Proceedings of Machine Learning Research 128:1–14, 2020 Conformal and Probabilistic Prediction and Applications

A Histogram based Betting Function for Conformal Martingales

Charalambos Eliades PAMBOSELIADES@HOTMAIL.COM
Computational Intelligence (COIN) Research Lab
Frederick University, Cyprus

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

Editor: Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov and Giovanni Cherubin

Abstract

This paper investigates the use of Conformal Martingales (CM) for providing a numerical indication of how likely it is that the exchangeability assumption holds on a set of data. Reliable and fast testing of exchangeability is an important challenge because many machine learning algorithms rely on this assumption. Therefore a technique with only a few parameters to tune, that is able to reject the exchangeability assumption with respect to a significance level should be very beneficial for enhancing the reliability of such machine learning models. Our approach consists of a CM whose betting function is estimated on the previously seen p-values, we compare its computational efficiency and its performance with a kernel betting function and the Kolmogorov-Smirnoff test. We test our approach on two benchmark data-sets, USPS and Statlog Satellite data.
Keywords: Conformal, Martingales, Exchangeability

1. Introduction

Many machine learning algorithms that deal with real life problems rely on the exchangeability assumption. If this assumption is violated it could lead to misleading results, hence a valid tool for testing exchangeability would be beneficial for verifying the suitability of the data used for some task.

By the term exchangeability of the data we mean that the examples are invariant to any permutation. This means that we can feed any subset of data as training set, validation set and test set to a machine learning algorithm that is based on the exchangeability assumption. Note that testing that the data is exchangeable is equivalent to testing the data for being independent and identically distributed (i.i.d) (Fedorova et al., 2012).

Vovk et al. (2003) proposed a way of testing exchangeability in an online manner based on Conformal Prediction (CP) and exchangeability martingales. This technique does not suffer from the difficulties that standard statistical approaches face, which are inappropriate for handling high dimensional data (Vapnik, 1998). The technique proposed by (Vovk et al., 2003) consists of calculating a sequence of p-values by applying conformal prediction. The p-values are calculated in an online manner where the p-value of each new example is calculated from the new example and the previously seen examples. After the p-values are calculated a betting function is applied on each p-value and the product of the betting function outputs is the value of the Martingale. When the value M of the Martingale, becomes large enough we can reject the exchangeability assumption with significance level 1/M. Note that a value of M within 20 and 100 is sufficient enough.

In this paper we are concerned with the computational inefficiency of existing betting functions for calculating the martingale. We address this by proposing a much more computationally efficient betting function based on approximating the distribution of the p-values using histograms. We test the proposed betting function on two benchmark data sets and compare it with a kernel betting function and the Kolmogorov–Smirnov test (KST). Our experimental results show that the time needed to compute the betting functions is much lower without sacrificing the rejection ability of the Martingale.

The rest of the paper starts with an overview of related work on CM in Section 2. Section 3 gives a brief overview of the ideas behind conformal martingales. Section 4 describes the proposed approach and defines the betting function used for our CM. Section 5 presents the experimental setting and performance measures used in our evaluation and reports the obtained experimental results. Finally, Section 6 gives our conclusions and plans for future work.

2. Related Work

Conformal Martingales introduced by Vovk et al. (2003) can be used as a tool for testing if a set of data satisfies the exchangeability assumption and for change point detection in time series.

Ho (2005) implemented a CM based on a simple betting mixture function extending the idea of detecting exchangeability online to detect concept changes in time-varying data streams. Two martingale tests were implemented based on: (i) martingale values and (ii) the martingale difference.

Fedorova et al. (2012) tested the exchangeability of data on two data-sets, USPS and Statlog Satellite data. The data is tested in an online manner i.e. the examples arrive one by one and then the value of the CM is calculated which is a valid measure indicating if the exchangeability assumption should be rejected. In this article the authors used a density estimator of the observed p-values as a betting function. The kernel density estimation has been employed and it has been shown that it outperforms the simple mixture betting function.

Volkhonskiy et al. (2017) implemented an Inductive version of Conformal Martingales to detect when a change occurs in a time series. In this study the underlying model is trained on the first observations of the time sequence. All the nonconformity scores are calculated via the underlying model. The authors tested their methods on synthetic data-sets and showed that their results are comparable with those of many other methods.

3. Conformal Martingales

For the needs of this study a CM is employed for testing exchangeability as described in (Vovk et al., 2003). The reason for using CM is their ability to handle multidimensional data, the fact that testing can be performed online and that the only assumption made is that the p-value sequence is stable, for more details refer to Fedorova et al. (2012). In this section we give a brief description of the exchangeability assumption and the main principles of Conformal Martingales. For more details see Vovk et al. (2005).

3.1 Exchangeability

Let $(Z_1, Z_2, ...)$ be an infinite sequence of random variables. Then the joint probability distribution $\mathbb{P}(Z_1, Z_2, ..., Z_N)$ (where N is natural number) is exchangeable if it is invariant under any permutation of those random variables. The joint distribution of the infinite sequence $(Z_1, Z_2, ...)$ is exchangeable if the marginal distribution of $(Z_1, Z_2, ..., Z_N)$ is exchangeable for every N. As mentioned in Section 1, testing if the data is exchangeable is equivalent to testing the data for being 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 $(S_1, S_2, S_3, ...)$ being equal to or greater than zero that keep the conditional expectation $\mathbb{E}(S_{n+1}|S_1, ..., S_n) = S_n$.

To illustrate how a martingale works consider a fair game were a gambler with infinite wealth follows a strategy that is 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 to have high profit(C) would be small, $\mathbb{P}\{\exists n : S_n \geq C\} \leq 1/C$.

According to Ville's inequality (Ville, 1939) for the case of the exchangeability assumption a large final value of the Martingale suggests rejection of the assumption with a significance level equal to the inverse of the Martingale value. Specifically a large Martingale value such as 20 or 100 is the rule of thumb for rejecting the hypothesis of exchangeability, respectively at 5% or 1% significance level.

3.3 Calculating p-values and Non-conformity scores

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 $y_i \in R$ the label of the corresponding input vector. The examples will be fed one by one so we want to examine how strange or unusual a new example z_j is compared to the previously seen examples $\{z_1, z_2, \dots, z_{j-1}\}$. To make this possible a numerical value will be assigned to each example called nonconformity score (NCS) denoted by $a_{i,j}$ and equal to $A(z_i, \{z_1, \dots, z_j\})$ with $i \in \{1 \dots j\}$. The calculation of the NCS is based on the underlying algorithm. A big value of the nonconformity measure indicates a strange example and a small value indicates a less strange example. Specifically when a new example z_j arrives a new NCS should be assigned to the previously seen examples, this is done by repeatedly training the model on the sequence $\{z_1, z_2, \dots, z_j\}$ and recalculating the NCS's.

For every new example z_j we calculate the sequence $H_j = \{a_{1,j}, a_{2,j}, \dots, a_{j-1,j}, a_{j,j}\}$ to find the p-value. Note that the NCS in H_j are calculated when the underlying algorithm is trained on $\{z_1, z_2, \dots, z_j\}$. Given the sequence H_j we can calculate the corresponding p-value (p_j) of the new example z_j with the function:

$$p_j = \frac{|\{\alpha_{i,j} \in H_j | \alpha_{i,j} < \alpha_{j,j}\}| + U_j.|\{\alpha_{i,j} \in H_j | \alpha_{i,j} = \alpha_{j,j}\}|}{j}, \tag{1}$$

where $\alpha_{j,j}$ is the NCS of the new example and $\alpha_{i,j}$ is the NCS of the i^{th} element in the example series set and U_j is a random number from the uniform distribution (0,1).

Given a pair (X_k, y_k) with a p-value less than or equal to δ, this means that this example will be generated with at most δ frequency, under the assumption that the examples are exchangeable, proven by Vovk et al. (2005).

For the two benchmark data-sets tested on this study we have used the 1 Nearest neighbor (1-NN) algorithm as in (Fedorova et al., 2012) and (Ho, 2005). Assuming that we have observed j examples then the nonconformity score of an example z_i is the difference of its smallest distance to the example with the same label and its smallest distance to the example with a different label:

$$a_{i,j} = min_{i \neq k: y_i = y_k} d(x_i, x_k) - min_{i \neq k: y_i \neq y_k} d(x_i, x_k) \tag{2}$$

where $d(x_i, x_k)$ is the euclidean distance of the examples z_i and z_k with $i, k \in \{1 \dots j\}$.

Note than when the number of examples observed is small it is possible for the quantity $min_{i \neq k: y_i = y_k} d(x_i, x_k)$ or $min_{i \neq k: y_i \neq y_k} d(x_i, x_k)$ to be undefined because there might not exist a k with $y_i = y_k$ or $y_i \neq y_k$ in that case we set $min_{i \neq k: y_i = y_k} d(x_i, x_k)$ and or $min_{i \neq k: y_i \neq y_k} d(x_i, x_k)$ say a big number C. Obviously when only one example has been observed then the p_1 would be a random number within $(0, 1)$.

3.4 Conformal Martingales

A Conformal Martingale is an exchangeability test Martingale (see subsection 3.2) which is calculated as a function of p-values as described in subsection 3.3.

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

$$S_n = \prod_{i=1}^{n} f_i(p_i) \tag{3}$$

where $f_i(p_i) = f_i(p_i | p_1, p_2, \dots, p_{i-1})$ is the betting function.

Equation 3 should satisfy the constraint: $\int_0^1 f_i(p)dp = 1$ and also it follows that $\mathbb{E}(S_{n+1} | S_0, S_1, \dots, S_n) = S_n$.

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

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

Using equation (3) it is easy to show that $S_n = S_{n-1}.f_n(p_n)$ which allows us to update the martingale online. Let's say that the value of S_n is equal to M then Ville's inequality (Ville, 1939) suggests that we can reject the exchangeability assumption with a significance level equal to $1/M$.

Algorithm 1: Calculating Martingale

Data: $z_1, z_2, \ldots, z_n$
Result: Martingale Value S_n
for *i=1,...,n* **do**
  input z_i;
  for *j=1,...,i* **do**
    $a_j = A(z_j, \{z1, \ldots, z_i\})$
  end
  $p_i = \frac{\#\{j:a_j>a_i\}+U_j\#\{j:a_j=a_i\}}{i}$
  Calculate density estimator $f_i = f(p_1, \ldots, p_{i-1})$
  $S_i = S_{i-1}.f_i(p_i)$
end

4. Proposed Approach

In this section we describe the proposed approach i.e. the betting functions used. In this study we examine if a sequence of betting functions equal to the probability density estimation of the p-values seen so far is sufficient enough for the rejection of the exchangeability assumption. The density function estimator $\hat{f}_n$ we used is the simple histogram.

4.1 Histogram Estimator

Our goal here is to calculate a density estimator $\hat{f}_n$ of the density distribution of the p-value series: $p_1, p_2, \ldots, p_{n-1}$. It should be beneficial that the calculation of this estimator is fast and it is desirable to have a small number of parameters 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 k this will partition $[0, 1]$ into $B_1 = [0, 1/k)$, $B_2[1/k, 2/k)$,..., $B_{k-1} = [(k-2)/k, (k-1)/k)$ and $B_k = [(k-1)/k, 1]$. When a p-value $p_n \in B_j$ then the density estimator will be equal to

$$\hat{f}_n(p_n) = \frac{n_j.k}{n-1}, \tag{4}$$

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 that $\exists x : \hat{f}_n(x) = 0$, in that case until a sufficient number of observations arrives we set $\hat{f}_n = 1$ so that the martingale value doesn't become zero.

4.2 Calculating the Martingales

To reject the exchangeability assumption for a pre-specified significance level δ the value of the Martingale must exceed $1/\delta$. In Algortihm 1 we summarize the process for calculating Martingales.

Algorithm 2: Kolmogorov Smirnoff

Data: $z_1, z_2, \ldots, z_n$
Result: p-value $\tilde{p_n}$
for i=1,...,n **do**
  input z_i;
  for j=1,...,i **do**
    $a_j = A(z_j, \{z1, \ldots, z_i\})$
  end
  $p_i = \frac{\#\{j:a_j>a_i\}+U_j\#\{j:a_j=a_i\}}{i}$
  Apply Kolmogorov Smirnoff test on $(p_1, \ldots, p_i)$ and output $\tilde{p_i}$
end

Now if the final value of the Martingale S_n exceeds 20 or 100 then we can reject the exchangeability assumption at a significance level equal to 5% and 1% respectively.

5. Experiments and Results

This section describes the performance of our proposed approach on two benchmark data-sets and on synthetic data. We compare the proposed approach with a similar method used by Fedorova et al. (2012) who used a kernel as a p-value density estimator. The kernel density estimator is calculated using the extended sample $\bigcup_{i=1}^{\infty}\{-p_i, p_i, 2-p_i\}$ where p_i is the p-value of the i^{th} example, but for simplicity reasons unlike Fedorova et al. (2012) we kept the default Matlab parameters. We also compare our approach with a Kolmogorov-Smirnov test described in section 5.1.

5.1 Kolmogorov-Smirnov Test

Kolmogorov Smirnov Test (KST) is a statistical test that can be used to compare a sample with a given distribution. This test is based on finding the supremum of the distance of the empirical cumulative distribution with the candidate cumulative distribution, for more details refer to (Smirnov, 1948). This test is asymptotically valid for large samples of data, which makes it suitable for our case since our data-sets have a big number of examples.

Rejecting the hypothesis that the p-values are uniformly distributed is equivalent re-jecting exchangeability assumption. Here we will compare the p-values $p_1, .., p_n$ with the uniform distribution. By applying KST on a sequence of p-values this will result in a p-value $\tilde{p_n}$ which is equivalent to saying that the exchangeability assumption is rejected with a significance level $\tilde{p_n}$. Small values of $\tilde{p_n}$ with $n \to \infty$ indicate strong evidence for rejecting exchangeability, in fact a $\tilde{p_n}$ less than 5% is considered sufficient enough.

In Algorithm 2 we summarize the process for calculating the $\tilde{p_n}$.

5.2 Datasets

The benchmark data-sets we used are the USPS and the Statlog Satellite data-sets as in (Fedorova et al., 2012) by merging the training and test examples.

The USPS data-set is one of the standard data-sets widely used in computer vision. It consists of 7291 training examples and 2007 test examples where each example is described by 256 attributes and labeled with the corresponding digit belonging to (0-9). Its worth mentioning that for both data-sets the exchangeability assumption does not hold in the original order (Vovk et al., 2003),(Fedorova et al., 2012).

The Statlog Satellite data-set (Frank and Asuncion, 2010) consists of 4435 training examples and 2000 test examples. Each example is described by 36 attributes and the labels are numbers belonging from 1 to 7, excluding 6.

Here while assessing results the process described in Algorithms 1 and 2 has been repeated ten times with the data-sets in the original order and ten times with data-sets shuffled. The reason we repeated our testing process ten times when using data-sets in the original order is to examine if the hypothesis tests obtained by each series of p-values are consistent and how the choice of a betting function affects this. Recall from equation 1 that while calculating p-values we add a random number drawn from the uniform distribution that takes place when two non-conformity scores are equal.

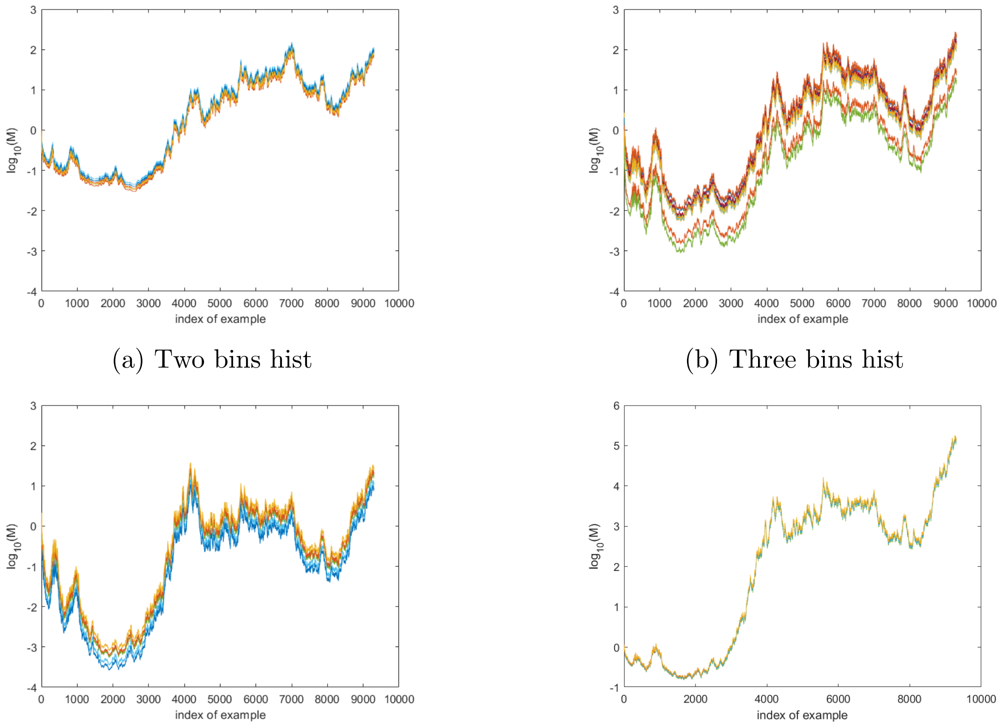
5.3 Empirical Results

For each data-set we compare the results obtained while implementing the CM when using the histogram and kernel betting function (Algorithm 1) as well as with the KST (Algorithm 2).

5.3.1 USPS Data-set

In this subsection we present the Martingale growth with (Algorithm 1) and the p-values obtained by KST on the USPS data-set when each example arrives in the original order and when we randomly shuffle their order.

Figure 1, presents the Martingale performance while using Algorithm 1 with the histogram betting function with a number of bins set to 2,3,4 and the kernel betting function. Since the final value of all Martingale sequences is greater than 10 the exchangeability assumption is rejected with a significance level of 10% in all cases. Particularly when using the histogram betting function with a number of bins equal to 2 the assumption is rejected at 1% significant level. This is due to the fact that when using a big number of bins the histogram betting function is more sensitive to statistical fluctuations. Also in Figure 1 we can observe that while performing our simulations using the histogram betting function for each example the generated Martingale values are closest when setting the number of bins to two or four. Overall, when comparing the Martingale values of the three different histogram functions the more robust results are obtained when setting the number of bins to two. While observing the Martingale growth when using Algorithm 1 and the kernel betting function the martingale value is bigger than when using the histogram betting function thus the exchangeability assumption can be rejected for smaller significance levels, here in each experiment for each index example the values of the Martingales are within a tiny region, but this benefit comes at the expense of the computational time needed to calculate the kernel which will be discussed in another section. Although the final Martingale value when using the histogram betting function is several hundred times smaller that the one obtained with the kernel betting function the usability of the histogram betting function is



(a) Two bins hist

(b) Three bins hist

(c) Four bins histogram

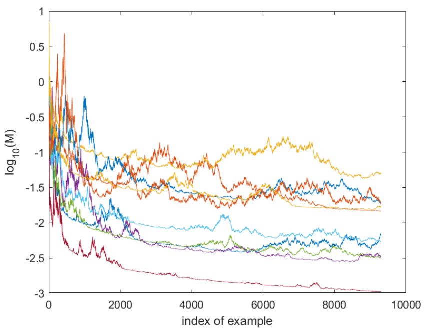
(d) Kernel Betting Function

Figure 1: Martingale growth for the USPS dataset when using Algorithm 1 with histogram and kernel betting function

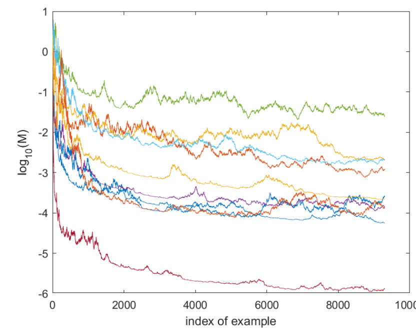
still practical due to the fact that the exchangeability assumption is rejected. Note than in this figure the scale of the graph for the kernel betting functions is different due to the fact the max Martingale value is much bigger than the ones obtained by the histogram betting function.

Figure 2 shows the Martingale behavior when using the histogram and the kernel betting function together with Algorithm 1 on randomly shuffled data. As we can see for all betting functions the exchangeability assumption practically cannot be rejected for any significance level because the value of the Martingales is less than 1.

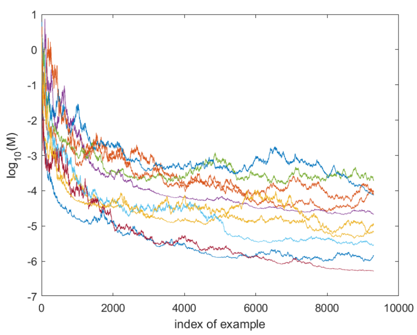
Figure 3 shows the p-values calculated by KST on the original dataset order and on the randomly shuffled one. In Figure 3a we can see that exchangeability assumption can be rejected at significance levels bigger than 0.1%. While for the case of the randomly shuffled data figure 3b suggests that the exchangeability assumption cannot be rejected for significance levels smaller than 20%. The results obtained here are consistent with the Martingale method.



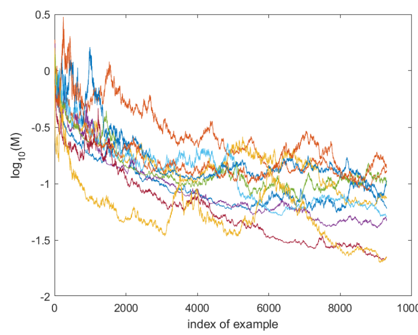
(a) Two bins hist-alg1



(b) Three bins hist-alg1

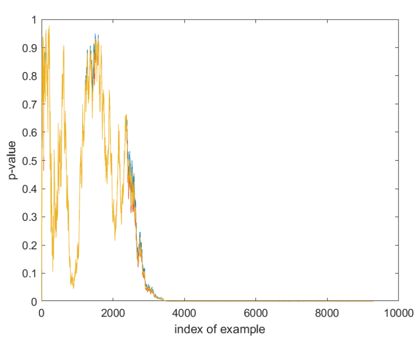


(c) Four bins histogram-alg1

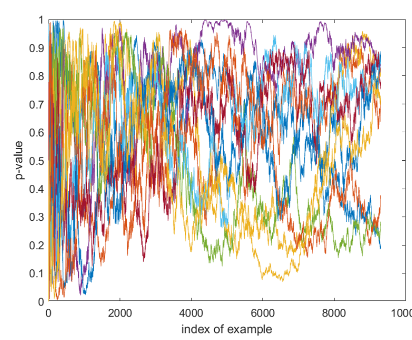


(d) Kernel Betting Function

Figure 2: Martingale growth for the shuffled USPS dataset when using Algorithm 1 with histogram and kernel betting function



(a) KST original order

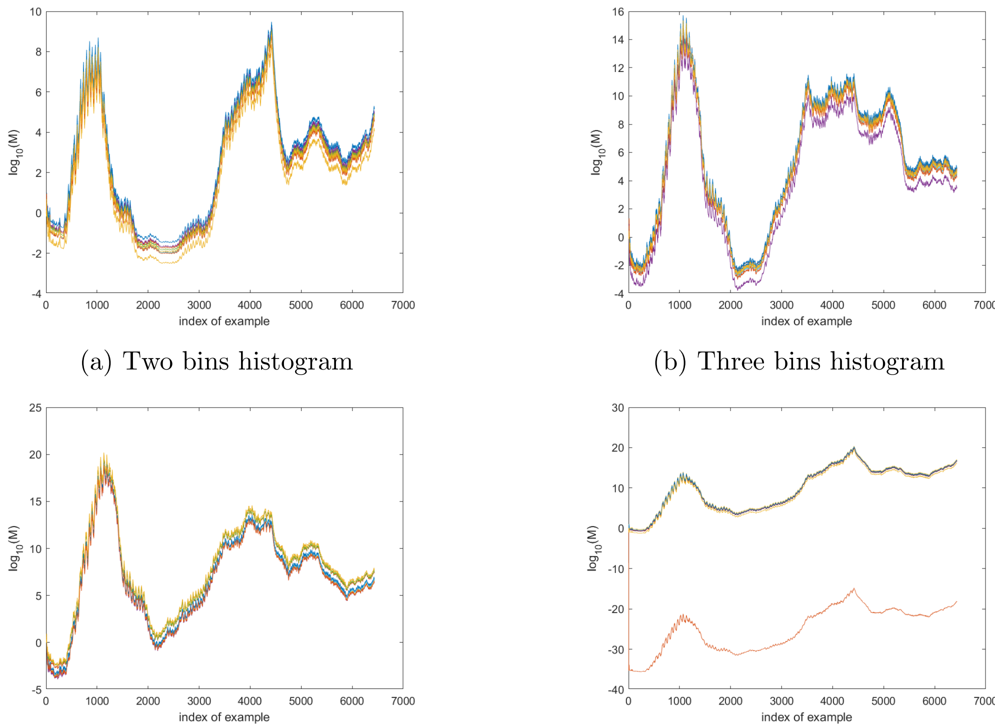


(b) KST shuffled order

Figure 3: p-values computed from KST on the USPS dataset when using Algorithm 2

5.3.2 Statlog Satellite Data-set

In this subsection we present the Martingale growth and the p-values obtained by KST on the Statlog Satellite data-set when each example is observed in the original order and when we randomly shuffle them.



(a) Two bins histogram

(b) Three bins histogram

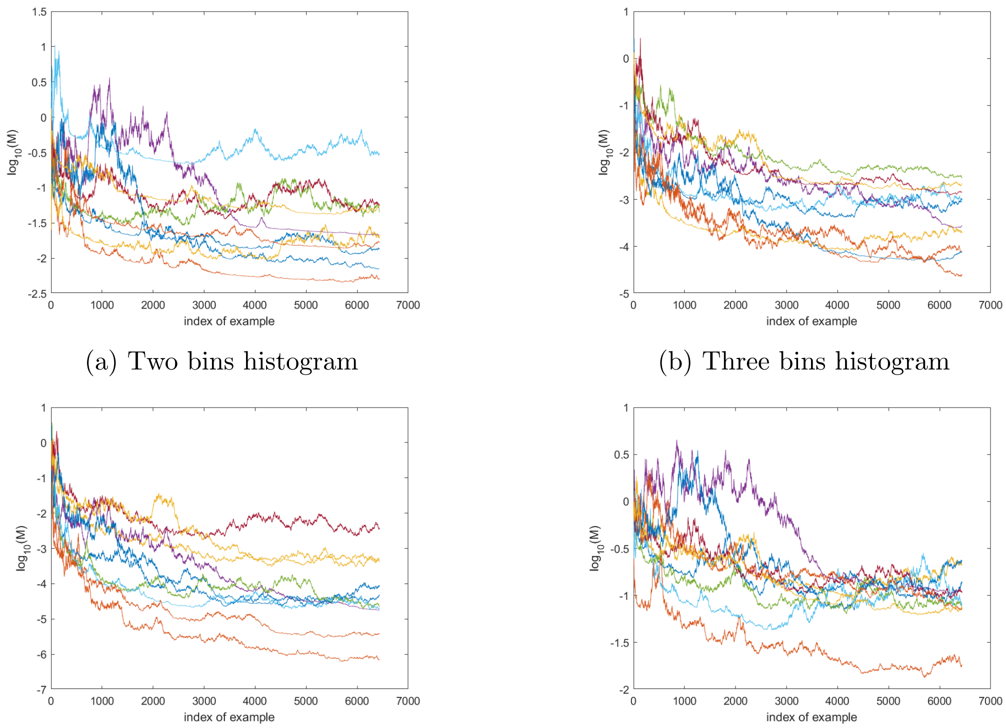
(c) Four bins histogram

(d) Kernel Betting Function

Figure 4: Martingale growth for the Stalog Satellite when using Algorithm 1 with histogram and kernel betting function

Figure 4 presents the Martingale performance while using Algorithm 1 and the histogram betting function with a number of bins equal to 2,3,4. In all cases the exchangeability assumption is rejected with a significance level less than 0.1%. Also in Figure 4 we can observe that the Martingale can reject the exchangeability assumption easier than in the case of the USPS data-set, this is due to the fact that the distribution of the p-values is not that 'complicated'. The Martingale growth when using Algorithm 1 with the kernel betting function suggests that the exchangeability assumption in most cases is rejected at very small significance levels of less than 0.1%, but there is one case where we cannot reject the exchangeability assumption. This is due to the fact that the martingale starts from a low value and fails to recover to a sufficiently big value that allows us to reject exchangeability. Perhaps a better tuning of the kernel density estimation function could overcome this difficulty. On the other hand, the histogram betting function for every bin selection does not suffer from a similar problem and successfully rejects exchangeability assumption in all ten simulations. It should be noted that again in this figure the scale of the four graphs is different due to the fact the max Martingale values are very different.

Figure 5 shows the Martingale behavior when using Algorithm 1 on the randomly shuffled data with the histogram betting function and the kernel betting function. As we can see when using Algorithm 1 for both betting functions the exchangeability assumption cannot



(a) Two bins histogram

(b) Three bins histogram

(c) Four bins histogram

(d) Kernel Betting Function

Figure 5: Martingale growth for the randomly shuffled Stalog Satellite dataset when using Algorithm 1 with histogram and kernel betting function

be rejected practically for any significance level because in each case the final value of the Martingale is less that 1.

Figure 6 shows the p-values calculated by the KST test on the original dataset order and on the randomly shuffled one. In Figure 6a we can see that exchangeability assumption can be rejected at significance levels less than 0.1%. While for the case of the randomly shuffled data Figure 6b suggests that the exchangeability assumption cannot be rejected for significance levels smaller than 10%.

5.4 Computational Efficiency

In this subsection we present the time (in seconds) needed to calculate the Martingale value for each data-set while using Algorithm 1 with the two different betting functions used in this study. We also present the time needed for running KST. Note that when the time is calculated we omit the time needed to compute the Nonconformities and p-values, thus the resulting time is the time needed for the betting function and the martingale sequence to be calculated. As for the case of the KST the resulting time is the time needed for the p-values sequence generated by KST. The times reported in this section were obtained on a machine with an i7 4790k cpu and 32gb ram.

Table 1: Computational time for the USPS data-set in seconds

Algorithm	Hist 2 bins	Hist 3 bins	Hist 4 bins	Kernel	KST
	27	29	29	420	800

Table 2: Computational time for the Statlog satellite data-set in seconds

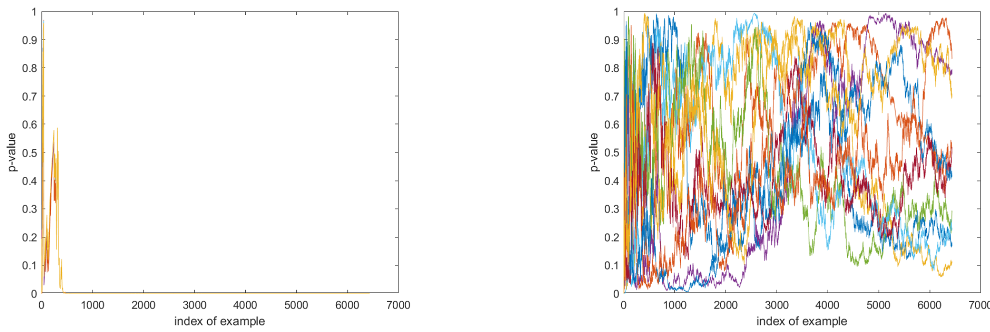
Algorithm	Hist 2 bins	Hist 3 bins	Hist 4 bins	Kernel	KST
	21	22	23	280	350

Table 1 shows the time needed for applying each algorithm on USPS data-set. The findings here suggest that the histogram betting function combined with Algorithm 1 outperforms in term of computational efficiency the kernel betting function and the KST.

Table 2 shows the time needed for applying each algorithm on the Statlog satellite data-set. As in the case of USPS data-set the findings here suggest that the Histogram betting function combined with Algorithm 1 outperforms in term of computational efficiency the kernel betting function and the KST. For the case of the USPS data-set using the histogram betting function is 14 to 15 times faster than the kernel betting function and 27 to 29 times faster than the KST, while for the case of the Statlog Satellite data-set the histogram betting function is 12 to 13 times faster than the kernel betting function and 15 to 16 times faster than the KST this due to the high computational cost of the kernel betting function and the KST.

6. Conclusions

In this study we have examined the use of the histogram density estimator as a betting function for testing exchangeability. We have tested this approach on two real life datasets. Our results are in line with the ones obtained when using the kernel density estimation as a betting function and they are consistent with the widely used statistical KST. Although in most cases the final value of the martingales is much lower than the one obtained with the kernel betting function the practical usability of the proposed betting function relies on the fact that the histogram betting function was able to reject the exchangeability assumption in every simulation with a significance level of 10% at most. Also while using this betting function for rejecting the exchangeability assumption the required amount of time is significantly reduced compared to using a kernel estimator or the KST test. This computational efficiency advantage is especially significant when dealing with change point detection or concept drift problems. Another advantage of the histogram function is that the only parameter required to tune is the number of bins, in which the tuning process can be completed fast. As a future work we intend to investigate the utilization of conformal martingale based approaches on challenges related to change point detection and concept drift.



(a) KST original order

(b) KST shuffled order

Figure 6: p-values computed from KST on the Statlog Satellite dataset when using Algorithm 2

References

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.

A. Frank and A. Asuncion. Uci machine learning repository. *http://archive.ics.uci.edu/ml*, 2010. URL `'https://ci.nii.ac.jp/naid/10031099995/en/'`.

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`.

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

N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19(2):279–281, 06 1948. doi: 10.1214/aoms/1177730256. URL `https://doi.org/10.1214/aoms/1177730256`.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

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, 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.