Proceedings of Machine Learning Research 266:1–18, 2025 Conformal and Probabilistic Prediction with Applications

# Testing Exchangeability for Multiple Sequences of P-values

**Henrik Boström** BOSTROMH@KTH.SE
*School of Electrical Engineering and Computer Science,*
*KTH Royal Institute of Technology, Sweden*

**Editor:** Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

## Abstract

Given a sequence of p-values, conformal test martingales can be used for signaling that the exchangeability assumption is violated, while the false alarm rate is controlled by a user-specified significance level. In some scenarios, multiple p-values are observed at each time step, e.g., p-values may be received from multiple conformal predictors for a single target, or p-values are obtained for multiple targets. In such cases, signaling whenever a violation is detected for any of the sequences, leads to an increased risk of false alarms. Bonferroni correction, which is a standard approach to controlling the error rate when testing multiple hypotheses, is shown to be dominated by the straightforward approach of forming a single conformal test martingale from the martingales generated from the individual sequences of p-values. In addition to testing exchangeability for the individual sequences, approaches for testing them jointly are also investigated. For the latter, the use of aggregation operators to transform multiple sequences of p-values into a single sequence is investigated, as well as a previously proposed approach for detecting covariate shift. Experimental results are presented, highlighting the potential strengths and weaknesses of the different approaches.
**Keywords:** Conformal test martingales, exchangeability, multiple sequences

## 1. Introduction

Conformal test martingales have been proposed as an approach for testing exchangeability in online scenarios, where p-values are observed over time (Volkhonskiy et al., 2017; Vovk, 2021; Vovk et al., 2021). The ability to test this property is crucial in many cases, where guarantees of employed methods, e.g., the validity of conformal predictors, are relying on an assumption that this property holds. The false alarm rate, i.e., the probability that the martingale incorrectly signals that the exchangeability assumption is violated, is in these approaches controlled by a user-specified significance level. In some scenarios, multiple p-values are observed at each time step, e.g., p-values may be received from multiple conformal predictors for a single target (Eliades and Papadopoulos, 2023), or the prediction task concerns multiple targets (Messoudi et al., 2020). In such cases, signaling whenever the assumption is violated for any of the sequences, leads to that the false alarm rate may exceed the user-specified significance level.

In this paper, we will investigate approaches to testing exchangeability for multiple sequences of p-values to control the false alarm rate when signaling for violations of the exchangeability assumption, either for any of the individual sequences or for the joint sequence. For testing sequences individually, we will consider the standard approach of employing Bonferroni correction as well as an approach that combines the martingales generated for each individual sequence of p-values by uniform weighting. We will show that

the former is an inadmissible approach, as it is dominated by the latter, i.e., the martingale obtained by Bonferroni correction is upper-bounded by the combined martingale. We will then continue with approaches for testing exchangeability of the joint sequence; these are all based on the idea of transforming the original sequences of p-values into a single sequence of scores, from which a single sequence of p-values, and consequently a single conformal test martingale, can be generated. We will consider a set of aggregation operators, including previously proposed approaches for averaging p-values (Vovk and Wang, 2020), as well as their combination. Finally, we will investigate an approach for detecting covariate shift using conformal test martingales, described in (Vovk et al., 2022, p. 247-248), here using the p-values for a given time step as covariates. This approach relies on using a reference set of objects to which the Euclidean distance is computed, but rather than choosing a fixed size of this reference set, which can be viewed as a hyperparameter of the approach, we consider a slight extension which forms a combined martingale for different values of this parameter.

The main contributions of the paper are:

- The uniformly weighted conformal test martingale is shown to dominate the Bonferroni corrected martingale, when testing sequences of p-values individually.
- It is shown that a joint sequence of exchangeable sequences may not be exchangeable, while a joint sequence formed from non-exchangeable sequences may be exchangeable.
- Novel algorithms for testing joint sequences of p-values are introduced.
- The effectiveness of the algorithms is empirically investigated when applied to p-values from multiple conformal predictors for both single and multiple targets.

In the next section, we provide the terminology employed in the paper. In Section 3, we will outline the different approaches to handling multiple sequences of p-values, which are evaluated experimentally in Section 4. Finally, we summarize the main findings and outline some directions for future work in Section 5.

## 2. Conformal Test Martingales

Given a sequence of examples $z_1, z_2, \ldots$, where each $z_i \in \mathbf{Z}$, for some example space $\mathbf{Z}$, a sequence of nonconformity scores $\alpha_1 = A(z_1)$, $\alpha_2 = A(z_2)$, $\ldots$ is obtained using some nonconformity function $A : \mathbf{Z} \to \mathbb{R}$. It should be noted that the example space $\mathbf{Z}$ may consist of, e.g., objects and/or labels, and the employed nonconformity function $A$, which here is assumed to be fixed prior to observing the sequence of examples, is assumed to handle the examples accordingly.

From the sequence of nonconformity scores $\alpha_1, \alpha_2, \ldots$ and a sequence of values $\tau_1, \tau_2, \ldots$ that are sampled IID and uniformly from the unit interval, a (smoothed) p-value for the $n$th nonconformity score is computed in the *semi-online* mode by:

$$p_n = \frac{|i : \alpha_i > \alpha_n| + \tau_n |i : \alpha_i = \alpha_n|}{n} \tag{1}$$

where $i \in \{1, \ldots, n\}$ (Vovk et al., 2022). If the examples $z_1, z_2, \ldots$ are IID, or if the examples are exchangeable and we consider a finite horizon, then the p-values are IID and distributed uniformly on $[0, 1]$ (Vovk, 2021).

For each p-value in the sequence, a *conformal test martingale* is computed by:

$$s_n = f_1(p_1) \cdots f_n(p_n) \tag{2}$$

where each $f_i$ is a *betting function* $f : [0, 1] \rightarrow [0, \infty]$ satisfying:

$$\int_0^1 f(u)du = 1 \tag{3}$$

If the sequence of examples $z_1, z_2, \ldots$ is exchangeable, then Ville's inequality implies (Vovk et al., 2022):

$$P(\{(z_1, z_2, \ldots) : \exists n : s_n \geq c\}) \leq 1/c \tag{4}$$

where $c$ is some positive constant.

It should hence be noted that the conformal test martingale can be used for signaling that the assumption of exchangeability does not hold, while controlling the false alarm rate (by choosing a suitable $c$). For example, a common choice is to set $c = 100$, which means that the probability of ever observing a conformal test martingale greater than or equal to this, for an exchangeable sequence of examples, is less than or equal to 1%.

In addition to choosing a suitable nonconformity function, also a betting function and a suitable formulation of the conformal test martingale have to be selected when implementing the framework. In this work, we have chosen the betting function (5), referred to by (9.32) in (Vovk et al., 2022, p. 282), and the Sleeper/Stayer algorithm (Vovk et al., 2022, p. 283). In Alg. 1, we present a slight modification of the original algorithm; investments are here distributed before computing the first martingale, which otherwise always will be 1.

$$f_{a,b}(p) = \begin{cases} \frac{b}{a} & \text{if } p \leq a \\ \frac{1-b}{1-a} & \text{otherwise} \end{cases} \tag{5}$$

**Algorithm 1:** Sleeper/Stayer

**Input:** $p_1, p_2, \ldots$ (p-values), $r \in (0, 1)$ (rate of investment), $g \in \mathbb{N}^+$ (grid size)

$G \leftarrow \{(i/g, j/g) : (i, j) \in \{1, \ldots, g-1\} \times \{1, \ldots, g-1\}\}$
$S_\square \leftarrow 1 - r$
**for** $(a, b) \in G$ **do** $S_{a,b} = r/|G|$
**for** $i = 1, 2, \ldots$ **do**
  **for** $(a, b) \in G$ **do** $S_{a,b} = S_{a,b} f_{a,b}(p_i)$
  $s_i \leftarrow S_\square + \sum_{(a,b)\in G} S_{a,b}$
  **for** $(a, b) \in G$ **do** $S_{a,b} = S_{a,b} + rS_\square/|G|$
  $S_\square = (1 - r)S_\square$
**end**
**return** $s_1, s_2, \ldots$

We will also need the following definition of a joint sequence, to denote a sequence of tuples formed from an ordered set of individual sequences.

**Definition 1 (Joint sequence)** *Given an ordered set of sequences $P_1, \ldots, P_m$, where $P_i = p_{i,1}, p_{i,2}, \ldots$, for $i = 1, \ldots, m$, the joint sequence, denoted $\otimes(P_1, \ldots, P_m)$, is*

$$\otimes(P_1, \ldots, P_m) = (p_{1,1}, \ldots, p_{m,1}), (p_{1,2}, \ldots, p_{m,2}), \ldots$$

## 3. Testing Multiple Sequences of P-values

In some scenarios, we may observe more than one sequence of p-values simultaneously, e.g., if we employ multiple nonconformity functions, e.g., based on different algorithms or data, or use multiple example spaces, e.g., with different sets of objects, labels, or both. Ensembles of conformal predictors are examples of the former, while conformal predictors for multitarget prediction tasks are natural examples of the latter.

More formally, we consider a scenario where we observe $m > 1$ aligned sequences of p-values $P_1, \ldots, P_m$, where $P_i = p_{i,1}, p_{i,2}, \ldots,$ for $i = 1, \ldots, m$, and where we at each time step $t \in \mathbb{N}^+$ observe the p-values $p_{1,t}, \ldots, p_{m,t}$. We consider the following two main questions:

i) Is the exchangeability assumption violated for any of the sequences $P_1, \ldots, P_m$?

ii) Is the exchangeability assumption violated for the joint sequence $\otimes(P_1, \ldots, P_m)$?

We will in this section first show why the two questions need to be answered separately and then consider approaches for answering them.

### 3.1. Testing Individual vs. Joint Sequences

We will here show why the two addressed questions, i.e., if the exchangeability assumption is violated for any of the individual sequences and for the joint sequence, respectively, need to be answered separately. This follows from the theorems below stating that a joint sequence may not be exchangeable even if this holds for the individual sequences, and that an exchangeable sequence may be obtained by aggregation, even if the individual sequences are not exchangeable.

**Theorem 2** *For an ordered set of exchangeable sequences, $P_1, \ldots, P_m$, it does not necessarily hold that the joint sequence $\otimes(P_1, \ldots, P_m)$ is exchangeable.*

**Proof** Assume that we with equal probabilities observe any of the following six pairs of series:

Case 1:
$$P_1 = \langle 1, 2, 3 \rangle$$
$$P_2 = \langle 1, 2, 3 \rangle$$
$$\otimes(P_1, P_2) = \langle (1,1), (2,2), (3,3) \rangle$$

Case 2:
$$P_1 = \langle 1, 3, 2 \rangle$$
$$P_2 = \langle 1, 3, 2 \rangle$$
$$\otimes(P_1, P_2) = \langle (1,1), (3,3), (2,2) \rangle$$

Case 3:
$$P_1 = \langle 2, 1, 3 \rangle$$
$$P_2 = \langle 2, 3, 1 \rangle$$
$$\otimes(P_1, P_2) = \langle (2,2), (1,3), (3,1) \rangle$$

Case 4:
$$P_1 = \langle 2, 3, 1 \rangle$$
$$P_2 = \langle 2, 1, 3 \rangle$$
$$\otimes(P_1, P_2) = \langle (2,2), (3,1), (1,3) \rangle$$

Case 5:
$$P_1 = \langle 3, 1, 2 \rangle$$
$$P_2 = \langle 3, 2, 1 \rangle$$
$$\otimes(P_1, P_2) = \langle (3,3), (1,2), (2,1) \rangle$$

Case 6:
$$P_1 = \langle 3, 2, 1 \rangle$$
$$P_2 = \langle 3, 1, 2 \rangle$$
$$\otimes(P_1, P_2) = \langle (3,3), (2,1), (1,2) \rangle$$

Then it follows that both $P_1$ and $P_2$ are exchangeable, i.e., all possible permutations are equally probable, while this does not hold for $\otimes(P_1, P_2)$, as, e.g., the probability of observing the sequence $\langle (3,3), (2,2), (1,1) \rangle$ is zero, while the probability of observing other permutations of this sequence, including $\langle (1,1), (2,2), (3,3) \rangle$, is non-zero.
■

**Theorem 3** *An exchangeable sequence may be obtained by aggregation from a joint sequence $\otimes(P_1, \ldots, P_m)$, even if not all the sequences $P_1, \ldots, P_m$ are exchangeable.*

**Proof** Assume that we employ the aggregation operator $F_{Double}$ (10) to obtain scores from the joint sequence $\otimes(P_1, P_2)$, where we with equal probabilities observe any of the following four pairs of series:

Case 1:
$$P_1 = \langle 1, 2 \rangle$$
$$P_2 = \langle 3, 4 \rangle$$
$$F_{Double}(\otimes(P_1, P_2)) = \langle 4, 6 \rangle$$

Case 2:
$$P_1 = \langle 1, 2 \rangle$$
$$P_2 = \langle 3, 4 \rangle$$
$$F_{Double}(\otimes(P_1, P_2)) = \langle 4, 6 \rangle$$

Case 3:
$$P_1 = \langle 1, 2 \rangle$$
$$P_2 = \langle 5, 2 \rangle$$
$$F_{Double}(\otimes(P_1, P_2)) = \langle 6, 4 \rangle$$

Case 4:
$$P_1 = \langle 2, 1 \rangle$$
$$P_2 = \langle 4, 3 \rangle$$
$$F_{Double}(\otimes(P_1, P_2)) = \langle 6, 4 \rangle$$

Then it follows that neither of $P_1$ and $P_2$ are exchangeable, i.e., the possible permutations are not equally probable, while this holds for $F_{Double}(\otimes(P_1, P_2))$.
■

### 3.2. Testing Individual Sequences

Maintaining one sequence of conformal test martingales $S_i = s_{i,1}, s_{i,2}, \ldots$ for each sequence of p-values $P_i$, for $i = 1, \ldots, m$, where $m > 1$, and using the same specified threshold ($c$), as when observing a single sequence, is clearly not a viable approach as the risk of a false alarm increases with the number of sequences ($m$). We consider two approaches for controlling the false alarm rate; employing Bonferroni correction and combining the individual conformal test martingales.

#### 3.2.1. Bonferroni Correction

A standard approach of avoiding an increase of the likelihood of making a type I error (incorrectly rejecting a null hypothesis) due to testing multiple hypotheses is to employ Bonferroni correction, i.e., divide the significance level with the number of hypotheses tested. In our case, where the original significance level is $\frac{1}{c}$ and the number of hypotheses $m$, the Bonferroni corrected significance level becomes $\frac{1}{mc}$ and an alarm should be raised whenever one of the martingales presents a value greater than or equal to $mc$. Equivalently, the original threshold ($c$) can be kept if we instead divide each martingale by $m$.

**Definition 4 (Bonferroni corrected martingale)**
*Given a set of conformal test martingales $\{S_1, \ldots, S_m\}$, the Bonferroni corrected martingale $S_B = s_{B,1}, s_{B,2}, \ldots$ is defined by:*

$$s_{B,t} = \max(\frac{s_{1,t}}{m}, \ldots, \frac{s_{m,t}}{m}) \tag{6}$$

*for all $t \in \mathbb{N}^+$.*

#### 3.2.2. Combining Conformal Test Martingales

As an alternative to the Bonferroni correct martingale, we may form a combined martingale by taking the weighted sum of the given individual martingales.

**Definition 5 (Uniformly weighted martingale)**
*Given a set of conformal test martingales $\{S_1, \ldots, S_m\}$, a uniformly weighted martingale $S_* = s_{*,1}, s_{*,2}, \ldots$ is defined by:*

$$s_{*,t} = \frac{1}{m} \sum_{i=1}^{m} s_{i,t} \tag{7}$$

*for all $t \in \mathbb{N}^+$.*

It is easy to see that the uniformly weighted martingale is a conformal test martingale.

**Theorem 6** *Given a set of conformal test martingales $\{S_1, \ldots, S_m\}$, the uniformly weighted martingale $S_* = s_{*,1}, s_{*,2}, \ldots$ is a conformal test martingale, i.e., it satisfies:*

$$\mathbb{E}(s_{*,t} | s_{*,1}, \ldots, s_{*,t-1}) = s_{*,t-1} \tag{8}$$

*for all $t \in \mathbb{N}^+$.*

**Proof** It holds that $\mathbb{E}(s_{*,t}|s_{*,1},\ldots,s_{*,t-1}) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(s_{i,t}|s_{i,1},\ldots,s_{i,t-1})$. Since each $S_i$ is a conformal test martingale, it follows that $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(s_{i,t}|s_{i,1},\ldots,s_{i,t-1}) = \frac{1}{m}\sum_{i=1}^{m} s_{i,t-1}$, which is equivalent to $s_{*,t-1}$ by definition. ■

It can also be easily seen that the uniformly weighted test martingale dominates the Bonferroni corrected martingale, i.e., the former will always output at least as high values as the latter.

**Theorem 7** *Given a set of conformal test martingales $\{S_1,\ldots,S_m\}$, a Bonferroni corrected martingale $S_B = s_{B,1}, s_{B,2},\ldots$, and a uniformly weighted martingale $S_* = s_{*,1}, s_{*,2},\ldots$, then it holds that*

$$s_{*,t} \geq s_{B,t} \tag{9}$$

*for all $t \in \mathbb{N}^+$.*

**Proof** This follows from the definitions of the Bonferroni corrected and the uniformly weighted martingales. Given some $t \in \mathbb{N}^+$, $s_{B,t} = \max(\frac{s_{1,t}}{m},\ldots,\frac{s_{m,t}}{m})$, while $s_{*,t}$ is the sum of a set of terms including $s_{B,t}$, i.e., $s_{*,t} = \frac{1}{m}\sum_{i=1}^{m} s_{i,t} = s_{B,t} + \frac{1}{m}\sum_{i\in\{1,\ldots,m\}\setminus\{b\}} s_{i,t}$, where $s_{b,t} = s_{B,t}$. ■

### 3.3. Testing Joint Sequences

As an alternative to testing exchangeability of each of the sequences of p-values $P_1,\ldots,P_m$, we may consider testing their joint sequence $\otimes(P_1,\ldots,P_m)$. In order to generate a conformal test martingale from such a sequence, we somehow need to obtain a single p-value for each element $(p_{1,t},\ldots,p_{m,t}) \in \otimes(P_1,\ldots,P_m)$, for $t = 1,2,\ldots$. A natural candidate for this is to combine the p-values at each time-step by averaging (Vovk and Wang, 2020). It should however be noted that although the resulting average is a valid p-value, it is not distributed uniformly on $[0,1]$. Conformal test martingales generated from such averages would hence almost immediately signal for a violation of the exchangeability assumption, as they are betting against the null hypothesis of the p-values being uniformly distributed. On the other hand, we could treat a sequence of such aggregated values as a new sequence of nonconformity scores, from which a sequence of p-values could be obtained. We will here consider two such approaches; one based on aggregation operators, including those that have been proposed for averaging p-values (Vovk and Wang, 2020), and another that has been proposed for detecting covariate shift (Vovk et al., 2022, p. 247-248).

#### 3.3.1. Aggregating P-values

Alg. 2 formalizes how to obtain conformal test martingales from a joint sequence using aggregation operators; each tuple in the sequence is converted to a nonconformity score using the operator $(F)$, allowing a conformal test martingale to be computed from the p-value obtained for the score, using some suitable (unspecified) algorithm, e.g., Sleeper/Stayer.

**Algorithm 2:** Aggregating martingale

**Input:** $x_1 = (p_{1,1}, \ldots, p_{m,1}), x_2 = (p_{1,2}, \ldots, p_{m,2}), \ldots$ (joint sequence),
$F$ (aggregation operator)
**for** $i = 1, 2, \ldots$ **do**
  $\alpha_i \leftarrow F(x_i)$
  $p_i \leftarrow$ p-value from $\alpha_1, \ldots, \alpha_i$ (1)
  $s_i \leftarrow$ conformal test martingale from $p_1, \ldots, p_i$
**end**
**return** $s_1, s_2, \ldots$

As operators for aggregating p-values, we consider the following selection of the most straightforward approaches to averaging p-values that are discussed in (Vovk and Wang, 2020):

$$F_{Double}(p_1, \ldots, p_m) = \frac{2}{m} \sum_{i=1}^{m} p_i \tag{10}$$

$$F_{Bonferroni}(p_1, \ldots, p_m) = m \min(p_1, \ldots, p_m) \tag{11}$$

$$F_{Geometric}(p_1, \ldots, p_m) = e \left( \prod_{i=1}^{m} p_i \right)^{\frac{1}{m}} \tag{12}$$

$$F_{Harmonic}(p_1, \ldots, p_m) = \frac{e\, m \ln m}{\sum_{i=1}^{m} \frac{1}{p_i}} \tag{13}$$

In addition, we also consider an operator that certainly does not result in a p-value, but which may be useful for detecting distribution shifts, namely the standard deviation of the p-values:

$$F_{Std\ dev}(p_1, \ldots, p_m) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (p_i - \bar{p})^2} \tag{14}$$

where $\bar{p}$ is the arithmetic mean of $p_1, \ldots, p_m$.

It should be noted that the constant multipliers in the above equations, e.g., $m$ and $e$, which are required for guaranteeing that the resulting average is a p-value (Vovk and Wang, 2020), can be ignored here as they have no effect on the order of the nonconformity scores and hence also no effect on the resulting p-values.

In addition to generating conformal test martingales using each of the above operators, we will also consider generating a uniformly weighted martingale (Def. 5) on top of them.

#### 3.3.2. Euclidean distance to nearest neighbor in a reference set

In (Vovk et al., 2022, p. 247-248), an approach to detecting covariate shift using conformal test martingales was proposed, where nonconformity scores were computed by the Euclidean distance of an object to the closest object in the proper training set, ignoring the labels. A

similar approach can be applied here to the task of testing exchangeability for joint sequences of p-values, where each object $x \in [0,1]^m$ is represented by the $m$ p-values observed at each time-point. However, in our case we may not have access to a (single) proper training set, e.g., as the sequences of p-values may come from multiple conformal predictors. Instead, the initial sequence of tuples is used as a reference set when computing the Euclidean distances.

Alg. 3 formalizes this idea. It takes as input a joint sequence of p-values and the size of the reference set. For each tuple that is included in the reference set, the conformal test martingale will be 1, while for subsequent tuples, conformal test martingales are computed from the derived p-values using some suitable (unspecified) algorithm, e.g., Sleeper/Stayer.

**Algorithm 3:** Nearest neighbor martingale

**Input:** $x_1 = (p_{1,1}, \ldots, p_{m,1}), x_2 = (p_{1,2}, \ldots, p_{m,2}), \ldots$ (joint sequence),
$n \in \mathbb{N}^+$ (size of reference set)

$R \leftarrow \emptyset$
**for** $i$ = *1, 2, ...* **do**
  **if** $i \leq n$ **then**
    $R \leftarrow R \cup \{x_i\}$
    $s_i \leftarrow 1$
  **else**
    $\alpha_i \leftarrow min_{x_j \in R} \|x_i - x_j\|$
    $p_i \leftarrow$ p-value from $\alpha_{n+1}, \ldots, \alpha_i$ (1)
    $s_i \leftarrow$ conformal test martingale from $p_{n+1}, \ldots, p_i$
  **end**
**end**
**return** $s_1, s_2, \ldots$

The size of the reference set, which is a hyperparameter of the approach, may need some tuning; a too low value will give a sample that is not very representative, resulting in high variability, while a too high value will reduce the possibility for detecting a distribution shift early on, since the nearest neighbor martingales for the selected reference tuples are always 1. Rather than fine-tuning this parameter, we will again employ the uniformly weighted martingale (Def. 5) on top of a number of nearest neighbor martingales, each generated with a different value for the size parameter.

## 4. Experimental Investigation

In this section, we will evaluate the different approaches for testing exchangeability for multiple sequences of p-values. We will consider two main scenarios; the first concerns a single target, for which multiple conformal predictors have been generated, and the second concerns multiple targets, where a conformal predictor has been generated for each target. To be able to determine whether test outcomes are false or not, we synthetically introduce violations to the exchangeability assumption and investigate how soon (if at all) these changes can be detected.

### 4.1. Experimental Setup

#### 4.1.1. Approaches

In the experiments, we will employ the Sleeper/Stayer algorithm (Alg. 1) with the parameter settings considered in (Vovk et al., 2022, p. 283), i.e., $r = 0.001$ and $g = 10$, in conjunction with the proposed betting function (5) (Vovk et al., 2022, p. 282), to generate conformal test martingales for both the individual and the joint sequences of p-values. For testing exchangeability of individual sequences, we will employ both the Bonferroni corrected (Def. 4) and the uniformly weighted martingales (Def. 5). For testing joint sequences, will consider two uniformly weighted martingales, formed from a set of aggregating martingales (Alg. 2) and a set of nearest neighbor martingales (Alg. 3), respectively. The former are generated from the five operators described in Sec. 2, while the latter are formed from 20 different sizes of reference sets, i.e., $n \in \{10, 20, \ldots, 200\}$.

We will throughout the experiments use the significance level 0.01, i.e., $c = 100$, to signal that the exchangeability assumption can be rejected.

#### 4.1.2. Multiple predictors for a single target

We here consider a regression task; predicting house sale prices for the King County, Washington, between May 2014 and May 2015. The employed dataset, named `house_sales` in the OpenML repository[1], consists of 21 613 instances represented with 19 features. The dataset is randomly split into a training set (3/4 of the full dataset) and a test set (1/4). The training set is further randomly split into a proper training set (2/3 of the training set) and a calibration set (1/3). The latter is repeated 100 times; each time using the proper training set to fit a random forest regressor (with default parameter settings as provided in `scikit-learn`[2]), which together with the calibration set is used to form a standard conformal predictive system, as implemented in `crepes`[3] (Boström, 2024). The 100 resulting conformal predictive systems are subsequently applied to each test instance in a semi-inductive setting; the p-value for the true target is first computed for each conformal predictive system, which is followed by an update of each calibration set, before proceeding to the next test instance. Furthermore, after having observed half of the 5404 test instances, we introduce a drift in two different ways; by gradually increasing the volatility and by introducing a trend.

Volatility is introduced by multiplying the true target with a factor that is uniformly sampled between a lower and higher rate, with the expected value of 1.0. The employed lower rate is gradually decreasing from 1.0 to 0.8 at equal decrements for the 2702 drifting test instances. Similarly, the higher rate is gradually increasing from 1.0 to 1.125 at equal increments. Hence, the drift does not change the expected value, but leads to an increased variance, initially hardly noticeable but quite substantial towards the end of the sequence.

Trend is introduced by multiplying the true target with a factor that is uniformly sampled between a lower and higher rate, with an expected value that is greater than 1.0. The employed lower rate is gradually increasing from 1.0 to 1.05 at equal increments for the 2702 drifting test instances. Similarly, the higher rate is gradually increasing from 1.05 to 1.1 at

1. www.openml.org
2. www.scikit-learn.org
3. http://www.github.com/henrikbostrom/crepes/

equal increments. Hence, the drift does not change the variance, but leads to a gradually increasing positive trend.

#### 4.1.3. One predictor for each of multiple targets

We here consider a multi-target binary classification task; predicting properties of chemical compounds as extracted from the `pcba` dataset from `DeepChem`[4]. The dataset consist of 128 binary targets in total, but labels are not available for all compounds, so the following strategy was employed to extract a sufficiently large dataset with known labels (positive or negative). The targets were considered in order according to the number of chemical compounds with a positive label, and selected if there were at least 100 positive instances and if after removing all instances for which the target label was missing, all previously selected targets still have at least 100 positive instances. The resulting dataset consists of 30 290 chemical compounds and 20 targets. The class distributions are quite imbalanced, with the relative frequency for the positive class ranging from 0.005 up to 0.03. The chemical compounds are represented by the simplified molecular-input line-entry system (SMILES) and the package `RDKit`[5] was employed to generate features from the SMILES strings, more specifically, *Morgan fingerprints* (binary vectors, all of length 1024).

Similarly to the previous scenario, the dataset was randomly split into a training set (3/4 of the full dataset) and a test set (1/4). For each of the 20 targets, a random forest classifier (again with default parameter settings) was fitted using the full training set. A conformal classifier was formed for each random forest model, each with an initially empty calibration set. The 20 conformal classifiers were subsequently applied to each test instance in a semi-inductive setting; the p-value for the true target was first computed for each conformal classifier, followed by an update of each calibration set, before proceeding to the next test instance. Furthermore, after having observed half of the 7573 test instances, we introduce drift by selecting from the remaining test instances in an informed way rather than randomly. We consider selecting test objects based on the highest uncertainty of underlying model predictions, as measured by the maximum Gini impurity for any of the 20 models, and the smallest average Euclidean distance to the ten closest training objects with at least one positive label.

### 4.2. Experimental Results

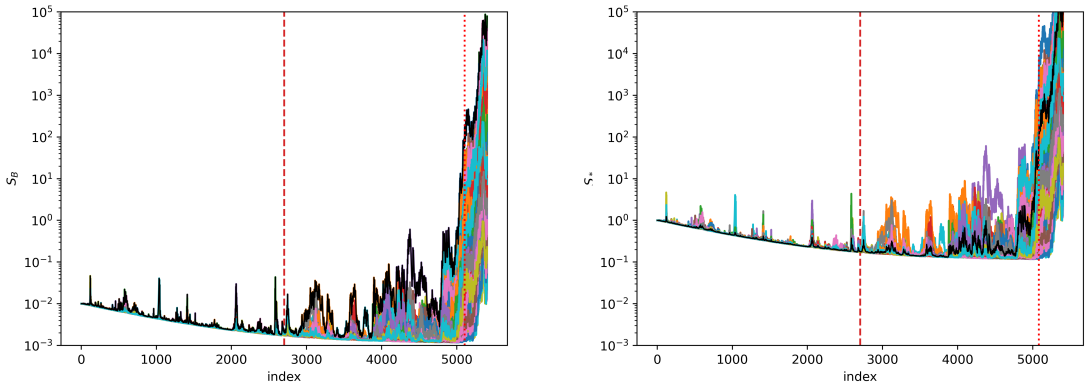
#### 4.2.1. Multiple predictors for a single target

In Fig. 1, the martingales for the four main approaches are displayed for the single target task with increased volatility. In Fig. 1a, the Bonferroni corrected martingale, obtained from 100 individual conformal test martingales, is indicated with a black line, with the lower-valued, individual martingales indicated with lines of other colors. The dashed red line indicates where the volatility starts to increase and the dotted red line indicates at which point in time the Bonferroni corrected value reaches $c = 100$; the latter occurs for the Bonferroni corrected martingale for test example no. 5105. In Fig. 1b, the uniformly weighted martingale of the 100 individual conformal test martingales is indicated with a

4. https://github.com/deepchem/deepchem
5. https://www.rdkit.org

black line, while the individual martingales are indicated using other colors, appearing both below and above the uniformly weighted martingale. As shown by the dotted red line, the uniformly weighted martingale reaches $c = 100$ for test example no. 5082, i.e., well before the Bonferroni corrected martingale. In Fig. 1c, each of the five aggregating martingales generated from the joint sequence of p-values (with each tuple consisting of 100 p-values) are displayed together with a uniformly weighted martingale generated on top of them (black line). Of the individual aggregating martingales, it is only the one based on the standard deviation operator that reaches a value above 100 (for test example no. 4802), something which is caught by the uniformly weighted martingale soon after, which signals after reaching test example no. 4868. In Fig. 1d, each of the 20 nearest neighbor martingales generated from the joint sequence of p-values are displayed together with a uniformly weighted martingale generated on top of them (black line). All of the individual martingales reaches 100 at various time points; the earliest one, using 180 tuples in the reference set, signals at test example no. 4874, and the last one, using 20 reference objects, signals at test example no. 5146. The uniformly weighted martingale over the 20 individual martingales reaches 100 at test example no. 5043.

In Fig. 2, the martingales for the four main approaches are displayed for the single target task with trend. In Fig. 2a, the results for the Bonferroni corrected martingale, obtained from 100 individual conformal test martingales, are shown; it reaches $c = 100$ at test example no. 3016. In Fig. 2b, the results for the uniformly weighted martingale of the 100 individual conformal test martingales is shown; it reaches $c = 100$ for test example no. 2940, i.e., clearly ahead of the Bonferroni corrected martingale. In Fig. 2c, each of the five aggregating martingales generated from the joint sequence of p-values (with each tuple consisting of 100 p-values) are displayed together with a uniformly weighted martingale generated on top of them (black line). Four of the five individual aggregating martingales provide a signal; it is only the one based on the standard deviation operator that does not reach a value above 100. The uniformly weighted martingale reaches 100 for test example no. 3016. Fig. 2d shows that none of the 20 nearest neighbor martingales generated from the joint sequence of p-values signals for a violation of the exchangeability assumption, and consequently the uniformly weighted martingale fails to detect this too.



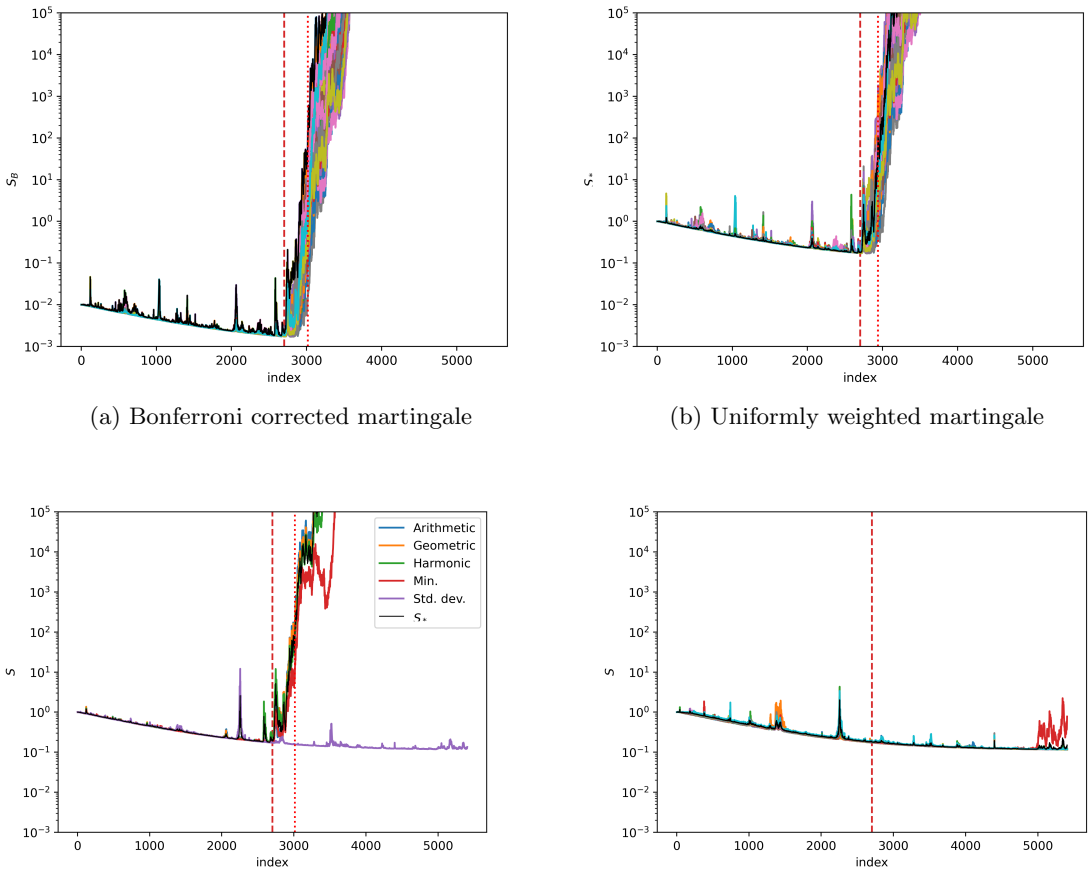

(a) Bonferroni corrected martingale

(b) Uniformly weighted martingale

(c) Aggregating martingale

(d) Nearest neighbor martingale

Figure 1: Single target; increased volatility



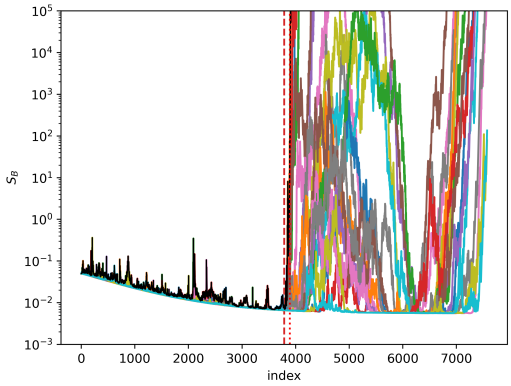

(a) Bonferroni corrected martingale

(b) Uniformly weighted martingale

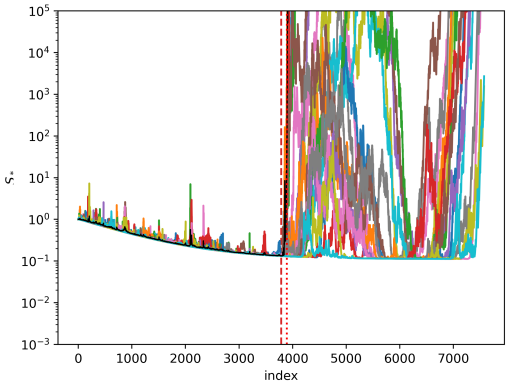
(c) Aggregating martingale

(d) Nearest neighbor martingale

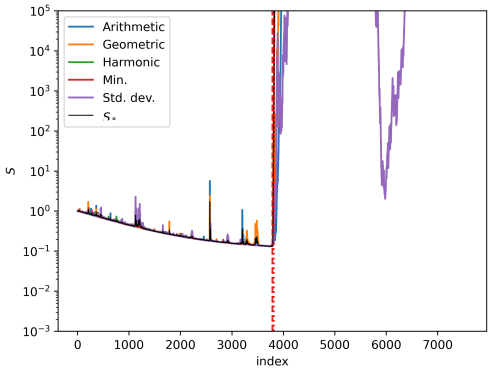
Figure 2: Single target; trend



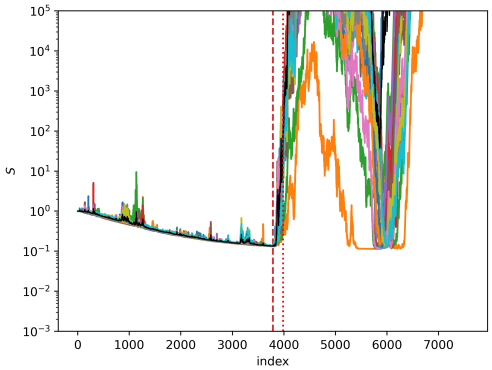

(a) Bonferroni corrected martingale




(b) Uniformly weighted martingale




(c) Aggregating martingale



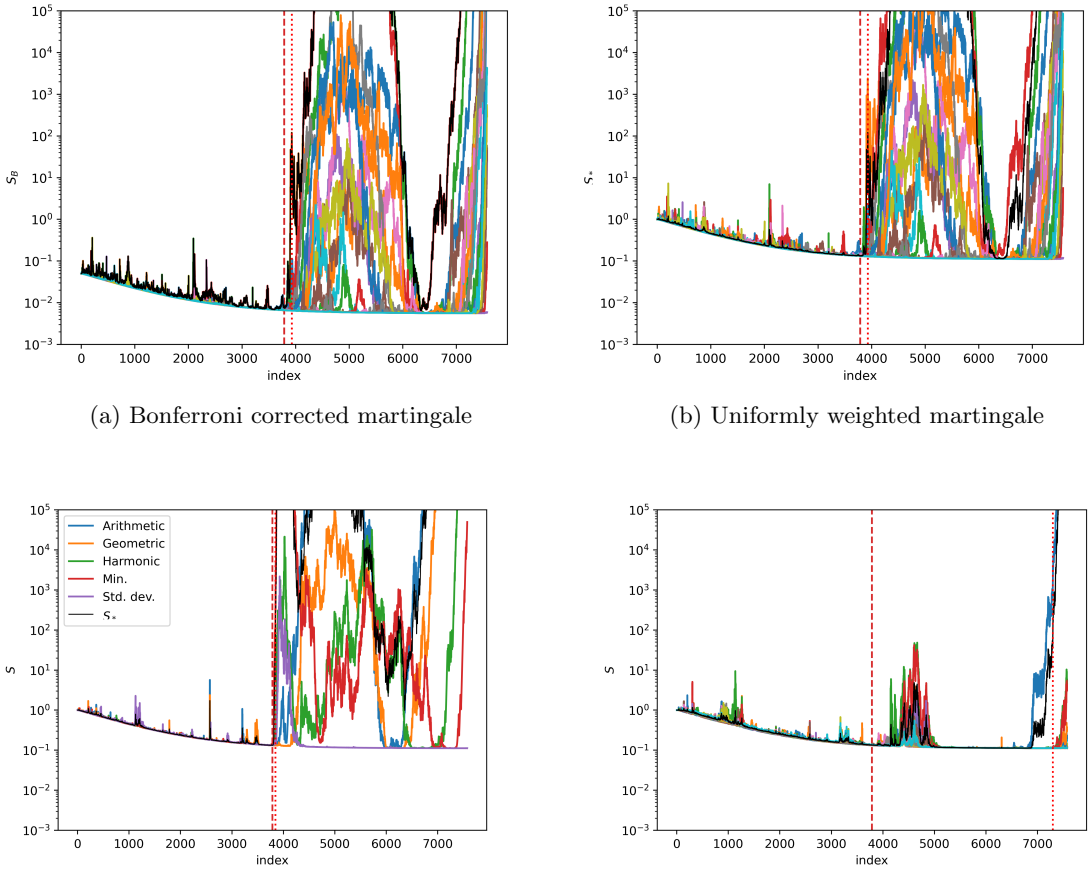

(d) Nearest neighbor martingale

Figure 3: Multiple targets; selection based on highest uncertainty

#### 4.2.2. One predictor for each of multiple targets

In Fig. 3, the martingales for the four main approaches are displayed for the multi-target task with where a distribution drift is introduced by selecting objects with the most uncertain predictions. The Bonferroni corrected martingale, obtained from 20 individual conformal test martingales, detects a deviation after 3894 test objects; it can be noted that the uniformly weighted martingale is signaling at the same time point. For the joint sequences, it can be observed that the combined aggregating martingale signals after 3812 test objects, while the uniformly weighted nearest neighbor martingale signals after 3985 test objects.

Finally, in Fig. 4, the martingales for the four main approaches are displayed for the multi-target task with where a distribution drift is introduced by selecting objects with the shortest Euclidean distance to objects in the training set. The Bonferroni corrected martingale, obtained from 20 individual conformal test martingales, detects a deviation after 3928 test objects, and it can again be noted that the uniformly weighted martingale is no more efficient, reaching a value of 100 after exactly the same number of test objects. For




(a) Bonferroni corrected martingale

(b) Uniformly weighted martingale

(c) Aggregating martingale

(d) Nearest neighbor martingale

Figure 4: Multiple targets; selection based on shortest distance

Table 1: Change points

| **Task** | **Actual** | **Bonf.** | **Unif. w.** | **Aggr.** | **Nearest n.** |
|---|---|---|---|---|---|
| Single target; volatility | 2703 | 5105 | 5082 | **4868** | 5043 |
| Single target; trend | 2703 | 3016 | **2940** | 3016 | - |
| Multiple targets; uncertainty | 3787 | 3894 | 3894 | **3812** | 3985 |
| Multiple targets; distance | 3787 | 3928 | 3928 | **3847** | 7298 |

the joint sequences, it can be observed that the combined aggregating martingale signals after 3847 test objects, while the uniformly weighted nearest neighbor martingale signals only after 7298 test objects when selecting the least distant objects.

In Table 1, we summarize the above results by showing the actual and detected change-points, indicating the earliest detection in boldface.

### 4.3. Discussion

The uniformly weighted martingale was observed to reject the null hypothesis well ahead of the of the Bonferroni corrected martingale for the two experiments on single target predictions. However, for the two multitarget cases (with a lower number of combined p-values), the two martingales signaled at the same time point. This illustrates that the theoretical guarantee of the former to signal as early as the latter (Theorem 7) may in some cases, but not always, lead to more effective testing.

When testing a joint sequence of p-values, the experiments indicate that forming a uniformly weighted aggregating martingale is more effective than forming a uniformly weighted nearest neighbor martingale. The former apparently benefits from the diversity of the included aggregating martingales, e.g., the one based on the standard deviation was the most effective for detecting drift due to increased volatility, while it was the least effective for detecting drift due to trend, in both cases however leading to that the uniformly weighted martingale on top of them managed to signal not far after the most effective individual martingale. The results for the nearest-neighbor based approach are more mixed; in two of the considered four cases it was either not at all or at a very late stage able to detect a violation of the exchangeability assumption.

## 5. Concluding Remarks

We have presented an investigation of approaches for testing exchangeability for multiple sequences of p-values, either by testing each sequence individually and controlling for multiple hypothesis testing or by testing the joint sequence of p-values, through obtaining nonconformity scores from the joint observations. We have shown that for the first task, a uniformly weighted martingale is guaranteed to signal at least as early as a Bonferroni corrected martingale. The experimental results confirm the theoretical finding and show that the difference in some cases can be substantial, while it may have no impact in other cases. We also proposed two approaches for the second task; aggregating martingales and nearest neighbor martingales. The experimental results highlight the benefits of employing a diverse set of operators to form a uniformly weighted aggregating martingale; using only one of the operators may lead to a failure to detect a deviation in some cases. For the considered scenarios, the results indicate that the aggregating martingale is clearly ahead of the nearest neighbor martingale.

There are several directions for future work. In the presented experiments, a single underlying algorithm for generating conformal test martingales was considered together with a specific betting function. The effect of these choices on the performance of the proposed approaches remains to be investigated. Other directions concern investigating alternative operators for the aggregating martingales and ways of computing nonconformity scores for the nearest neighbor martingales. One important direction for future work is to compare the proposed approaches to the use of e-values (Vovk and Wang, 2021); in contrast to when considering p-values, multiple hypothesis testing can be conducted straightforwardly by averaging the e-values.

## Acknowledgements

HB was partly funded by Vinnova (RAPIDS, grant no. 2021-02522) and Digital Futures. The author would like to thank the anonymous reviewers for their insightful remarks and suggestions.

## References

Henrik Boström. Conformal prediction in python with crepes. In *Proc. of the 13th Symposium on Conformal and Probabilistic Prediction with Applications*, pages 236–249. PMLR, 2024.

Charalambos Eliades and Harris Papadopoulos. A conformal martingales ensemble approach for addressing concept drift. In Harris Papadopoulos, Khuong An Nguyen, Henrik Boström, and Lars Carlsson, editors, *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 328–346. PMLR, 13–15 Sep 2023.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Conformal multi-target regression using neural networks. In Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Giovanni Cherubin, editors, *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 65–83. PMLR, 09–11 Sep 2020.

Denis Volkhonskiy, Evgeny Burnaev, Ilia Nouretdinov, Alexander Gammerman, and Vladimir Vovk. Inductive conformal martingales for change-point detection. In Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos, editors, *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 132–153. PMLR, 13–16 Jun 2017.

Vladimir Vovk. Testing randomness online. *Statistical Science*, 36(4):595–611, 2021.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4): 791–808, 2020.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3), 2021.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: conformal test martingales for change-point detection. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 191–210. PMLR, 08–10 Sep 2021.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World (2nd ed.)*. Springer, 2022.