Gaussian compression model with variance 1”).
- There is no doubt a similar example can be constructed for the IID model [to do].

Plan

1. Different ways of testing
2. Need for forgetting
3. Positive results

Proposition for a finite horizon

Let $N \in \{1, 2, \dots\}$ be a finite horizon (i.e., we are only interested in test martingales $S_0, \dots, S_N$ w.r.t. to a filtration $\mathcal{F}_0, \dots, \mathcal{F}_N$).

Proposition

Let (S^θ) be a family of test martingales w.r. to the same filtration (perhaps not natural) and a statistical model (P_θ). Then there exists a family of natural test martingales $(\tilde{S}^\theta)$ such that

$$\inf_{\theta \in \Theta} \tilde{S}^\theta_N = \inf_{\theta \in \Theta} S^\theta_N.$$

Proof: by backward induction.

Proposition for an infinite horizon

Proposition

Suppose that the parameter set Θ *is finite and that different* P_θ *in the statistical model (our null hypothesis)* (P_θ) *are mutually singular. Let* (S^θ) *be a family of test martingales w.r.t. to the same filtration and* (P_θ)*, and let* $\epsilon > 0$ *(be arbitrarily small). Then there exists a family of natural test martingales* $(\tilde{S}^\theta)$ *such that*

$$\liminf_{n\to\infty} \inf_\theta \tilde{S}_n^\theta \ge (1-\epsilon) \limsup_{n\to\infty} \inf_\theta S_n^\theta$$

a.s. under any probability measure P_θ *from the null hypothesis.*

Change detection problem (1)

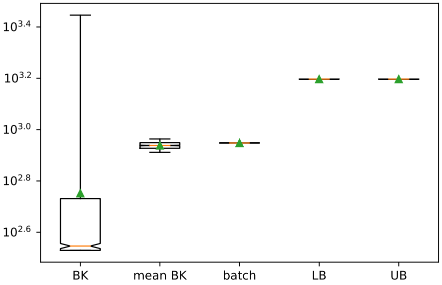
- In the paper I apply the idea of backward induction to a change detection problem with a finite horizon to obtain a new test martingale (“batch benchmark”).
- The observation space is $\mathbf{Z} := \{0, 1\}$, the null hypothesis is the IID model $(\mathbf{B}_\theta^{20} \mid \theta \in [0, 1])$, the alternative hypothesis Q is that 10 observations are generated from $\mathbf{B}_{0.1}$, and another 10 observations are generated from $\mathbf{B}_{0.9}$. All 20 observations are generated independently.

Change detection problem (2)

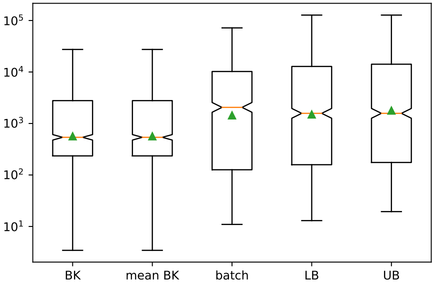
Previous work:

- BK (Bayes–Kelly conformal test martingale),
- mean BK (not a test martingale),
- LB (lower benchmark, valid but works only in toy cases),
- UB (upper benchmark, not valid).

fixed dataset:



random dataset:



References

Vladimir Vovk, Alex Gammerman, and Glenn Shafer.
Algorithmic learning in a random world.
New York: Springer, 2022.
Part III.

My paper in the Proceedings.

Thank you for your attention!