Proceedings of Machine Learning Research 230:1–19, 2024 Conformal and Probabilistic Prediction with Applications

Reliable Change Point Detection for ACGH data

Charalambos Eliades ST009072@STUD.FREDERICK.AC.CY
Computational Intelligence (COIN) Research Lab, Frederick University, Cyprus

Harris Papadopoulos H.PAPADOPOULOS@FREDERICK.AC.CY
Computational Intelligence (COIN) Research Lab, Frederick University, Cyprus

Editor: Editor's name

Abstract

This study introduces two algorithms based on the Inductive Conformal Martingale (ICM) approach to address the change point (CP) detection problem in array-based Comparative Genomic Hybridization (aCGH) data. The ICM, a distribution-free approach with minimal assumptions, is particularly suitable for this application. We have implemented two ICM-based algorithms; the first utilizes nonconformities from preprocessed data, while the second incorporates the label conditional distribution and the labels' distribution to enhance detection accuracy. This approach significantly improves our results, demonstrating the potential of ICM in complex genomic data analysis.

Keywords: ICM, aCGH, Change Point

1. Introduction

Detecting DNA sequence alterations is essential for understanding the genetic underpinnings of various disorders and cancers. Comparative Genomic Hybridization (CGH), first introduced by Kallioniemi A (1992), provides a comprehensive method for analyzing chromosomal variations across the genome. This technique saw significant advancements with the introduction of Array CGH (aCGH) by Pinkel et al. (1998), which uses microarrays to improve resolution. aCGH enables the accurate identification of genomic gains or losses by measuring fluorescence ratios between test and reference DNA on microarrays, a critical step for developing targeted therapeutic strategies.

The ideal case would be for the log2 ratio to be zero; however, a positive log2 ratio indicates a gain in the genomic sequence, while a negative log2 ratio indicates losses. While observing a problematic aCGH sequence, the data-generating mechanism may change at some point, i.e., to start showing high gains or losses. It is also possible that the aCGH sequence will switch from gains to losses and vice versa; thus, we do not deal only with the absolute value of these changes but also take into account their sign, i.e. whether it is a gain or a loss. We aim to identify these change points (CP), which is a complex and time-consuming tasks. Thus, implementing a reliable method of low computational complexity that allows for the analysis of each person's aCGH data and detects at which point a change occurs, will contribute to understanding the kind of genetic disorder medical practitioners are dealing with, enabling treatments tailored to specific genetic variations (Doudican et al., 2015).

Our work introduces a CP detection algorithm tailored for aCGH data by testing the exchangeability assumption at a prespecified significance level. Our methodology is based

© 2024 C. Eliades & H. Papadopoulos.

on Inductive Conformal Martingales (Volkhonskiy et al., 2017) and draws upon the properties of Conformal Prediction, as introduced by Vovk et al. (2005), to evaluate the data's exchangeability. Conformal Martingales is a robust framework for identifying distribution changes, which makes it suitable for genomic data where statistical properties undergo significant shifts.

Most previous works rely on assumed distributions for hypothesis testing or fit a model on the data. However, these assumptions might not always hold; the model might not fit on the data well, or the assumed distribution might be wrong.

The proposed approach requires only data preprocessing and tests the exchangeability assumption without assuming a particular distribution. Furthermore, we don't assume any structure in the data, e.g. constant variance or mean within segments. These properties allow us to apply this CP framework across various genomic data types without the need of any model on the data. These minimal assumptions ensure our method's probabilistic validity.

A CP occurs when the underlying data distribution shifts, given a data stream $S = \{(x_0, y_0), (x_1, y_1), \dots\}$ consisting of feature vectors x_i and labels y_i. A CP at timestamp t occurs when S can be segmented into two sets $S_{0,t} = \{(x_0, y_0), \dots, (x_t, y_t)\}$ and $S_{t+1,\dots} = \{(x_{t+1}, y_{t+1}), \dots\}$, such that $S_{0,t}$ and $S_{t+1,\dots}$ are generated by different distributions.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of the existing literature on change detection. In Section 3, we introduce the principles upon which our methodology is based. Section 4 details our proposed methodology, followed by Section 5, which presents our experimental framework, the performance metrics used, and the results of our evaluations. We conclude in Section 6 by summarizing our findings and suggestions for future research.

2. Related work

Given the vast amount of research on this topic we will present only the most prominent works related to the method we follow.

2.1. Related Work on ACGH Change Point Detection

In this section, we explore the contributions of various researchers towards CP detection in aCGH data.

Circular Binary Segmentation (CBS), introduced by Olshen et al. (2004), assumes no specific data distribution but employs permutation-based reference distributions. Assuming that the data has a constant number of the log2 ratio in each segment, it identifies regions with a statistically significant change in the mean log2 ratio of the aCGH data.

The Energy Divisive (ED) Algorithm proposed by Matteson and James (2014) is a permutation-based nonparametric method. This method employs hierarchical clustering through divisive and agglomerative algorithms to estimate the number and locations of change points.

Chen and Wang (2009) introduced the Mean and Variance CP Model (MVCM), a novel approach for identifying copy number variations in aCGH data based on hypothesis testing for mean and variance. This model, which assumes a Gaussian distribution of log2 ratios, employs a Binary Segmentation Procedure alongside the Schwarz Information Criterion

to accurately detect CNVs. Through iterative refinement of genomic segments, MVCM efficiently identifies significant changes in mean and variance.

In their work, Hyun et al. (2021) advanced the field of CP detection within copy number variation (CNV) data analysis by tailoring post-selection inference methods to this context, under the assumption of normally distributed (Gaussian) data variables. They managed to produce uniformly distributed p-values, which is a crucial aspect for determining the statistical significance of the identified CP's.

While the previously discussed methods effectively analyse aCGH data, they depend heavily on specific statistical assumptions about the data distribution. For instance, MVCM presupposes a Gaussian distribution of log2 ratios, an assumption which may not hold under specific biological conditions where such assumptions are invalid. Meanwhile, even though CBS and ED do not assume a specific distribution, they rely on permutation-based references, which can be computationally intensive if a high number of permutations is required. In contrast, the approach proposed in this work does not rely on heavy computations or assumptions about the aCGH data, enhancing its effectiveness across diverse genomic datasets where verifying statistical assumptions can be challenging.

2.2. Related Work on Conformal Martingales

In this section, we explore the contributions of various researchers towards testing the Exchangeability Assumption (EA) using Conformal Martingales (CM). A notable challenge in this domain has been the reliance on specific distributional assumptions for test statistics, a limitation adeptly addressed by the introduction of CM. This innovative approach, detailed in (Vovk et al., 2005), offers a robust framework for EA testing without the constraints of predefined distributional assumptions on the test statistics.

One notable contribution by Vovk et al. (2003) in this field involved a method for online exchangeability testing based on Conformal Prediction and Conformal Martingales. This method involves computing a sequence of p-values using conformal prediction online, where each new example's p-value is determined using new and previously seen examples. Following this, a Betting Function (BF) is applied to each p-value, and the product of these BF outputs forms the Martingale's value. When the Martingale's value M becomes sufficiently large, the EA can be rejected at a significance level of $1/M$. This approach is valuable for testing if a dataset satisfies the EA and detecting CPs in time series, aiding in Change Detection.

Further developing the concept introduced by Vovk et al. (2003), another study (Ho, 2005) introduced an enhanced Conformal Martingale (CM) that utilizes a straightforward betting mixture function. This adaptation is tailored to identify concept shifts within dynamic data streams. The authors of this work formulated two types of martingale tests: one predicated on the values of the martingale itself and the other on the differences observed in successive martingale values. Both types of tests were derived using the mixture Betting Function (BF) as their core computational element.

Extending these principles, another investigation Fedorova et al. (2012) applied them to test the exchangeability of data in two datasets, USPS and Statlog Satellite. The approach involves online testing, where data is processed sequentially, and the CM value is computed as a valid measure for assessing the EA. They utilized a density estimator for the observed

p-values as a BF, with kernel density estimation showing superior performance to the simple mixture BF.

In addition, Volkhonskiy et al. (2017) introduced an Inductive version of CM for detecting changes in time series. In their study, the initial observations of the time sequence are used to train the underlying model, and all nonconformity scores are calculated via this model. They experimented with several BFs and found that the pre-computed kernel BF yields the most efficient results, evidenced by the lowest mean delay in their tests on synthetic datasets. Their findings are comparable with other methods like CUSUM, Shiryaev-Roberts, and Posterior Probability statistics.

In their work Ho and Wechsler (2012) leverage CM to identify concept changes in data streams by examining the EA. Their innovative approach, grounded on Doob's Maximal Inequality, establishes a robust framework for hypothesis testing within time-varying data environments. They rigorously tested their methodology on synthetic and real-world datasets, showcasing its applicability and effectiveness in detecting concept changes.

A novel real-time martingale-based approach is proposed by Ho et al. (2019) using Gaussian Process Regression (GPR) to predict and detect anomalous flight behaviour as data arrives sequentially. They implemented multiple CM tests to reduce the number of false alarms and the detection delay time, again utilizing the mixture BF for Martingale calculation.

In our study presented in Eliades and Papadopoulos (2021), we explored the integration of ICM with a histogram betting function. This novel combination is specifically designed to detect violations of the EA and, as a result, identify CD in data streams. Notably, our approach is distribution-free, distinguishing it from other methods that often presuppose a specific distribution in their drift detection metrics.

Eliades and Papadopoulos (2022) introduced the 'Cautious' betting function to improve the detection of Concept Drift by preventing the martingale values from going near zero when no changes occur. This method was tested with kernel and histogram functions across five datasets, demonstrating enhanced detection and model accuracy.

To conclude, this section lays the groundwork for our study, in which we employ Inductive Conformal Martingales (ICM) to tackle Change Detection (CD) challenges. Inspired by the seminal studies reviewed, our methodology tailors ICM to specifically address the dynamic characteristics of changing data streams.

3. Inductive Conformal Martingales

In this section we describe the basic concepts of ICM and how our nonconformity scores and p-values are calculated.

3.1. Data Exchangeability

Let $(Z_1, Z_2, \dots)$ be an infinite sequence of random variables. Then the joint probability distribution $\mathbb{P}(Z_1, Z_2, \dots, Z_N)$ is exchangeable if it is invariant under any permutation of these variables. The joint distribution of the infinite sequence $(Z_1, Z_2, \dots)$ is exchangeable if the marginal distribution of $(Z_1, Z_2, \dots, Z_N)$ is exchangeable for every $N \in \mathbb{N}$. Testing if the data is exchangeable is equivalent to testing if it is independent and identically distributed

(i.i.d.); this is an outcome of de Finetti's theorem (Schervish, 1995): any exchangeable distribution on the data is a mixture of distributions under which the data is i.i.d.

3.2. Exchangeability Martingale

A test exchangeability Martingale is a sequence of random variables $(S_1, S_2, S_3, \dots)$ being equal to or greater than zero that keep the conditional expectation $\mathbb{E}(S_{n+1}|S_1, \dots, S_n) = S_n$.

To give an idea how a martingale works, consider a fair game where a gambler with infinite wealth follows a strategy based on the distribution of the events in the game. The gain acquired by the gambler can be described by the value of a Martingale, specifically Ville's inequality (Ville, 1939) indicates that the probability of having high profit (C) is small, $\mathbb{P}\{\exists n : S_n \geq C\} \leq 1/C$.

According to Ville's inequality for the case of the EA, a large final value of the Martingale suggests rejection of the assumption with a significance level equal to the inverse of the Martingale value, i.e. a Martingale value of 10 or 100 rejects the hypothesis of exchangeability at 10% or 1% significance level, respectively.

3.3. Calculating Non-conformity Scores and Pvalues

As mentioned in Section 1, this study eliminates the necessity of assuming a model for calculating Non-Conformity Scores (NCS); instead, it requires only preprocessing of the data. Here, we begin with an overview of Inductive Conformal Martingales (ICM) and demonstrate how this approach can be adapted to our context, which is free from model dependencies. We will show how we calculate pvalues, label conditional pvalues and label pvalues.

Let $\{z_1, z_2, \dots\}$ be a sequence of examples, where $z_i = (x_i, y_i)$ with x_i an object given in the form of an input vector, and y_i the label of the corresponding input vector. The CM approach generates a sequence of pvalues corresponding to the given sequence of examples and then calculates the martingale as a function of these p-values. ICM uses the first k examples $\{z_1, z_2, \dots, z_k\}$ in the sequence to train a classification algorithm, which it then uses to generate the p-values for the next examples. Consequently, it starts checking for violations of the EA from example z_{k+1} on, i.e. the sequence $\{z_{k+1}, z_{k+2}, \dots\}$.

Our aim is to examine how strange or unusual a new example $z_j \in \{z_{k+1}, z_{k+2}, \dots\}$ is. For this purpose, we define a function $A(z_i, \{z_1, \dots, z_k\})$, where $i \in \{k+1 \dots\}$, called a nonconformity measure (NCM) that assigns a numerical value α_i to each example z_i, called nonconformity score (NCS). The NCM is based on the trained underlying classification algorithm. The bigger the NCS value of an example, the less it conforms with $\{z_1, \dots, z_k\}$ according to the underlying algorithm.

For every new example z_j we generate the sequence $H_j = \{\alpha_{k+1}, \alpha_{k+2}, \dots, \alpha_{j-1}, \alpha_j\}$ to calculate its p-value. Note that the NCSs in H_j are calculated with the underlying algorithm trained on $\{z_1, z_2, \dots, z_k\}$. Given the sequence H_j we can calculate the corresponding p-value (p_j) of the new example z_j with the function:

$$p_j = \frac{|\{\alpha_i \in H_j | \alpha_i > \alpha_j\}| + U_j \cdot |\{\alpha_i \in H_j | \alpha_i = \alpha_j\}|}{j - k}, \tag{1}$$

where α_j is the NCS of the new example, α_i is the NCS of the i^{th} element in the example sequence set and U_j is a random number from the uniform distribution (0,1). Here we calculate the p-value p_j for a new example z_j by comparing its nonconformity score α_j relative to all previous scores in the sequence set H_j. For more information, refer to (Vovk et al., 2003).

As described in (Vovk, 2020) it is possible to calculate label conditional pvalues given by the slightly modified function:

$$p_j = \frac{|\{\alpha_i \in H_j | \alpha_i > \alpha_j \wedge y_i = y_j\}| + U_j \cdot |\{\alpha_i \in H_j | \alpha_i = \alpha_j \wedge y_i = y_j|}{|\{\alpha_i \in H_j | y_i = y_j\}|}, \tag{2}$$

In this adaptation, the p-value calculation specifically focuses on examples with the same label as the new example z_j. By conditioning the comparison on the label y_j, the p-value p_j is determined by comparing the nonconformity score α_j of the new example against only those scores from the sequence H_j that belong to the same label category. This adaptation is crucial in datasets where each example's distribution depends on the label.

In our study we also use label-pvalues as described in (Vovk, 2020). The label nonconformity measure we use is:

$$\tilde{a}_j = median\{a_i \in \{\alpha_1, \ldots, \alpha_j\} | y_j = y_i\} \tag{3}$$

Then we generate the sequence $\tilde{H}_j = \{\tilde{\alpha}_{k+1}, \tilde{\alpha}_{k+2}, \ldots, \tilde{\alpha}_{j-1}, \tilde{\alpha}_j\}$ and the produced label pvalues are calculated by

$$p_j = \frac{|\{\tilde{\alpha}_i \in \tilde{H}_j | \tilde{\alpha}_i > \tilde{\alpha}_j\}| + U_j \cdot |\{\tilde{\alpha}_i \in \tilde{H}_j | \tilde{\alpha}_i = \tilde{\alpha}_j\}|}{j - k}, \tag{4}$$

The nonconformity score a_j is transformed to $\tilde{a}_j$ to measure the strangeness of the label of each new example z_j. The label p-value p_j is then calculated by comparing this transformed nonconformity $\tilde{a}_j$ with all other label nonconformity scores in the sequence $\tilde{H}_j$.

However, in this study, no model is required, thus $k = 0$. Therefore, it is possible to start checking for violations of the EA from example z_1 onward, i.e., the sequence $\{z_1, z_2, \ldots\}$. The function $A(z_i, \{z_1, \ldots, z_k\})$ simplifies to $A(z_i, M)$, where M denotes both the molecular cytogenetic method used to produce the aCGH data and the preprocessing applied to this data. Furthermore, the label y_j used in (2), (3) and (4) is determined as follows: it takes the value GAIN if $z_j > 0$ and LOSS if $z_j < 0$.

3.4. Inductive Conformal Martingales

An ICM is an exchangeability test Martingale (see Subsection 3.2), which is calculated as a function of p-values such as the ones described in Subsection 3.3.

Given a sequence of p-values $(p_1, p_2, \ldots)$ the Martingale S_n is calculated as:

$$S_n = \prod_{i=1}^{n} f_i(p_i) \tag{5}$$

where $f_i(p_i) = f_i(p_i | p_1, p_2, \ldots, p_{i-1})$ is the betting function (Vovk et al., 2003).

The betting function should satisfy the constraint: $\int_0^1 f_i(p)dp = 1, f_i(p) \geq 0$ and also the S_n must keep the conditional expectation: $\mathbb{E}(S_{n+1}|S_0, S_1, \ldots, S_n) = S_n$.

The integral $\int_0^1 f_i(p)dp$ equals to 1 because $f_i(p)$ is the p-values $(p_1, p_2, \ldots, p_{i-1})$ density estimator. We also need to prove that $\mathbb{E}(S_{n+1}|S_0, S_1, \ldots, S_n) = S_n$ under any exchangeable distribution.

Proof $\mathbb{E}(S_{n+1}|S_0, S_1, \ldots, S_n) = \int_0^1 \prod_{i=1}^n f_i(p_i) \cdot f_{n+1}(p)dp = \prod_{i=1}^n f_i(p_i) \cdot \int_0^1 f_{n+1}(p)dp = \prod_{i=1}^n f_i(p_i) = S_n$ ■

Using (5), it is easy to show that $S_n = S_{n-1} \cdot f_n(p_n)$, which allows us to update the Martingale online. Let us say that the value of S_n equals M, then Ville's inequality (Ville, 1939) suggests that we can reject the EA with a significance level equal to $1/M$.

When calculating p-values using (2), we obtain a label conditional conformal Martingale, denoted as S_n^c. This tests whether the label conditional distribution, in our case the amplitude of the log2 ratio for specific classes of labels ('GAINS' or 'LOSSES'), changes. Conversely, when using (4), the result is a label Conformal Martingale, denoted as S_n^l, which tests if the distribution of labels ('GAINS' or 'LOSSES') changes. The product of these two Martingales, $S_n^c \cdot S_n^l$, forms an exchangeability Martingale as demonstrated by Vovk (2020), which tests whether the joint distribution of amplitude and label has changed. Thus, if this product reaches a value equal to M, Ville's inequality (Ville, 1939) allows us to reject the Exchangeability Assumption (EA) with a significance level equal to $1/M$. This product of Martingales contains valuable information about both the amplitude changes and label distribution in the dataset,which, in our case, helps to detect more CP.

Note that we can calculate equation (5) in the logarithmic scale to deal with precision issues.

4. Proposed Approach

This section presents our approach for detecting CP in aCGH data. Our method identifies CPs by evaluating the exchangeability assumption (EA) against a predefined significance level. A violation of the EA, indicating a shift in the data-generating mechanism, signals the presence of a CP. One of the key advantages of our methodology is its independence from traditional model fitting and the assumption-free nature of our hypothesis testing concerning the EA.

Our process begins with applying median filters to the data, followed by standardization through the use of sliding windows. This preprocessing step enables us to define nonconformities in the data by calculating the absolute values of the filtered data. Subsequently, for each instance, we determine the labels based on the sign of the log2 ratio (GAIN or LOSS). We then compute p-values and assess the Exchangeability Assumption (EA) using Inductive Conformal Martingales (ICM).

To implement CM with the derived sequence of p-values, we must employ a betting function. The 'Cautious' betting function (Eliades and Papadopoulos, 2022), integrated with density estimation techniques such as the Simple Histogram or the Kernel Density Estimator (KDE), has been utilized.

In the following subsections, we describe the 'Cautious' betting function, the density estimators integrated with it, the preprocessing techniques applied to our dataset, and the details of our CP detection algorithm.

4.1. Preprocessing

As previously mentioned, our methodology does not depend on fitting any specific model to the data; thus, nonconformities can be calculated directly after preprocessing. The preprocessing steps are as follows:

- **Missing Values Removal:** All missing values are removed as their accurate estimation is challenging due to the lack of technical details.
- **Outlier Removal:** Our dataset includes a dummy variable for outlier identification, simplifying this task. We eliminate outliers because extreme values can lead to false alarms of CPs or even miss CPs.
- **Median Filtering:** This step smoothes the data by replacing each instance with the median of its neighbouring values within a predefined window. This process helps eliminate undetected outliers, reduces noise, and preserves the data's dynamics.
- **Data Normalization:** We normalized the filtered data using a moving standard deviation. Each data point was divided by the standard deviation of its neighbours, utilizing the same window size as in the median filtering step.
- **Absolute Value Calculation:** Given that high deviations from zero indicate significant losses or gains in the genomic data, our final step consists of calculating the absolute value of the normalized data.

This study uses the absolute values obtained from the normalization process as *nonconformity scores* a_i. Ideally, all nonconformity scores would be zero for a perfectly conforming subject, indicating no deviation from a reference subject. High nonconformity scores suggest significant deviations, potentially indicating anomalies or changes in the genomic sequence, whereas low scores suggest minimal deviations.

However, using absolute values alone results in a loss of information about the direction of change, whether a gain or a loss in the genomic sequence. To illustrate this, consider the sequence $\{0.1, 0.1, 0.1, 0.1, \ldots, -0.1, -0.1, -0.1, -0.1, \ldots\}$ with corresponding labels $\{G, G, G, G, \ldots, L, L, L, L, \ldots\}$. Although the magnitude of the scores remains consistent, there is an evident shift in the label distribution from gains to losses, signifying a CP in the genomic data. Consequently, each instance is labelled: 'Gain' if the log2 ratio is positive and 'Loss' if negative.

Additionally, for calculations required to determine the value of the label Martingale, denoted as $\tilde{a}_i$, we take the median of the sequence $\{a_1, \ldots, a_i\}$, such that $\tilde{a}_i = \text{median}\{a_1, \ldots, a_i\}$. This measure serves as the nonconformity score used in Martingale computations to assess the exchangeability of the label sequences.

4.2. Cautious Betting Function

Here, we describe the Cautious Betting Function proposed in (Eliades and Papadopoulos, 2022). An issue of the CM and ICM is that they might need much time to recover from a value very close to zero (Volkhonskiy et al., 2017). This betting function avoids betting (i.e. $h_n = 1$) when insufficient evidence is available to reject the EA, thus keeping the value

of S_n from getting close to zero and reducing the time needed to detect a CD. Theorem 1 establishes that under a uniform distribution of p-values, any betting function diverging from constant unity leads to S_∞ identically equaling zero. Theorem 2 extends this, showing that for any sequence of betting functions converging uniformly to a function other than constant unity, S_∞ will also converge to zero. These principles underscore the selection of the Cautious Betting Function for its strategic avoidance of unnecessary bets, ensuring more stable and timely CP detection.

Theorem 1 (Eliades and Papadopoulos (2022)) *When the distribution of the p-values is uniform then for any betting function f other than $f = 1$, it follows that $S_\infty \equiv 0$*

Theorem 2 (Eliades and Papadopoulos (2024)) *When the distribution of the p-values is uniform, for any sequence of betting functions f_i that converges uniformly to a function f other than $f = 1$, it follows that $S_\infty \equiv 0$.*

Before defining the mathematical formulation of our Cautious Betting Function, let us consider the strategic interplay between two hypothetical players in a game of probability. Player 1 evaluates the performance of Player 2, who employs a variable betting strategy based on the density estimator f_n. This evaluation guides Player 1's decision to bet or abstain. The following equation formalizes this strategic evaluation, where the decision to bet hinges on a comparison of recent and past performance metrics:

$$h_n(x) = \begin{cases} 1 & \text{if } S1_{n-1}/\min\limits_k S1_{n-k} \leq \epsilon \\ f_n & \text{if } S1_{n-1}/\min\limits_k S1_{n-k} > \epsilon \end{cases} \tag{6}$$

with $S1_n = \prod_{i=1}^{n} f_i(p_i)$ representing the cumulative product of betting functions applied to p-values, and k spanning the range $\{1, \ldots, W\}$, the parameters $\epsilon > 0$ and $W \in \{1, \ldots, n-1\}$. ϵ, set at 10, serves as a critical threshold, beyond which betting is deemed justified based on the evidence against the exchangeability assumption. Contrary to initial intuition, a higher ϵ implies a more cautious approach, requiring stronger evidence for betting, thus enhancing the model's precision by betting only when substantial evidence is present. Meanwhile, W, fixed at inf, determines the breadth of historical data considered, enabling a broad analysis of past performance to guide current betting decisions.

In (Eliades and Papadopoulos, 2022), the parameters ϵ and W are set to $\epsilon = 100$ and $W = 5000$, here we set ϵ to 10 and W to inf. This modification is necessitated by the specific characteristics of our aCGH data sequence, which comprises fewer than 200 instances per chromosome.

4.3. Density estimators

Here we describe the density estimators we have used to combine with the Cautious betting function namely the: Histogram Density and the Kernel Density Estimators.

4.3.1. Kernel Density Estimator

This betting function is based on the kernel density estimate (KDE), which is a non parametric method, for approximating the p-value distribution. One drawback of the kernel

density estimator is that it is computationally expensive. Another drawback is that in some cases the estimation of the optimum bandwidth is a very time consuming task, for this reason we have used Silverman's "rule of thumb" (Silverman, 1986) for bandwidth selection. The KDE will be equal to:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=n-L+1}^{n} k(\frac{|x - x_i|}{h}) \tag{7}$$

where h is a bandwidth parameter and k is the simple Gaussian:

$$k(z) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}z^2). \tag{8}$$

Note that while calculating the KDE we have used the reflection method as in (Fedorova et al., 2012) to improve performance for points that are near the bounds [0,1]. Also to eliminate the risk of having an x_0 with $\hat{f}_n(x_0) = 0$ and that would lead to Martingale values equal to zero we add a negligible constant to the $\hat{f}_n(x)$. Because this constant is set to be extremely small (10^{-10}) it does not disturb the performance of KDE. The integral of $A = \int_0^1 (f_n(x) \hat{+} 10^{-10})dx \approx 1$ thus practically there is no need to multiply the integral with any constant to force equality to 1.

4.3.2. Histogram Density Estimator

Compared to KDE the histogram estimator is faster (Eliades and Papadopoulos, 2021) and needs less computational effort to tune. The p-values $p_i \in [0, 1]$, so we partition $[0, 1]$ into a predefined number of bins k and calculate the frequency of the observations that lie in each bin. Dividing these frequencies by the total number of observations and multiplying it by the number of bins gives us the histogram estimator.

Let us take a fixed number of bins κ this will partition $[0, 1]$ into $B_1 = [0, 1/\kappa)$, $B_2 = [1/\kappa, 2/\kappa)$,..., $B_{\kappa-1} = [(\kappa - 2)/\kappa, (\kappa - 1)/\kappa)$ and $B_\kappa = [(\kappa - 1)/\kappa, 1]$. Then for a p-value $p_n \in B_j$ the density estimator will be equal to:

$$\hat{f}_n(p_n) = \frac{n_j.\kappa}{n - 1}, \tag{9}$$

where $n-1$ is the number of p-values seen so far and n_j is the number of p-values belonging to B_j. Note that when n is small it is possible that $\exists x : \hat{f}_n(x) = 0$, in that case until a sufficient number of observations arrives we reduce the number of bins κ by 1, the reduction of κ is repeated until $\nexists x : \hat{f}_n(x) = 0$.

4.4. Detecting CP using ICM

In order to detect a CP at a pre-specified significance level δ, the Martingale value must exceed $1/\delta$, which leads to the rejection of the EA. This process is summarized in Algorithm 1. Specifically, if the Martingale value S_k at a given point k exceeds 100, a CP is detected at a significance level of 1%, where L denotes the number of p-values that our estimator uses.

Another method to detect CPs utilizes the label conditional conformal martingale (S_n^c) and the label conformal martingale (S_n^l).

Algorithm 1: Detect CP using ICM

1. Require set $\{z_1, z_2, \ldots, z_k\}$, significance level δ
2. Initialize $S_1 = 1$, $H = \{\}$
3. Apply preprocessing steps (see Section 4.1) to $\{z_1, z_2, \ldots, z_k\}$ to obtain $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$
4. For $i = 1$ to k
   (a) Append (α_i) to H
   (b) $p_i = \frac{|\{\alpha_j \in H | \alpha_j < \alpha_i\}| + U_j \cdot |\{\alpha_j \in H | \alpha_i = \alpha_j\}|}{|H|}$
   (c) Calculate betting function $h_i = h(p_{i-L}, \ldots, p_{i-1})$
   (d) $S_i = S_{i-1} \cdot h_i(p_i)$
   (e) If $S_i > \frac{1}{\delta}$
      i. Raise an Alarm
      ii. $H = \{\}$

In Algorithm 1, we calculated p-values using (1). To compute the label conditional conformal martingale (S_n^c) and the label conformal martingale (S_n^l), the process is slightly modified, and we employ (2) and (4) respectively for p-value calculations. Here, S_n^c tests for changes within the conditional distribution of the data, specifically analyzing the amplitude of the log2 ratio for distinct classes of labels (such as 'GAINS' or 'LOSSES'), while S_n^l tests for changes in the overall distribution of labels.

The product of these martingales must exceed $1/\delta$ to lead to the rejection of the EA at significance level δ. This approach is detailed in Algorithm 2.

4.5. Computational Efficiency

Here we examine the computational complexity of Algorithms 1 and 2. We focus on the preprocessing, pvalue calculation, and betting function calculation.

For algorithm 1 we have:

- Preprocessing steps: Removing Missing Values and Outliers: Each of these steps is implemented on k instances. Thus their combined complexity is $O(k)$
- Median filtering and standartization: for each point we examine w neighboring instances. Thus the overall complexity is $O(k \cdot w)$.
- pvalues calculation: The pvalue calculation for each new instance starts by comparing 1 instance gradually to k instances, making the overall complexity $O(k^2)$.
- Betting function calculation: The betting function is evaluated k times, based on L instances each time, thus the overall complexity is $O(L \cdot k)$

Algorithm 2: Detect CP using ICM in aCGH data

1. Require set $\{z_1, z_2, \ldots, z_k\}$, significance level δ
2. Initialize $S_1^c = 1$,$S_1^l = 1$
3. Apply preprocessing steps (see Section 4.1) to $\{z_1, z_2, \ldots, z_k\}$ to obtain $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$
4. $H = \{\}$
5. $\tilde{H} = \{\}$
6. For $i = 1$ to k
   (a) Append(α_i) to H
   (b) $\tilde{a_i} = median(a_i | a_i \in H \wedge y_i = y_j)$
   (c) Append($\tilde{a_i}$) to $\tilde{H}$
   (d) $p_i = \frac{|\{\alpha_j \in H | \alpha_i > \alpha_j \wedge y_i = y_j\}| + U_j \cdot |\{\alpha_j \in H | \alpha_i = \alpha_j \wedge y_i = y_j|}{|\{\alpha_i \in H | y_i = y_j\}|}$
   (e) $\tilde{p_i} = \frac{|\{\tilde{\alpha}_j \in \tilde{H} | \tilde{\alpha}_i > \tilde{\alpha}_j\}| + U_j \cdot |\{\tilde{\alpha}_j \in \tilde{H} | \tilde{\alpha}_i = \tilde{\alpha}_j\}|}{\tilde{a_i} \in \tilde{H}}$
   (f) Calculate betting function $h_i = h(p_{i-L}, \ldots, p_{i-1})$
   (g) Calculate betting function $\tilde{h_i} = h(\tilde{p_{i-L}}, \ldots, \tilde{p_{i-1}})$
   (h) $S_i^c = S_{i-1}^c \cdot h_i(p_i)$
   (i) $S_i^l = S_{i-1}^l \cdot h_i(\tilde{p})$
   (j) If $S_i^c \cdot S_i^l > \frac{1}{\delta}$
      i. Raise an Alarm
      ii. $H = \{\}$,$\tilde{H} = \{\}$
      iii. $S_1^c = 1$,$S_1^l = 1$

Algorithm 2 has many steps identical to Algorithm 1, including data preprocessing, calculation of p-values, and calculation of betting functions. However, the difference here is that specific calculations are performed twice, specifically, the calculation of two different sets of p-values and the calculation of two separate betting functions.

However, here, we use the median to find the label nonconformity measure where the calculation for each instance starts by comparing 1 instance gradually to k instances, making the overall complexity $O(k^2)$.

5. Experiments and Results

In this section, we conduct experiments on the proposed approaches, using aCGH data. Identifying CP in this context is challenging due to the absence of a definitive ground truth to confirm false alarms or missed CP. Nonetheless our approach is probabilistically valid, making no assumptions about the data distribution. We present our results within the context of a 1% significance level. We start by describing our dataset.

5.1. Dataset

The aCGH (Stransky et al., 2006) dataset used in this study includes data from 57 patients diagnosed with bladder tumours. It features the log2 ratio of DNA quantities between tumorous cells and a healthy reference. A negative log2 ratio indicates losses in the genomic sequence, while a positive log2 ratio implies gains. To be able to calculate conditional p-values (refer to (2)), label nonconformities (refer to (3)), and consequently label p-values (refer to (4)), a categorical variable was introduced, with values 'GAINS' and 'LOSSES', based on the sign of the log2 ratio. Figure 1 illustrates a heatmap representing the distribution of the remaining aCGH sequence elements for each patient by chromosome after removing all NaNs and outliers, we also omit chromosomes associated with gender from this analysis. As observed, there are cases, such as chromosome 22, where the number of remaining instances is very low, ranging from 11 to 27. For CP detection, we handle each chromosome itself; however, detecting CPs under such conditions poses a significant challenge.

5.2. Experimental Setting

This section details the configurations used in our experiments, where we employed the ICM to test the exchangeability assumption and thus detect CP. As previously discussed, our method does not presuppose any specific model. For instance, as a model, we consider the molecular method that quantifies the log2 ratio of gains or losses in the DNA sequence. After preprocessing, as discussed in Section 4.1, the nonconformities a_i are prepared for calculating label-conditional p-values. During the preprocessing stage, while applying the median filter and standardization, the frame window size is set to 15. A separate sequence of p-values is computed for each label. The nonconformity measure used is $\tilde{a}_i = \text{median}\{a_1, \ldots, a_i\}$. When the Martingale value exceeds $1/\delta$, where $\delta = 0.01$, a CP is detected, and the process is then restarted from the subsequent point.

In this study, the Cautious Betting function is employed with the parameter ϵ set to 10 and W set to ∞. These settings are chosen because each chromosome's aCGH time series is

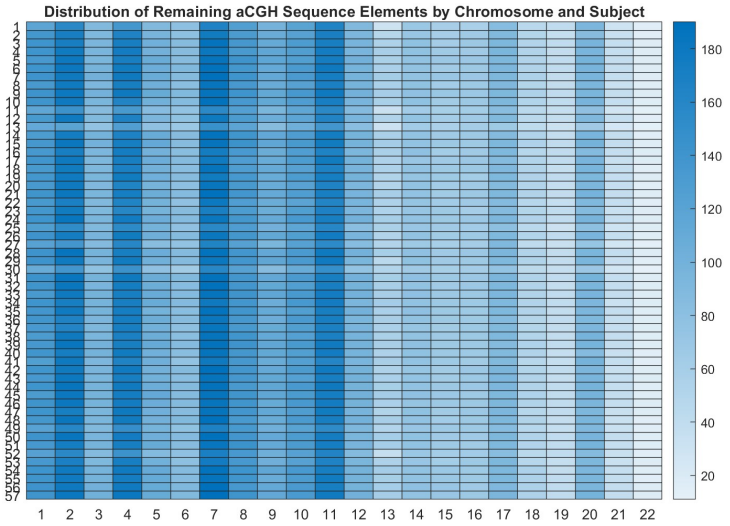


Figure 1: Distribution of Remaining aCGH Sequence Elements by Chromosome and Subject

relatively short, and a high value of ϵ would unnecessarily delay detection. Setting W to ∞ allows the utilization of all available data, aiding the decision-making process for betting. This function is integrated with both the Histogram estimator and the Kernel estimator, yielding the following betting functions:

a) **Cautious-Hist**: This betting function combines the Cautious Betting approach with a Histogram density estimator featuring 15 bins. The parameter L is set to ∞, and the number of bins κ to 15.

b) **Cautious-Kernel**: This betting function integrates the Cautious Betting function with the Kernel betting function, where $L = 100$ and the number of bins κ is set to 15.

The parameter ϵ is set to 100 in both betting functions.

In the forthcoming subsection, we will conduct simulations on the aCGH dataset described above. Given that there is no definitive ground truth to determine if a CP detection is a false alarm or delayed, we will present several figures to illustrate the average number of CPs per subject and per chromosome.

5.3. Results

In this subsection, we evaluate the performance of Algorithms 1 and 2, comparing them with the CBS and ED algorithms as they are applied to detect CPs in DNA sequences. For Algorithms 1 and 2, we assess their efficacy using both the Cautious-Kernel and Cautious-Histogram betting functions. Descriptive statistics for all algorithms, including the mean, median, standard deviation, and range of change points per person, are presented in Table 1.

Table 1: Comparative Descriptive Statistics for CP Detection Across Algorithms 1, 2, CBS, and ED

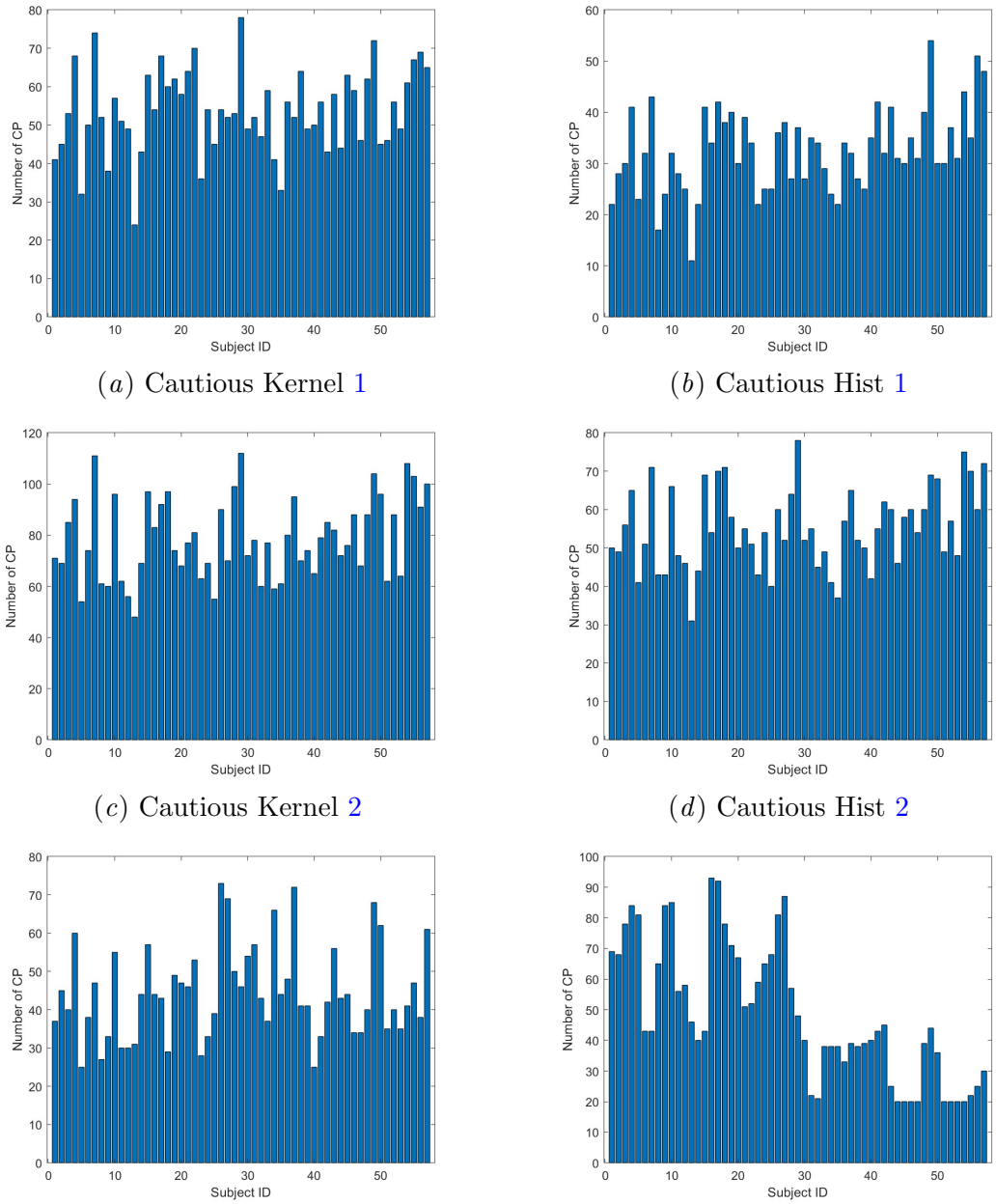
	Algorithm 1		Algorithm 2		CBS	ED
Betting function	Caut Ker	Caut Hist	Caut Ker	Caut Hist	-	-
Mean	53.70	32.49	78.63	55.11	44.37	46.43
Standard Deviation	11.02	8.21	15.90	10.51	12.04	20.91
Median	53	32	77	54	43	43
Range	24-78	11-54	48-112	31-78	25-73	21-91

While observing Table 1, we note that both the mean and median values of Algorithm 2 exceed those obtained with Algorithm 1 for both betting functions. Furthermore, the maximum values for Algorithm 2 consistently surpass those observed with Algorithm 1. In terms of minimum values, for the Cautious-Kernel betting function, Algorithm 2 exhibits slightly lower values compared to Algorithm 1, yet, when employing the Cautious-Histogram betting function, the minimum values are higher for Algorithm 2. Additionally, Algorithm 2 exhibits a consistently larger standard deviation, indicating greater variability around its mean. This suggests that while Algorithm 2 tends to detect more CPs on average, its performance across different genomic sequences or conditions is more variable.

When comparing the performance metrics of the CBS algorithm with those of Algorithms 1 and 2, we notice that Algorithm 2, particularly when utilizing the Kernel betting function, outperforms the CBS algorithm. Specifically, Algorithm 2 with the Kernel betting function has a higher mean and median CP detection rate than the CBS algorithm. The range of CP detections for Algorithm 2 is broader, with a higher minimum and maximum values. Despite the larger standard deviation, which suggests a greater variability in CP detection outcomes, this variability highlights the algorithm's adaptability to a wide range of genomic sequences. The same observations hold true when comparing Algorithm 2 to the ED algorithm.

Figure 2 comprises six subfigures illustrating the distribution of detected CPs using the Cautious-Kernel and Cautious-Histogram betting functions with Algorithms 1 and 2, as well as with the CBS and ED algorithms.

Figure 3 shows the differences in detected CPs across subjects using various betting functions and algorithms. Subfigure 3(a) compares CP detection between Algorithms 1 and 2 using the Cautious Histogram betting function. Subfigure 3(b) presents differences in CP detection between the same algorithms using the Cautious Kernel betting function. Additionally, Subfigure 3(c) displays the CP detection differences between Algorithm 2 with the CBS algorithm, using the Cautious Kernel betting function. Subfigure 3(d) displays the CP detection differences between Algorithm 2 with the ED algorithm, using the Cautious Kernel betting function. Consistent with the data in Table 1, Algorithm 2 typically detects more CPs than Algorithm 1 with both betting functions and more CPs than the CBS algorithm while using the Cautious Kernel. The better performance of Algorithm 2 in detecting more CP is because it utilizes numerical information from both the label distribution and the label conditional distribution of the data, enhancing its ability to detect more CPs by effectively rejecting the Exchangeability Assumption.



(a) Cautious Kernel 1 (b) Cautious Hist 1

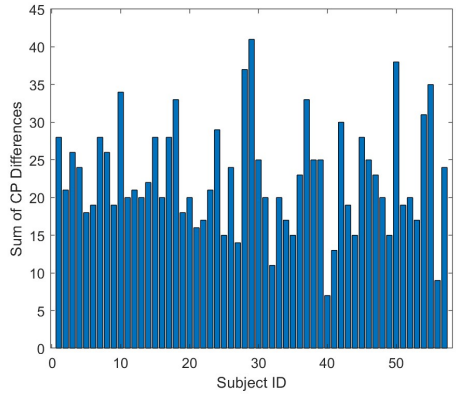
(c) Cautious Kernel 2 (d) Cautious Hist 2

(e) CBS (f) ED

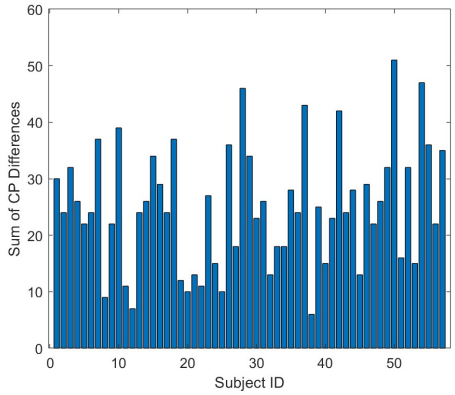
Figure 2: Distribution of Detected CP Across Subjects

6. Conclusion

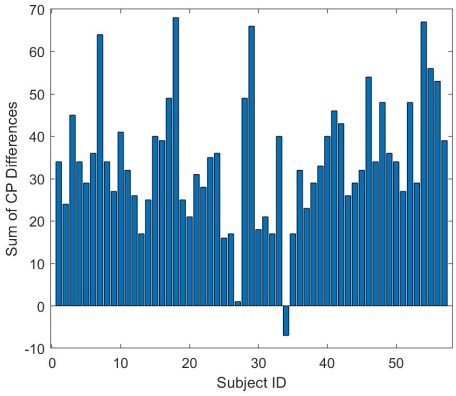
In this study, we implemented two ICM-based algorithms to address the challenge of change point (CP) detection in aCGH data. These algorithms require only data preprocessing and do not assume any specific model or distribution when testing the exchangeability assump-



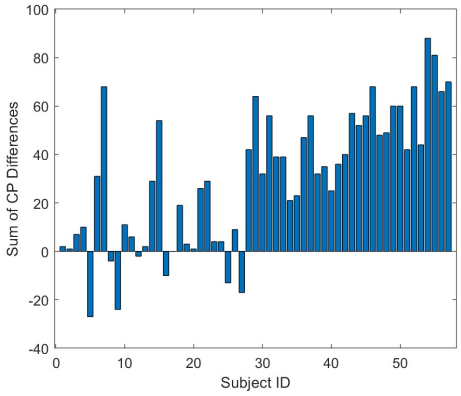
(*a*) Cautious Hist 2-Cautious Hist 1



(*b*) Cautious Kernel 2-Cautious Kernel 1



(*c*) Cautious Kernel 2-CBS



(*d*) Cautious Kernel 2-ED

Figure 3: Distribution of the difference of Detected CP Across Subjects

tion to detect CPs. Furthermore, our results significantly improve when we incorporate label distribution and label conditional data distribution into our calculations. Using the Cautious Kernel betting function, we surpass the performance of the CBS algorithm in terms of the number of changes detected. Our future plans include utilizing more datasets and combining multiple tests to enhance the robustness and accuracy of our detection methods.

References

J. Chen and Y.P. Wang. A statistical change point model approach for the detection of dna copy number variations in array cgh data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):529–541, Oct-Dec 2009. doi: 10.1109/TCBB.2008.129.

Nicole A. Doudican, Ansu Kumar, Neeraj Kumar Singh, Prashant R. Nair, Deepak A. Lala, Kabya Basu, Anay A. Talawdekar, Zeba Sultana, Krishna Kumar Tiwari, Anuj Tyagi, Taher Abbasi, Shireen Vali, Ravi Vij, Mark Fiala, Justin King, MaryAnn Perle, and Amitabha Mazumder. Personalization of cancer treatment using predictive sim-

ulation. *Journal of Translational Medicine*, 13(1):43, 2015. ISSN 1479-5876. doi: 10.1186/s12967-015-0399-y. URL https://doi.org/10.1186/s12967-015-0399-y.

Charalambos Eliades and Harris Papadopoulos. Using inductive conformal martingales for addressing concept drift in data stream classification. 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 171–190, 08–10 Sep 2021. URL https://proceedings.mlr.press/v152/eliades21a.html.

Charalambos Eliades and Harris Papadopoulos. A betting function for addressing concept drift with conformal martingales. In Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson, editors, *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pages 219–238. PMLR, 24–26 Aug 2022. URL https://proceedings.mlr.press/v179/eliades22a.html.

Charalambos Eliades and Harris Papadopoulos. Icm ensemble with novel betting functions for concept drift, 2024. URL https://arxiv.org/abs/2406.15760.

Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 923–930, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Shen-Shyang 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, page 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.

Shen-Shyang Ho and Harry Wechsler. On the detection of concept changes in time-varying data stream by testing exchangeability. *CoRR*, abs/1207.1379, 2012. URL http://arxiv.org/abs/1207.1379.

Shen-Shyang Ho, Matthew Schofield, Bo Sun, Jason Snouffer, and Jean Kirschner. A martingale-based approach for flight behavior anomaly detection. In *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pages 43–52, 2019. doi: 10.1109/MDM.2019.00-75.

Sangwon Hyun, Kevin Z. Lin, Max G'Sell, and Ryan J. Tibshirani. Post-Selection Inference for Changepoint Detection Algorithms with Application to Copy Number Variation Data. *Biometrics*, 77(3):1037–1049, 01 2021. ISSN 0006-341X. doi: 10.1111/biom.13422. URL https://doi.org/10.1111/biom.13422.

Sudar D Rutovitz D Gray JW Waldman F Pinkel D. Kallioniemi A, Kallioniemi OP. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821, October 1992. doi: 10.1126/science.1359641.

David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109 (505):334–345, 2014.

Adam B. Olshen, Ennapadam S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5 4:557–72, 2004. URL https://api.semanticscholar.org/CorpusID:5871867.

Daniel Pinkel, Richard Segraves, Damir Sudar, Steven Clark, Ian Poole, David Kowbel, Colin Collins, Wen-Lin Kuo, Chira Chen, Ye Zhai, Shanaz H. Dairkee, Britt-marie Ljung, Joe W. Gray, and Donna G. Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 20(2): 207–211, 1998. doi: 10.1038/2524.

Mark J. Schervish. *Theory of Statistics*. Springer, New York, 1995.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986. URL https://cds.cern.ch/record/1070306.

Nicolas Stransky, Céline Vallot, Fabien Reyal, Isabelle Bernard-Pierrot, Stéphane Grégoire de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, A. Graham, J. Southgate, B. Asselain, Y. Allory, C. C. Abbou, D. G. Albertson, J. P. Thiery, D. K. Chopin, D. Pinkel, and F. Radvanyi. Regional copy number-independent deregulation of transcription in cancer. *Nat Genet*, 38(12):1386–1396, Dec 2006. doi: 10.1038/ng1923. Epub 2006 Nov 12. Erratum in: Nat Genet. 2008 Mar;40(3):373.

J. Ville. Étude critique de la notion de collectif. by j. ville. pp. 144. 75 francs. 1939. monographies des probabilités, calcul des probabilités et ses applications, publiées sous la direction de m. Émile borel, fascicule iii. (gauthier-villars, paris). *The Mathematical Gazette*, 23(257):490–491, 1939. doi: 10.2307/3607027.

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, Stockholm, Sweden, 13–16 Jun 2017. PMLR. URL http://proceedings.mlr.press/v60/volkhonskiy17a.html.

Vladimir Vovk. Testing for concept shift online. *ArXiv*, abs/2012.14246, 2020. URL https://api.semanticscholar.org/CorpusID:229678222.

Vladimir Vovk, Ilia Nouretdinov, and Alexander Gammerman. Testing exchangeability on-line. pages 768–775, 01 2003.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. 01 2005. doi: 10.1007/b106715.