Supplementary materials for: FDR control for Online Anomaly Detection

Etienne Krönert[1], **Alain Célisse**[2] **and Dalila Hattab**[1]

[1]*FS Lab, Worldline, France, e-mail:* etienne.kronert@worldline.com; dalila.hattab@worldline.com

[2]*SAMM, Paris 1 Panthéon-Sorbonne University, France, e-mail:* alain.celisse@univ-paris1.fr

1. Supplementary theoretical results

1.1. Proof of Theorem 4

Theorem 4 (FDR control with BH [1])**.** *Let m be a positive integer and $(X_i)_1^m$ be* independent *random variables such that $X_i \sim \mathcal{P}_0$, $1 \leq i \leq m_0$, and $X_i \sim \mathcal{P}_1$, $m_0 + 1 \leq i \leq m$. Let us also define the set of true p-values, for all $1 \leq i \leq m$ by $p_i = \mathbb{P}_{X\sim P_0}(X \geq X_i) \in [0, 1]$, and assume that each $p_i \sim U([0, 1])$. Then for every $\alpha \in]0, 1]$, BH_α applied to $p = (p_i)_{1\leq i\leq m}$ yields the exact FDR control at the prescribed level α that is,*

$$FDR_1^m(\varepsilon_{BH_\alpha}, p) = \frac{m_0\alpha}{m}.$$

Proof of Theorem 4. Let R be a random variable describing the number of rejections made by BH_α that is, $R = \sum_{i=1}^m D_i$, where $D_i = 1$ if hypothesis $\mathcal{H}_{0,i}$ is rejected. Let also FP be the number of false positives made by BH_α. Then, $FP = \sum_{i=1}^m A_i D_i = \sum_{i=1}^{m_0} D_i$, where A_i is a random variable equal to 1 if hypothesis $\mathcal{H}_{0,i}$ is true and 0 otherwise. Furthermore

$$\begin{aligned} FDP = \frac{FP}{R} &= \sum_{i=1}^{m_0} \frac{\mathbb{1}[p_i \leq \frac{\alpha R}{m}]}{R} \qquad (\text{since } D_i = \mathbb{1}[p_i \leq \tfrac{\alpha R}{m}]) \\ &= \sum_{k=1}^{m} \sum_{i=1}^{m_0} \frac{\mathbb{1}[p_i \leq \frac{\alpha k}{m}]\mathbb{1}[R = k]}{k}. \end{aligned} \tag{1.1}$$

Let us now introduce the random variables $R(i)$ that are the number of rejections generated by BH when p_i is replaced by the value 0 that is, $R(i) = BH_\alpha(p_1, .., p_{i-1}, 0, p_{i+1}, \ldots, p_m)$. It results that

$$\mathbb{1}[p_i \leq \frac{\alpha k}{m}]\mathbb{1}[R = k] = \mathbb{1}[p_i \leq \frac{\alpha k}{m}]\mathbb{1}[R(i) = k],$$

since, on the event $\{p_i \leq \frac{\alpha k}{m}\}$, p_i is rejected and therefore $R = R(i)$. Let us also notice that the independence between the p-values is already used at this stage since modifying the value of p_i does not affect that of the others.

By combining the previous argument and the independence between $R(i)$ and the other p-values, the expectation on both sides yields

$$\begin{aligned} FDP &= \sum_{k=1}^{m} \sum_{i=1}^{m_0} \frac{\mathbb{1}[p_i \leq \frac{\alpha k}{m}]\mathbb{1}[R(i)=k]}{k} \\ \Rightarrow \quad FDR = \mathbb{E}[FDP] &= \sum_{k=1}^{m} \sum_{i=1}^{m_0} \frac{\mathbb{P}[p_i \leq \frac{\alpha R}{m}]\mathbb{P}[R(i)=k]}{k} \\ &= \sum_{k=1}^{m} \sum_{i=1}^{m_0} \frac{\frac{\alpha k}{m}\mathbb{P}[R(i)=k]}{k} \\ &= \frac{m_0 \alpha}{m}, \end{aligned}$$

where the last equality results from the fact that the true p-values follow a uniform distribution on $[0,1]$. The result finally follows from noticing that for each $1 \leq i \leq m_0$, $\sum_{k=1}^{m} \mathbb{P}[R(i)=k]$, since $R(i) \geq 1$ by definition. □

1.2. FDR control with disjoint subseries

1.2.1. Preliminary discussion: Disjoint and Overlapping subseries

In the context of online anomaly detection, the main focus in what follows is put on two situations where the data-driven thresholds $(\hat{\varepsilon}_t)_{t\geq 0}$ can be defined from a set of m empirical p-values: (i) the *disjoint* case where disjoint subseries of length m are successively considered, and (ii) the *overlapping* case where the subseries (of length m) successively considered share $m-1$ common observations at each step.

Let us start with a subseries of length m where each observation is summarized by its corresponding empirical p-value, and let us assume that there exists a function $f_m : [0,1]^m \to [0,1]$ that is mapping a set of m empirical p-values onto a real-valued random variable. This random variable corresponds to the data-driven threshold that is applied to the subseries of length m to detect potential anomalies. This function f_m is called the *local threshold* function since it outputs a threshold which applies to a subseries of length m.

Given the above notations, the threshold sequences $\hat{\varepsilon}_d = (\hat{\varepsilon}_{d,t})_t$ and $\hat{\varepsilon}_o = (\hat{\varepsilon}_{o,t})_t$ can be defined as follows.

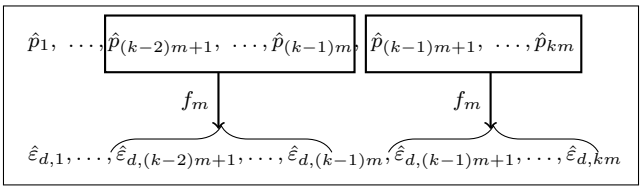
- **Disjoint subseries:** $\hat{\varepsilon}_d : t \mapsto \hat{\varepsilon}_{d,t}$ is given by

$$\forall k \geq 0, \quad \forall t \in [\![km+1, (k+1)m]\!], \quad \hat{\varepsilon}_{d,t} = f_m(\hat{p}_{km+1}, \ldots, \hat{p}_{(k+1)m}) \tag{1.2}$$

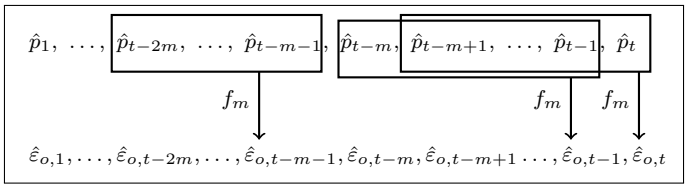
- **Overlapping subseries:** $\hat{\varepsilon}_o : t \mapsto \hat{\varepsilon}_{o,t}$ is given by

$$\forall t \geq m, \qquad \hat{\varepsilon}_{o,t} = f_m(\hat{p}_{t-m+1}, \ldots, \hat{p}_t). \tag{1.3}$$

Figure 1 illustrates these two situations. In Figure 1a, the full time series is split into small disjoint subseries of length m. f_m is applied to each such subseries and the threshold is the same for all observations within a given subseries. Figure 1b displays the situation where overlapping subseries are successively considered. Because two successive subseries differ from each other by two observations, the thresholds are different at each time step unlike the disjoint case. Furthermore



(a) Illustration of disjoint windows.



(b) Illustration of Overlapping sliding windows.

Figure 1: Comparison of disjoint window and overlapping window for the threshold function.

the sequences $\hat{\varepsilon}_d$ and $\hat{\varepsilon}_o$ do not enjoy the same dependence properties. Figure 1a illustrates that all thresholds $\hat{\varepsilon}_{d,(k-1)m+1}, \ldots, \hat{\varepsilon}_{d,km}$ are computed by applying f_m to the same subseries $\hat{p}_{(k-1)m+1}, \ldots, \hat{p}_{km}$. Therefore only thresholds computed from different subseries are independent, while all thresholds from the same subseries are equal. In other words, $\hat{\varepsilon}_{d,t_1}$ and $\hat{\varepsilon}_{d,t_2}$ are independent if and only if t_1 and t_2 belong to different subseries that is, $\lfloor t_1/m \rfloor \neq \lfloor t_2/m \rfloor$, where $\lfloor \cdot \rfloor$ denotes the integer part. By contrast Figure 1b shows that the variables $\hat{\varepsilon}_{o,t}$ and $\hat{\varepsilon}_{o,t-1}$ are dependent because they share $m-1$ common observations. But all of them are still different and, for each t, $\hat{\varepsilon}_{o,t}$ is independent from $\hat{\varepsilon}_{o,t-m-1}$. This can be reformulated as $\hat{\varepsilon}_{o,t_1}$, $\hat{\varepsilon}_{o,t_2}$ are independent if and only if $|t_1 - t_2| > m$.

In the present online anomaly detection context, considering the overlapping case sounds more convenient since the detection threshold can be updated at each time step (as soon as a new observation has been given), which makes the anomaly detector more versatile. However for technical reasons, next Section 1.2.2 still focuses disjoint subseries as a means to introduce important

notions without introducing too many technicalities, while the main paper in Section 4.1 gives a more general and realistic case of overlapping subseries.

1.2.2. FDR control with disjoint subseries

As illustrated in Section 4 in the main paper, controlling FDR on each subseries of length m (locally) is not equivalent to controlling FDR (globally) on the full time series. However in online anomaly detection, a decision has to be made at each time step regarding the potential anomalous status of each new observation. (This is a typical instance of a local decision since at step t, the decision making process ignores what will be observed at the next step.) This requires a criterion to be optimized locally (on subseries) in such a way that the resulting global FDR value (the one of the full time series) can be proved to be controlled at the desired level α.

This requirement for a local criterion justifies the introduction of the modified FDR criterion, denoted by mFDR [10, 5], which is defined as follows.

Definition 1 (mFDR)**.** *With the previous notations, the mFDR expression of the subseries from $t-m+1$ to t is given by*

$$mFDR_{t-m+1}^{t}(\hat{\varepsilon},\hat{p}) = \frac{\mathbb{E}\left[\sum_{u\in\mathcal{H}_0,t-m+1\leq u\leq t}\mathbb{1}[\hat{p}_u\leq\hat{\varepsilon}_u]\right]}{\mathbb{E}\left[\sum_{u=t-m+1}^{t}\mathbb{1}[\hat{p}_u\leq\hat{\varepsilon}_u]\right]},$$

where $\hat{\varepsilon}=(\hat{\varepsilon}_u)_{t-m+1\leq u\leq t}$ denotes a sequence of thresholds, $\hat{p}$ is a sequence of empirical p-values evaluated at each observation of the subseries from $t-m+1$ to t.

Mathematically the difference between the mFDR and the FDR is that the expectation is no longer on the ratio but independently on the numerator and the denominator. The main interest for mFDR is clarified by Theorem 5, which establishes its connection to FDR. To be more specific, the control of the latter at the α level provides a global control of the FDR at the same level under simple conditions.

Theorem 5 (Global FDR control with disjoint subseries)**.** *Assume that $\hat{\varepsilon}_d : t\mapsto\hat{\varepsilon}_{d,t}$ is given by $\hat{\varepsilon}_{d,t}=f_m(\hat{p}_{km+1},\ldots,\hat{p}_{(k+1)m})$, for any $t\in[\![km+1,(k+1)m]\!]$ $(k\geq 0)$ and any integer $m\geq 1$ (see Eq. (1.2)). Let us also assume that the p-value random process $\hat{p}=(\hat{p}_t)_{t\geq 1}$ follows the scheme detailed in Definition 3 of the main paper. Then, the global FDR value of the full (infinite) time series is equal to the local mFDR value of the any subseries of length m from $t=km+1, k\in\mathbb{N}^*$. More precisely,*

$$FDR_1^{\infty}(\hat{\varepsilon}_d,\hat{p}) = mFDR_{km+1}^{(k+1)m}(\hat{\varepsilon}_d,\hat{p}) = mFDR_{km+1}^{(k+1)m}(\hat{\varepsilon}_d,\hat{p}).$$

Since the full time series is assumed to be infinite, Theorem 5 is an asymptotic result. It gives rise to a strategy for controlling the (asymptotic) FDR criterion

at level α by means of successive local controls of mFDR on small subseries of length m. According to the asymptotic nature of Theorem 5, there is no particular constraint on the integer m. However when dealing within time series of a finite length T, the Theorem 5 proof suggests that choosing an m "not too large" would be better since then, $k = T/m$ would take large values making the LLN applicable (see for instance Eq. (1.4)). Actually in the online anomaly detection context, practitioners only have a limited freedom regarding the choice of m. Therefore, for a given fixed m, the control of the FDR value of the full time series given by Theorem 5 will be all the more accurate as T will be large. Fortunately this is not a limitation in the online anomaly detection context. The main limitation of Theorem 5 lies in the use of disjoint subseries, which sounds somewhat restrictive (at least from a practical perspective). This limitation is be overcome by Theorem 2 in the main paper.

Proof of Theorem 5. Let $k \geq 1$ denote an integer and $T = mk$. Then, the FDP definition and the A_t variables introduced in Definition 3, in the main paper, justify that

$$FDP_1^T(\hat{\varepsilon}_d, \hat{p}) = \frac{FP_1^T(\hat{\varepsilon}_d, \hat{p})}{R_1^T(\hat{\varepsilon}_d, \hat{p})} = \frac{\sum_{t=1}^T \mathbb{1}[\hat{p}_t < \hat{\varepsilon}_{d,t}](1 - A_t)}{\sum_{t=1}^T \mathbb{1}[\hat{p}_t < \hat{\varepsilon}_{d,t}]},$$

where $R_1^T(\hat{\varepsilon}_d, \hat{p})$ and $FP_1^T(\hat{\varepsilon}_d, \hat{p})$ respectively denote the number of rejections (resp. false positives) at the threshold $\hat{\varepsilon}_d$ for the subseries $\hat{p}$.

Using the partitioning into k subseries of length m, its first comes that $FP_1^T(\hat{\varepsilon}_d, \hat{p}) = \sum_{i=1}^k FP_{im-1}^{(i+1)m}(\hat{\varepsilon}_d, \hat{p})$. It is also noticeable that the k random variables $\{FP_{im-1}^{(i+1)m}(\hat{\varepsilon}_d, \hat{p})\}_{1 \leq i \leq k}$ are independent and identically distributed since the thresholds $\hat{\varepsilon}_{d,i}$ remain unchanged within each subseries, they are identically distributed from one block to another, and the empirical p-values from different blocks are independent and identically distributed as well. Therefore the random variables $(FP_{im-1}^{(i+1)m}(\hat{\varepsilon}_d, \hat{p}))_{1 \leq i \leq k}$ are independent and identically distributed, which implies (Law of Large Numbers theorem) that, almost surely,

$$\lim_k \frac{1}{k} \sum_{i=1}^k FP_{im-1}^{(i+1)m}(\hat{\varepsilon}_d, \hat{p}) = \mathbb{E}[FP_1^m(\hat{\varepsilon}_d, \hat{p})], \tag{1.4}$$

where the expectation is taken over all sources of randomness. (Here it is implicitly assumed that T can go to $+\infty$.) Repeating the argument for $R_1^T(\hat{\varepsilon}_d, \hat{p})$, it also comes that

$$\mathbb{E}[R_1^m(\hat{\varepsilon}_d, \hat{p})] = \lim_k \frac{1}{k} \sum_{i=1}^k R_{im-1}^{(i+1)m}(\hat{\varepsilon}_d, \hat{p}), \qquad a.s..$$

The conclusion then results from noticing that

$$mFDR_1^m(\hat{\varepsilon}_d, \hat{p}) = \frac{E[FP_1^m(\hat{\varepsilon}_d, \hat{p})]}{E[R_1^m(\hat{\varepsilon}_d, \hat{p})]} = \lim_k \frac{\sum_{i=1}^k FP_{im-1}^{(i+1)m}(\hat{\varepsilon}_d, \hat{p})}{\sum_{i=1}^k R_{im-1}^{(i+1)m}(\hat{\varepsilon}_d, \hat{p})} = \lim_T FDP_1^T(\hat{\varepsilon}_d, \hat{p}).$$

□

2. Supplementary experiments

2.1. Comparison of p-values estimators

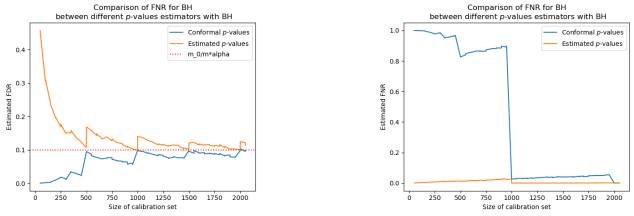
The control of the FDR is not achievable using classical multiple testing [2, 9] since the empirical p-value estimator, shown in Section 2.2 of the main paper, is not super-uniform. Conformal p-value estimator $\tilde{p}$, shown in Equation 2.1, verifies the super-uniform property.

$$\hat{p}_e(s_t, \mathcal{S}_t^{cal}) = \frac{1}{|\mathcal{S}_t^{cal}| + 1} \left(1 + \sum_{s \in \mathcal{S}_t^{cal}} \mathbb{1}[s_t > s] \right) \tag{2.1}$$

However, this estimator $\tilde{p} \geq \frac{1}{m+1}$ has lower power because zero anomalies are detected with thresholds below $\frac{1}{m+1}$.

Figure 2 displays the comparison between estimated p-values and conformal p-values using the BH-procedure. As shown in Figure 2a, the conformal p-values ensure an upper bound on the FDR at level $\frac{m_0}{m}\alpha$, while the estimated p-values ensure only a lower bound at the same level. Moreover, perfect control are reached for $n = 1000$ and $n = 2000$ with conformal p-values while the control is reached for $n = 999$ and $n = 1999$ with estimated p-values. As shown in Figure 2b, the FNR for conformal p-values estimator is always larger than the one for estimated p-values. However for the n points that control the FDR, the FNR values are close.

To conclude, the choice between conformal p-values and estimated p-values depends on the calibration set size. Indeed, for calibration set $n = 1000$ the performances are similar. But for other calibration set sizes as $n = 1500$ the FDR control are similar but the FNR is better for estimated p-values.



(a) FDR (b) FNR

Figure 2: Comparison between p-value estimators using Benjamini-Hochberg

2.2. Disjoint subseries vs overlapping subseries

Experiment Description Theorems 5 and 2 theoretically prove the control of the FDR over the full time series throw control of the mFDR over disjoint subseries or overlapping subseries. According Corollary 5 in the main paper, the procedure mBH_α allows the control of the mFDR over subseries under assumption **Heuristic** and **Power** that are hard to verify. Empirical results from Section 5.2 of the main paper show that control of mFDR for the disjoint subseries can be obtained for scenarios where the level of atypicity δ is high enough. It still unknown whether these results hold true in cases where the subseries overlap In this section FDR control throw disjoint and overlapping subseries are compared.

For each scenario, the quantities $mFDR_1^m$ and FNR_1^m are estimated two times, using disjoint subseries and using overlapping subseries. All subseries are extracted from the same time series of size $T = 10^4$. The distribution of these estimations is obtained by repeating the experiment across $B = 100$ time series. Thus, the two estimations of $mFDR_1^m$ and FNR_1^m quantities can be compared. The experimental design is described as follows:

1. With b in $[\![1, B]\!]$ and t in $[\![1, T]\!]$, the time series is generated from a mixture model:
   - $A_{b,t} \sim Ber(\pi)$
   - If $A_{b,t} = 0$, $p_{b,t} \sim U([0, 1])$
   - Otherwise: $p_{b,t} \sim U([0, 1/\delta])$
2. The thresholds of mBH are estimated on each subseries $p_{b,k,t+1}, \ldots, p_{b,k,t+m_0+m_1}$:
   - $\hat{\varepsilon}_{b,t} = mBH_\alpha(p_{b,k,t+1}, \ldots, p_{b,k,t+m_0+m_1})$.
3. The numbers of rejections, false positives and false negatives are calculated, according the different cases.
   (a) In the disjoint subseries case, the quantities are computed using only thresholds on the form $\hat{\varepsilon}_{b,km}$ over disjoint subseries
   For $1 \leq b \leq L$ and $1 \leq k \leq K = T/m$:
   - $R_{b,k,d} = \sum_{t=km+1}^{(k+1)m} \mathbb{1}[p_{b,k,t} \leq \hat{\varepsilon}_{b,km}]$,
   - $FP_{b,k,d} = \sum_{t=km+1}^{(k+1)m} \mathbb{1}[p_{b,k,t} \leq \hat{\varepsilon}_{b,km}](1 - A_t)$,
   - $FN_{b,k,d} = \sum_{t=km+1}^{(k+1)m} \mathbb{1}[p_{b,k,t} > \hat{\varepsilon}_{b,km}]A_t$.

   The mFDR and FNR are estimated:
   - $mFDR_{b,d} = \frac{1}{K} \sum_{k=1}^{K} FP_{b,k,m,d} \frac{1}{K} \sum_{k=1}^{K} R_{b,k,m,d}$,
   - $FNR_{b,d} = \frac{1}{K} \sum_{k=1}^{K} \frac{FN_{b,k,m,d}}{m_1}$.

(b) In the overlapping subseries case, the quantities are computed using the thresholds from all overlapping subseries $\hat{\varepsilon}_{b,t}$:
For $1 \leq b \leq L$ and $1 \leq k \leq K = T/m$:

- $R_{b,k,o} = \sum_{t=km+1}^{(k+1)m} \mathbb{1}[p_{b,t-m+1,t,o} \leq \hat{\varepsilon}_{b,t}]$,
- $FP_{b,k,o} = \sum_{t=km+1}^{(k+1)m} \mathbb{1}[p_{b,t-m+1,t} \leq \hat{\varepsilon}_{b,t}](1 - A_t)$,
- $FN_{b,k,o} = \sum_{t=km+1}^{(k+1)m} \mathbb{1}[p_{b,t-m+1,t} > \hat{\varepsilon}_{b,t}]A_t$.

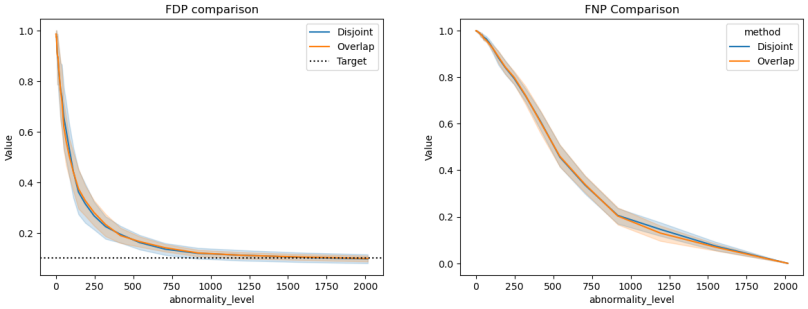
Notice the difference with disjoint windows case, all p-values of a subseries are compared to different thresholds and not to the same $\hat{\varepsilon}_{b,km}$.

The mFDR and FNR are estimated:

- $mFDR_{b,o} = \frac{1}{K}\sum_{k=1}^{K} FP_{b,k,m,o} \frac{1}{K}\sum_{k=1}^{K} R_{b,k,m,o}$,
- $FNR_{b,o} = \frac{1}{K}\sum_{k=1}^{K} \frac{FN_{b,k,m,o}}{m_1}$.

Different scenarios are generated by varying the proportion of anomalies π and the atypicity level δ.

Results and analysis As shown in Figure 3, disjoint and overlapping subseries control give similar results in mFDR and FNR for considered cases. Indeed, the curves are indistinguishable and decrease at the same rate.



(a) mFDR, $\alpha = 0.1$, $\pi = 0.01$ (b) FNR, $\alpha = 0.1$, $\pi = 0.01$

Figure 3: Comparison of mFDR and FNR control with disjoint and overlapping windows method.

Conclusion The FDR control quality are similar for both strategies, overlapping windows and disjoint windows. This imply that performances of the anomaly detector to not decrease by using overlapping windows instead of disjoint windows. This is a practical result that allows to do real time detection without having to wait to complete disjoint windows.

2.3. Example on real data

In this section, the anomaly detector is applied to one real data example: the musk dataset [3]. The Musk Anomaly Detection dataset describes a set of molecules, and the objective is to detect musks (normal) from non-musks (anomalies). This dataset contains 3000 instances of 166 features with 3% anomalies. This is tabular data, but it is assumed that the data arrives sequentially. The goal is to detect anomalies online and check that the FDR remains less than $\alpha = 0.1$.

Our data set is divided into three sets.

- The training set is used to learn the atypicity score. In this section, the atypicity score is estimated using Isolation Forest [7].
- The calibration set, which is used to calculate the p values. It contains normal data only. In this section, the same calibration set is used for all data.
- The test set, which contains data arriving sequentially and for which anomalies are to be detected.

Anomalies are detected using the following steps:

1. First, the scoring function is trained on the training set. Here, the "fit" method from the "IForest" class in the "PyOD" [11] Python library is used. The atypicity score resulting from this training is noted by $\hat{a}$.
2. The atypicity scores of the calibration set data points are calculated. This is essential for estimating the p-values of the observations.
3. To apply Theorem 3, the parameter α' must be specified, so that $\frac{\mathbb{E}R^{*,m}_{1,\alpha'}}{\mathbb{E}R^{m}_{1,\alpha'}}\alpha' \leq \alpha$. In order to achieve this, the ratio $\frac{\mathbb{E}R^{*,m}_{1,\alpha'}}{\mathbb{E}R^{m}_{1,\alpha'}}$ can be estimated for different values of $\tilde{\alpha}$, as described in Appendix A.1.2. In this section the training set data is reused. The function a is applied to each data point in the training set, then these score values are compared with those in the calibration set to find out their empirical p-value. For each value of $\tilde{\alpha}$, to estimate $\mathbb{E}R^{m}_{1,\alpha'}$ the average number of rejections is counted when BH is applied to subseries of length $m = 100$. Similarly, to estimate $\mathbb{E}R^{*,m}_{1,\alpha'}$ one of the p-values is replaced by 0. The results are shown in Figure 4. The x-axis represents the $\tilde{\alpha}$ parameter used in BH, and the y-axis represents an upper bound on the mFDR. The largest $\tilde{\alpha}$ ensuring $\frac{\mathbb{E}R^{*,m}_{1,\tilde{\alpha}}}{\mathbb{E}R^{m}_{1,\tilde{\alpha}}}\tilde{\alpha} \leq \alpha$, is $\alpha' = 0.074$ according to Figure 4.

   If the proportion of π anomalies were known then α' could be estimated from the heuristic described in Appendix A.1.1, $\alpha' = \alpha\left(1 + \frac{1-\alpha}{m\pi}\right)^{-1} = 0.075$. The results are therefore very similar.

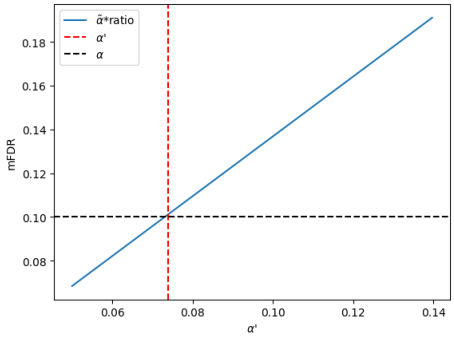


Figure 4: Illustration of α' selection method

Threshold estimation	FDR	FNR
our ($BH_{\alpha'}$)	0.08	0.1
Fixed Threshold	0.54	0.0

TABLE 1
Anomaly detection, according to the threshold selection procedure on musk dataset.

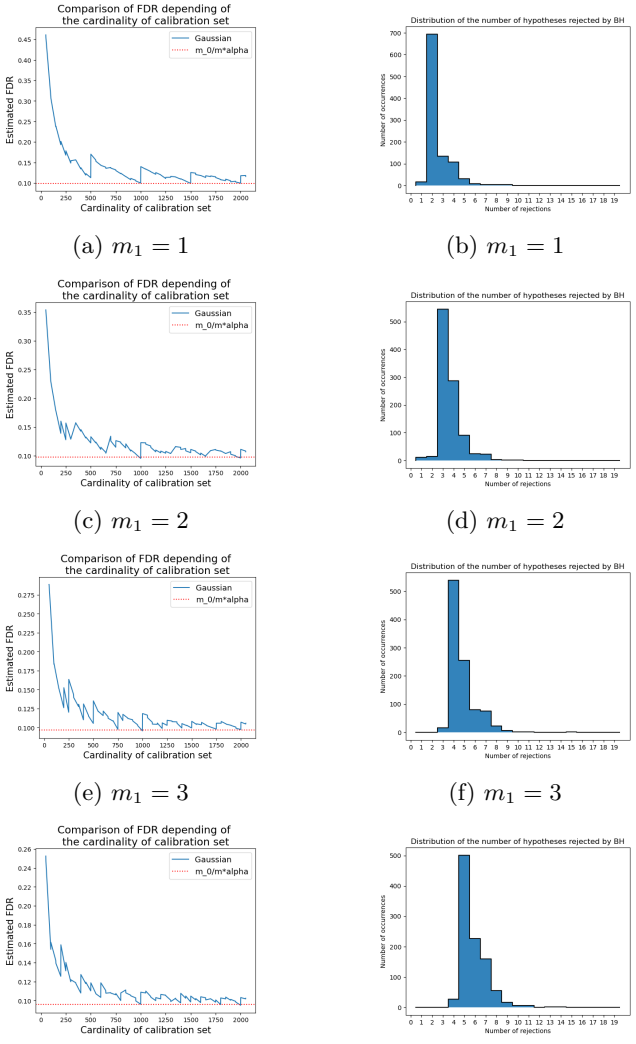
4. In the final step, the anomaly detector is applied to the test set data. For each data point X_t coming from the test set, the score function a is applied to obtain s_t. Then s_t is compared to the scores of the calibration set $\mathcal{S}^{cal}$ to obtain an estimate of the $p-value$ $\hat{p}_t$. In order to control the FDP, the detection threshold $\hat{\varepsilon}_t$ is estimated by applying the Benjamini-Hochberg procedure at the α' level on $\hat{p}_{t-m}, \ldots, \hat{p}_t$. The status of X_t is obtained by comparing $\hat{p}_t$ and $\hat{\varepsilon}_t$.

The results of the detection are presented in Table 1. The columns represent the FDR and FNR for the detection on the musk dataset. For comparison, the result using a fixed threshold ($\varepsilon_t = 0.05$) is used, as in Conformal Anomaly Detection [6].

Table 1 shows that when a fixed threshold is used, the proportion of false detections is not controlled. In this example, the FDR is 0.54. The FDR is controlled to a value of 0.08 using the multiple testing procedure developed in this paper.

3. Supplementary Figures

3.1. *Effect of the number detections by BH on the intermediate drops for the FDR control in Section 5.1*



(a) $m_1 = 1$ (b) $m_1 = 1$

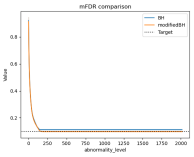
(c) $m_1 = 2$ (d) $m_1 = 2$

(e) $m_1 = 3$ (f) $m_1 = 3$

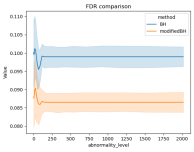
(g) $m_1 = 4$ (h) $m_1 = 4$

Figure 5: Effect of the number detections by BH on the intermediate drops for the FDR control

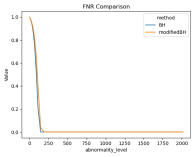
3.2. Figures related to experiment of Section 5.2



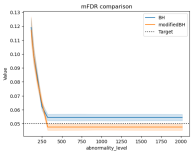
(a) mFDR, $\alpha = 0.1$, $\pi = 0.07$



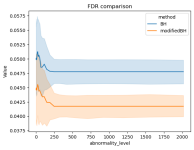
(b) FDR, $\alpha = 0.1$, $\pi = 0.07$



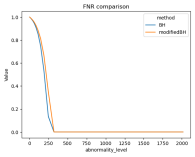
(c) FNR, $\alpha = 0.1$, $\pi = 0.07$



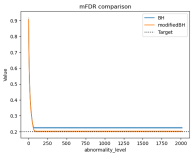
(d) mFDR, $\alpha = 0.05$, $\pi = 0.07$



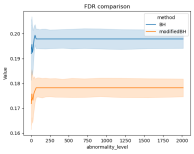
(e) FDR, $\alpha = 0.05$, $\pi = 0.07$



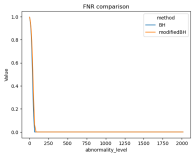
(f) FNR, $\alpha = 0.05$, $\pi = 0.07$



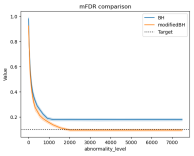
(g) mFDR, $\alpha = 0.2$, $\pi = 0.07$



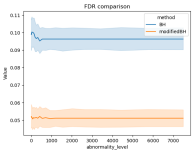
(h) FDR, $\alpha = 0.2$, $\pi = 0.07$



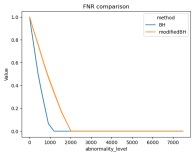
(i) FNR, $\alpha = 0.2$, $\pi = 0.07$



(j) mFDR, $\alpha = 0.1$, $\pi = 0.01$



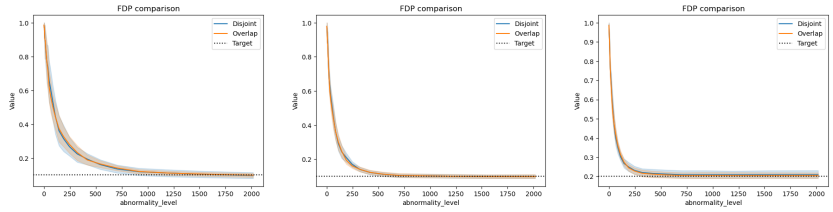
(k) FDR, $\alpha = 0.1$, $\pi = 0.01$



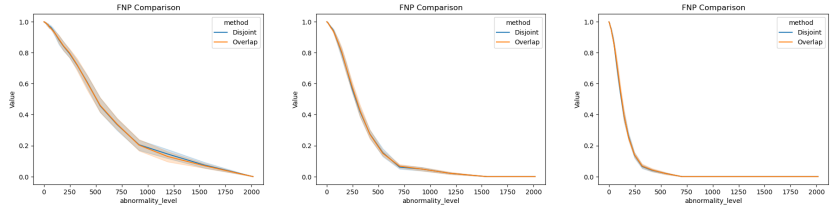
(l) FNR, $\alpha = 0.1$, $\pi = 0.01$

Figure 6: Effect of the atypicity level on the mFDR, FDR and FNR, according to different multiple testing procedures.

3.3. Figures related to experiment of Section 2.2



(a) mFDR, $\alpha = 0.1$, $\pi = 0.01$ (b) mFDR, $\alpha = 0.1$, $\pi = 0.02$ (c) mFDR, $\alpha = 0.2$, $\pi = 0.02$



(d) FNR, $\alpha = 0.1$, $\pi = 0.01$ (e) FNR, $\alpha = 0.1$, $\pi = 0.02$ (f) FNR, $\alpha = 0.2$, $\pi = 0.02$

Figure 7: Effect of atypicity level on mFDR and FNR, depending on whether detection is on disjoint or overlapping subseries

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57** 289–300.

[2] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 1165–1188.

[3] Chapman, D. and Jain, A. Musk (version 1), UCI Machine Learning Repository (1994). *DOI: https://doi. org/10.24432/C5ZK5B*.

[4] Fisher, R. (1951). The Design of Experiments, volume 6th Ed. *Hafner, New York, NY*.

[5] Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **70** 429–444.

[6] Laxhammar, R. (2014). Conformal anomaly detection: Detecting abnormal trajectories in surveillance applications, PhD thesis, University of Skövde.

[7] Liu, F. T., Ting, K. M. and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining* 413–422. IEEE.

[8] Phipson, B. and Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology* **9**.
[9] Ramdas, A., Zrnic, T., Wainwright, M. and Jordan, M. (2018). SAFFRON: an adaptive algorithm for online control of the false discovery rate. In *International conference on machine learning* 4286–4294. PMLR.
[10] Xu, Z. and Ramdas, A. (2022). Dynamic algorithms for online multiple testing. In *Mathematical and Scientific Machine Learning* 955–986. PMLR.
[11] Zhao, Y., Nasrullah, Z. and Li, Z. (2019). PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research* **20** 1-7.