Venn-Abers Testing of Exchangeability

Ilia Nouretdinov I.R.NOURETDINOV@RHUL.AC.UK
Centre for Reliable Machine Learning, Royal Holloway University of London, Egham Hill, Egham, Surrey, TW20 OEX, United Kingdom.

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

Abstract

A recurrent problem in many domains is the accurate and rapid detection of a change in the distribution of observed variables. This is important since our algorithms have been trained for a certain data distribution, and if the distribution has changed, the results will not be accurate and/or valid any longer. Instances of this problem, which are generally referred to as change-point detection, are found in fault detection in vehicle control systems, detection of the onset of an epidemic, and many other applications.

Recently, new methods based on reliable machine learning have shown important advantages of this statistical task. Conformal Test Martingales (CTM) allow one to avoid this limitation and obtain valid results without information about used distributions. This is done with the assumption that the data are i.i.d. (or exchangeable) in online mode, and the corresponding martingale accumulates evidence against this assumption.

This work aims to extend the conformal framework and consider the other family of reliable machine learning methods, the Venn-Abers method of probabilistic prediction, to test the data for change points. This work shows how Venn-Abers testing of exchangeability (VATE) can be founded on the ground of e-value theory, including recently developed e-pseudomartingales, and studies its advantages and drawbacks, compared to CTM. Our conclusion is that the efficiency of this approach is related to the type of causality in the data set.

Keywords: reliable machine learning, exchangeability testing, change-point detection, probabilistic prediction, e-values

1. Introduction

In this work, we study data deviations from i.i.d. in the online mode. The *i.i.d.* (independent, identically distributed), also known as the *power* assumption about the data, means that the data instances are generated independently by the same distribution. Typically, it is tested by testing the *exchangeability* of the data sequence or by the assumption that its instances are in random order. A frequent example of i.i.d. violation is a *change point*, when the data generation mechanism changes at some step(s). The contribution of this work is to apply the Venn-Abers method in combination with e-prediction, for testing of exchangeability.

The general principle behind this way of testing is that changes can be detected by the number of errors made in the generation of predictions after a change occurs. This happens because the training that was done on the preceding data, before the change, followed a different distribution. However, nonpredictability due to the distribution changes could be separated from nonpredictability due to the randomness present in the data. The testing approach developed in Vovk et al. (2023) allows us to take randomness into account by

providing the prediction with some measure of confidence - a reliable prediction. The principle can be formulated in this way: the cause for alarm is not the low accuracy itself, but having a wrong prediction combined with a high claimed confidence in these predictions.

The work reported in Vovk et al. (2023) presents two families of reliable machine learning methods: conformal prediction for classification with confidence; Venn prediction for well-calibrated probabilistic classification. The second was later developed further into the Venn-Abers prediction framework Vovk and Petej (2014). These methods have proven validity properties with the assumption that the data-generating process is i.i.d. Therefore, a change point that breaks the i.i.d. assumption can be reflected as a break in the the validity properties of conformal or Venn-Abers methods, which ensure correct coverage in i.i.d. case.

However, to date, such studies have concentrated on the conformal version of this idea, such as the conformal Test Martingale (CTM) Ho (2005), which accumulates evidence against the i.i.d. hypothesis during online analysis of the conformal prediction output. Practically no attention has been paid to using reliable probabilistic predictions. However, as probabilistic prediction produced by the Venn-Abers framework may be more informative than conformal prediction, it may be worth attention in the context of testing as well. The goal of this work is to fill this gap.

We found support in the theory of *e-prediction* Vovk (2023a,b) which also has an application to the change point detection task Vovk (2023a). An analogue of CTM is called *e-pseudomartingales*, based on the recently developed theory of *e*-prediction Vovk (2023a); Vovk et al. (2024). Compared to the conformal prediction scheme, this approach has the advantage of simplicity without loss of efficiency. Initial comparisons on artificial and real data sequences were performed in Purtle (2020). In this work, we show how to use this foundation to build the *Venn-Abers approach to test exchangeability (VATE)*. Some initial results of this work were previously presented as an extended conference abstract Nouretdinov and Gammerman (2023).

The plan of the work is as follows. Sect. 2 defines the necessary notion of e-prediction, following the preceding work. In Sect. 3 we describe another necessary element (Venn-Abers prediction) and how it can be put into the e-prediction framework. In Sect. 4, we study some artificial models of the data, why we expect VATE to be a more powerful testing method than CTM, at least for some kinds of data set, depending on their causality structure. In Sect. 5 we present the results of its usage on real data. Sect. 6 concludes the work.

2. Background and Related Work

2.1. E-predictor

The notion of *e*-value (with 'e' for expectation) is an alternative to *p*-value (with 'p' for probability) in statistics. In the following description of *e*-value as a function, we follow recent works Vovk (2023a); Vovk et al. (2024, 2025).

In Vovk et al. (2024, 2025) this technique was called *e-pseudomartingale* and contrasted to Conformal Test Martingales. The term pseudomartingale is used because this technique shares most of the martingale properties but not all of them.

Let Z be a measurable data example space. Typically in machine learning, an example $z_i \in Z$ is a pair of the object (typically, in the form of a feature vector) $x_i \in X$ and the

label $y_i \in Y$, where X is a finite-dimensional vector space, Y is the label space that can be finite or infinite.

An *e-predictor* is a series of measurable functions

$$f_n : Z^n \to \mathbf{R}^n,$$

$$(\alpha_1, \dots, \alpha_n) = f_n(z_1, \dots, z_n),$$

with the properties of:

1. non-negativeness, $\alpha_i \geq 0$;
2. fairness, $\alpha_1 + \dots + \alpha_n = n$, meaning that α_i give 1 in average in exchangeability assumption;
3. equivariance $f_n\left(z_{s(1)}, \dots, z_{s(n)}\right) = \left(\alpha_{s(1)}, \dots, \alpha_{s(n)}\right)$

In the last line, s means any permutation of elements $1, \dots, n$. So, the property means: if order of the original examples is changed, then the order of the scores has to be changed in the same way.

This approach of *e-values* exists as an alternative to p-values (Chebysheff's inequality implies that the inverse of e-value can be used as p-value if needed, although the opposite is not always true).

In the context of conformal prediction Vovk et al. (2023), the e-values are also analogs of the nonconformity scores. Both types of score are measures of nonconformity (strangeness) of an instance. In machine learning, they quantify the degree of disagreement between the true value of the label and its predicted value, found after training on the other examples.

Based on Vovk (2023a), an algorithmic prediction scheme for the detection of multiple change points with sample applications to real data was initially presented in Purtle (2020). That work has shown some advantage in the efficiency of change point detection. However, it did not use the new possibilities given by the e-value scheme for the definition of the testing function.

2.2. Venn-Abers predictor scheme

Venn-Abers predictor (VAP) framework for reliable probabilistic prediction was developed in Vovk and Petej (2014), as a special version of the multi-probabilistic Venn predictor Vovk et al. (2023).

Like conformal predictors, Venn machines can be linked to an underlying method. In the Venn-Abers version, this is done in a uniform way by means of isotonic regression.

To date, Venn or Venn-Abers machines have never been used for creating test martingales, in part because their validity properties are more complex than in the case of conformal prediction.

Venn predictor as it is, adopts calibration properties of the conformal prediction framework to the predictions made in a probabilistic form instead of p-values. This has the advantage of better quantification of risk, when a confident prediction is impossible. However, it is less sensitive to individual anomalies.

Venn-ABERS version has the main advantage of proposing a uniform way of linking an underlying scoring method, without extra parameters such as a pre-defined number of categories. Its general scheme for binary classification is represented in Algorithm 1. It includes solving a quadratic optimisation problem with linear constraints that can be implemented with a standard package for such tasks.

Algorithm 1 Venn-Abers prediction scheme

INPUT: training data sequence $Z = (x_1, y_1), \ldots, (x_{N-1}, y_{N-1})$
INPUT: new example (x_N)
INPUT: underlying scoring function $\mathcal{S}$
FOR $\hat{y} := 0, 1$
FOR $j := 1, \ldots, N$
$s_j := \mathcal{S}\left((x_j, y_j), Z \cup \{(x_N, y_N := \hat{y})\} \setminus \{(x_j, y_j)\}\right)$
END FOR
find $(g_1, \ldots, g_N)$ s.t. $\sum_{i=1}^{N}(g_1 - y_i)^2 \to \min$ wrt. $(s_i \leq s_j) \Rightarrow (g_i \leq g_j)$
$\hat{p}_y := g_N$
END FOR
OUTPUT: $(\hat{p}_0, \hat{p}_1)$

The **underlying scoring function** $\mathcal{S}$ can be used to link the framework to any prediction technique such as SVM, Nearest Neighbours, Neural Networks, Random Forest, and others. It is defined as a function of a (training) set and an example, which outputs some quantitative evidence for the example belonging to the positive class.

It can be said that the Venn-Abers framework plays the calibration role, transforming an unreliable scoring function to a function with calibrated probabilistic meaning. The probability here refers to the estimation of the chance that $y = 1$.

Venn-Abers predictor produces a pair $(\hat{p}_0, \hat{p}_1)$ of estimates instead of one, which means that the truth is either one or the other. However, for the purpose of the current paper, only the element $p_{\hat{y}}$ assigned to the true value $\hat{y} = y_N$ is essential and relevant for the data testing task.

2.3. E-prediction in online testing

Following Vovk (2023a); Vovk et al. (2024), Alg. 2 (MUSUC, which is short for Cumulative Sum backwards) and Alg. 3 (modified MUSUC) give change point detection algorithms based on e-prediction in the form of e-pseudomartingale testing. In this approach, the evidence against exchangeability is accumulated within e-values by continuous play in online machine learning, and measured with the martingale-type value.

Online machine learning means that prediction is made step by step. In step n, the label y_n is predicted for a new unlabelled example x_n after training on the previous $n - 1$ examples with known labels.

Player can bet that the new example $z_n = (x_n, y_n)$ does not follow the same distribution as the training set, and his capital is multiplied with the corresponding score α_n assigned to the new example. To distinguish from the other steps, the value of α_n calculated in the n-th step is called E_n.

Algorithm 2 MUSUC procedure of e-pseudomartingale testing

INPUT: data sequence $((x_1, y_1), \dots, (x_N, y_N), \dots)$ where $(x_i, y_i) \in Z$
INPUT: threshold C
INPUT: e-predictor f_n.
$i := 0$
$\sigma_0 := 0$
FOR $n := 1, \dots, N, \dots$
$E_n :=$ the last dimension of $f_n((x_{\sigma_i+1}, y_{\sigma_i+1}), \dots, (x_n, y_n))$
IF $E_{\sigma_i+1} \times \dots \times E_n > C$
$i := i + 1$
$\sigma_i := n$
END IF
END FOR
OUTPUT: the sequence of change points $\sigma_1, \sigma_2, \dots, \sigma_i, \dots$

E-predictor takes the form of betting on the examples. Under the exchangeability assumption, the amount of capital 1 invested by a Player following scores generated by a randomly chosen set of N examples is paid back. Reciprocally, if the assumption of exchangeability is broken and the data sequence is not in the random order, the Player can make a profit.

Basically, Player's capital after n steps is $E_1 E_2 \dots E_n$. If it reaches a high value (above a large number C) then the i.i.d. hypothesis about the data sequence is discarded at the significance level $1/C$, as follows from Ville's inequality.

The *MUSUC* procedure suggested in Vovk (2023a) is a way to solve the task of detecting multiple change points. Following the above example, when the level of C is reached, Player's capital again becomes 1 and the game starts from the beginning.

The *Modified MUSUC* procedure has one more advantage in that it allows us to ignore the initial loss of Player's capital. To alert the change point, it is sufficient to detect that there existed at least some continuous time period so that Player's capital at the end is at least C times as large as it was in the beginning of the period. This period must start after the preceding change point, but not necessarily follows immediately afterwards (this is the difference from non-modified MUSUC).

2.4. Conformal Test Martingale

Conformal Test Martingale (CTM), or testing exchangeability based on conformal prediction, was initially suggested in Vovk et al. (2003). It was motivated by the observation of a disorder in the USPS data set Repository (1994).

The Conformal Test Martingale that is focused on change detection was initially discussed in Ho (2005). In this work, the rapid growth of the martingale when the i.i.d. assumption was violated was used as a basis for change-point detection. The particular statistic used to detect the change was the value of the martingale itself (Martingale Test 1) or the difference between its values at neighbouring points (Martingale Test 2).

Normally, the aim is to capture the violation of exchangeability in a data sequence, but it does not have to be restricted to that. An example of how assumptions other than

Algorithm 3 Modified MUSUC procedure of e-pseudomartingale testing

INPUT: data sequence $((x_1, y_1), \ldots, (x_N, y_N), \ldots)$ where $(x_i, y_i) \in Z$
INPUT: threshold C
INPUT: e-predictor f_n.
$i := 0$
$\sigma_0 := 0$
FOR $n := 1, \ldots, N, \ldots$
$E_n :=$ the last dimension of $f_n((x_{\sigma_i+1}, y_{\sigma_i+1}), \ldots, (x_n, y_n))$
IF $\max\{E_j \times \cdots \times E_n | j \in \{\sigma_i + 1, \ldots, n\}\} > C$
$i := i + 1$
$\sigma_i := n$
END IF
END FOR
OUTPUT: the sequence of change points $\sigma_1, \sigma_2, \ldots, \sigma_i, \ldots$

exchangeability are tested can be found in Fedorova et al. (2023). The work Fedorova et al. (2012b) demonstrated a non-exchangeability (Gaussian) version for the same problem. In some aspects, the approach is similar to Ho (2005), however, here the examples were intentionally ordered by one of the attributes instead of the time. The work in Fedorova et al. (2012b) also highlighted the challenge of detecting the *second* (and next) change points and how best to detect it correctly after the first one. Normally, a martingale gives reliable information only about the first change point, while its further behaviour is more difficult to analyse within the same setting.

A further development of change point detection by means of conformal martingales was made in Volkhonskiy et al. (2017). One of the innovations was adaptation of the approaches *CUSUM* Shiryaev (2010) and *Shiryaev-Roberts* Shiryaev (1963), in order to analyse martingale growth more efficiently, thus increasing the significance of the results.

The details of the Conformal Test Martingale (CTM) can be found in Vovk et al. (2003) or Vovk et al. (2024). In the current work, we will use it only as a baseline method for comparison. Therefore, we give only a brief description.

The approach is game-theoretic. It consisted of a player known as Gambler who was playing against a second player called the Predictor. The violation of the exchangeability assumption would lead to a cumulative increase in the Gambler's capital. The growth of this capital is defined by a conformal martingale which measures deviation of the p-values output with a conformal predictor from a uniform distribution, which is expected if the data is exchangeable.

Formally, CTM consists of the following elements. The first two steps follow onine conformal prediction, but are limited to assigning the p-value to the true hypotheses only. Note that the NCM function here plays only the ranking role: it just orders the examples by a measure of their relative strangeness. Unlike e-prediction, exact numerical values are not important.

1. Nonconformity scores of a *nonconformity measure (NCM)* are defined in a way analogous to

$$(\alpha_1, \ldots, \alpha_n)$$

in the e-values, although not required to sum to n.

2. The *p-values* are calculated using the formula

$$p = \frac{|\{i : \alpha_i > \alpha_n\}| + \theta_n |\{i : \alpha_i = \alpha_n\}|}{n}$$

where θ_n is distributed uniformly on [0,1].

3. A *betting function* $b(p) = b_n(p) \geq 0$ is defined for $0 \leq p \leq 1$ and must have the average of 1 to be fair. It may change as n changes, but should be fixed on any step n.

4. The *martingale* grows as follows:

   - $C_0 = 1$;
   - $C_n = C_{n-1} b(p_n)$ for $n > 0$.

5. In the change point detection, we apply the CUSUM mode (Shiryaev (2010); Vovk (2022)) that is a full analogue of Modified MUSUC for the e-values: a new change point is detected once the capital becomes C times larger than at any other point after the previous change point.

3. Methodology

3.1. Venn-Abers predictor as an e-predictor

Let us now explain the Venn-Abers testing of exchangeability (VATE) approach.

As we discussed earlier, the e-value scheme is known mostly for its simplicity, compared to the p-values used in conformal prediction. Now we exploit its advantage of being a more generous scheme. We show how it can be used to detect change points by mistakes made by *multiprobabilistic (Venn)* prediction.

To create the martingale, we need to define the outputs $\alpha_1, \ldots, \alpha_n$ for the functions f_n in the MUSUC or Modified MUSUC procedure (Alg. 2 and 3).

In basic form, this is done as follows:

$$\alpha_i = (1 - \varepsilon) + \varepsilon \times \frac{|\{j = 1, \ldots, n : g_j = g_i\}|}{|\{j : g_j = g_i, y_j = B_n\}|} \times 1_{\{y_i = B_n\}}$$

In this model, Player's strategy is defined as betting on a fixed label value $y_n = B_n$. Player makes a choice of the value B_n of the label y_n, and bets on it with a part of this capital. His suspicion is that the deviation of exchangeability caused an underestimation of this label's probability made by the Venn-ABERS prediction and that in the real post-change data stream it is really higher. If the suspicion is true, then Player gains by using this strategy.

The parameters of the strategy are the following.

1. The choice of B_n, the label value (0 or 1) for which Player is betting for on n-th step;

2. ε is the proportion of the capital which Player puts at risk that is the maximal part allowed to be lost if $y_n \neq B_n$.

The Player may have a special advanced strategy for the choice of these parameters, e.g. based on his inside knowledge. In this paper, we consider a simple example, where *varepsilon* is fixed (although variable). B_n is either fixed (in artificial examples where the best choice is obvious), or we consider both of them averaged, as if the capital is shared equally into betting for 1 and betting for 0 (in real data examples).

The notation 1 means just the characteristic function (1 for True, 0 for False).

The notation g_i refers to the Venn-Abers machine (Alg. 1). In terms of the general Venn prediction framework, it is called a *taxon (category)* with the meaning of a marker for similar grouped objects. It is possible to use only a single taxon, although in this case the probabilistic prediction would suffer from a lack of diversity among the examples. However, this may be used in simple models or when the data size is small. For testing, this approach is applied to the training data sequence

$$(x_{\sigma_i+1}, y_{\sigma_i+1}), \ldots, (x_{n-1}, y_{n-1})$$

and the new example x_n, with the same scoring function $\mathcal{S}$, but g_i is calculated only for the true hypothesis $\hat{y} = y_n$.

3.2. Online testing

By analogy to conformal martingales Vovk et al. (2003); Fedorova et al. (2012a), this procedure allows a way to test the data for exchangeability. If the prediction is valid (which is proven for Venn-Abers), but Player is gaining capital by gambling, this is considered as a symptom of the data being non-exchangeable. The change is reflected in the sudden growth of Gambler's capital.

We will use *modified MUSUC* approach for online change point detection (Alg.3 from Sec.2).

4. Motivating and artificial examples

4.1. When VATE may perform better than CTM?

Let us start with a simple data example showing what kind of gap in p-value techniques (CTM) is filled with e-values (VATE), due to some difference between the e-value E_n and the betting function applied to p-value $b_n(p_n)$.

Tab. 1 describes a distribution on sequences of length 3. We assume that the data distribution is equally shared between 4 concrete sequences (a–d), and it is known that other sequences have zero probability. Here we consider a very concrete question. What can be reached by CTM and VATE techniques in the third step, on the assumption that the distribution is known in advance and can be fully used as a hint? To compare the methods, we measure the expected value $\log b(p_n)$, that is, the logarithmic increase given to the capital for $n = 3$.

In **CTM**, we assume that the last example is the strangest in (a,b), and the last two examples are the strangest in (c,d). We assume them to be equally strange, and it is easy to check that a preference to one of them would not change the final result. So, the p-value will be uniformly distributed on $[0; \frac{1}{3}]$ for (a,b) and $[0; \frac{2}{3}]$ for (c,d) taken together.

About the betting function b, let us assume that it is:

Prob.	Ref.	Ex.1	Ex.2	Ex.3	p	CTM	e
0.25	(a)	(-1,-1)	(-2,-1)	**(+1,+1)**	un.$[0;\frac{1}{3}]$	b_1	3
0.25	(b)	(-2,-1)	(-1,-1)	**(+1,+1)**	un.$[0;\frac{1}{3}]$	b_1	3
0.25	(c)	(-1,-1)	(+1,+1)	**(+2,+1)**	un.$[0;\frac{2}{3}]$	un.$\{b_1, 3-b_1\}$	1.5
0.25	(d)	(-1,-1)	(+2,+1)	**(+1,+1)**	un.$[0;\frac{2}{3}]$	un.$\{b_1, 3-b_1\}$	1.5

Table 1: A data example for Sect. 4.1

- stepwise, with a constant value (denoted as b_r) on each of the intervals $[\frac{r-1}{n}; \frac{r}{n}]$;
- monotonically non-increasing (i.e. betting is done for *high* strangeness of the new example).

We use the following notation for the betting function:

- for $p \in [0; \frac{1}{3}]$, $b(p) = b_1 \leq 3$;
- for $p \in [\frac{1}{3}; \frac{2}{3}]$, $b(p) = b_2 = 3 - b_1$;
- for $p \in [\frac{2}{3}; 1]$, $b(p) = b_3 = 0$, these values of p are never reached on the sequences (a–d).

In **VATE**, the best e-values are calculated by VATE with a single category, betting the entire capital on $y = +1$. In terms of Sect. 3, we apply $B_n = 1$, a trivial (constant) g_i, and the parameter value of $\varepsilon = 1$.

To measure the multiplicative contribution of Step 3 to the overall growth of the martingale, it is natural to measure it on the logarithmic scale. On average, this makes the following comparison.

$$0.75 \log(b_1) + 0.25 \log(3 - b_1) < 0.5 \log(3) + 0.5 \log(1.5)?$$

The last inequality is true for any positive $b_1 \leq 3$, with the maximum reached at $b_1 = 2.25$.

This example has shown how the usage of the betting function on p-values in CTM is limited by the necessity of applying the same function to the cases of the same size, while the e-value technique is free from this restriction.

4.2. Causality hypothesis

The example above shows a principal issue of CTM: the uniformity of the betting function on the sequences of the same size but of different nature. VATE has shown more flexibility. But how does it reflect real data analysis problems? What are the prerequisites for replacing CTM with VATE?

Prob.	Ref.	Ex.1	Ex.2	Ex.3	p	CTM	e
0.5	(a)	0	**0**		un.$[0;1]$	1	1
0.5	(b)	0	**1**		un.$[0;\frac{1}{2}]$	2	2
0.5	(a)	0	0	**1**	un.$[0;\frac{1}{3}]$	$\times b_1 = b_1$	$\times$3=3
0.5	(b)	0	1	**1**	un.$[0;\frac{2}{3}]$	$\times$un.$\{b_1, 3-b_1\}$	$\times$1.5=3

Table 2: A data example for Sect. 4.3.

A specific property of the probabilistic framework (Venn-ABERS) is the prediction of the *conditional* probability $P(y|x)$ where y is for the label and x is for the object (typically a feature vector). For the CP framework, the division of the example into x and y is more conventional. How can this be reflected in the testing applications?

Our working hypothesis is the relationship with the *causality* in the data set. It is inspired with the example from Vovk et al. (2023) (p.234) the terms 'causal' and 'anticausal' are used for classification problems, in relation to the efficiency of different types of label-conditional CTM.

For better transparency, we explore the terms 'causal' and 'anticausal' as 'object-to-label' and 'label-to-object' causality: whether the label y is generated from the object x, or the object x is generated from the label y.

These terms refer to a natural mechanism for generating data rather than a property of the distribution in the pairs (x, y), because any distribution $P(x, y)$ can be formally decomposed into $P(x)P(y|x)$ or $P(y)P(x|y)$. However, in the presence of a change point, we can make the meaning of these terms more concrete, if we may assume that the shift also follows a type of causality. So, in the case of the object-to-label causality, the conditional distribution $P(y|x)$ is shifted at the change point, while $P(x|y)$ does not change. In contrast, a complementing model of the label-to-object causality needs a point where $P(x|y)$ changes, while $P(y|x)$ does not.

4.3. Change points in Bernoulli sequences

Consider a simplistic example of object-to-label causality in data sequences. Assume now that the object x is trivial ($|X| = 1$), $y \in Y = \{0, 1\}$, and the sequence is of type $\mathbf{000\ldots000111\ldots1111}$ (k zeros followed by $n-k$ ones). What is modelled here? The object does not depend on the label, while the label is determined by the object, and the mechanism of generation of the label from the object is changed at some steps.

Like in the previous example, here the same simple one-category VATE strategy betting strictly for 1, will make

$$\frac{k+1}{1}\frac{k+2}{2}\cdots\frac{n}{n-k} = C_n^k$$

Can the same be reached by a CTM, assuming that 1 is stranger than 0? Yes, but only if k is known in advance. Otherwise, it will be corrupted by the non-adaptive betting function.

For example, let us take only the distribution shared between **001** and **011** (Tab. 2). In the second step, it is easy to make x1 on 001 and x2 on 011. In the third step, we will have the choice: either to bet for $p \in [0; \frac{1}{3}]$ to multiply by 3 on 001, or for $p \in [0; \frac{2}{3}]$ to multiply by 1.5 on the third step of 011, and by 3 on the whole sequence. Unlike perfect fitting to one of them, we will have a loss on the other. So, if we set $b_1 = 1$ to achieve 3 on 001, then the capital on 011 will be either 6 or 0, which is worse than 3 (in average logarithms). The best way for 011 is $b_1 = 1.5$ that makes 3 anyway, but this will decrease the gain on 001 from 3 to 1.5.

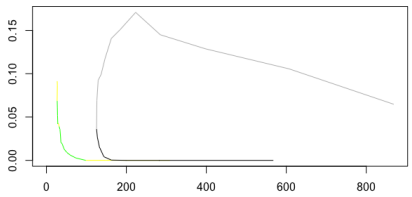
Now, let us consider the model of a change points in a long Bernoulli sequence. There are no feature vectors (x is trivial), and the label is binary (0 or 1). Generalizing the example above, we assume that the probability of 1 was δ before the change point and $1-\delta$ after the change point, while the position of the change point is unknown. For our running, we generate binary sequences of length 1000, with a change point on $1, 2, \ldots, 100$. For each change point, we make 500 random repetitions of the experiment.

Like in the examples above, we apply VATE in the single-category setting, with betting for 1 against 0, in the Modified MUSUC mode. In order to apply the Venn-Abers e-predictor, a value ε of the risk factor must be chosen as a parameter. As a baseline, we use the CTM with the non-conformity score $\alpha_i = y_i$ (1 is always stranger than 0) and the standard betting function $b(p) = \varepsilon p^{\varepsilon-1}$ with variable ε. All other details are the same.

We observe the running until either the change-point alarm is raised for the first time or the time horizon is reached. Then, we compare the time with the true change point and record how late (or how early) it was, measured in time steps. For example, if the true change point is 50 and the alarm was raised at step 79, the score of early detection is 0, and the score of late detection is 29; on the contrary, if the alarm is raised at step 21, there is 29 in early detection and 0 in late detection. Note that in some cases only late detection is observed.

The compared results are presented in Tab. 3 for $\delta = 0.05$ and Tab. 4 for $\delta = 0.20$. In the supplementary plots, the horizontal axis is for the late detection, the vertical axis is for early detection, the green/yellow colour is for VATE and the black/gray is for CTM. We set the threshold for the detection of change points at $C = 1000$. We measure the early and late detection rates averaged over 100 change point positions from 1 to 100, and over 1000 random seeds. For both methods (VATE and CTM), the risk factor is variable from 0 to 1 with the step of 0.05. The stars in the table (as well as gray/yellow colour in the graphs) mark *inadmissible* results, which can be improved in both directions simultaneously (lowering both the early and late detection scores) by replacing the parameter. The remaining black figures show how one of these two scores can be improved at the cost of the other.

The main tables are supported by the plots. Extra plots (Fig. 1–4) are presented after running with other parameter settings: the possible change point location range is extended to 100 to 200, while the threshold is lowered from 1000 to 50. This is done in such a way that the ranges for the late- and early-detection scores are numerically close to each other. In this setting, the CTM and VATE curves are closer, but the advantage of the second is



Risk	VATE-early	VATE-late	CTM-early	CTM-late
0.05	*0.00000	*309.07560	*0.06468	*868.87906
0.10	*0.00008	*149.63342	*0.10548	*607.66668
0.15	0.00000	98.54752	*0.12878	*397.24344
0.20	0.00268	74.07202	*0.14492	*285.02424
0.25	0.00592	60.23724	*0.17074	*223.31232
0.30	0.00902	51.37352	*0.15100	*184.85250
0.35	0.01278	44.71540	*0.14062	*162.44364
0.40	0.01884	39.66642	*0.11706	*146.50678
0.45	0.02100	36.37944	*0.09932	*137.06972
0.50	0.03582	33.97952	*0.09232	*129.54504
0.55	0.03914	31.92256	*0.06918	*126.37102
0.60	0.04280	29.89726	0.03608	125.61090
0.65	0.04186	28.57624	0.02666	127.59618
0.70	0.05210	27.87650	0.01534	132.71852
0.75	0.06124	27.49064	0.00416	143.86996
0.80	0.06846	27.09910	0.00032	162.61128
0.85	0.06862	26.94158	0.00000	199.58800
0.90	*0.08664	*27.47360	*0.00000	*282.35210
0.95	*0.09136	*27.10576	*0.00000	*566.80730

Table 3: VATE and CTM on artificial data with $\delta = 0.05$: (hor.) late detection score; (ver.) early d. score

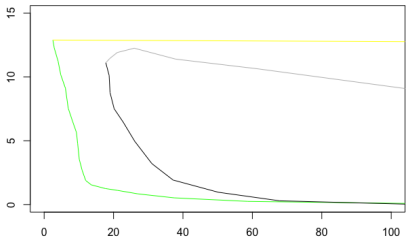
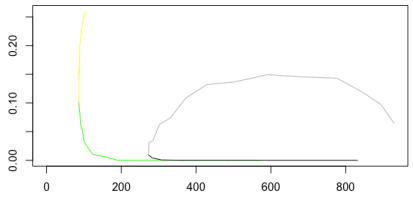


Figure 1: VATE and CTM on artificial data with $\delta = 0.01$ with the change point location up to 200, low threshold $C = 50$: (hor.) late detection score; (ver.) early d. score

still very visible. For these graphs, we also increased the range of values of δ (including 0.01 and 0.40) to show the robustness of the comparison results.



Risk	VATE-early	VATE-late	CTM-early	CTM-late
0.05	0.00000	571.84500	*0.06456	*929.64058
0.10	0.00000	277.92180	*0.09696	*895.02962
0.15	0.00026	188.06258	*0.11848	*845.87960
0.20	0.00778	146.07636	*0.14322	*777.09876
0.25	0.01014	123.44104	*0.14546	*686.15340
0.30	0.02342	109.06224	*0.14940	*590.60830
0.35	0.03204	100.64346	*0.13660	*501.98150
0.40	0.05408	94.85930	*0.13200	*428.18330
0.45	0.06072	90.76998	*0.10872	*372.47154
0.50	0.08302	88.75808	*0.07370	*330.73480
0.55	0.10198	85.78144	*0.06292	*302.22276
0.60	0.10038	86.29856	*0.03388	*283.82496
0.65	*0.11500	*86.43350	*0.03096	*273.67250
0.70	*0.16050	*86.50686	0.00936	272.33138
0.75	*0.15244	*87.48102	0.00408	282.69684
0.80	*0.20212	*88.77330	0.00054	307.11092
0.85	*0.21230	*91.54318	0.00000	362.55860
0.90	*0.24548	*97.06346	*0.00000	*488.12380
0.95	*0.25976	*104.48710	*0.00000	*831.4357

Table 4: VATE and CTM on artificial data with $\delta = 0.2$: (hor.) late detection score; (ver.) early d. score

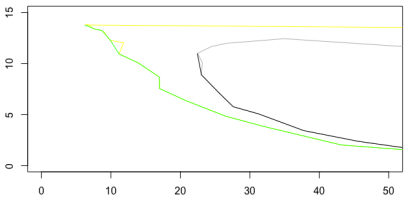


Figure 2: VATE and CTM on artificial data with $\delta = 0.05$ with the change point location up to 200, low threshold $C = 50$: (hor.) late detection score; (ver.) early d. score

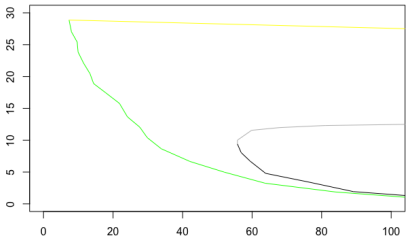


Figure 3: VATE and CTM on artificial data with $\delta = 0.2$ with the change point location up to 200, low threshold $C = 50$: (hor.) late detection score; (ver.) early d. score

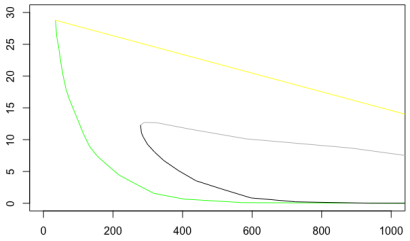


Figure 4: VATE and CTM on artificial data with $\delta = 0.4$ with the change point location up to 200, low threshold $C = 50$: (hor.) late detection score; (ver.) early d. score

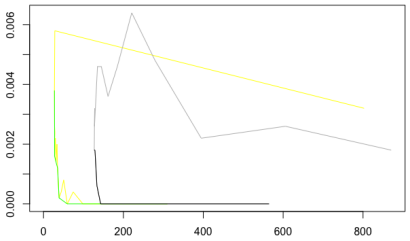


Figure 5: VATE and CTM on artificial data with $\delta = 0.05$:
(hor.) late detection score; (ver.) false alarms

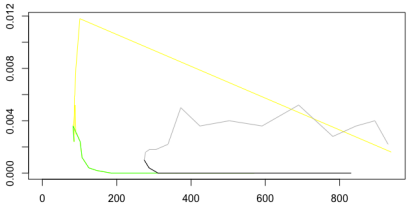


Figure 6: VATE and CTM on artificial data with $\delta = 0.2$:
(hor.) late detection score; (ver.) false alarms

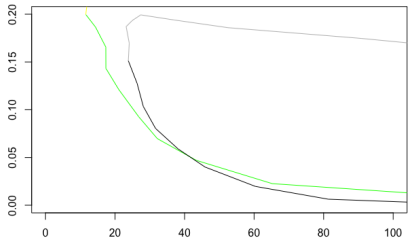


Figure 7: VATE and CTM on artificial data with $\delta = 0.05$ with the change point location up to 200, low threshold $C = 50$: (hor.) late detection score; (ver.) false alarms

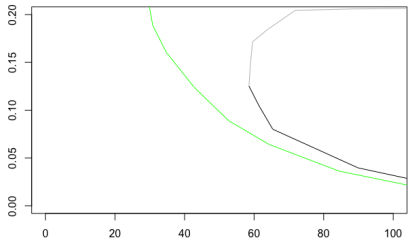


Figure 8: VATE and CTM on artificial data with $\delta = 0.2$ with the change point location up to 200, low threshold $C = 50$: (hor.) late detection score; (ver.) false alarms

In addition (Fig. 5,6,7,8), we include modelling a typical decision-making process in practice. In this setting, the running does not always stop at the first alarm, but multiple alarms are allowed before the real alarm is reached. This means that the delay, shown on the horizontal axis, is always positive. Instead of an early detection score, the vertical axis is for the average *number* of false alarms, raised before the real change point.

Together, these results confirm that the VATE approach is more efficient in an object-to-label problem.

5. Experiments on real data

5.1. Data sets

The data set **Wine Quality** from Repository (2019) consists of two separate parts: 4,898 White Wine instances and 1,599 Red Wine instances. If they are concatenated sequentially, the overall data sequence will have a known change point at 4,899. Originally, the quality of the wine was measured on a scale of 1 to 10. We reduce this to binary classification: 1 if the original value is 5 at least, and 0 otherwise. To model the case of the known change point, we take 500 randomly selected White Wine examples followed by 500 randomly selected Red Wine examples. This data set is taken as an example of object-to-label causality.

For the alternative model, we use the **United States Postal Service (USPS)** dataset from Repository (1994) that is known to be non-exchangeable and initially motivated conformal testing in Vovk et al. (2003). One of the possible causes of the deviation from

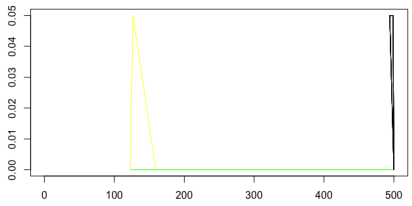


Figure 9: Results on Wine Data Set ($C = 1000$):
(hor.) late detection score; (ver.) false alarms

exchangeability may be a change in handwriting style. Each USPS example is a 16x16 grayscale image representing a hand-written digit. For our example, we take two classes: digit 5 (as the class $y = 0$) and digit 8 (as the class $y = 1$). There are 1,424 such examples in total. To model the case of the known change point, we take 500 first examples (in random order) followed by 500 last examples (not changing their order).

In both data sets, we generate data sequences (of 1000 examples) where the location of the first change points is known (at the middle position 500). In particular, in both of them we reshuffle the order of first 500 examples, so that this part of the data can play the role of a control set for validation, where no real change points are expected to exist. This allows measuring the accuracy according to the same criterion mentioned above: delay in the detection of this change point, or too early detection. We focus on the detection of one change point in a data sequence.

The data generation scheme includes randomisation at different stages. We reshuffle only the first half before the change point for USPS, and both the halves for Wine. In addition, randomization appears in CTM due to the use of p-values. We repeat each experiment 20 times.

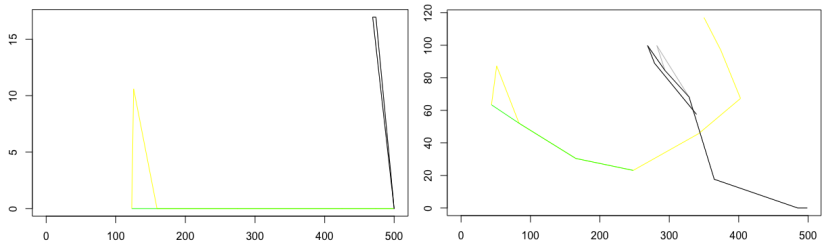
5.2. Results and discussion

We apply a modified MUSUC procedure (Alg. 3). The threshold is set to $C = 1000$ because a smaller one is too likely to be achieved occasionally on a data sequence of this length.

As an underlying algorithm, we use the 1-nearest-neighbour method with Euclidean distance. In the Wine data set, it was applied after each feature is standalised (linearly rescaled to the variance of 1). In USPS, rescaling is skipped because the features are uniform. The scoring function of an example is defined as the ratio of the distance to the nearest neighbour within the other class divided by the distance to the nearest neighbour within the same class.

This is the opposite of the NCM function defined for the CTM baseline: the distance to the nearest neighbor of the same class divided by the distance to the nearest neighbors of the other class. For CTM, a betting function is also needed, and we reuse the standard betting function $b(p) = \varepsilon p^{\varepsilon-1}$.

The main results are presented in Tab. 5 and Tab. 6 for the examples from the Wine and USPS data, as mentioned above. We provide them with the figures where this makes sense for informativeness. As in the previous section, we use the horizontal axis for the late detection score and the vertical axis for the early detection score (alternatively, for the



Method	Risk ε	$C = 1000$ av.early	$C = 1000$ av.late	$C = 50$ av.early	$C = 50$ av.late
VATE	0.1	0	122.9	63.4	43.8
VATE	0.2	*10.60	*125.55	*87.3	*51.3
VATE	0.3	0	159.2	52.05	83.6
VATE	0.4	0	274.85	30.45	165.25
VATE	0.5	0	425.15	23.1	248
VATE	0.6	0	473.7	*45.70	*342.35
VATE	0.7	0	500	*67.3	*402.6
VATE	0.8	0	500	*97.05	*374.1
VATE	0.9	0	500	*117	*350
CTM	0.1	0	500	57.55	339.15
CTM	0.2	16.95	469.05	88.95	278.75
CTM	0.3	16.95	473.75	99.7	268.8
CTM	0.4	16.95	473.75	84.45	293.85
CTM	0.5	0	500	*99.8	*282.05
CTM	0.6	0	500	68.2	328.5
CTM	0.7	0	500	17.65	364.80
CTM	0.8	0	500	0	484.95
CTM	0.9	0	500	0	498.75

Table 5: Results on Wine Data Set (averaged over 20 random seeds), delay of the change point detection (for $C = 1000$ and $C = 50$): (hor.) late detection score; (ver.) early d. score.

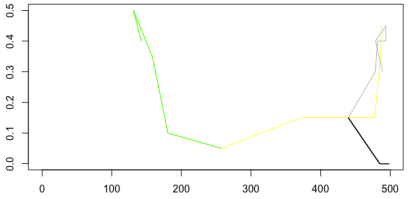
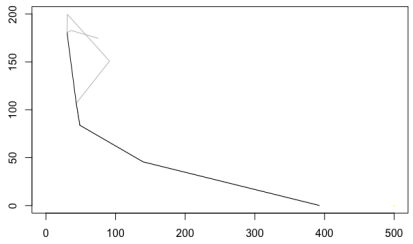


Figure 10: Results on Wine Data Set (low threshold $C = 50$): (hor.) late detection score; (ver.) false alarms



Method	Risk ε	$C = 1000$ av.early	$C = 1000$ av.late	$C = 50$ av.early	$C = 50$ av.late
VATE	any	0	500	0	500
CTM	0.1	0	450.2	*174.6	*75.4
CTM	0.2	0	383.8	*182.8	*36.4
CTM	0.3	0	184.85	180.65	30.30
CTM	0.4	0	184.6	*199.85	*30.60
CTM	0.5	0	184.6	*150.8	*91.4
CTM	0.6	0	218.3	106.85	43.55
CTM	0.7	0	466.85	83.85	48.75
CTM	0.8	0	483.4	45.6	139.5
CTM	0.9	0	500	0	392.8

Table 6: Results on USPS Data Set (averaged over 20 random seeds), delay of the change point detection (for $C = 50$): (hor.) late detection score; (ver.) early d. score.

number of false alarms). The black colour represents admissible CTM points and the green colour represents admissible VATE points.

For both approaches, we observe the dynamics and how the results depend on the risk factor ε. The results are presented in the tables as a *difference* between the first alarm of the change point raised by an algorithm and the true position of the change point (that is, always 500). So, the possible range is between -500 (meaning that the change point is detected immediately) and +500 (meaning that it was either not detected at all, or only at the last moment). The desirable result is a positive number slightly above zero, which means a small delay.

VATE algorithm also needs a setting of $B_n \in \{0, 1\}$ (whether the betting is for 1 or for 0). For symmetry, we consider the average (the mixture) of these two e-pseudomartingales as the final result. The results in Wine could be further improved by taking only $B_i = 0$, but we consider it an unfair hint. The best results with the VATE approach were achieved with a relatively low risk level (0.1). If $C = 1000$, the early detection rate is 0 in most cases. Therefore, we also present the results for $C = 50$, to show how the results would look when both types of error are possible. In general, VATE still shows an advantage where the results are comparable.

Overall, with the optimised parameters, VATE has shown its advantage on the Wine data set, while CTM looks better on the USPS data set. This matches our expectations because these two data sets show object-to-label and label-to-object causation, respectively.

A possible advantage of CTM here is the ability to assign low p values to all hypotheses. In a machine learning context, such predictions are called low-credibility. In case of a concept drift at a change point, the shifted classes first look as new classes which have some difference from all the classes observed before. Venn machine by its nature needs to assign a high probability at least to one of them, therefore it is hardly suitable for anomaly detection tasks, and for the same reason it is less suitable for a change in a handwriting style. This is why VATE may be to be efficient on such kinds of data.

6. Conclusion

In this work, we study the advantages of the e-prediction martingale testing technique for the change point detection. Compared with conformal martingales, e-prediction martingales are simpler and more transparent because they do not include the additional step of checking a given p-value for being distributed uniformly. This allows us to create a generous scheme for creating martingales. In particular, in this work we have studied involving Venn-Abers (probabilistic) prediction within this approach.

We initially checked the causality hypothesis on the applicability of the new technique. It is worth further investigation, especially in combined cases. For example, in a typical medical area, one can distinguish the factors (increasing the risk of a disease) and the symptoms (that are caused by a disease) in the same feature vector. Another important direction for future work may be the detection of multiple change points.

7. Acknowledgements

The authors thank Alex Gammerman and Volodymir Vovk for very useful discussions and to the University of Stockholm for support of the research.

References

V. Fedorova, A. Gammerman, I. Nouretdinov, and V. Vovk. Plug-in martingales for testing exchangeability on-line. ICML'12, pages 923–930, Madison, WI, USA, 2012a. Omnipress.

V. Fedorova, E. Ivin, I. Nouretdinov, and A. Gammerman. Testing and clustering with gauss linear assumption for a household data, 2012b. URL https://pure.royalholloway.ac.uk/en/publications/testing-and-clustering-with-gauss-linear-assumption-for-a-househo.

V. Fedorova, I. Nouretdinov, and A. Gammerman. Testing the gauss linear assumption for on-line predictions. 1(3):205–213, 2023. URL https://doi.org/10.1007/s13748-012-0022-x.

S.-S. Ho. A martingale framework for concept change detection in time-varying data streams. In *Proceedings of the 22nd International Conference on Machine Learning*,

ICML '05, pages 321–327, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102392. URL https://doi.org/10.1145/1102351.1102392.

I. Nouretdinov and A. Gammerman. The Venn-Abers testing for change-point detection. In *Proceedings of Machine Learning Research: 12th Symposium on Conformal and Probabilistic Prediction with Applications*, 2023.

M. Purtle. Conformal anomaly detection, August 2020. Master's thesis.

Repository. Handwritten digits usps dataset, 1994. URL https://www.kaggle.com/bistaumanga/usps-dataset.

Repository. Wine quality data set, 2019. URL https://archive.ics.uci.edu/dataset/186/wine+quality.

A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8:22–46, 1963. URL https://api.semanticscholar.org/CorpusID:120941074.

A. N. Shiryaev. Quickest detection problems: Fifty years later. *Sequential Analysis*, 29: 345–385, 2010. URL https://api.semanticscholar.org/CorpusID:122295674.

D. Volkhonskiy, E. Burnaev, I. Nouretdinov, A. Gammerman, and V. Vovk. Inductive conformal martingales for change-point detection. In *International Symposium on Conformal and Probabilistic Prediction with Applications*, 2017. URL https://api.semanticscholar.org/CorpusID:7698983.

V Vovk. Testing randomness, 2022. at https://arxiv.org/abs/1906.09256.

V. Vovk. Conformal e-prediction for change detection, 2023a. URL https://arxiv.org/abs/2006.02329.

V. Vovk. Cross-conformal e-prediction, 2023b. URL https://arxiv.org/abs/2001.05989.

V. Vovk and I. Petej. Venn-Abers predictors, June 2014. URL https://arxiv.org/abs/1211.0025.

V. Vovk, I. Nouretdinov, and A. Gammerman. Testing exchangeability on-line. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 768–775. AAAI Press, 2003.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, New York, 2023.

V. Vovk, I. Nouretdinov, and A. Gammerman. Validity and efficiency of the conformal cusum procedure, 2024. URL https://arxiv.org/pdf/2412.03464.

V. Vovk, I. Nouretdinov, and A. Gammerman. Conformal e-testing. *Pattern Recognition*, 2025.