FDR control for Online Anomaly Detection

Etienne Krönert, Alain Célisse, Dalila Hattab

FDR control for Online Anomaly Detection

Etienne Krönert[1,2], **Alain Célisse**[2,3] **and Dalila Hattab**[1]

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

[2]*Modal Team, Inria, France*

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

Abstract: The goal of anomaly detection is to identify observations generated by a process that is different from a reference one. An accurate anomaly detector must ensure low false positive and false negative rates. However in the online context such a constraint remains highly challenging due to the usual lack of control of the False Discovery Rate (FDR). In particular the online framework makes it impossible to use classical multiple testing approaches such as the Benjamini-Hochberg (BH) procedure. Our strategy overcomes this difficulty by exploiting a local control of the "modified FDR" (mFDR). An important ingredient in this control is the cardinality of the calibration set used for computing empirical p-values, which turns out to be an influential parameter. It results a new strategy for tuning this parameter, which yields the desired FDR control over the whole time series. The statistical performance of this strategy is analyzed by theoretical guarantees and its practical behavior is assessed by simulation experiments which support our conclusions.

MSC2020 subject classifications: Anomaly detection, Time series, Multiple testing.

1. Introduction

1.1. Context

By observing indicators along the time to check the system health, anomaly detection aims at raising an alarm if abnormal patterns are detected [1, 39]. A motivation for automatic anomaly detection is to reduce the workload of operations teams by allowing them to prioritize their efforts where necessary. This is usually made possible by using statistical and machine learning models [14, 10, 7]. However when badly calibrated an anomaly detector leads to alarm

fatigue. An overwhelming number of alarms desensitizes the people tasked responding to them, leading to missed or ignored alarms or delayed responses [15, 8]. One of the reasons for alarm fatigue is the high number of false positives which take time to be managed [54, 35]. The main goal of the present work is to design a new (theoretically grounded) strategy allowing to control the number of false positives when performing automatic anomaly detection in sequential context.

1.2. Related works

Anomaly Detection in time series According to [24] an anomaly is "An observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism." Actually the features of what is usually called an "anomaly" is more diverse depending on the context. For example an anomaly can refer to a value that appears larger than the mean of other values, or to a sequence of values having a smaller variance than other neighboring sequences of observations. Anomaly detection in time series is applied in various domains such as fault [45] or attack [46] detection in information systems, fault detection in industrial equipment [36, 43], fault detection in vehicles [29, 23], medical diagnostics [40, 37, 62] and astronomy [26]. The question of designing anomaly detectors has been considered in different contexts as explained by [14]: supervised, unsupervised and semi-supervised. The present work focuses on the unsupervised framework which is the most widespread one since labeling is often difficult and costly. Sometimes, specific pattern of anomalies are known, and specific detectors can be designed to detect them [9, 25, 31]. But most of the time the patterns have to be discovered as well. Review papers [14, 7, 51] reveal the diversity of existing approaches. Some of the most important categories of anomaly detectors are: distance-based anomaly detectors identifying anomalies far from other past observations [38], or probability-based anomaly detectors identifying anomalies within a low probability area of a statistical model fitted on past observations [2, 34]. Prediction-based anomaly detector aims at building a forecast model on training data. Anomalies are defined with high prediction error on streaming data [50], [48], [11], [44]. And reconstruction-based anomaly detection uses dimension reduction and anomalies correspond to high reconstruction errors on streaming data [32],[43], [59].

The present work describes a versatile anomaly detector that is sensitive to all categories of anomalies depending on the underlying *score function* (see Section 2.1).

A high diversity of abnormality score functions can be used to detect different abnormality patterns [17]. Abnormality scores are often not easily interpretable if the score distribution is unknown. Therefore, it is impossible to make a judicious choice of the detection threshold. The Conformal Anomaly Detection was introduced to alleviate this issue.

Conformal Anomaly Detection Conformal Anomaly Detection (CAD) introduced in [33] is a method derived from Conformal Prediction [3]. The goal of CAD is to give a probabilistic interpretation of the score using empirical p-values. Inductive Conformal Anomaly Detection (ICAD) introduced in [33] improves the CAD linear complexity in time and adapts it for Online Anomaly Detection by introducing the concept of *calibration set*. CAD can be used with a wide variety of anomaly score functions. For instance [52] presents an anomaly detector based kernels combined with CAD. [12] combined distance and density based scoring function with CAD. CAD gives the opportunity to control the expected number of false positives within a time period. But its main limitation is that it yields no control over the false alarm rate on the whole time series that is, the proportion of false positive among all detections. By contrast the present work aims at having a control over it (to avoid alarm fatigue [8, 15]), more precisely on the False Discovery Rate (FDR).

FDR Control Benjamini-Hochberg (BH) procedure [5, 6] is a multiple testing procedure that controls the proportions of false positives among rejections that is the False Discovery Rate. The BH procedure can be improved by estimating the proportion of anomalies in the dataset [13, 21]. Most procedure based on Benjamini-Hochberg assume that the true p-values are known. When the distribution of the scores under the null hypothesis is unknown it is generally not possible to ensure the FDR control with BH. For instance the Monte-Carlo Multiple-Testing has been suggested by [22, 19, 64] to overcome this difficulty. An alternative method for controlling FDR is based on the so-called "local FDR" [58, 60]. Unfortunately this approach relies on a Gaussian assumption. In the context of ICAD, FDR can be controlled using conformal p-values with BH as shown in [4, 61, 42]. Moreover the FDR control can be achieved simultaneously with upper and lower bound as suggested in [41].

Online FDR Control In online multiple-testing, the decision of new observed value as an anomaly has to be done instantaneously. If the BH procedure is applied on the current time series, the time complexity will increase with the size of the data. To tackle this problem, recent papers advocate different methods for the online control of the FDR [60, 30, 49, 63]. In [60] the author suggests using the principle of local FDR. At each observation, a decision is taken depending on the estimation of the local FDR. The [30, 49, 63] introduce a method based on alpha-investing. The p-value is compared to an adaptive threshold depending on the previous decisions. But this method is not applicable for conformal p-values because of its low detection power.

Controlling false positives for online anomaly detection remains a difficult task. In particular two challenges arise with online anomaly detection:

- The true p-values are unknown and need to be estimated.
- The decisions are made in an online context, whereas most of the multiple testing methods are done in the offline context.

The main contributions are to tackle these challenges. More precisely it is established that it is possible to design online anomaly detectors controlling the FDR of the time series.

- This paper study the relationship between the FDR and the cardinality of the calibration set used to estimate p-values. To guarantee FDR control, a calibration set cardinality tuning method is proposed.
- This paper describes an online calibration strategy for anomaly detection based on multiple testing ideas to control the False Discovery Rate (FDR).
  - It explains how global control of the time series FDR can be obtained from local control of a modified version of subseries FDR. This makes it possible to control the FDR within an online context.
  - A modified version of the Benjamini-Hochberg procedure is suggested to achieve local control of the modified FDR.

1.3. Description of the paper

First, the problem is explained and important objects are introduced in Section 2. Second Section 3 deals with conditions on p-values estimations to ensure local control of FDR is controlled at a desired level. Third this paper develops algorithms that allows global control the FDR time series and studies them in Section 4. Finally our solution is evaluated against one competitor from the literature in Section 5.

2. Statistical framework

2.1. The Anomaly Detector

Let $(\mathcal{X}, \Omega, \mathbb{P})$ be a probability space and assume a realization of the random variables $(X_t)_{t\geq 1}$, with X_t taking values in $\mathcal{X}$ for all t, is observed at equal time steps. $T \in \mathbb{N} \cup \{\infty\}$ is the size of the time series. Let $\mathcal{P}_0$ be a probability distribution, called reference distribution, on the space $\mathcal{X}$. For each instant t, the observation X_t is called "normal" if $X_t \sim \mathcal{P}_0$. Otherwise, X_t is an "anomaly". The aim of an online anomaly detector is to find all anomalies among the new observations along the time series $(X_t)_{t\geq 1}$: for each instant $t > 1$, a decision is taken about the status of X_t based on past observations: $(X_s)_{1\leq s\leq t}$.

In this paper, the following general online anomaly detector description is suggested. It uses multiple testing ideas from [42] and the online context from [33]. It is based on the three following notions:

1. **Atypicity score**: A score $a : \mathcal{X} \to \mathbb{R}$ is a function reflecting the atypicity of an observation X_t. To be more specific, the further $\mathcal{P}_{X_t}$ from $\mathcal{P}_0$, the larger $a(X_t)$.

2. p**-value**: It is the probability of observing $a(X)$ higher than $a(X_t)$ when $X \sim \mathcal{P}_0$:
$$p_t = \mathbb{P}_X\left(a(X) \geq a(X_t)\right).$$
The p-value enables an interpretable criterion measuring how much unlikely $X_t \sim \mathcal{P}_0$. It is estimated using empirical p-value, given in Definition 2, and it allows to tackle the problem of low interpretability of an unknown distribution of the atypicity score.

3. **Detection threshold**: $\varepsilon \in [0, 1]$, it discriminates observations considered as abnormal from others. The observations considered as anomalies are x_t whose (estimated) p-value is lower than the threshold ε,
$$\mathbb{P}_X\left(a(X) \geq a(X_t)\right) < \varepsilon.$$

This general online anomaly detector is formalized in the pseudo-code given by Algorithm 1.

Algorithm 1 Generic Online Anomaly Detector with fixed threshold ε

Require: $T > 0$, $(X_t)_1^T$ time series, a an abnormality score, ε a threshold
  for $1 \leq t \leq T$ **do**
    $S_t \leftarrow a(X_t)$
    $p_t \leftarrow p$-value(S_t)
    if $p_t < \varepsilon$ **then**
      $d_t = 1$
    else
      $d_t = 0$
    end if
  end forOutput: $(d_t)_1^T$ detected anomalies boolean

The threshold ε allows to control the "detection frequency": a smaller threshold will generate fewer detections. This is equivalent to defining anomalies as points above a quantile in the tail of the score distribution. Nevertheless, in practice, the calibration of the threshold is difficult. Since ε affects directly the number of false detections (false positives), it is not possible to know in advance the number of false positives due to the choice of ε.

One of the main contributions of this work consists in developing a data-driven rule allowing to choose a threshold $\hat{\varepsilon}_t$ at each time step t. This rule has the advantage of ensuring a global control of the false discovery rate on the complete set of observations. In this paper, it is suggested to replace the last step of the Generic Online Anomaly Detector (Algorithm 1) with an adaptive threshold that is computed in real time before any decision would be made.

2.2. Control of false positives and multiple testing

Since the present goal is to use FDR, a natural strategy is to rephrase the online anomaly detection problem as a multiple testing problem: At each step

$1 \leq t \leq T$, a statistical test is performed on the hypotheses:

$$\mathcal{H}_{0t}, \text{“}X_t \text{ is not an anomaly ”} \qquad \text{against} \qquad \mathcal{H}_{1t} \text{“}X_t \text{ is an anomaly”}.$$

A natural criterion controlling the proportion of type I errors (False Positives) of the whole time series is FDR [5]. For a given data-driven threshold $\hat{\varepsilon}$ and a set of *estimated* p-values $\hat{p} = (\hat{p}_t)_{t \geq 1}$, the FDR criterion of the sequence from 1 to T is given by

$$FDR_1^T(\hat{\varepsilon}, \hat{p}) = \mathbb{E}[FDP_1^T(\hat{\varepsilon}, \hat{p})],$$
$$\text{with} \quad FDP_1^T(\hat{\varepsilon}, \hat{p}) = \frac{\sum_{t \in \mathcal{H}_0} \mathbb{1}[\hat{p}_t \leq \hat{\varepsilon}_t]}{\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t \leq \hat{\varepsilon}_t]},$$

with the convention that $0/0 = 0$. In the above expression, FDP_1^T denotes the *False Discovery Proportion* (FDP) of the time series from 1 to T. Also $\mathcal{H}_0 = \{t \in \mathbb{N}^* | \mathcal{H}_{0t} \text{ is true}\}$ is called the set of null hypotheses. Let us emphasize that the anomalies (according to Algorithm 1) satisfy $\mathbb{1}[\hat{p}_t \leq \hat{\varepsilon}_t] = 1$. The main objective of the present work is to define a data-driven sequence $\hat{\varepsilon} : t \mapsto \hat{\varepsilon}_t$ such that, for a given control level $\alpha \in [0, 1]$, under weak assumptions on the sequence $\hat{p} : t \mapsto \hat{p}_t$,

$$FDR_1^T(\hat{\varepsilon}, \hat{p}) \leq \alpha \tag{2.1}$$

The control is said exact when “$\leq$” is replaced with “$=$”. Such a control would imply that for a level $\alpha = 0.1$, at most 10% of the detected anomalies along the whole time series are false positives.

The detection power of the anomaly detector is measured by means of the *False Negative Rate* defined, for the sequence from 1 to T, by

$$FNR_1^T(\hat{\varepsilon}, \hat{p}) = \mathbb{E}[FNP_1^T(\hat{\varepsilon}, \hat{p})], \tag{2.2}$$
$$\text{with} \quad FNP_1^T(\hat{\varepsilon}, \hat{p}) = \frac{\sum_{t \in \mathcal{H}_1} \mathbb{1}[\hat{p}_t \leq \hat{\varepsilon}_t]}{|\mathcal{H}_1|}, \tag{2.3}$$

where FNP_1^T denotes the *False Negative Proportion* (FNP) of the sequence from 1 to T and $\mathcal{H}_1 = \{t \in \mathbb{N}^* | \mathcal{H}_{1t} \text{ is true}\}$ is the set of alternative hypotheses.

However a crucial remark at this stage is that controlling FDR on the complete time series is a highly challenging task in the present online context for at least two reasons:

- The main existing approaches for controlling FDR are described in an “offline” framework where the whole series is observed first, and decisions are taken afterwards [5, 41]. This makes these approaches useless in the present context.
- The already existing approaches designed in the online context [30, 63] are difficult to parameterize and hard to apply with *estimated p-values*. Let us emphasize that realistic scenarios usually exclude the knowledge of the true probability distribution of the test statistics, leading to approximating or estimating the related p-values in practice.

2.3. FDR control with Empirical p-value

A classical (offline) strategy for controlling FDR is the so-called Benjamini-Hochberg (BH) multiple testing procedure [5]. Exact control relies on the knowledge of true p-values, which is usually not realistic. Actually since the true reference distribution is unknown in practical anomaly detection scenarios, there is no true p-values available.

2.3.1. Empirical p-value

The atypicity level of an observation is quantified by an atypicity score. The underlying scoring function assigns each observation with a real value such that *the more atypical the observation, the higher the score value.*

Definition 1 (Scoring function)**.** *The scoring function is introduced:* $a : \mathcal{X} \to \mathbb{R}$. *The abnormality score* u *at the point* x *is defined by* $u = a(x)$.

The interpretation is that the higher the score (value) at an observation, the more unlikely the corresponding observation has been generated from a reference distribution implicitly encoded in the scoring function.

Examples (Examples of scoring functions)**.** *Let* $x \in \mathcal{X}$ *and* $z_1^\ell = \{z_1, \ldots, z_\ell\}$ *be a training set generated from* $\mathcal{P}_0$.

1. *Z-score [65, 11]: Let* μ *and* σ *be estimators of mean and standard deviation of* z_1^ℓ,
$$a(x) = a_Z(x, z_1^\ell) = |(x - \mu)/\sigma|$$

2. *kNN score [53, 12]: Let* d *be a metric on* $\mathcal{X}$ *and* $k > 0$ *and* $kNN(x, z_1^\ell)$ *is the* k*-th nearest neighbor of* x *in* z_1^ℓ.
$$a(x) = a_{kNN}(x; z_1^\ell, k) = \frac{1}{k} \sum_{z \in kNN(x, z_1^\ell)} d(x, z)$$

3. *KDE-score [53, 12]: let* K *a kernel function.*
$$\begin{aligned} a(x) &= a_{KDE}(x, K, z_1^\ell) \\ &= \frac{1}{\ell} \sum_{z \in z_1^\ell} (-K(x, z)) \end{aligned}$$

4. *Auto-Encoder score [65, 43]. Considers having* $\mathcal{X} = \mathbb{R}^n$, *a compression function* $f_\theta : \mathbb{R}^n \to \mathbb{R}^d$ *with* $d < n$, *a reconstruction function* $g_\phi : \mathbb{R}^d \to \mathbb{R}^n$ *and the hypothesis spaces* Θ *and* Φ.
$$(\hat{\theta}, \hat{\phi}) \in \arg \min_{\theta \in \Theta, \phi \in \Phi} \sum_{x \in z_1^\ell} ||g_\phi(f_\theta(x)) - x||^2$$

Then the abnormality score is given by

$$\forall x \in \mathbb{R}^n, a(x) = a_{AE}(x, z_1^\ell) = ||g_{\hat{\phi}}(f_{\hat{\theta}}(x)) - x||^2.$$

The choice of the abnormality score depends on the structure of the time series and the type of anomalies one is looking for. Intuitively a desirable scoring function should assign a high abnormality score to any true anomaly. For example, Z-score is only able to detect anomalies that are in the tail of the distribution. By contrast it is not effective to detect abnormal point between two modes of data with a bimodal distribution [65, 11]. kNN or KDE scores are more suited for multi-modal data because they raise a high score for points far from the observations of the training set. The intuition behind such a scoring function is that normal data should have a low distance from the training set. Auto-Encoders are often used for multidimensional data with complex distributions. It relies on the possibility to compress the input data in a low dimensional embedding space without losing too much information. This enables to reconstruct the input with low reconstruction error, at the computational price of training first a deep neural network [65, 43]. To the best of our knowledge, there does not exist any scoring function suitable for detecting all types of anomalies.

Defining a meaningful threshold from a score is the classical strategy for deciding that an observation is anomalous or not. This requires to know the true distribution of these scores, which is not realistic in general. The induced estimation step is usually made by two means. On the one hand, one can assume a parametric Gaussian distribution for the scores [55, 11, 27]. On the other hand, one can estimate the score distribution by use of sampling techniques [52, 12]. Since the Gaussian assumption can cause some troubles when it is violated, the present work rather focuses on the second strategy by considering anomaly detection relying on empirical p-values. By contrast to the Gaussian assumption, a strong asset of empirical p-value is that they can be used no matter the true score distribution or the scoring function.

Definition 2 (Empirical p-value)**.** *Let a be a scoring function (Definition 1). Let $\{x_1, ..., x_n\} \subset \mathcal{X}$ be a set of data called the calibration set. The empirical p-value is a function defined by*

$$\forall x \in \mathcal{X}, \quad \hat{p}\text{-}value(x; \{x_1, ..., x_n\}) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(a(x_i) \geq a(x)). \tag{2.4}$$

Let us emphasize that Definition 2 describes an estimator of the p-value under $\mathcal{P}_0$ provided the calibration set is composed of points generated from the reference distribution $\mathcal{P}_0$. However it is well known that the main difficulty with this p-value estimator is that it is not itself a p-value [42] since the so-called *super-uniformity property* is violated. More precisely, super-uniformity means that, for all $u \in [0, 1]$,

$$\mathbb{P}_{X, X_1, ..., X_m \sim \mathcal{P}_0}(\hat{p}\text{-value}(X; \{X_1, \ldots, X_m\}) \leq u) \leq u.$$

Therefore empirical p-values are usually replaced by an other p-value estimator called the conformal p-value[41, 33], given by

$$\tilde{p}\text{-value}(x; \{x_1, ..., x_n\}) = \frac{1}{m+1}\left(1 + \sum_{i=1}^{m} \mathbb{1}(a(x_i) \geq a(x))\right). \tag{2.5}$$

This definition implies the p-value property for all u in $[0, 1]$. But this estimator is less powerful, as illustrated by Figure 11 in Appendix A.1 where the FNR resulting from the use of conformal p-values is always larger than that of empirical p-values.

As a consequence, an important remark is that the present work focuses on empirical p-values (and not on conformal ones). However another motivation for this choice is provided in Section 3.2.2 where it is proved that the *super-uniformity* property also holds true with empirical p-values under some specific conditions that will be detailed later.

2.3.2. BH-procedure does not control FDR with empirical p-values

The present section starts by describing the behavior of the BH-procedure as well as establishing the resulting FDR control. An illustration is provided that the BH-procedure does not control FDR are the prescribed level when empirical p-values are used. This illustration is then theoretically justified, which shows that straightforwardly using the BH-procedure in our online context is prohibited.

Definition 3 (Benjamini-Hochberg ([5, 61]))**.** *Let m be an integer and $\alpha \in [0, 1]$. Let $(p_i)_{1 \leq i \leq m} \in [0, 1]^m$ be a family of p-values. The Benjamini-Hochberg (BH) procedure, denoted by BH_α, is given by*

- *a data-driven threshold:*

$$\varepsilon_{BH_\alpha} = max\{\frac{\alpha k}{m}; p_{(k)} \leq \frac{\alpha k}{m}, k \in [\![1, m]\!]\},$$

- *a set of rejected hypotheses:*

$$BH_\alpha\left((p_i; i \in [\![1, m]\!])\right) = \{i; p_i \leq \varepsilon_{BH_\alpha}, i \in [\![1, m]\!]\}.$$

The intuition behind this procedure consists in drawing the ordered statistics $i \mapsto p_{(i)}$ (Figure 1) with $p_{(1)} \leq \cdots \leq p_{(n)}$ and the straight line $i \mapsto \frac{\alpha i}{m}$. Then the BH-procedure amounts to rejecting all hypotheses corresponding to p-values smaller than the last crossing point between the straight line and the ordered p-values curve.

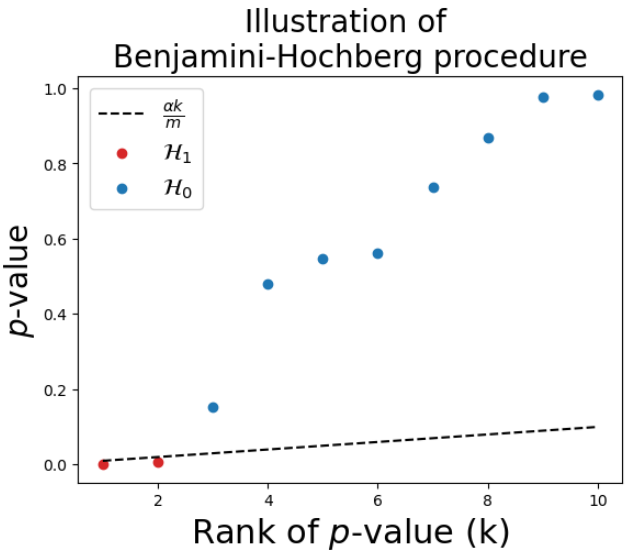


Figure 1: Illustration of the Benjamini-Hochberg procedure. p-values are sorted by increasing order. The threshold is the greatest p-value that is lower than $\alpha k/m$, when k is the rank of the p-value.

The striking property of this procedure is to yield the desired control of the FDR at the prescribed level α as stated by the next result.

Theorem 1 (FDR control with BH [5])**.** *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,*

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

The proof of the theorem is deferred to Appendix A.3. In particular, the FDR control results from the fact that under $\mathcal{H}_0$, the true p-values follow a uniform distribution. The equality could be replaced by an upper bound if the uniform distribution assumption were weakened by the super-uniform property.

By contrast with the previous framework, when performing anomaly detection, the abnormality score is computed using a scoring function (Definition 1), and the true p-value is given by

$$p_t = \mathbb{P}_{X \sim \mathcal{P}_0}\left(a(X) \geq a(X_t)\right),$$

where the notation clearly emphasizes the dependence with respect to the *unknown* reference distribution. This justifies why empirical p-values are now substituted to true ones as earlier explained (see Eq. 2.4).

A difficulty resulting from using empirical p-values in the BH-procedure is that the FDR control does no longer hold true as illustrated by Figure 2. This figure displays the actual FDR value (plain blue curve) versus the cardinality of the calibration set used to computed the empirical p-values (see Definition 2) in the specific situation of Gaussian data. Except for some a few values of the calibration set cardinality, the FDR control is no longer achieved (red horizontal line). Furthermore the actual FDR value is higher than the desired $m_0/m\alpha$. This results from the fact that the super uniform property is violated when using empirical p-values as established by Proposition 1 below.

Proposition 1 (Distribution of empirical p-value under H_0)**.** *Let $X \sim \mathcal{P}_0$ where $\mathcal{P}_0$ is the probability distribution under $\mathcal{H}_0$, the calibration set cardinality is denoted by n, and $\{X_1, \ldots, X_n\} \sim \mathcal{P}_0^n$ is the calibration set. If one further assumes that there are no ties among $a(X_1), \ldots, a(X_n)$, then the empirical p-value at X is denoted by $\hat{p}$-value$(X; \{X_1, \ldots, X_n\})$ and follows the discrete uniform distribution*

$$U(0, \frac{1}{n}, \frac{2}{n}, \ldots, 1).$$

Let us mention that under $\mathcal{H}_0$ the empirical p-value has a different distribution from that of the conformal p-value [41] which follows $U(1/(n+1), \ldots, 1)$. The conformal p-value is never smaller than $1/(n+1)$, which raises issues in terms of the detection power with lots of false negatives (see the right panel of Figure 11 and the discussion in Appendix A.1). With empirical p-values, it can be easily checked that

$$\mathbb{P}\left(\hat{p}\text{-value}(X; \{X_1, \ldots, X_n\}) \leq 0\right) = \frac{1}{n+1} > 0,$$

which violates the super uniformity property. As a consequence, FDR is no longer controlled by the BH-procedure [6] applied to empirical p-values. Other consequences owing to the use of empirical p-values violating super uniformity are illustrated in Section 3.1.

The assumption of no ties are allowed among the scores $a(X_i)$s is quite mild and fulfilled most of the time as supported by Example 2.3.1 as long as the reference distribution is continuous (admits a density).

Proof of Proposition 1. Since $X \sim \mathcal{P}_0$ and $X_1, \ldots, X_n \sim \mathcal{P}_0^n$ are independent, the empirical p-value $\hat{p}$ is a random variable computed from X and X_1^n and satisfies for any $0 \leq \ell \leq n$ that

$$\begin{aligned}
&\mathbb{P}(\hat{p} = \ell/n) \\
&= \mathbb{P}(a(X_{(\ell+1)}) < a(X) \leq a(X_{(\ell)})) \\
&= \mathbb{P}(\{\text{rank of } a(X) \text{ among } \{a(X), a(X_1), \ldots, a(X_n)\} \text{ is } \ell + 1\}),
\end{aligned}$$

where $a(X_{(n)}) < a(X_{(n-1)}) < \cdots < a(X_{(1)})$. The conclusion comes from noticing that the probability distribution of $\{X, X_1, \ldots, X_n\}$ is exchangeable, and assumption of no ties in scores $a(X_i)$ □

Let us also notice that Figure 2 shows that there exist particular values of the calibration set cardinality for which FDR is still controlled at the prescribed level. This perspective is further explored in Section 3.2.2, where a new multiple testing procedure yielding the desired FDR control for the whole time series is devised.

3. FDR control with Empirical *p*-values

The goal here is to describe a strategy achieving the desired FDR control for a time series of length m when using empirical p-values. A motivating example is first introduced for emphasizing the issue in Section 3.1. Then a theoretical understanding is provided along Section 3.2 which results in a new solution which applies to independent empirical p-values. An extension is then discussed to thz non-independent setup in Section 3.3.1. Finally experimental results are reported in Section 3.4 to (empirically) assess the validity of our previous theoretical conclusions.

3.1. Motivating example

The purpose here is to further explore the effect of the calibration set cardinality on the actual FDR control when using empirical p-values. This gives us more insight on how to find mathematical solutions.

Let us start by generating observations using two distributions. The reference distribution is $\mathcal{P}_0 = \mathcal{N}(0, 1)$ and the alternative distribution is $\mathcal{P}_1 = \mathcal{N}(4, 10^{-4})$. The anomalies are located in the right tail of the reference distribution. The length m of the signal is $m = 100$. The number of observations under $\mathcal{P}_0$ is $m_0 = 99$. The experiments have been repeated $K = 10^4$ times.

Figure 2 displays the actual value of FDR as a function of the cardinality n of the calibration set $\{x_1, ..., x_n\}$ used to compute the empirical p-values (see Definition 2). One clearly see that FDR is not uniformly controlled at level $m_0/m\alpha$. However there exist particular values of n for which this level of control is nevertheless achieved. As long as n has become large enough ($n \geq 500$), repeated picks can be observed with a decreasing height as n grows.

3.2. FDR control: main results for i.i.d. *p-values*

The present section aims at first explaining the shape of the curve displayed in Figure 2. This will help getting some intuition about how to design an online procedure achieving the desired FDR control for the full time series.

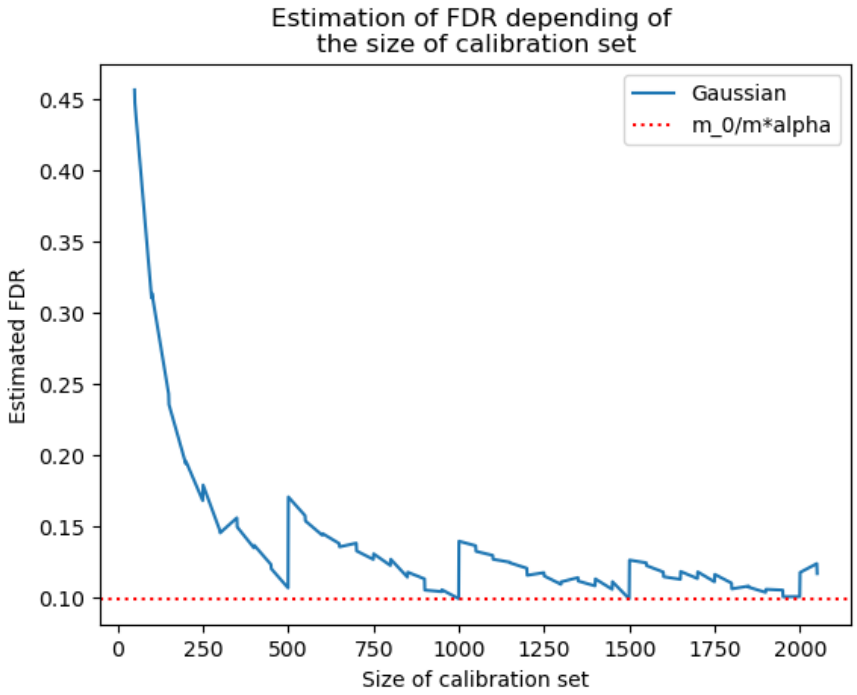


Figure 2

3.2.1. Proof of FDR control by BH revisited

The main focus is now given to independent p-values. In what follows, the classical proof (Proof A.3) of the FDR control by the BH-procedure is revisited then leading to the next result. Its main merit is to provide the mathematical expression of the plain blue curve observed in Figure 2.

Theorem 2. *Let n be the cardinality of the calibration set and m be that of the set of tested hypotheses where $\{X_1, \dots, X_m\}$ denotes a set of random variables. Let $m_0 \leq m$ be the cardinality of the random variables from the reference distribution $\mathcal{P}_0$. Let the empirical p-value be denoted, for any $i \in [\![1, m]\!]$, by $\hat{p}_i = \hat{p}\text{-value}(X_i, \{Z_{i,1}, \dots, Z_{i,n}\})$, where the calibration set is $\{Z_{i,1}, \dots, Z_{i,n}\}$ and each $Z_{i,j} \sim \mathcal{P}_0$. Let the random variables $R(i)$ be the number of detections raised by BH_α when replacing X_i with 0, as defined along the proof detailed in Appendix A.3. Then for every $\alpha \in]0, 1]$, the FDR value over the sequence from 1 to m is given by*

$$FDR_1^m(\hat{\varepsilon}_{BH_\alpha}, \hat{p}) = m_0 \sum_{k=1}^{m} \frac{\frac{\lfloor \frac{\alpha k n}{m} \rfloor + 1}{n+1}}{k} \mathbb{P}(R(i) = k),$$

where $\hat{\varepsilon}_{BH_\alpha}$ denotes the BH_α threshold from Definition 3 when the BH-procedure is applied to the empirical p-values $\hat{p} = (\hat{p}_i; 1 \leq i \leq m)$.

In general it is not possible to compute the exact value of the FDR without knowing the distribution of the random variables $R(i)$. This is in contrast with the case of true p-values where $\mathbb{P}(p_i \leq \frac{\alpha k}{m}) = \frac{\alpha k}{m}$, whereas with empirical p-values $\mathbb{P}(\hat{p}_i \leq \frac{\alpha k}{m}) = \frac{\lfloor \frac{\alpha k n}{m} \rfloor + 1}{n+1}$, which prevents from any simplification of the final bound. Nevertheless this value still suggests a solution to circumvent this difficulty: requiring conditions on α, m, and n such that $\frac{\lfloor \frac{\alpha k n}{m} \rfloor + 1}{n+1} = \frac{\alpha k}{m}$, for all k. This is precisely the purpose of next Corollary 1.

Proof of Theorem 2. When applying Proof A.3, the only modification is that $\mathbb{P}(\hat{p}_i \leq \frac{\alpha k}{m})$ is not equal to $\frac{\alpha k}{m}$ since $\hat{p}_i$ now follows the discrete uniform distribution

$$\mathbb{P}(\hat{p}_i \leq \frac{\alpha k}{m}) = \sum_{\ell=0}^{\lfloor n\alpha k/m \rfloor} \mathbb{P}(n\hat{p}_i = \ell) = \frac{\lfloor \frac{\alpha k n}{m} \rfloor + 1}{n+1}.$$

Plugging this in the FDR expression, it gives

$$FDR_1^m(\hat{\varepsilon}_{BH_\alpha}, \hat{p}) = m_0 \sum_{k=1}^{m} \frac{\frac{\lfloor \frac{\alpha k n}{m} \rfloor + 1}{n+1}}{k} \mathbb{P}(R(i) = k).$$

Recall that $\hat{p}_i$ follows $U(0, 1/n, 2/n, \ldots, 1)$ entails that $n\hat{p}_i$ follows $U(0, 1, 2, \ldots, n)$. □

3.2.2. Tuning of the calibration set cardinality

Corollary 1. *Under the same notations and assumptions as Theorem 2, the next two results hold true.*

1. *Assume that there exists an integer $1 \leq \ell$ such that $\frac{\ell m}{\alpha}$ is an integer. If the cardinality n of the calibration set satisfies $n = n_\ell - 1 = \ell m/\alpha - 1$, then*

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

2. *For every $\alpha \in]0, 1]$, assume that the cardinality of the calibration set satisfies $n = n_\ell - 1 = \lceil \frac{\ell m}{\alpha} \rceil - 1$, for any integer $\ell \geq 1$. Then,*

$$\frac{n}{(n+1)} \frac{m_0 \alpha}{m} \leq FDR_1^m(\hat{\varepsilon}_{BH_\alpha}, \hat{p}) \leq \frac{m_0 \alpha}{m}.$$

The proof is postponed to Section A.4. The first statement in Corollary 1 establishes that recovering the desired control of FDR at the exact prescribed level α is possible on condition that the calibration set cardinality is large enough

and more precisely that $n = \ell m/\alpha - 1$. This (mild) restriction on the values of α reflects that the empirical p-values do not satisfy the super-uniformity property. By contrast, the second statement yields the desired control at the level $\alpha m_0/m$ by means of lower and upper bounds. In particular, the lower bound tells us that the FDR value can be not lower than the desired level $\alpha m_0/m$ up to a multiplicative factor equal to $1-1/n$, which goes 1 as n grows. For instance with $\alpha = 0.1$ and $m = 100$, $n_\ell = 1000$ would yield that $FDR_1^m(\hat{\varepsilon}_{BH_\alpha}, \hat{p}) \cdot m/(m_0\alpha) \in [0.999, 1]$. This small lack of control is the price to pay for allowing any value of $\alpha \in]0, 1]$. It is also important to recall that in the anomaly detection field, abnormal events are expected to be rare. As a consequence $\frac{m_0}{m}$ is close to 1 and the actual FDR level is close to the desired α. However in situations where m_0/m could depart from 1 too strongly, then incorporating an estimator of m_0/m would be helpful.

3.3. Extension to dependent p-values

In Section 3.2, Corollary 1 states that FDR is controlled at a prescribed level with empirical p-values for which the super-uniformity property is not fulfilled. A key ingredient in the proof was the independence property across empirical p-values. One purpose of the present section is to extend these results to non-independent p-values.

Towards this extension, the concept of positive regression dependency (referred to as PRDS) [6] turns out to be useful. The PRDS property is a form of positive dependence between p-values where all pairwise p-value correlations are positive. It results that a small p-value for a given observation makes other p-values for all considered observations simultaneously small as well, and vice-versa [4].

3.3.1. Theoretical results to dependent p-values

A classical result established in [6] proves that FDR is upper bounded by $\alpha m_0/m$ provided the p-value family satisfies the PRDS and super-uniformity properties. It turns out that this result can be extended to our estimator with the same choice of calibration set cardinality as the one discussed in Corollary 1. Another important achievement is the fact that FDR can be also lower bounded in the case where the calibration set is the same for all (empirical) p-values (see Definition 2). This results originally proved by [41] is extended here to empirical p-values computed with a calibration set cardinality tuned as suggested in Corollary 1.

In Section 3.2, considering the size of the calibration set is correctly chosen, it has been proved that the control of the FDR can be achieved with estimated p-values for which the super-uniformity property is not fulfilled. The results obtained for i.I.d. p-values will be extended in this section for non i.i.d. p-values.

For this extension, the concept of positive regression dependency on each one from a subset called PRDS [6] is introduced. The PRDS property is a form of positive dependence of p-values where all pairwise p-value correlations are positive. Larger scores in the calibration set make the p-values for all test points simultaneously smaller, and vice-versa [4].

Definition 1 (PRDS property)**.** A family of p-values $\hat{p}_1^m$ is PRDS on a set $I_0 \subset \{1, \dots, m\}$ if for any $i \in I_0$ and any increasing set A, the probability $\mathbb{P}[\hat{p}_1^m \in A | \hat{p}_i = u]$ is increasing in u.

A classical result in [6] asserts that the FDR is upper bounded by $\frac{m_0}{m}\alpha$ in the case where the p-value family is PRDS and super-uniform. This result can be extended our estimator with the same choice of calibration set cardinality than in Theorem 1.

Corollary 2 (Corollary of Theorem 1.2 in [6])**.** *Suppose the family of p-values $\hat{p}_1^m$ is PRDS on the set $\mathcal{H}_0$ of true null hypotheses and suppose that $\hat{p}_1^m$ respects super-uniformity an all thresholds that can may resulting from BH*

$$\forall k \in [\![1, m]\!] \qquad \mathbb{P}(\hat{p}_i < \frac{\alpha k}{m}) \leq \frac{\alpha k}{m},$$

Then, the FDR is upper-bounded by α

$$FDR(\hat{p}, \hat{\varepsilon}_B H) \leq \frac{m_0 \alpha}{m}$$

Unique calibration set More over, in the case where the calibration set is the same for all p-values, the FDR can also be lower bounded as shown in [41]. This result can be extended to empirical p-values given in Definition 2 with a calibration set cardinality tuned as proposed in Theorem 1.

Corollary 3 (Corollary of Theorem 3.4 in [41])**.** *Assuming the following conditions: Let n be the cardinality of the calibration set, m be the cardinality of the active set and m_0 the number of normal observations. Let $\mathcal{P}_0$ be the reference distribution. Let Z_i for i in $[\![1, m]\!]$ independents random variables, following $\mathcal{P}_0$. Let X_i for i in $[\![1, m]\!]$ be random independents variables and independents from (Z_j). There are exactly m_0 random variables following the $\mathcal{P}_0$ distribution. Let a be a scoring function. For all i in $[\![1, m]\!]$, let $\hat{p}_i$ be the empirical p-values associated with the random variables X_i and computed as follows, $\hat{p}_i = p\text{-}value(X_i, \{Z_1, ..., Z_m\}, a)$.*

If the cardinality of the calibration set is a multiple of $n = n_\ell = \ell m/\alpha - 1$, then the FDR using $\hat{BH}_\alpha$ on $(\hat{p}_i)_{1 \leq i \leq m}$ is equal to $\frac{m_0 \alpha}{m}$:

$$FDR_1^m(\hat{\varepsilon}_{BH_\alpha}, \hat{p}) = \frac{m_0 \alpha}{m}$$

Overlapping calibration set In the context of online anomaly detection, moving windows are classically used to capture and process the incoming data.

This is why the calibration sets of the p-value family will overlap. To have a perfect control of the FDR, an upper and lower bounds is needed.

According Proposition 2, p-values with overlapping calibration sets are PRDS.

Proposition 2 (PRDS property for overlapping calibration sets)**.** *Let X_i for i in $[\![1, m]\!]$ be random independents variables. There are exactly m_0 random variables following the $\mathcal{P}_0$ distribution, with belong to $\mathcal{H}_0$. Let $\boldsymbol{Z}$ be the random vector that combine all calibration set, all elements of $\boldsymbol{Z}$ are generated from $\mathcal{P}_0$. The set of n indices defining the elements of calibration set related to $\hat{p}_i$ in $\boldsymbol{Z}$ is noted $\mathcal{D}_i$. The calibration set related to X_1 is noted $\boldsymbol{Z}_{\mathcal{D}_1} = (\boldsymbol{Z}_{i_1}, ..., \boldsymbol{Z}_{i_n})$. For all i in $[\![1, m]\!]$: $\hat{p}_i = p\text{-}value(X_i, \boldsymbol{Z}_{\mathcal{D}_i})$.*

Under these conditions, the set of p-values is PRDS on $\mathcal{H}_0$

The proof of the proposition is in delayed to Appendix A.2. Since such p-values are are PRDS, it gives an upper bound control of the FDR using Corollary 4.

Corollary 4 (PRDS property for overlapping calibration sets)**.** *Under the same conditions as Proposition 2 and the condition on calibration set cardinality satisfy $\exists \ell \geq 1, n = \ell \frac{m}{\alpha} - 1$:*

$$FDR(\hat{p}, \hat{\varepsilon}_{BH}) \leq \frac{m_0 \alpha}{m}$$

3.3.2. Calibration set and impact of the overlap

In the context of online anomaly detection, moving windows are usually used to capture and process the incoming observations. In this context, the calibration set coincides with the observations within this window. Since successive windows are overlapping each other depending on the size of the shift, the resulting calibration sets used for computing the successive empirical p-values are also overlapping. To have a perfect control of the FDR, an upper and lower bounds are needed. Since the Appendix A.2 proves that such p-values are PRDS, it gives an upper bound control of the FDR. No theoretical results exists to compute the lower bound, indeed the existing proof in [41] did not extend to overlapping calibration sets. Therefore the next discussion suggests to establish this lower bound empirically.

The following experiments aims at drawing a comparison between the FDR values in three scenarios: independent calibration sets, partially overlapping calibration sets with an overlap size driven by the value of sn (size of the shift), and the same calibration set for all empirical p-values. To be more specific, the calibration sets (and corresponding empirical p-values) were generated according to the following scheme. Each calibration set is of cardinality n. When moving from one calibration set to the next one, the shift size is equal to sn, where s in $[0, 1]$ is the proportion of independent data between calibration sets, resulting

in an overlap of cardinality $(1-s)n$. Therefore an overlap occurs as long as $s < 1$. All these ways to build the calibration sets are called "calibration sets strategies".

1. The independent p-values (iid Cal.) are generated according to

$$\forall i \in [\![1, m]\!], \quad \mathbf{Z}_i \sim \mathcal{P}_0^n, \quad \hat{p}_{1,i} = \hat{p}\text{-value}(X_i, \mathbf{Z}_i). \tag{3.1}$$

2. The p-values with the same calibration set (Same Cal.) are generated by

$$\forall i \in [\![1, m]\!], \quad \mathbf{Z} \sim \mathcal{P}_0^n, \quad \hat{p}_{2,i} = \hat{p}\text{-value}(X_i, \mathbf{Z}). \tag{3.2}$$

3. The p-values with overlapping calibration sets (Over. Cal.) are generated given, for $0 < s < n$, by

$$\begin{aligned} \forall i \in [\![1, m]\!], \quad \mathbf{Z}_i = \{Z_{\lfloor isn \rfloor + 1}, \ldots, Z_{\lfloor isn \rfloor + n}\}, \quad \hat{p}_{3,i} = \hat{p}\text{-value}(X_i, \mathbf{Z}_i), \\ \text{and} \qquad \{Z_{s+1}, \ldots, Z_{\lfloor 2sn \rfloor + 1}, \ldots, Z_{\lfloor msn \rfloor + 1}, \ldots, Z_{\lfloor msn \rfloor + n}\} \sim \mathcal{P}_0^{ms+n}. \end{aligned} \tag{3.3}$$

According to these three scenarios, as s increases, the overlap cardinality decreases, which results in more and more (almost) independent calibration sets. This is illustrated by the empirical results collected in Table 1. For each calibration set strategy, presented in row, and for each calibration set cardinality in column, the estimated FDR is shown. In this experiment, the reference distribution $\mathcal{P}_0$ is the Gaussian $\mathcal{N}(0, 1)$ and the anomalies are equal to $\Delta = 4$. The number of tested p-values, noted m, is equal to 100 and m_1, the number of anomalies, is equal to 1. On each sample, BH-procedure is applied with $\alpha = 0.1$ and the FDP is computed. Each FDR is estimated over 10^3 repetitions.

TABLE 1
FDR results with overlapping calibration sets

n	249	250	499	500	749	750	999	1000
Same Cal.	0.164	0.175	0.112	0.183	0.137	0.131	0.097	0.154
Over Cal. (s=0.1%)	0.167	0.174	0.100	0.156	0.138	0.125	0.093	0.140
Over Cal. (s=0.2%)	0.162	0.176	0.095	0.170	0.124	0.127	0.109	0.143
Over Cal. (s=0.5%)	0.163	0.166	0.110	0.170	0.116	0.132	0.111	0.149
Over Cal. (s=1%)	0.151	0.180	0.094	0.177	0.127	0.128	0.099	0.143
Over Cal. (s=2%)	0.164	0.180	0.108	0.179	0.133	0.140	0.097	0.143
Over Cal. (s=5%)	0.168	0.172	0.108	0.169	0.125	0.130	0.096	0.144
Over Cal. (s=10%)	0.165	0.181	0.104	0.185	0.122	0.140	0.105	0.146
Over Cal. (s=20%)	0.173	0.207	0.109	0.171	0.136	0.149	0.101	0.140
Over Cal. (s=50%)	0.180	0.187	0.103	0.183	0.121	0.128	0.094	0.143
iid Cal.	0.171	0.188	0.115	0.174	0.138	0.143	0.104	0.132

The values of n in the columns of Table 1 are chosen such that, for each pair of columns, the FDR value is smaller for the left column and larger for the right column (see Figure 2 for a visual illustration of this phenomenon). Table 1

illustrates that, in the context of the present numerical experiments, the FDR estimation is not too strongly impacted by the value of s (proportion of the overlap). To assert that the observed differences between FDR estimations in each column are not significant, permutation tests [16, 47] are performed. Under $\mathcal{H}_{0n}$ hypothesis, the FDR are the same across all calibration set strategies for the calibration set cardinality n. Under $\mathcal{H}_{1n}$ there are at least two calibration strategies leading to different FDR. The FDP samples that have been used to estimate the FDR are reused. The maximal gap between sample means is used as statistic. The test is performed using the function "permutation_test" from the Python library called Scipy. The significance level is fixed at 0.05. Since multiple tests are performed over the different cardinalities, the threshold for rejecting a hypothesis is 0.00625, according Bonferroni correction. The results are display in Table 2. All tested hypotheses have a p-values greater than the threshold 0.00625. There are no significant difference in the resulting FDR between the different proportions of overlapping in calibration sets. This would suggest that considering overlapping calibration sets should not worsen too much the control of false positives and negatives.

TABLE 2
p-values resulting from permutations test

n	249	250	499	500	749	750	999	1000
p-value of the test	0.300	0.0326	0.572	0.313	0.588	0.435	0.735	0.690

3.4. Empirical Results: Assessing the FDR control

The purpose of the present section is to compute the actual FDR value when empirical p-value are used instead of true ones. The question raised here is to check whether the FDR of the full time series is truly controlled at a prescribed level α. The empirical results must be compared with the theoretical FDR expression that has been established in Theorem 2.

In what follows, Section 3.4.1 describes the simulation design that has been considered, Section 3.4.2 details the criteria used for the assessment, and Section 3.4.3 discusses the experimental results.

3.4.1. Simulation design

Two scenarios have been considered to explore how much the thickness of the distribution tails can influence the results.

1. **Thin tails:**

   The reference probability distribution is $\mathcal{P}_0 = \mathcal{N}(0,1)$ for normal observations and $\mathcal{P}_1 = \delta_{\Delta_\mathcal{N}}$ for anomalies, where $\Delta_\mathcal{N} \in \mathbb{R}$ is a parameter encoding the strength of the shift. Here $\delta_{\Delta_\mathcal{N}}$ denotes the Dirac measure such that $\delta_{\Delta_\mathcal{N}}(z) = 1$ if $z = \Delta_\mathcal{N}$ and 0 otherwise. A Gaussian reference distribution

and anomalies generated from a Dirac distribution in the right tail. $\Delta_{\mathcal{N}}$ is the size of the abnormal spike in the Gaussian distribution.

2. **Thick tails:**

   $\mathcal{P}_0 = \mathcal{T}(5)$ is a Student probability distribution with 5 degrees of freedom and $\mathcal{P}_1 = \delta_{\Delta_{\mathcal{T}}}$ denotes the alternative distribution of anomalies, where $\Delta_{\mathcal{N}} \in \mathbb{R}$ is a parameter encoding the strength of the shift.

Regarding the value of the shift strength in Scenarios 1 and 2, two values of $\Delta_{\mathcal{N}}$ have been considered 3.5 and 4. The values of $\Delta_{\mathcal{T}}$ have been chosen such that

$$\mathbb{P}_{X\sim\mathcal{N}(0,1)}(X > \Delta_{\mathcal{N}}) = \mathbb{P}_{X\sim\mathcal{T}(5)}(X > \Delta_t)$$

for each choice of $\Delta_{\mathcal{N}}$. This avoids any bias in the comparison of the detection power of the considered strategy depending on the ongoing scenario.

Different cardinalities have been considered for the calibration set following the mathematical expression

$$n \in \{k \cdot 10, k \in [\![1, 200]\!]\} \cup \{\ell \cdot 10 - 1, \ell \in [\![1, 200]\!]\}.$$

In particular all integers between 10 and 2 000 are explored with a step size equal to 10 as well as all integers between 9 and 1 999 with a step size of 10. This choice is justified by the particular expression of the FDR value provided by Theorem 2.

All the n elements of the calibration set are generated from the reference distribution that is, $\{Z_1, \ldots, Z_n\} \sim \mathcal{P}_0$. All the m observations corresponding to the tested hypotheses $\{X_1, \ldots, X_m\}$ are generated according to a mixture of $m_1 = 1$ anomalies from $\mathcal{P}_1$ and $m_0 = m - m_1$ normal observations from $\mathcal{P}_0$. Here $m = 100$ and $m_0 = 99$.

Each simulation condition has been repeated $B = 10^4$ times. For each repetition $1 \leq b \leq B$, the observations are indexed by b such that $X_{b,j} \sim \mathcal{P}_1$ for each $j \in [\![1, m_1]\!]$, and $X_{b,j} \sim \mathcal{P}_0$ for $j \in [\![m_1 + 1, m]\!]$.

3.4.2. Criteria for the performance assessment

In the present scenarios, anomalies are all located in the right tail of the reference probability distribution. Therefore the empirical p-value are computed according to Definition 2 with the scoring function $a(x) = x$. For each repetition $1 \leq b \leq B$,

$$\forall 1 \leq j \leq m, \quad \hat{p}_{b,j} = p\text{-value}(X_{b,j}, \{Z_{b,1}, \ldots, Z_{b,n}\}).$$

After computing the empirical p-values, the BH_α procedure (see Definition 3) is applied in such a way that, for any $1 \leq j \leq m$,

$$d_{b,j} = \mathbb{1}_{BH_\alpha(\hat{p}_{b,1},\ldots,\hat{p}_{b,m})}(j),$$

where $\mathbb{1}_I$ denotes the indicator function of the index set I. The FDP value of the sequence from 1 to m is computed from the knowledge of the true label of the observations as "normal" or "anomaly". For each repetition $1 \leq b \leq B$,

$$(FDP_1^m)_b = \frac{\sum_{j=m_1+1}^{m} d_{b,j}}{\sum_{j=1}^{m} d_{b,j}}.$$

The results obtained after the B repetitions are averaged within the FDR estimate of the sequence from 1 to m as

$$FDR_1^m = \frac{1}{B}\sum_{b=1}^{B}(FDP_1^m)_b.$$

The FNR value of the sequence from 1 to m (Equation 2.2) is estimated by

$$(FNP_1^m)_b = \frac{1}{m-m_0}\sum_{j=m_0+1}^{m} d_{b,j}, \quad \text{and} \quad FNR_1^m = \frac{1}{B}\sum_{b=1}^{B}(FNP_1^m)_b.$$

3.4.3. Experimental discoveries

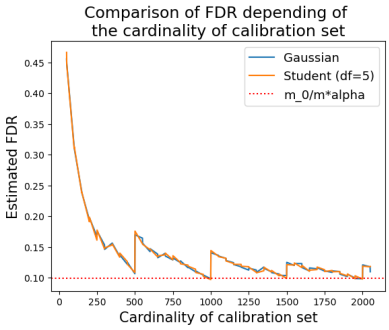
Figure 3 displays the FDR value (left panel) and the FNR value (right panel) as a function of the calibration set cardinality for the two scenarios (Gaussian and Student) described in Section 3.4.1. The blue (respectively orange) curve corresponds to the Gaussian (resp. Student) reference distribution. The horizontal line is the prescribed level $\alpha = 0.1$ at which FDR should be controlled with true p-values (Theorem 1). Figures 3a and 3b are obtained with $\Delta_{\mathcal{N}} = 4$, while Figures 3c and 3d result from $\Delta_{\mathcal{N}} = 3.5$.

According to these plots, the behavior of both FDR and FNR does not exhibit any strong dependence with respect to the reference probability distribution. The results are very close for both Gaussian and Student distributions.

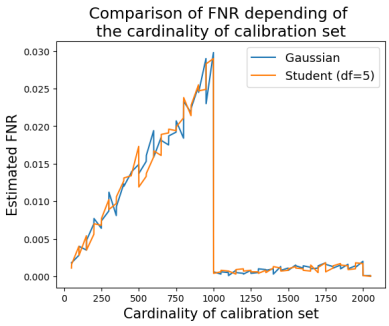
As illustrated by Figures 3a and 3c, the FDR control at the prescribed level is achieved for particular values of the calibration set cardinality. These values coincide with the ones exhibited by Theorem 1, which are multiples of $\alpha/m = 10^3$ (up to a downward shift by 1).

A striking remark is that the FNR curve sharply increases from 1 to $n = 999$. This reflects that although the FDR value becomes (close to) optimal as n increases from 1 to $n = 999$, the proportion of false negatives simultaneously increases leading to a suboptimal statistical performance (because of too many false negatives). Fortunately a larger cardinality n of the calibration set, for instance $n = 1999$, would greatly improve the results at the price of a larger calibration set, which also increases the computational cost.

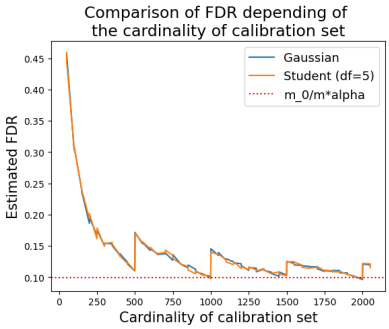
Consistently with what is established in Theorem 1, the FDR value does not depend on the strength of the distribution shift Δ as illustrated by Figures 3a



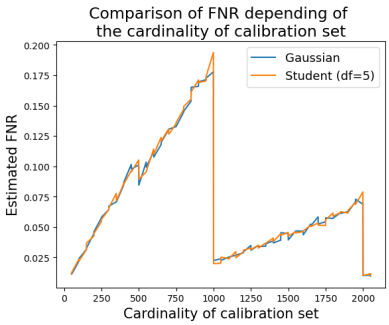
(a) FDR depending on n with 4 sigmas anomalies



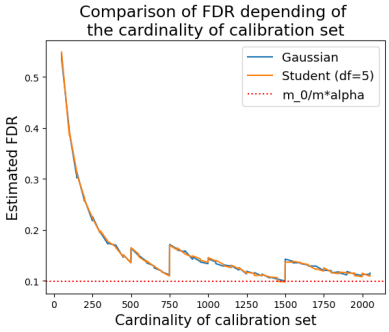
(b) FNR depending on n with 4 sigmas anomalies



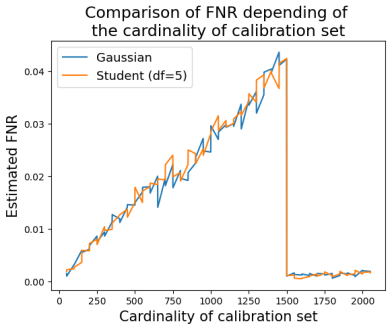
(c) FDR depending on n with 3.5 sigmas anomalies



(d) FNR depending on n with 3.5 sigmas anomalies



(e) FDR depending on n with 4 sigmas anomalies and $m = 150$



(f) FNR depending on n with 4 sigmas anomalies and $m = 150$

Figure 3: Effect of calibration set cardinality and abnormality score on the FDR control and the FNR

and 3c. As long as FDR is concerned (which is an expectation), the shift plays

no role. Let us mention that focusing of the expectation does not say anything about the probability distribution of FDP, which can be influenced by the shift strength. By contrast, the comparison of Figures 3b and 3d clearly shows the impact of the shift strength on the FNR value. As the shift strength becomes lower, anomalies are more difficult to be detected which inflates the FNR value.

The best cardinality n of the calibration set depends on the number m of tested hypotheses according to Theorem 1. For instance, Figure 3a shows the value $n = 999 = 100/0.1 - 1$ as a good candidate since it achieves the desired FDR control while reducing both the number of false negatives and the computation cost. By contrast, Figure 3e rather exhibits the value $n = 1499 = 150/0.1-1$ as the smallest n allowing a perfect FDR control and a small number of false negatives.

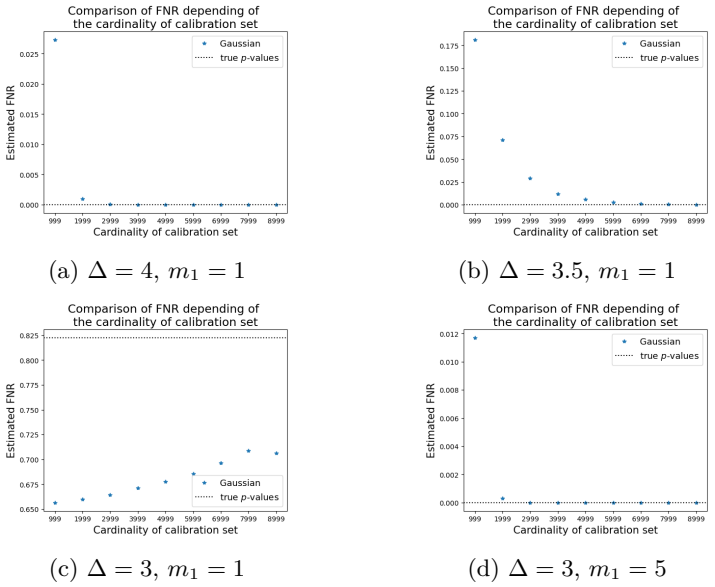
Figure 3a shows other intermediate values calibration set cardinalities yielding the FDR control. For instance $n = 1499$ (between $n = 999$ and $n = 1999$) is predicted by Theorem 1. However complementary experiments (summarized by Figure 12 in Appendix B.1) illustrate that these intermediate values of n allowing the FDR control actually depend on the number of anomalies m_1. Their existence can be explained by the distribution of the number of detections. For example, Figure 12d shows a high probability of detecting 3 anomalies. Assuming there exists $k^* \in [\![1, m]\!]$ such that $\mathbb{P}(R(i) = k^*) \approx 1$, Theorem 2 justifies that

$$\begin{aligned} FDR_1^m(\hat{\varepsilon}_{BH_\alpha}, \hat{p}) &= \frac{n}{n+1} \cdot \alpha \frac{m_0}{m} + \frac{m_0}{n+1} \sum_{k=1}^{m} \frac{(1 - q_{n,k})}{k} \mathbb{P}(R(i) = k) \\ &\approx \frac{n}{n+1} \cdot \alpha \frac{m_0}{m} + \frac{m_0}{n+1} \frac{(1 - q_{n,k^*})}{k^*}. \end{aligned}$$

Then the proof detailed in Appendix A.4 yields that $1 - q_{n,k^*} = \frac{\alpha}{m(n+1)}$ can be reached for all $\ell \geq 1$, such that $n = \ell \frac{m}{\alpha k^*} - 1$. This allows to conclude that $FDR_1^m(\hat{\varepsilon}_{BH_\alpha}, \hat{p}) \approx \frac{m_0 \alpha}{m}$.

3.4.4. How to choose the right cardinality of the calibration set?

Intuitively an optimal choice of the cardinality n of the calibration set should enable the FDR control while minimizing the number of false negatives and avoiding any excessive computation time. To achieve this objective, the first part of Corollary 1 explains that n must be chosen from the set $\mathcal{N} = \{\ell m/\alpha - 1, \quad \ell \geq 1\}$. Using the simulation scenarios described in Section 3.4.1, the aim is to visualize the relationship between the calibration set cardinality and FNR when $n \in \mathcal{N}$. The results are summarized by Figure 4 where the FNR value is displayed versus n.



(a) $\Delta = 4$, $m_1 = 1$ (b) $\Delta = 3.5$, $m_1 = 1$

(c) $\Delta = 3$, $m_1 = 1$ (d) $\Delta = 3$, $m_1 = 5$

Figure 4: FNR as a function of the calibration set cardinality constrained to belong to $\mathcal{N}$.

For all the considered scenarios (Fig. 4a, 4b, 4c, 4d), the FNR value converges to the value reached with true p-values (horizontal dashed line) as n grows. From Figures 4a and 4b, the convergence speed depends on the "difficulty" of the problem.

In practice, generating experiments similar to the ones illustrated by Figure 3.4.1 does not require to know the true reference distribution of the time series since is uses empirical p-values. However, the lack of labeled observations prevents us from computing the abnormality score and the actual FNR value, making the choice of the optimal value of n highly challenging. To tackle this challenge our suggestion is to choose the largest possible value of n that does not exceed the computation time limit. Doing that would output a value of n minimizing the FNR criterion while meeting the computational constraints. However following this suggestion does not prevent us from computational drawbacks as illustrated by Figure 4a where the FNR optimal value is reached for $n = 3999$ while choosing a larger n does not bring any gain (but still increases the computational costs).

4. Global FDR control over the full time series

While working with streaming time series data, the anomaly detection problem requires to control the FDR value of the full time series to make sure that the

global false alarm rate (FDR) remains under control at the end of the iterative process. The final criterion that is to be controlled is then the global FDR criterion given by

$$FDR_1^{\infty}(\hat{\varepsilon}, \hat{p}),$$

where $\hat{\varepsilon} = (\hat{\varepsilon}_t)_{t\geq 1}$ denotes a sequence of data-driven thresholds, and $\hat{p}$ stands for a sequence of empirical p-values (see Section 3 for further details). By contrast with this global objective, anomaly detection nevertheless requires making decision at each time step that is, for each new observation, without knowing what the next ones look like. This justifies the need for another (local) criterion that will be used to make a decision at each iteration, leading to the sequence of data-driven thresholds $\hat{\varepsilon} = (\hat{\varepsilon}_t)_{t\geq 1}$. One additional difficulty results from the connection one needs to create between this local criterion and the (global) FDR of the full time series.

To this end, Section 4.1 starts by showing that controlling the FDR criterion for subseries of the full time series does not provide the desired *global* FDR control. Here "global" means "on the full time series" by contrast with the *local* FDR control, corresponding to controlling FDR for a strict subseries of the full one. Then Section 4.2 explores the connection between FDR for the full time series and the so-called modified-FDR (mFDR) for subseries. In particular, it turns out that controlling the mFDR value for all subseries of a given length m yields the desired FDR control for the full time series. Section 4.3 then explains how the classical BH-procedure can be modified to get the mFDR control for subseries of length m, while Section 4.4 illustrates the practical behavior of the considered strategies on simulation experiments.

4.1. Local and global FDR controls are not equivalent

Let us consider a time series partitioned into 4 subseries as illustrated in Figure 5. The normal points are displayed in black and the anomalies in white. The surrounded points are those that have been detected as anomalies by the procedure. When computing the number of rejections, false positives and the

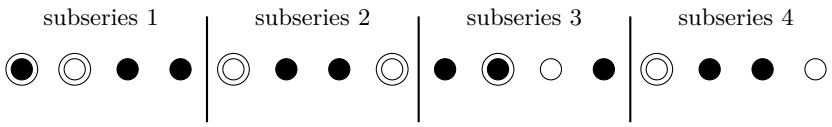


Figure 5: Illustration of anomaly detection in subseries.

False positive rate for each subseries, it comes

- Subseries 1 : 2 rejections, 1 false positive. $FDP_1^4 = 0.5$
- Subseries 2 : 2 rejections, 0 false positive. $FDP_5^9 = 0$

- Subseries 3 : 1 rejections, 1 false positive. $FDP_{10}^{14} = 1$
- Subseries 4 : 1 rejections, 0 false positive. $FDP_{15}^{19} = 0$

If one assumes that the same probability distribution has generated the observations within each subseries, the estimated (local) FDR can be defined as average of the successive FDP values for each subseries that is, $FDR_1^4 = 0.375$. Let us notice that the notation emphasizes that this FDR value is the average over subseries of respective length $m = 4$. If one reproduces the same reasoning for the full time series, it comes: 6 rejections, two false positives, so that $FDP_1^{16} = 1/3 = 0.333$. This example highlights that the FDP of the full time series is not equal to that of smaller subseries. This phenomenon gives some intuition on possible reasons why applying the classical BH-procedure on local windows of length m (subseries) does control the FDR criterion for the individual subseries, but does not yield the desired global FDR control for the full time series. This intuition is confirmed by the boxplots of Figure 6, where BH_α has been applied on subseries of length $m = 100$. The left boxplot shows that BH_α

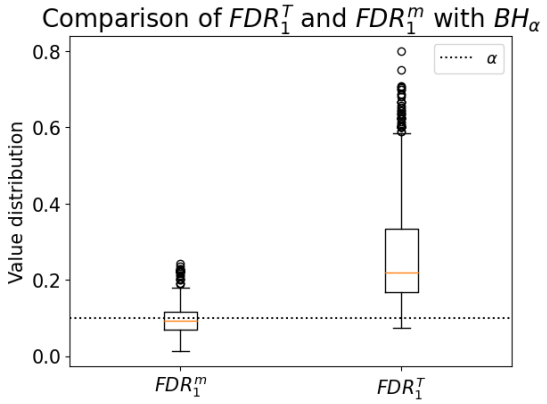


Figure 6: Comparison of the calculation of the FDR computed locally on a subseries and the FDR computed globally on the whole time series with Benjamini-Hochberg procedure applied on a subseries. This result is obtained by cutting a series of cardinality 1000 into 10 subseries of cardinality 100. Then the Benjamini-Hochberg procedure is applied on each subseries.

provides the desired control at level $\alpha = 10\%$ for each individual subseries of length m. However the right boxplot clearly departs from α, meaning that the actual FDR value for the full times series of length 1 000 is strongly larger than α (more than 20% on average) leading to more false positives at the level of the full time series. The boxplots represent the quantile of FDP over 100 repetitions.

4.2. How mFDR can help in controlling the FDR of the full time series

4.2.1. Mixture model and time series

In this section one assumes that the time series is generated from a mixture process between a reference distribution $\mathcal{P}_0$ and an alternative distribution $\mathcal{P}_1$. The anomaly positions are supposed to be independent and generated by a Bernoulli distribution. This is a common assumption usually in the literature [28, 56] for simplification purposes.

Definition 4 (Time series process with anomalies)**.** *Let $\pi \in [0,1]$ be the anomaly proportion and $\mathcal{P}_0$ and $\mathcal{P}_1$ be two probability distributions on the observation domain $\mathcal{X}$. $\mathcal{P}_0$ is the reference distribution and $\mathcal{P}_1$ denotes the alternative distribution. The generation process of a time series containing anomalies $(A_t)_{t\geq 0}$ is given, for every $t \geq 0$, by*

- $A_t \sim B(\pi)$ *(Bernoulli distribution)*
- *if $A_t = 0$, then $X_t \sim \mathcal{P}_0$.*
- *if $A_t = 1$, then $X_t \sim \mathcal{P}_1$.*

Moreover given the above scheme, $(X_t)_{t\geq 0}$ is a random process with independent and identically distributed random variables $X_t \sim (1-\pi)\mathcal{P}_0 + \pi\mathcal{P}_1$.

This definition details the way anomalies are generated. In particular it assumes that anomalies are independent from each other. Let us mention that this does not prevent us from observing a sequence of successive anomalies along the time series. However this scheme substantially differs from the case analyzed by [20] where specific patterns with successive anomalies are looked for.

4.2.2. 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.

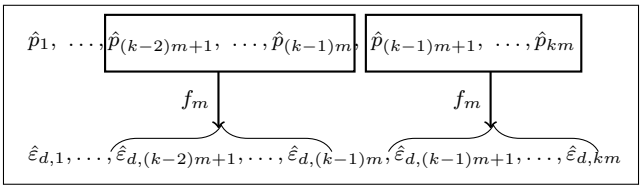
- **Disjoint subseries:** $\hat{\varepsilon}_d : t \mapsto \hat{\varepsilon}_{d,t}$ is given by

$$\forall k \geq 0, \quad \forall t \in [\![km+1, (k+1)m]\!], \quad \hat{\varepsilon}_{d,t} = f_m(\hat{p}_{km+1}, \ldots, \hat{p}_{(k+1)m}) \tag{4.1}$$

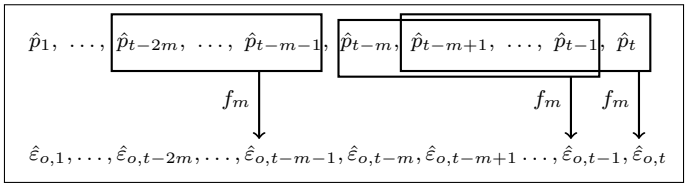
- **Overlapping subseries:** $\hat{\varepsilon}_o : t \mapsto \hat{\varepsilon}_{o,t}$ is given by

$$\forall t \geq m, \qquad \hat{\varepsilon}_{o,t} = f_m(\hat{p}_{t-m+1}, \ldots, \hat{p}_t). \tag{4.2}$$

Figure 7 illustrates these two situations. In Figure 7a, 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 7b 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 7: 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 7a 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 7b 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 4.2.3 still focuses disjoint subseries as a means to introduce important notions without introducing too many technicalities, while Section 4.2.4 extends the previous results to the more realistic case of overlapping subseries.

4.2.3. FDR control with disjoint subseries

As illustrated in Section 4.1, 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 [63, 18], which is defined as follows.

Definition 5 (mFDR)**.** *With the previous notations, the mFDR expression of the subseries from $t-m+1$ to t is given by*

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

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 3, 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 3 (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. (4.1)). Let us also assume that the p-value random process $\hat{p} = (\hat{p}_t)_{t \geq 1}$ follows the scheme detailed in Definition 4.*

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 3 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 3, there is no particular constraint on the integer m. However when dealing within time series of a finite length T, the Theorem 3 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. (4.3)). 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 3 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 3 lies in the use of disjoint subseries, which sounds somewhat restrictive (at least from a practical perspective). This limitation will be overcome by next Theorem 4.

Proof of Theorem 3. Let $k \geq 1$ denote an integer and $T = mk$. Then, the FDP definition and the A_t variables introduced in Definition 4 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,

$$\lim_k \frac{1}{k} \sum_{i=1}^k FP_{im-1}^{(i+1)m}(\hat{\varepsilon}_d, \hat{p}) = \mathbb{E}[FP_1^m(\hat{\varepsilon}_d, \hat{p})], \tag{4.3}$$

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

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

□

4.2.4. FDR control with overlapping windows

Unlike previous Theorem 3, following Theorem 4 establishes a similar control of the global FDR criterion on the full time series by means of successive local controls of the mFDR criterion on subseries that are allowed to overlap each other. This is closer to the practical situation arising in online anomaly detection where one new observations is collected at each time step, inducing a shift by one of the set of observations for which a decision has to be made.

Theorem 4 (Global FDR control with overlapping subseries)**.** *Assume that* $\hat{\varepsilon}_o : t \mapsto \varepsilon_{o,t}$ *is given by* $\hat{\varepsilon}_{o,t} = f_m(\hat{p}_{t-m+1}, \ldots, \hat{p}_t)$*, for any* $t \geq 1$*, with* $f_m : [0,1]^m \to [0,1]$ *permutation invariant (see Eq. (4.1)). Let us also assume that the p-value random process* $\hat{p} = (\hat{p}_t)_{t \geq 1}$ *follows the scheme detailed in Definition 4 and there exists n such that* $|t_1 - t_2| > n$ *implies that* $\hat{p}_{t_1}$ *and* $\hat{p}_{t_2}$ *are independent. 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 computed at time* $t \in \mathbb{N}^*$*. More precisely*

$$FDR_1^{\infty}(\hat{\varepsilon}_o, \hat{p}) = mFDR_{t-m+1}^t(\hat{\varepsilon}_o, \hat{p}) = mFDR_1^m(\hat{\varepsilon}_o, \hat{p})$$

Theorem 4 gives a similar result to the one of Theorem 3 but in a more realistic framework corresponding to the real time anomaly detection context. In particular the main improvement lies in that a threshold can be recomputed at each time step from a (shifted) subseries of length m. An important consequence is that the desired control for the FDR of the full (infinite) time series at level α can be achieved provided one can control the successive mFDR of all (shifted) subseries of length m at level α. This point is not obvious at all and constitutes the main concern of Section 4.3 where a new multiple testing procedure is designed to yield the desired control of the mFDR criterion. The main limitation of Theorem 4 is the requirement that f_m has to be permutation invariant. Let us emphasize that this property holds true with the BH-procedure for instance. Let us also mention that the empirical p-values for instance computed as $\hat{p}_t = \hat{p}\text{-value}(X_t, \{X_{t-n}, \ldots, X_{t-1}\})$ actually satisfy the requirements of Theorem 4 regarding the independence and the stationarity.

Proof of Theorem 4. Let us start with the FDP expression for a time series of length T.

$$\begin{aligned} FDP_{t=1}^{T} &= \frac{\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < \hat{\varepsilon}_{o,t}](1 - A_t)}{\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < \hat{\varepsilon}_{o,t}]} \\ &= \frac{\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < f_m(\hat{p}_{t-m+1}^{t})](1 - A_t)}{\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < f_m(\hat{p}_{t-m+1}^{t})]}, \end{aligned}$$

where $f_m(\hat{p}_i^j) = f_m(\hat{p}_i, \ldots, \hat{p}_j)$, for $i < j$. Since the decision process $(\mathbb{1}[\hat{p}_t < \hat{\varepsilon}_{o,t}])_t$ and the false positives process $(\mathbb{1}[\hat{p}_t < \hat{\varepsilon}_{o,t}]A_t)_t$ are not independent.

Therefore it is not possible to use the Law of Large Numbers as in the proof of Theorem 3. The alternative strategy consists first in splitting the numerator and denominator into several disjoint subseries corresponding to independent and identically distributed processes. Then partitioning the times series of length $T = T'(n+m)$ into T' subseries, each of length $n+m$, it results that

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < f_m(\hat{p}_{t-m+1}^{t})](1 - A_t) \\ &= \frac{1}{n+m}\sum_{k=1}^{n+m}\left(\frac{1}{T'}\sum_{t=0}^{T'-1} \mathbb{1}[\hat{p}_{t(n+m)+k} < f_m(\hat{p}_{t(n+m)+k-m+1}^{t(n+m)+k})](1 - A_{t(n+m)+k})\right). \end{aligned} \tag{4.4}$$

Interestingly for each k from 1 to $m+n$, the summands within the brackets do all belong to different subseries, which makes the sum over t a sum of independent and identically distributed random variables. It results that, for each $1 \leq k \leq n+m$, the average within the brackets is converging to its expectation by the LLN theorem.

Since the limit of a (finite) sum is equal to the sum of the limits, the average in Eq. (4.4) is converging and

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < \hat{\varepsilon}_{o,t}](1 - A_t) &= \sum_{k=1}^{m} \mathbb{E}\left[\mathbb{1}[\hat{p}_k < \hat{\varepsilon}_{o,k}](1 - A_k)\right] \text{ a.s.} \\ &= m\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}](1 - A_m)\right] \text{ a.s..} \end{aligned}$$

Then after applying the same reasoning on the denominator, it gives:

$$FDP_1^{\infty}(\hat{\varepsilon}_o, \hat{p}) = \frac{\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}](1 - A_m)\right]}{\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}]\right]}$$

What remains to show is to proof that:

$$\frac{\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}](1 - A_m)\right]}{\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}]\right]} = mFDR_1^m(\hat{\varepsilon}_o, \hat{p})$$

This result comes from the stationarity of $\hat{p}$ and the permutation invariance of f_m. These properties imply that all p-values inside a subseries have the same probability to be rejected.

$$\forall t \in [\![1, m]\!], \quad \mathbb{E}[\mathbb{1}[\hat{p}_t < f_m(\hat{p}_1^m)](1 - A_t)] = \mathbb{E}[\mathbb{1}[\hat{p}_m < f_m(\hat{p}_1^m)](1 - A_m)] \quad (4.5)$$

Which imply

$$\mathbb{E}[\sum_{t=1}^{m} \mathbb{1}[\hat{p}_t < f_m(\hat{p}_1^m)](1 - A_t)] = m\mathbb{E}[\mathbb{1}[\hat{p}_m < f_m(\hat{p}_1^m)](1 - A_m)] \quad (4.6)$$

Using the same argument for the denominator, is gives:

$$\frac{\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}](1 - A_m)\right]}{\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}]\right]} = \frac{m\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}](1 - A_m)\right]}{m\mathbb{E}\left[\mathbb{1}[\hat{p}_m < \hat{\varepsilon}_{o,m}]\right]} \quad (4.7)$$

$$= \frac{\mathbb{E}[\sum_{t=1}^{m} \mathbb{1}[\hat{p}_t < f_m(\hat{p}_1^m)](1 - A_t)]}{\mathbb{E}[\sum_{t=1}^{m} \mathbb{1}[\hat{p}_t < f_m(\hat{p}_1^m)]]} \quad (4.8)$$

$$= mFDR_1^m(\hat{\varepsilon}_o, \hat{p}) \quad (4.9)$$

□

4.3. Modified BH-procedure and mFDR control

As shown in Section 4.2, controlling the FDR value of the full time series is possible. The strategy then consists in first controlling the mFDR criterion of all successive subseries of length m along the full time series at level α. The main challenge addressed in the present section is to design a new multiple testing procedure that controls the local mFDR criterion at a prescribed level α.

In Section 4.3.1, it is proved that applying the classical BH-procedure on a time series of length m does not yield the control of mFDR at level α. However the proof of this result gives rise to a strategy for modifying the classical BH-procedure (Section 4.3.3) in a such a way that applying the so-called modified BH-procedure provides the desired mFDR control at level α.

4.3.1. mFDR control with the BH-procedure

Next Proposition 3 establishes the actual mFDR level achieved by the BH-procedure.

Proposition 3. *Let $(X_i)_1^m$ satisfy the requirements detailed by Definition 4. Let $(p_1, \ldots, p_m)$ be the true p-values corresponding to a subseries of length m that is, for any $1 \leq i \leq m$, $p_i = \mathbb{P}_{X \sim \mathcal{P}_0}(X \geq X_i) \in [0, 1]$. Then for every $\alpha \in [0, 1]$, applying BH_α on the p-values $(p_i)_{1 \leq i \leq m}$ leads to*

$$mFDR_1^m(p) = \frac{\alpha m_0}{m} \frac{\mathbb{E}R_{1,\alpha}^m(1)}{\mathbb{E}R_{1,\alpha}^m},$$

where $R^m_{1,\alpha}(1) = |BH_\alpha(0, p_2, .., p_m)|$ *and* $R^m_{1,\alpha} = |BH_\alpha(p_1, p_2, .., p_m)|$*, with* $|S|$ *denoting the cardinality of the set* S*.*

If the ratio $\mathbb{E}R(1)^m(1)/\mathbb{E}R(1)^m$ were known, it could be possible to control of the $mFDR$ criterion at level α by simply applying the BH-procedure with a preliminary level $\alpha' = \frac{m}{m_0}\frac{\mathbb{E}R_{1,\alpha}}{\mathbb{E}R_\alpha}\alpha$. Unfortunately at this stage, this ratio is not known and the latter strategy cannot be straightforwardly applied. Deriving such a modified BH-procedure is the purpose of the next sections. Let us also recall that in the anomaly detection context, m_0 is unknown but expected to be close to m since only a few anomalies are usually expected. Therefore the main challenge remains to compute $\mathbb{E}R(1)^m(1)/\mathbb{E}R(1)^m$.

Proof of Proposition 3. The mFDR formula is given by

$$mFDR^m_1(p) = \frac{\mathbb{E}[FP^m_{1,\alpha}(p)]}{\mathbb{E}[R^m_{1,\alpha}(p)]}.$$

Let us compute the numerator $\mathbb{E}[FP^m_1(p)]$ value after applying the BH_α. For simplification purposes, let $R = R^m_{1,\alpha}(p)$ (respectively $FP^m_1(p) = FP^m_{1,\alpha}(p)$) denote the number of rejections (resp. of false positives) resulting from BH_α. Introducing $D_i = \mathbb{1}[p_i \leq \frac{R}{m}\alpha]$, it appears that

$$FP^m_1(p) = \sum_{i=1}^{m}(1 - A_i)D_i.$$

Now focusing on D_i and using $R(i)$, the (random) number of rejections output by BH_α when the p_i is replaced by 0, it comes that

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

since, on the event $\{R = k\}$, p_i is already rejected that is, $p_i \leq \frac{k}{m}\alpha$. Taking the conditional expectation given p except p_i on both sides, it results

$$\mathbb{E}[D_i|p\backslash p_i] = \sum_{k=1}^{m}\mathbb{P}[p_i \leq \frac{k}{m}\alpha]\mathbb{1}[R(i) = k] = \frac{\alpha}{m}\sum_{k=1}^{m}k\mathbb{1}[R(i) = k]$$

since the true p-value p_i follows a uniform distribution on $[0, 1]$. Now integrating over all remaining p-values yields that

$$\mathbb{E}[D_i] = \frac{\alpha}{m}\sum_{k=1}^{m}k\mathbb{P}[R(i) = k] = \frac{\alpha}{m}\mathbb{E}[R(i)].$$

Since A_i and D_i are independent random variables, it results that

$$\begin{aligned}\mathbb{E}[FP_1^m(p)] &= \sum_{i=1}^{m} \mathbb{E}[(1-A_i)\mathbb{E}[D_i \mid A_i]] = \sum_{i=1}^{m} \mathbb{E}[(1-A_i)\mathbb{E}[D_i]] \\ &= \frac{\alpha}{m}\mathbb{E}[R(1)] \sum_{i=1}^{m} \mathbb{E}[1-A_i] \\ &= \alpha\frac{m_0}{m}\mathbb{E}[R(1)],\end{aligned}$$

where the last-but-one equality stems from the fact that all random variables $R(i)$ are identically distributed. □

4.3.2. Evaluating the ratio of rejection numbers

Previous Section 4.3.1 raises the importance of the ratio $\mathbb{E}R_{1,\alpha}^m(1)/\mathbb{E}R_{1,\alpha}^m$ of rejection numbers. The present section aims at deriving a numeric approximation to this ratio. In a first step, a first result details the value of the denominator. In a second step, an approximation to the numerator is derived based on a heuristic argument and also empirically justified on simulation experiments.

Calculating the expected number of rejections When $mFDR$ is assumed to equal α, the expected number of rejections can be made explicit.

Proposition 4. *With the previous notation, let $(X_1, \ldots, X_m)$ be given by Definition 4, where π denotes the unknown proportion of anomalies, and assume that $mFDR_1^m = \alpha$ and $FNR_1^m = \beta \in [0,1]$. Then*

$$\mathbb{E}[R_{1,\alpha}^m] = \frac{m\pi(1-\beta)}{1-\alpha}. \tag{4.10}$$

The proof is postponed to Appendix A.5. For instance, Eq. (4.10) establishes that the expected number of rejection output by BH_α increases with π, the unknown proportion of anomalies along the signal. This makes sens since the more anomalies, the more expected rejections. The expected number of rejection is also increasing with α: the larger α, the less restrictive the threshold, and the more rejections should be made. However the number of rejection decreases with the FNR value β. As β increases, the proportion of false negatives grows meaning that fewer alarms are raised, which results in a smaller number of rejections.

Heuristic arguments In what follows, the assumption is made that anomalies are easy to detect, meaning that the FNR_1^m value β is negligible compared to 1. In this context, Proposition 4 would yield that

$$\mathbb{E}[R_{1,\alpha}^m] \approx \frac{m\pi}{1-\alpha}. \tag{Power}$$

An another assumption is also made about the relationship between $\mathbb{E}[R^m_{1,\alpha}]$ and $\mathbb{E}[R^m_{1,\alpha}(i)]$. This assumption is based on a heuristic argument supported by the results of numerical experiments as reported in Table 3. In what follows, it is assumed that

$$\mathbb{E}[R^m_{1,\alpha}(i)] = \mathbb{E}[R^m_{1,\alpha}] + 1. \qquad \textbf{(Heuristic)}$$

No mathematical proof of this statement is given in the present paper. However, Table 3 displays numerical values which empirically support this approximation, whereas further analyzing the connection between these quantities should be necessary.

BH_α	0.05	0.1	0.2
$\mathbb{E}[R^m_{1,\alpha}]$	2.14	2.32	2.78
$\mathbb{E}[R^m_{1,\alpha}(i)]$	3.18	3.44	3.99

TABLE 3
Numerical evaluations for different values of α (10^3 repetitions)

Let us emphasize that Table 3 has been obtained with Gaussian data (generated similarly to those detailed in Section 3.4). For all the three considered values of α, one observes that $\mathbb{E}[R(1)]$ remains close to (but also slightly larger than) $\mathbb{E}[R] + 1$.

From now on and in all what follows, (**Heuristic**) is assumed to hold true. In particular, (**Heuristic**) gives rise to a strategy for computing the ratio $\frac{\mathbb{E}[R(1)]}{\mathbb{E}[R]}$. So all ingredient to build a procedure that control the mFDR are given.

4.3.3. Modified BH

From previous Sections 4.3.1 and 4.3.2, it is now possible to suggest and analyze the new modified BH-procedure (mBH in the sequel).

Definition 6 (Modified BH-procedure (mBH))**.** *Let m be an integer and $\alpha \in [0,1]$. Let us introduce the level $\alpha' = \alpha(1 + \frac{1-\alpha}{m\pi})^{-1}$, where $\pi \in [0,1]$ denotes the unknown proportion of anomalies (see Definition 4. Let us further assume* (**Power**) *and* (**Heuristic**) *hold true. Then the modified BH-procedure, denoted by mBH_α, is given for all true p-values $(p_1, \ldots, p_m) \in [0,1]^m$ by,*

$$mBH_\alpha(p_1, \ldots, p_m) = BH_{\alpha'}(p_1, \ldots, p_m).$$

The related mBH_α threshold at level α is defined as

$$\varepsilon_{mBH_\alpha} = \varepsilon_{BH_{\alpha'}},$$

when computed with true p-values, and $\hat{\varepsilon}_{mBH_\alpha} = \hat{\varepsilon}_{BH_{\alpha'}}$ when used with empirical p-values.

The above definition defines the mBH_α in terms of the BH-procedure by simply changing the level of control α'. This new level value depends on the

unknown proposition π of anomalies. Since in realistic anomaly detection scenarios observations are not labeled, [57] provides guidelines on how π could be estimated. From Proposition 4, it also appears that $\alpha' = \alpha(1 + \frac{1-\alpha}{m\pi(1-\beta)})^{-1}$ should arise as the ideal threshold. However since β is unknown, (**Power**) leads to the approximation suggested within the above definition.

Corollary 5 (Control of the FDR using mBH)**.** *Under the same notations and assumptions as Theorem 4. Let m and ℓ be two integers and assuming* (**Power**) *and* (**Heuristic**) *hold, let n and α' be defined by*

$$\alpha' = \frac{\alpha}{1 + \frac{1-\alpha}{m\pi}} \quad \text{and} \quad n = \ell m/\alpha' - 1.$$

the next two results hold true.

1. *If the FDR can be controlled at level α' on subseries of size m, using $BH_{\alpha'}$: $FDR_1^m(\hat{\varepsilon}_{BH_{\alpha'}}, \hat{p}) = (1-\pi)\alpha'$, then the FDR of the whole time series can be controlled at the level α using mBH_α*

$$FDR_1^\infty(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) = (1-\pi)\alpha \tag{4.11}$$

2. *If $FDR_1^m(\hat{\varepsilon}_{BH_{\alpha'}}, \hat{p}) \leq (1-\pi)\alpha'$. Then, the FDR of the whole time series can be controlled at the level α*

$$FDR_1^\infty(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) \leq (1-\pi)\alpha \tag{4.12}$$

Proof of Corallary 5. All conditions being satisfied Theorem 4 gives that:

$$FDR_1^\infty(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) = mFDR_1^m(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) \tag{4.13}$$

By definition of mBH_α in Definition 6 and with Proposition 3:

$$\begin{aligned} mFDR_1^m(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) &= mFDR_1^m(\hat{\varepsilon}_{BH_{\alpha'}}, \hat{p}) \\ &= \frac{\mathbb{E}[R(1)]}{\mathbb{E}[\mathbb{R}]} FDR_1^m(\hat{\varepsilon}_{BH_{\alpha'}}, \hat{p}) \end{aligned}$$

Under the assumptions **Heuristic** and **Power**, it gives:

$$\frac{\mathbb{E}[R(1)]}{\mathbb{E}[R]} = 1 + \frac{1-\alpha}{m\pi}$$
$$mFDR_1^m(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) = (1 + \frac{1-\alpha}{m\pi}) FDR_1^m(\hat{\varepsilon}_{BH_{\alpha'}}, \hat{p})$$

From hypothesis $FDR_1^m(\hat{\varepsilon}_{BH_{\alpha'}}, \hat{p}) = (1-\pi)\alpha'$. Replacing the value of α' with its expression it gives:

$$mFDR_1^m(\varepsilon_{mBH_\alpha}) = (1 + \frac{1-\alpha}{m\pi})(1-\pi)(1 + \frac{1-\alpha}{m\pi})^{-1}\alpha = (1-\pi)\alpha$$

Plugin this result into Eq. 4.13 this gives desired result.

$$FDR_1^{\infty}(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) = (1-\pi)\alpha$$

□

The next result shows how mBH_α can be applied to reach a global FDR control for the full time series at the desired level. It applies Corollary 5 by specifying different ways to estimate p-values on the time series.

Theorem 5 (Global FDR control using mBH_α)**.** *Let (X_t) be a mixture process introduced in Definition 4, with π denoting the anomaly proportion. Let $\alpha \in [0,1]$ be the desired FDR level for the full time series. For m and ℓ two integers and assuming* (**Power**) *and* (**Heuristic**) *hold, let n and α' be defined by*

$$\alpha' = \frac{\alpha}{1 + \frac{1-\alpha}{m\pi}} \quad \text{and} \quad n = \ell m/\alpha' - 1.$$

If $(\hat{p}_t)$ are given following one of the schemes

1. $\hat{p}_t = 1 - \mathbb{P}_{X\sim\mathcal{P}_0}(a(X) > a(X_t))$
2. $\hat{p}_t = \hat{p}\text{-}value(X_t, \{Z_{t,1}, \ldots, Z_{t,n}\})$ *with* $Z_{t,i} \sim \mathcal{P}_0$

then,

$$FDR_1^{\infty}(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) = (1-\pi)\alpha.$$

Otherwise, if $(\hat{p}_t)$ are given by

4. $\hat{p}_t = \hat{p}\text{-}value(X_t, \{(1-A_{t-n+1})X_{t-n+1} + A_{t-n+1}Z_{t,1}, \ldots, (1-A_t)X_t + A_tZ_{t,n}\})$ *with* $Z_{t,i} \sim \mathcal{P}_0$,

then,

$$FDR_1^{\infty}(\hat{\varepsilon}_{mBH_\alpha}, \hat{p}) \leq (1-\pi)\alpha.$$

The main merit of Theorem 5 is to establish the actual level of control for the global FDR of the full time series depending on the type of empirical p-value used in the anomaly detection process. This level of control depends on the unknown proportion π of anomalies. Therefore if π is close to 0 (only a few anomalies are expected), then the FDR control is close to α. However, since alpha' is close to 0 when pi is close to 0, an estimation of π (or expert knowledge) is required to detect anomalies. Some guidelines are provided in [57]. The last type of empirical p-values is (almost) the one which is used in practice in the present work. More precisely Section 5.2.3 describes empirical p-values based on a "Sliding Calibration Set".

Proof of Theorem 5. The Corollary 5 gives the two properties that the p-values families has to verify to control the FDR of the time series:

- The p-values are stationary and independent when time distance is larger than n.
- The FDR is controlled at level α' on subseries of size m: $FDR(\hat{\varepsilon}BH_1^{\alpha'}, \hat{p}) = (1-\pi)\alpha'$

In the following, these properties are verified for the different p-values.

1. The true p-value family is i.i.d., because the time series mixture is i.i.d. Also the Theorem 1 gives the control for subseries. Using Corollary 5 the FDR of the whole time series is controlled at level $(1-\pi)\alpha$.
2. This p-value family is i.i.d., because the time series mixture is i.i.d. and the calibration sets are also i.i.d. Using the Theorem 1, it gives the FDR control on subseries. With Corollary 5 the FDR of the whole time series is controlled at level $(1-\pi)\alpha$.
3. This p-value family is not i.i.d. However, because the calibration are build using a sliding window of size n, two p-values p_{t_1} and p_{t_2} are independent when $|t_1 - t_2| > m$. The calibration sets of the p-values overlapped, Then Corollary 4 ensures the upper bound of the FDR for subseries of size m. With Corollary 5 the FDR of the whole time series is upper bounded controlled at level $(1-\pi)\alpha$.

□

4.4. Empirical results

In this section, the abilities to get local control of the mFDR and the global control of the FDR, using mBH are assessed empirically. Corollary 5 and Theorem 5 give theoretical results about the control of the mFDR for subseries under the assumptions **Power** and **Heuristic**. However, these assumptions are hard to ensure in practice. In Section 4.4.1, the assessment is done on simulated data where the level of atypicity of the anomalies varies from one sample to another. Different scenarios are tested to verify if the mFDR control hold. Theorem 3 and Theorem 4 give FDR control over the full time series. In Section 4.4.2, the abilities of thresholds computed on disjoint and overlapping subseries to control the mFDR using are compared. Theorem 3 and Theorem 4 give asymptotic FDR control over the full time series. But there is no result about the speed of convergence, which is necessary when used on finite time series. In Section 4.4.3, the FDR for the full time series is calculated across different situations, as a function of time series size. It is possible to figure out when the entire series reaches control of the FDR.

4.4.1. Control of the mFDR on disjoint subseries

Experiment description From Corollary 5 the mBH_α controls the $mFDR_1^m$ only if the **Power** assumption is satisfied. Since the power of the anomaly detector depends on how easy it is to detect anomalies, the level of atypicity δ is

introduced. To quantifies the atypicity of a data point X_t, the true p-value is computed as $p_t = \mathbb{P}_{X\sim\mathcal{P}_0}(X > X_t)$, and the atypicity level is defined as the inverse of the p-value: $\delta_t = 1/p_t$. The atypicity level is preferred over the p-values because it is easier to show on the x-axis of the chart, when the p-value is small. To evaluate the effect of power, for each sample all anomalies have their level of atypicity lower bounded a given parameter δ. Therefore, it is possible to observe the effect of a variation in the level of atypicity on the $mFDR_1^m$, FDR_1^m and FNR_1^m.

For a given scenario—meaning a proportion of anomalies π, a level of atypicity δ, and a desired level of mFDR noted α—the actual mFDR, FDR, and FNR are estimated. These quantities are estimated using $K = 50$ samples of m data points. To control the estimation error made when estimating on a finite number of samples, each estimation is repeated $B = 100$ times. The estimation proceeds as follows:

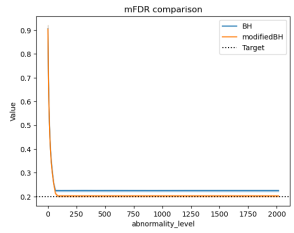
1. With $1 \leq b \leq B$, and $1 \leq k \leq K$, m data point are generated.
   - m_0 normal data $p_{b,k,1}, \ldots, p_{b,k,m_0}$ are generated according the reference law $U([0,1])$.
   - m_1 abnormal data $p_{b,k,m_0+1}, \ldots, p_{b,k,m}$ are generated using the alternative law $U([0,1/\delta])$, with δ the level of atypicity of the anomalies.
2. Then, for each sample, the thresholds are estimated with BH and mBH procedures:
   - $\hat{\varepsilon}_{b,k,BH} = BH_\alpha(p_{b,k,1}, \ldots, p_{b,k,m})$,
   - $\hat{\varepsilon}_{b,k,mBH} = mBH_\alpha(p_{b,k,1}, \ldots, p_{b,k,m})$.
3. The number of rejections, false positives and false negatives are computed on each sample and according each threshold. Using $M \in \{mBH, BH\}$:
   - $R_{b,k,M} = \sum_{i=1}^{m} \mathbb{1}[p_{b,k,i} \leq \hat{\varepsilon}_{b,k,M}]$,
   - $FP_{b,k,M} = \sum_{i=1}^{m_0} \mathbb{1}[p_{b,k,i} \leq \hat{\varepsilon}_{b,k,M}]$,
   - $FN_{b,k,M} = \sum_{i=m_0+1}^{m} \mathbb{1}[p_{b,k,i} > \hat{\varepsilon}_{b,k,M}]$.
4. The FDR, mFDR and FNR are estimated by averaging results over the K samples:
   - $FDR_{b,M} = \frac{1}{K}\sum_{k=1}^{K} \frac{FP_{b,k,m}}{R_{b,k,m}}$,
   - $mFDR_{b,M} = \frac{\sum_{k=1}^{K} FP_{b,k,m}}{\sum_{k=1}^{K} R_{b,k,m}}$,
   - $FNR_{b,M} = \frac{1}{K}\sum_{k=1}^{K} \frac{FN_{b,k,m}}{m_1}$.

These steps are then repeated over the different scenarios.

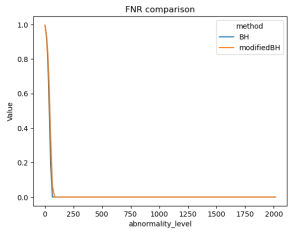
Results and Analysis The results are shown in Figure 8 by varying δ, α and m_1. In Figure 8, the level of atypicity δ in represented in the abscissa. The ordinate represents the estimated mFDR (in Figure 8a or 8c) or FNR (in Figure 8b or 8d). Different colors are used to distinguish between BH and mBH procedures.

For a low level of atypicity δ, the FNR and the mFDR are high because the anomalies are difficult to detect. By increasing δ, the FNR and the mFDR decrease. As shown in Figure 8b, with values of δ around 100, the FNR is equal to 0 which can also generate a constant mFDR as shown in Figure 8a. For the mBH-procedure, the mFDR is constant and equal to α. This is consistent with Theorem 4, which guarantees the control at level α when all anomalies are detected.

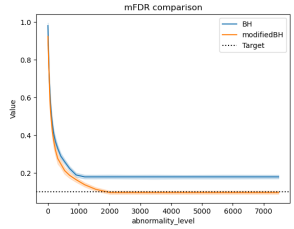
Figure 8d shows the totality of the anomalies detected for $\delta = 2000$. The same result in figure 8b with $\delta = 100$. This is explained by the different parameters of the experiment. The easier the anomalies are detected, faster the $FNR = 0$ is reached for a small δ and therefore the easier it is to guarantee $mFDR = \alpha$.



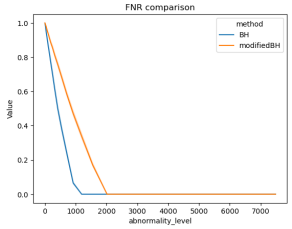
(a) mFDR, $\alpha = 0.2$, $\pi = 0.07$



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



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



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

Figure 8: mFDR and FNR as a function of level of atypicity across different scenarios

Conclusion In order to control the mFDR at the desired level α using mBH, the FNR has to be equal to 0. The capacity of mBH to control the mFDR depends of the difficulty of the problem. When abnormality proportion and level of atypicity are lower, the power of mBH decreases and the mFDR is harder to control. The results of this experiment gives an idea of the atypicity level that

the detector can find.

4.4.2. Disjoint subseries vs overlapping subseries

Experiment Description Theorems 3 and 4 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, 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 4.4.1 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 $1 \leq k \leq K = T/m$:

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

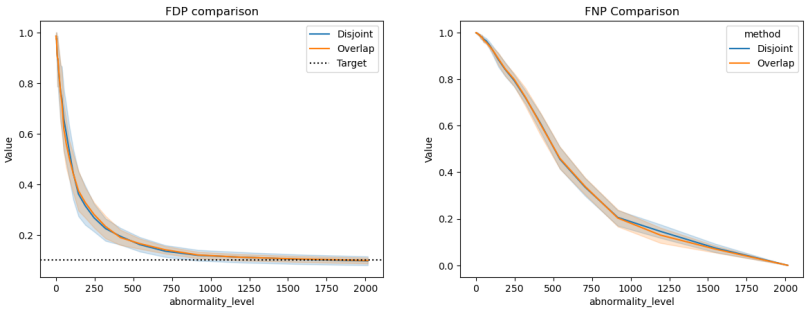
Notice the difference with disjoint windows case, all p-values of a subseries are compared to different thresholds and not to the same $\hat{\varepsilon}_{b,km}$.

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 9, 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 9: 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.

4.4.3. Convergence of false discovery rate control

This section studies the convergence rate of the FDR over the full time series using mBH_α.

Experiment Description The theoretical results obtained in Theorem 4 only guarantee an asymptotic control of the FDR on the whole time series. In practice, it is more useful to have a control of the FDR at any time, i.e. on subseries of finite size. The question is empirically studied by observing the speed of convergence of the false discovery rate towards the level α. The FDR of the full time series is calculated across different scenarios, as a function of time series size. In order to get the distribution of the FDR, the experiment is repeated on $B = 100$ time series. The maximal time series size explored is $T = 10^4$.

1. For $1 \leq b \leq B$ and for $1 \leq t \leq T$:
   - $A_{b,t} \sim Ber(\pi)$
   - If $A_{b,t} = 0$, $p_{b,t} \sim U([0,1])$
   - Otherwise: $p_{b,t} \sim U([0, 1/\delta])$
2. The thresholds are estimated with mBH_α:

$$\hat{\varepsilon}_{b,t,\alpha} = \hat{\varepsilon}_{mBH_\alpha}(p_{b,t-m+1}, \ldots, p_{b,t})$$

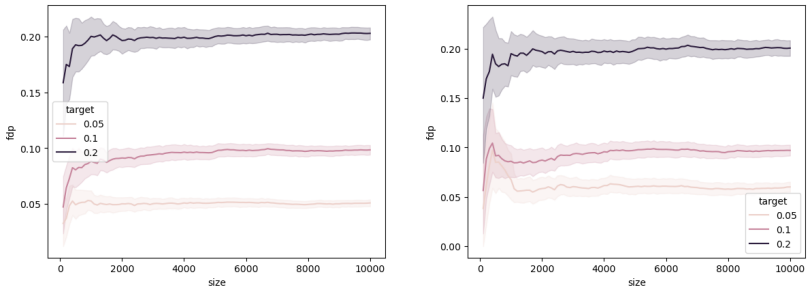
3. The proportion of false discovery (FDP) on the partial time series are calculated:

$$FDP_{b,t,\alpha} = \frac{\sum_{u=1}^{t}(1 - A_{b,t})\mathbb{1}[p_{b,t} \leq \hat{\varepsilon}_{b,t,\alpha}]}{\sum_{u=1}^{t}\mathbb{1}[p_{b,t} \leq \hat{\varepsilon}_{b,t,\alpha}]}$$

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

Results and analysis In Figure 10, the false discovery proportion is represented in the ordinate according to the size of the time series given in the abscissa. The different levels of α used to compute mBH threshold are experimented with the results of the median FDP and its 95% band is shown in different colors. Different scenarios are represented by varying the proportion of anomalies between the sub figures.

It can be observed that the convergence is quite fast from a size of 2000 data points, since for a α of 0.05, is has 95% chance to have a false positive rate between 0.04 and 0.06, on Figure 10. Thus, the control of the false positive rate, can be ensured with a high probability, for a series of one data point per minute recorded over a few days, .



(a) $\delta = 1000$, $\pi = 0.02$ (b) $\delta = 1000$, $\pi = 0.01$

Figure 10: FDR over the full time series as a function of the time series size.

Conclusion This ensures that the control at level is reached not only for infinite time series but also for finite time series which allows our model to be used in practice.

5. Empirical simulation against competitor

The control of the FDR with p-values estimated empirically has been studied at Section 3. Theorem 2 ensure the control of the FDR_1^m when the p-values are estimated on calibration set having particular cardinality value. Theorems 3 and 4 ensure the control of the FDR of the full time series throw control of the $mFDR_1^m$ of the subseries. Corollary 5 enables to deduce that the mBH_α procedure can be apply to control the FDR of the full time series under the **Heuristic** and **Power** assumptions. Even though these assumptions are hard to ensure theoretically, the experiment at Section 4.4.1 shows that the $mFDR_1^m$ is controlled for tested scenario, provided that anomalies are sufficiently atypical. Experiment from Section 4.4.3 shows that the control of the FDR is possible even the time series is not infinite as required by Theorem 4.

These different results provide the conditions for building an anomaly detector that controls the FDR of the time series through control of the mFDR on the subseries and the p-empirical value. Our anomaly detector is evaluated under different scenarios by varying the generated anomalies and the targeted FDR. To understand the source of the difficulties that the anomaly detector may encounter, different sequences of p-values with oracle information are introduced. Our anomaly detector is compared against Levels based On Recent Discovery (LORD) which is a online multiple testing procedure, introduced in [30] to control the FDR.

5.1. Data

The synthetic data are generated from Gaussian distribution. With the use of the empirical p-value estimator there are no need to evaluate on other data distribution. Only anomaly proportion and the distribution shift associated to anomalies impact the performances of the anomaly detector. Data are generated accordingly with Definition 4 with Gaussian reference distribution and anomaly spike like in Section 3.4.1. The strength of the distribution shift noted by Δ takes value in $\{3\sigma, 3.5\sigma, 4\sigma\}$ and the abnormality proportion noted π is equal to 0.01. Each generated time series contain $T = 10^4$ data points. Each experiment is repeated over 100 time series.

For t in $[\![1, T]\!]$:

- $A_t \sim B(\pi)$
- $X_t = \begin{cases} \mathcal{N}(0, \sigma^2) & \text{if } A_t = 0 \\ \Delta\sigma & \text{else} \end{cases}$

The value of Δ represents the atypicity score of the anomalies. Anomalies with higher Δ are easier to detect. In this experiment, the standard deviation σ is set to 1.

5.2. Threshold and p-value estimators description

5.2.1. Our proposal mBH on overlapping subseries

Using the p-value with the empirical estimator, the anomalies are detected by using mBH as the threshold estimator on overlapping subseries in the Algorithm 1. For each time t, the threshold is computed as: $\hat{\varepsilon}_{mBH_\alpha,t} = f_m(\hat{p}_{t-m}, \ldots, \hat{p}_t)$, where f_m is the mBH-procedure. To ensure FDR control according to Theorem 1, the cardinality of the calibration set to be equal to $n = \frac{m}{\alpha} - 1$. In this experiment m is equal to 100 and α takes values 0.1 and 0.2 depending the tested scenario. So the calibration set takes values 999 or 1999.

5.2.2. LORD

LORD introduced in [30] is based on alpha-investing rules to define a threshold on p-values. For each time t the threshold is computed from according to the alpha-investing rules, depending on previous decision made by the algorithm. For more precision refer to the original article [30]. The empirical p-value specified in Definition 2.4 does not respect this property while the conformal p-value, defined in Equation 2.3.1 respects this property. Using conformal p-values to apply LORD algorithm leads to a weak power detecting anomalies. The issue is that $\tilde{p} \geq \frac{1}{n+1}$ is always verified and the threshold sequence $\hat{\varepsilon}_t$ decreases quickly when no rejection are made. No anomaly can be detected. For these reasons, the

empirical p-value introduced in Equation 2.4 is used while applying LORD and mBH. In this experiment LORD3 from [30] is used with the same parameters as in the original paper.

5.2.3. p-value estimation

Different sequences of p-values are used to understand the limitations of our anomaly detector. The true p-values are used to evaluate the case where the only limitation comes from the multiple testing procedure. One can thus understand how the estimation of the p-values affects the detection of anomalies. One way to estimate p-values in practice is to use the same calibration set for all p-values. This is referred as the fixed calibration set. However, the p-values may be biased in that particular calibration set. In practice, the usual way to implement the estimated p-values is to use a sliding calibration set. To evaluate the p-value of a data point X_t, n preceding data points are used as a calibration set. To a bias in the estimation, the points detected as abnormal cannot be part of the calibration set. However, the calibration set can be biased by undetected anomalies. To evaluate this impact, the sliding calibration set-$\star$ is introduced, where the knowledge oracle of the labels is used to construct the calibration set from the previous data points.

The different p-value sequences are computed as follows:

- **Oracle**: The true p-value is used instead of the estimated one.

$$\forall t \in [\![1, T]\!], \hat{p}_t = \Phi(X_t)$$

- **Fixed calibration set (Fixed Cal.)**: The p-value is estimated using the same calibration set $\{Z_i, i \in [1, n]\}$ for all observations.

$$\forall t \in [\![1, T]\!], \hat{p}_t = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[Z_i > X_t]$$

- **Sliding Calibration set-$\star$ (Sliding Cal.-$\star$)**: The p-value is estimated using a calibration that is a sliding windows containing the n previous true normal data.

$$\forall t \in [\![1, T]\!], \hat{p}_t = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[X_{h(t,i)} > X_t]$$

  With h the function that select observation that respect $\mathcal{H}_0$. For each t and i, $h(t, i)$ gives the i-th observation lower than t and that respect $\mathcal{H}_0$ hypothesis.

- **Sliding Calibration set (Sliding Cal.)**: The calibration set is a sliding windows containing the n previous estimated normal data.

$$\forall t \in [\![1, T]\!], \hat{p}_t = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[X_{\hat{h}(t,i)} > X_t]$$

With $\hat{h}$ the function that estimates the function h. For each t and i, $\hat{h}(t,i)$ give the i-th observation lower than t and $d_{\hat{h}(t,i)} = 0$.

5.3. Performance metrics

The anomaly detector are evaluated using their ability to control the FDR and minimize the FNR of the full time series. Therefore, the two applied metrics are the FDP and the FNP computed as:

$$FDP = \frac{\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < \hat{\varepsilon}_t](1 - A_t)}{\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < \hat{\varepsilon}_t]}$$

and

$$FNP = \frac{\sum_{t=1}^{T} \mathbb{1}[\hat{p}_t < \hat{\varepsilon}_t](1 - A_t)}{\sum_{t=1}^{T} A_t}$$

where $\hat{\varepsilon}_t$ is estimated using mBH or LORD and $\hat{p}_t$ is estimated using one of the estimator defined in Section 5.2.3.

5.4. Results

The box plots shown in Figures 15-18 represent the FDP and FNP distribution for 1000 repetitions. Inside each sub figure (a, b, c, d, e,..), the box plot distributions are displayed according to:

1. the multiple testing method mBH or LORD,
2. the p-value estimation model set to Oracle PV, Fixed Cal., Sliding Cal.-$\star$ or Sliding Cal.
3. and the distribution shift between the normal data and anomalies, noted Δ, varying from 4σ to 3σ.

Table 4 gives a summary using FDR and FNR estimations. It enables easily the comparison between these values coming from the different strategies combining:

1. the multiple testing method mBH or LORD,
2. the choice of the level α varying from 0.1 to 0.2,
3. the p-value estimation model set to Oracle PV, Fixed Cal., Sliding Cal.-$\star$ or Sliding Cal.
4. and the distribution shift between the normal data and anomalies, varying from 4σ to 3σ.

FDR, $\alpha = 0.1$	$\Delta = 4\sigma$	$\Delta = 3.5\sigma$	$\Delta = 3\sigma$
mBH with Oracle PV	0.101	0.113	0.281
mBH with Fixed Cal.	0.100	0.109	0.348
mBH with Sliding Cal.-$\star$	0.100	0.113	0.256
mBH with Sliding Cal.	0.335	0.222	0.346
LORD with Oracle PV	0.106	0.115	0.367
LORD with Fixed Cal.	0.111	0.277	0.736
LORD with Sliding Cal.-$\star$	0.070	0.190	0.841
LORD with Sliding Cal.	0.075	0.098	0.627

(a) FDR, $\alpha = 0.1$

FNR, $\alpha = 0.1$	$\Delta = 4\sigma$	$\Delta = 3.5\sigma$	$\Delta = 3\sigma$
mBH with Oracle PV	0.020	0.151	0.793
mBH with Fixed Cal.	0.026	0.135	0.669
mBH with Sliding Cal.-$\star$	0.019	0.140	0.669
mBH with Sliding Cal.	0.040	0.217	0.694
LORD with Oracle PV	0.033	0.260	0.905
LORD with Fixed Cal.	0.070	0.340	0.896
LORD with Sliding Cal.-$\star$	0.781	0.845	0.978
LORD with Sliding Cal.	0.052	0.327	0.907

(b) FNR, $\alpha = 0.1$

FDR, $\alpha = 0.2$	$\Delta = 4\sigma$	$\Delta = 3.5\sigma$	$\Delta = 3\sigma$
mBH with Oracle PV	0.200	0.208	0.277
mBH with Fixed Cal.	0.206	0.211	0.301
mBH with Sliding Cal.-$\star$	0.210	0.219	0.283
mBH with Sliding Cal.	0.833	0.815	0.761
LORD with Oracle PV	0.211	0.216	0.290
LORD with Fixed Cal.	0.210	0.263	0.665
LORD with Sliding Cal.-$\star$	0.061	0.149	0.625
LORD with Sliding Cal.	0.117	0.133	0.321

(c) FDR, $\alpha = 0.2$

FNR, $\alpha = 0.2$	$\Delta = 4\sigma$	$\Delta = 3.5\sigma$	$\Delta = 3\sigma$
mBH with Oracle PV	0.009	0.062	0.395
mBH with Fixed Cal.	0.014	0.045	0.355
mBH with Sliding Cal.-$\star$	0.008	0.059	0.339
mBH with Sliding Cal.	0.003	0.018	0.101
LORD with Oracle PV	0.016	0.117	0.610
LORD with Fixed Cal.	0.04	0.144	0.689
LORD with Sliding Cal.-$\star$	0.805	0.835	0.941
LORD with Sliding Cal.	0.026	0.168	0.692

(d) FNR, $\alpha = 0.2$

TABLE 4

Comparison of mBH versus LORD for online anomaly detection in Gaussian white noise with different abnormality levels.

5.5. Analysis

5.5.1. Effect of the strength of the distribution shift Δ

According the assumption **Power** from Theorem 5, mBH_α enables control of the FDR at level α if all anomalies are detected.
To test this assertion, the different columns of the Table 4a, are compared. In the row "mBH with Oracle PV", with $\Delta = 4\sigma$ the FDR is estimated at 0.101 which is close to the desired level $\alpha = 0.1$. While, when $\Delta = 3\sigma$ the FDR level is estimated at 0.281 which is almost three times the desired level α. The FNR results in Table 4b needs to be taken into consideration. When $\Delta = 4\sigma$, the FNR is close to 0, while when $\Delta = 3\sigma$ the FNR is equal to 0.793. Similar results are obtained with other test configurations in Table 4c and Table 4d. The FDR control at the desired level need the FNR to be close to 0.

5.5.2. Effect of p-value estimation

To understand how the p-value estimation can prevent the control of the FDR, the first four rows in Table 4a are compared. In the column "4σ", the FDR values for the configurations "Oracle PV", "Fixed Cal." and "Sliding Cal.-$\star$" are very close to the desired level $\alpha = 0.1$. This control is enabled by Theorem 5, since the p-values verify all hypotheses, in particular all data in the calibration sets are generated according to the reference distribution. However, in the case of "Sliding Cal.", the FDR increases at a value of 0.335. For the same configurations, the FDR remains low, between 0.019 and 0.040 as shown at Table 4b. The increase of FDR when using "Sliding Cal." instead of "Sliding Cal.-$\star$" is a consequence of calibration set contamination. Indeed, according to the procedure used to build the calibration sets, described in Section 5.2.3, all detected anomalies are removed from calibration sets used in the estimation of next p-values. When an observation is wrongly detected as an anomaly, this data point cannot be part of the calibration set at future steps of the online detection. Instead, it is replaced by an other data point having statistically a lower atypicity score. Indeed false positives have high atypicity score to be (wrongly) detected as anomalies. As a result, the calibration set contains data points with lower scores than if it had been generated under $\mathcal{P}_0$. It leads to underestimate the p-values and to increase the number of false positives. This illustrates the major drawback of mBH: it is highly sensitive to the non robustness of the p-value estimator. Figure 15a shows that using fixed calibration instead of sliding calibration-$\star$ gives a larger variance on the FDP while the FDR is the same. Using a single calibration set for the entire time series means that the FDP is highly dependent on the start of the time series. By modifying the calibration set at each time step, the statistical fluctuations in the FDP are smoothed over the course of the time series analysis.

5.5.3. Comparison with LORD

In this section, the results found using mBH and the ones using LORD are compared. As known from the literature, LORD controls the FDR of super-uniform p-values. In this experiment, the question is in the capacity of LORD method to control the FDR of empirical p-values that have no theoretical guaranties. It can be noticed in Figure 4a that LORD is able to ensure the control of the FDR for all calibration set definitions when anomalies are easier to detect as for $\Delta = 4\sigma$ or $\Delta = 3.5\sigma$. In particular, unlike mBH, LORD is able to control the FDR in the case of the sliding calibration set. However, mBH method has a lower FNR compared to the LORD method, as shown in 4a and 4b. For example, Table 4b shows that the FNR is equal to 0.019 with mBH while it is equal to 0.781 with LORD, in the case using Sliding Calibration set-$\star$ on data having $\Delta = 3\sigma$. Nevertheless, with the Sliding Calibration set case, the LORD method has quite the same FNR but with lower FDR (0.335 against 0.075). The contamination issue of mBH offsets the superior performance observed in the Sliding Calibration set-$\star$.

6. Conclusion

In this article, an online anomaly detector that aims to have a better control of the FDR at a given level α has been proposed. The research has been developed to tackle two issues:

- the empirical p-values: it ensures conditions on the calibration cardinality to ensure FDR control when using Benjamini-Hochberg.
- and the online detection: it ensures a global control of the FDR through local control of the mFDR of subseries, using a modified version of the BH-procedure.

The results of our research is the assessment of our proposal from the theoretical point of view and from empirical experiments. Our method has been compared with a method from the state of the art. It shows the strong capability for ensuring control of the FDR even in the case of empirical p-values. The major drawback and improvement path of our method is it relies on non-robust p-value estimation.

7. Acknowledgments

The authors would like to thank Cristian Preda, director of the MODAL team at Inria, for valuable discussions.

References

[1] Ahmed, M., Mahmood, A. N. and Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications* **60** 19–31.

[2] Ahmed, T. (2009). Online anomaly detection using KDE. In *GLOBECOM 2009-2009 IEEE Global Telecommunications Conference* 1–8. IEEE.

[3] Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

[4] Bates, S., Candès, E., Lei, L., Romano, Y. and Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics* **51** 149–178.

[5] 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.

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

[7] Blázquez-García, A., Conde, A., Mori, U. and Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)* **54** 1–33.

[8] Blum, J. M. and Tremper, K. K. (2010). Alarms in the intensive care unit: too much of a good thing is dangerous: is it time to add some intelligence to alarms? *Critical care medicine* **38** 702–703.

[9] Bos, H. and Huang, K. (2006). Towards software-based signature detection for intrusion prevention on the network card. In *Recent Advances in Intrusion Detection: 8th International Symposium, RAID 2005, Seattle, WA, USA, September 7-9, 2005. Revised Papers 8* 102–123. Springer.

[10] Braei, M. and Wagner, S. (2020). Anomaly detection in univariate time-series: A survey on the state-of-the-art. *arXiv preprint arXiv:2004.00433*.

[11] Buda, T. S., Caglayan, B. and Assem, H. (2018). Deepad: A generic framework based on deep learning for time series anomaly detection. In *Pacific-Asia conference on knowledge discovery and data mining* 577–588. Springer.

[12] Burnaev, E. and Ishimtsev, V. (2016). Conformalized density-and distance-based anomaly detection in time-series data. *arXiv preprint arXiv:1608.04585*.

[13] Celisse, A. and Robin, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference* **140** 3132–3147.

[14] Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)* **41** 1–58.

[15] Cvach, M. (2012). Monitor alarm fatigue: an integrative review. *Biomedical instrumentation & technology* **46** 268–277.

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

[17] Foorthuis, R. (2021). On the nature and types of anomalies: a review of deviations in data. *International journal of data science and analytics* **12** 297–331.

[18] 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.

[19] Gandy, A. and Hahn, G. (2014). MMCTest—a safe algorithm for implementing multiple Monte Carlo tests. *Scandinavian Journal of Statistics* **41** 1083–1101.

[20] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research* **13** 723–773.

[21] Guedj, M., Robin, S., Celisse, A. and Nuel, G. (2009). Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC bioinformatics* **10** 1–12.

[22] Guo, W. and Peddada, S. (2008). Adaptive choice of the number of bootstrap samples in large scale multiple testing. *Statistical applications in genetics and molecular biology* **7**.

[23] Han, X., Zhou, Y., Chen, K., Qiu, H., Qiu, M., Liu, Y. and Zhang, T. (2022). ADS-lead: Lifelong anomaly detection in autonomous driving systems. *IEEE Transactions on Intelligent Transportation Systems* **24** 1039–1051.

[24] Hawkins, D. M. (1980). *Identification of outliers* **11**. Springer.

[25] Hindy, H., Brosset, D., Bayne, E., Seeam, A., Tachtatzis, C., Atkinson, R. and Bellekens, X. (2018). A taxonomy and survey of intrusion detection system design techniques, network threats and datasets.

[26] Hönes, C. J., Miller, B. K., Heras, A. M. and Foing, B. H. (2021). Automatically detecting anomalous exoplanet transits. *arXiv preprint arXiv:2111.08679*.

[27] Hu, W., Gao, J., Li, B., Wu, O., Du, J. and Maybank, S. (2018). Anomaly detection using local kernel density estimation and context-based regression. *IEEE Transactions on Knowledge and Data Engineering* **32** 218–233.

[28] Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution* 492–518. Springer.

[29] Hundman, K., Constantinou, V., Laporte, C., Colwell, I. and Soderstrom, T. (2018). Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* 387–395.

[30] Javanmard, A. and Montanari, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics* **46** 526–554.

[31] Khraisat, A., Gondal, I., Vamplew, P. and Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity* **2** 1–22.

[32] Kundu, A., Sahu, A., Serpedin, E. and Davis, K. (2020). A3D: Attention-based auto-encoder anomaly detector for false data injection attacks. *Electric Power Systems Research* **189** 106795.

[33] Laxhammar, R. (2014). Conformal anomaly detection: Detecting ab-

normal trajectories in surveillance applications, PhD thesis, University of Skövde.

[34] Laxhammar, R., Falkman, G. and Sviestins, E. (2009). Anomaly detection in sea traffic-a comparison of the gaussian mixture model and the kernel density estimator. In *2009 12th international conference on information fusion* 756–763. IEEE.

[35] Lewandowska, K., Mędrzycka-Dąbrowska, W., Tomaszek, L. and Wujtewicz, M. (2023). Determining Factors of Alarm Fatigue among Nurses in Intensive Care Units—A Polish Pilot Study. *Journal of Clinical Medicine* **12** 3120.

[36] Li, D., Chen, D., Jin, B., Shi, L., Goh, J. and Ng, S.-K. (2019). MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks* 703–716. Springer.

[37] Li, H. and Boulanger, P. (2020). A survey of heart anomaly detection using ambulatory electrocardiogram (ECG). *Sensors* **20** 1461.

[38] Li, Y., Fang, B., Guo, L. and Chen, Y. (2007). Network anomaly detection based on TCM-KNN algorithm. In *Proceedings of the 2nd ACM symposium on Information, computer and communications security* 13–19.

[39] Li, Y., Zhang, X., He, S., Chen, Z., Kang, Y., Liu, J., Li, L., Dang, Y., Gao, F., Xu, Z. et al. (2022). An Intelligent Framework for Timely, Accurate, and Comprehensive Cloud Incident Detection. *ACM SIGOPS Operating Systems Review* **56** 1–7.

[40] Malhotra, P., Vig, L., Shroff, G., Agarwal, P. et al. (2015). Long Short Term Memory Networks for Anomaly Detection in Time Series. In *Esann* **2015** 89.

[41] Marandon, A., Lei, L., Mary, D. and Roquain, E. (2022). Machine learning meets false discovery rate. *arXiv preprint arXiv:2208.06685*.

[42] Mary, D. and Roquain, E. (2022). Semi-supervised multiple testing. *Electronic Journal of Statistics* **16** 4926–4981.

[43] Miele, E. S., Bonacina, F. and Corsini, A. (2022). Deep anomaly detection in horizontal axis wind turbines using graph convolutional autoencoders for multivariate time series. *Energy and AI* **8** 100145.

[44] Munir, M., Siddiqui, S. A., Dengel, A. and Ahmed, S. (2018). DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *Ieee Access* **7** 1991–2005.

[45] Notaro, P., Cardoso, J. and Gerndt, M. (2021). A survey of aiops methods for failure management. *ACM Transactions on Intelligent Systems and Technology (TIST)* **12** 1–45.

[46] Pang, G., Li, J., van den Hengel, A., Cao, L. and Dietterich, T. G. (2022). ANDEA: anomaly and novelty detection, explanation, and accommodation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 4892–4893.

[47] 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**.

[48] Pincombe, B. (2005). Anomaly detection in time series of graphs using arma processes. *Asor Bulletin* **24** 2.

[49] 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.

[50] Roberts, S. (2000). Control chart tests based on geometric moving averages. *Technometrics* **42** 97–101.

[51] Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G. and Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE* **109** 756–795.

[52] Safin, A. M. and Burnaev, E. (2017). Conformal kernel expected similarity for anomaly detection in time-series data. *Advances in Systems Science and Applications* **17** 22–33.

[53] Smith, J., Nouretdinov, I., Craddock, R., Offer, C. and Gammerman, A. (2014). Anomaly detection of trajectories with kernel density estimation by conformal prediction. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10* 271–280. Springer.

[54] Solet, J. M. and Barach, P. R. (2012). Managing alarm fatigue in cardiac care. *Progress in Pediatric Cardiology* **33** 85–90.

[55] Son, S., Gil, M.-S. and Moon, Y.-S. (2017). Anomaly detection for big log data using a Hadoop ecosystem. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)* 377–380. IEEE.

[56] Staerman, G. (2022). Functional anomaly detection and robust estimation, PhD thesis, Institut polytechnique de Paris.

[57] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **66** 187–205.

[58] Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* **102** 901–912.

[59] Vahdat, A. and Kautz, J. (2020). NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems* **33** 19667–19679.

[60] Wang, W., Liu, Z., Shi, X. and Pierce, L. (2019). Online fdr controlled anomaly detection for streaming time series. In *5th Workshop on Mining and Learning from Time Series (MiLeTS)*.

[61] Weinstein, A., Barber, R. and Candes, E. (2017). A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*.

[62] Wulsin, D., Blanco, J., Mani, R. and Litt, B. (2010). Semi-supervised anomaly detection for EEG waveforms using deep belief nets. In *2010 Ninth international conference on machine learning and applications* 436–441. IEEE.

[63] Xu, Z. and Ramdas, A. (2022). Dynamic algorithms for online multiple testing. In *Mathematical and Scientific Machine Learning* 955–986. PMLR.
[64] Zhang, M., Zou, J. and Tse, D. (2019). Adaptive monte carlo multiple testing via multi-armed bandits. In *International Conference on Machine Learning* 7512–7522. PMLR.
[65] Zhou, Z.-G. and Tang, P. (2016). Continuous anomaly detection in satellite image time series based on z-scores of season-trend model residuals. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* 3410–3413. IEEE.

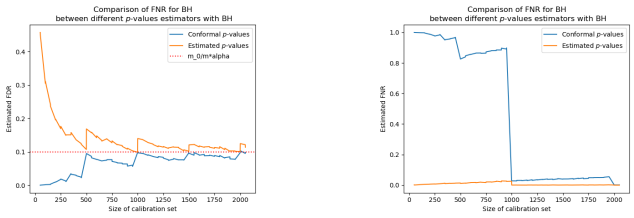
Appendix A: Proofs

A.1. Comparison of p-values estimators

The control of the FDR is not achievable using classical multiple testing [6, 49] since the p-value estimator, shown in Definition 2, is not super-uniform. Conformal p-value estimator $\tilde{p}$, shown in Equation 2.3.1, verifies the super-uniform property. 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 11 displays the comparison between estimated p-values and conformal p-values using the BH-procedure. As shown in Figure 11a, 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 11b, 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 11: Comparison between p-value estimators using Benjamini-Hochberg

A.2. PRDS property for p-values having overlapping calibration set

The following construction is used to describe a family of p-values with overlapping calibration set. Let Z the vector that combine all calibration set, the Z_i are i.i.d. with marginal probability $\mathcal{P}_0$. The set of the n indices defining the elements of the calibration set related to $\hat{p}_i$ in Z is noted $\mathcal{D}_i$. The calibration related to X_1 is noted $Z_{\mathcal{D}_1} = (Z_{i_1}, \ldots, Z_{i_n})$. For all i in $[\![1, m]\!]$: $\hat{p}_i = p\text{-value}(X_i, Z_{\mathcal{D}_i})$.

To proof that p-values with overlapping calibration set are PRDS as described in Definition 1, the methodology used in [4] to be extended in the case of overlapping calibration set. For i in $[\![1, m]\!]$ the calibration set associated to X_i is noted $Z_{\mathcal{D}_i}$. The law of total probabilities gives:

$$\begin{aligned}\mathbb{P}\left[\hat{p}_1^m \in A | \hat{p}_i = u\right] &= \int \mathbb{P}\left[\hat{p}_1^m \in A | \hat{p}_i = u | Z_{\mathcal{D}_i} = z\right] \mathbb{P}\left[Z_{\mathcal{D}_i} = z|\right] dz \\ &= \mathbb{E}_{Z_{\mathcal{D}_i}|\hat{p}_i=u} \mathbb{P}\left[\hat{p}_1^m \in A | \hat{p}_i = u | Z_{\mathcal{D}_i} = z\right]\end{aligned}$$

If these two lemma are suppose to be true, the PRDS property is verified.

Lemma 5.1. *For non-decreasing set A and vectors z, z' such that $z \succeq z'$, then*

$$\mathbb{P}\left[\hat{p}_1^m \in A | Z_{\mathcal{D}_i} = z\right] \geq \mathbb{P}\left[\hat{p}_1^m \in A | Z_{\mathcal{D}_i} = z'\right] \tag{A.1}$$

Lemma 5.2. *For $u \geq u'$, if i belongs to the set of inliers, the exists $Z_{\mathcal{D}_i,1} \sim Z_{\mathcal{D}_i}|\hat{p}_i = u$ and $Z_{\mathcal{D}_i,2} \sim Z_{\mathcal{D}_i}|\hat{p}_i = u'$ such that $\mathbb{P}[Z_{\mathcal{D}_i,1}] \succeq \mathbb{P}[Z_{\mathcal{D}_i,2}]$*

Indeed, take $i \in [\![1, m]\!]$ and $u \geq u'$ and define $Z_{\mathcal{D}_i,1}$ and $Z_{\mathcal{D}_i,2}$ as in the statement of Lemma 5.2.

$$\begin{aligned}\mathbb{P}[\hat{p}_1^m \in A | p_i = u] &= \mathbb{E}_{Z_{\mathcal{D}_i,1}}[\mathbb{P}[\hat{p}_1^m \in A | Z_{\mathcal{D}_i} = Z_{\mathcal{D}_i,1}]] \text{ (Lemma 5.2)} \\ &\geq \mathbb{E}_{Z_{\mathcal{D}_i,2}}[\mathbb{P}[\hat{p}_1^m \in A | Z_{\mathcal{D}_i} = Z_{\mathcal{D}_i,2}]] \text{ (Lemma 5.1)} \\ &\geq \mathbb{P}[\hat{p}_1^m \in A | p_i = u'] \text{ (Lemma 5.2)}\end{aligned}$$

It shows that, when $u \geq u'$ then $\mathbb{P}[\hat{p}_1^m \in A | p_i = u] \geq \mathbb{P}[\hat{p}_1^m \in A | p_i = u']$, which means $\mathbb{P}[\hat{p}_1^m \in A | p_i = u]$ is increasing in u. The PRDS property is satisfied. To complete the proof, the introduced lemmas are proven.

Proof of Lemma 5.1. Let be i in $[\![1, m]\!]$ and vectors z, z' and $\overline{z}$ vectors such that $z \succeq z'$. The vectors z, z' are used to define the calibration set related to the p-values $\hat{p}_i$ and $\overline{z}$ is used to define elements of calibrations sets that are not in the calibration set of $\hat{p}_i$. By conditioning on the calibration sets defined by $(z, \overline{z})$ and $(z', \overline{z})$ it gives:

$$\mathbb{P}\left[\hat{p}_1^m \in A | Z_{\mathcal{D}_i} = z, Z_{\overline{\mathcal{D}_i}} = \overline{z}\right] \geq \mathbb{P}\left[\hat{p}_1^m \in A | Z_{\mathcal{D}_i} = z', Z_{\overline{\mathcal{D}_i}} = \overline{z}\right] \tag{A.2}$$

This result comes from the decomposition the following decomposition, for all j in $[\![1, m]\!]$

$$\begin{aligned}\hat{p}_j &= \frac{1}{n}\sum_{k\in\mathcal{D}_j} \mathbb{1}[a(Z_k) \geq a(X_j)] \\ &= \frac{1}{n}\left(\sum_{k\in\mathcal{D}_j\cap\mathcal{D}_i} \mathbb{1}[a(Z_k) \geq a(X_j)] + \sum_{k\in\mathcal{D}_j\setminus\mathcal{D}_i} \mathbb{1}[a(Z_k) \geq a(X_j)]\right)\end{aligned}$$

The conclusion comes from $Z_{\mathcal{D}_i} \succeq Z'_{\mathcal{D}_i}$ which implies $Z_{\mathcal{D}_i\cap\mathcal{D}_j} \succeq Z'_{\mathcal{D}_i\cap\mathcal{D}_j}$.

Since $Z_{\mathcal{D}_j\setminus\mathcal{D}_i} \perp Z_{\mathcal{D}_i}$, Eq. A.2 can be integrated over $Z_{\overline{\mathcal{D}_i}}$ to give:

$$\mathbb{P}\left[\hat{p}_1^m \in A | Z_{\mathcal{D}_i} = z\right] \geq \mathbb{P}\left[\hat{p}_1^m \in A | Z_{\mathcal{D}_i} = z'\right] \tag{A.3}$$

□

Proof of Lemma 5.2. Let $S'_{i,(1)} \leq S_{i,(2)} \leq \ldots \leq S_{i,(n)}$ the order statistics of $(a(Z_{\mathcal{D}_i,1}), \ldots, a(Z_{\mathcal{D}_i,n}))$. Let $S'_{i,(1)} \leq S'_{i,(2)} \leq \ldots \leq S'_{i,(n+1)}$ the order statistics of $(a(Z_{\mathcal{D}_i,1}), \ldots, a(Z_{\mathcal{D}_i,n}), a(X_i))$. And R_i the rank of $a(X_i)$ among these.

$$\left\{(S_{(1)}, \ldots, S_{(n)}) | R_i = k, S'_{i,(1)}, \ldots, S'_{i,(n+1)}\right\} = (S'_{(1)}, \ldots, S'_{(k-1)}, S'_{(k+1)}, \ldots, S'_{(n+1)}) \tag{A.4}$$

Using that R_i is independent of $S'_{i,(1)}, \ldots, S'_{i,(n+1)}$:

$$\left\{(S_{(1)}, \ldots, S_{(n)}) | R_i = k\right\} = (S'_{(1)}, \ldots, S'_{(k-1)}, S'_{(k+1)}, \ldots, S'_{(n+1)}) \tag{A.5}$$

The right-hand side is not increasing with k and $\hat{p}_i = \frac{R_i - 1}{n}$ □

A.3. Proof of Theorem 1

Proof of Theorem 1. 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

$$\begin{aligned}FDP = \frac{FP}{R} &= \sum_{i=1}^{m_0} \frac{\mathbb{1}[p_i \leq \frac{\alpha R}{m}]}{R} \qquad (\text{since } D_i = \mathbb{1}[p_i \leq \tfrac{\alpha R}{m}]) \\ &= \sum_{k=1}^{m}\sum_{i=1}^{m_0} \frac{\mathbb{1}[p_i \leq \frac{\alpha k}{m}]\mathbb{1}[R = k]}{k}.\end{aligned} \tag{A.6}$$

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, $R(i) = BH_\alpha(p_1, .., p_{i-1}, 0, p_{i+1}, \dots, p_m)$. 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. □

A.4. Proof of Corollary 1

Proof of Corollary 1. To get a deeper understanding of the FDR expression obtained in Theorem 2, $q_{n,k}$ the fractional part of $\frac{\alpha k n}{m}$ is introduced:

$$q_{n,k} = \frac{\alpha k n}{m} - \left\lfloor \frac{\alpha k n}{m} \right\rfloor$$

When plugged into the FDR expression, it gives:

$$\begin{aligned} FDR &= m_0 \sum_{k=1}^{m} \frac{\frac{\frac{\alpha k n}{m} + 1 - q_{n,k}}{n+1}}{k} \mathbb{P}(R(1) = k) \\ FDR &= \frac{m_0\alpha}{m}\frac{n}{n+1} + \frac{m_0}{n+1}\sum_{k=1}^{m} \frac{1 - q_{n,k}}{k}\mathbb{P}(R(1) = k) \end{aligned} \tag{A.7}$$

In order to get lower and upper bounds of the FDR, the value of $q_{n,k}$ should be expressed as a function of α, k, n and m.

For the next part of the proof, it is useful to express the relation between $q_{n,k}$ and $q_{n+1,k}$. It gives the effect of increasing the cardinality of the calibration by one. Using the definition of the fractional part:

$$q_{n+1,k} - q_{n,k} = \frac{\alpha k(n+1)}{m} - \left\lfloor \frac{\alpha k(n+1)}{m} \right\rfloor - \frac{\alpha kn}{m} + \left\lfloor \frac{\alpha kn}{m} \right\rfloor$$
$$q_{n+1,k} - q_{n,k} = \frac{\alpha k}{m} - \left\lfloor \frac{\alpha k(n+1)}{m} \right\rfloor + \left\lfloor \frac{\alpha kn}{m} \right\rfloor$$

Which can be expressed as a congruence relation:

$$q_{n+1,k} - q_{n,k} \equiv \frac{\alpha k}{m} \pmod 1 \tag{A.8}$$

Two cases are studied:

1. Particular case: there exists an integer $1 \leq \ell$ such that $\frac{\ell m}{\alpha}$ is an integer. the notation $n_\ell = \frac{\ell m}{\alpha}$ is introduced. Since: $\frac{\alpha k n_\ell}{m} = \frac{\alpha k \ell m/\alpha}{m} = k\ell$ is an integer, then the fractional part is null:

   $$q_{n_\ell,k} = 0$$

   If the calibration set cardinality n is equal to $n = n_\ell - 1 = \frac{\ell m}{\alpha} - 1$. Then, the congruence relation in Eq. A.8 gives:

   $$q_{n_\ell - 1,k} \equiv q_{n_l,k} - \alpha k/m \pmod 1$$
   $$q_{n_\ell - 1,k} \equiv 0 - \alpha k/m \pmod 1$$

   Using the fact that fractional part of a number belongs to $[0, 1[$, the only possible value to $q_{n_\ell - 1,k}$ is:

   $$q_{n_\ell - 1,k} = 1 - \alpha k/m$$

   Plugging the value of $q_{n_\ell - 1,k}$ into Eq. A.7, it gives:

   $$FDR = \frac{m_0 \alpha n}{m(n+1)} + \frac{m_0}{n+1} \sum_{k=1}^{m} \frac{\alpha k}{km} \mathbb{P}(R(1) = k)$$

   Simplifying by k and using that $\sum_{k=1}^{m} \mathbb{P}(R(i) = k) = 1$, the result is obtained:

   $$FDR = \frac{m_0 \alpha n}{m(n+1)} + \frac{m_0 \alpha}{(n+1)m}$$
   $$FDR = \frac{m_0 \alpha}{m}$$

2. General case: With $\alpha \in]0,1]$, for each ℓ the notation $n_\ell = \lceil \frac{\ell m}{\alpha} \rceil$ is introduced. Notice that this definition is consistent with the particular case. The ceiling function definition gives:

$$\left\lceil \frac{\ell m}{\alpha} \right\rceil - 1 < \frac{\ell m}{\alpha} \leq \left\lceil \frac{\ell m}{\alpha} \right\rceil$$

Multiplying by αk on each side and the n_ℓ notation:

$$\frac{\alpha k(n_\ell - 1)}{m} < k\ell \leq \frac{\alpha k(n_\ell)}{m}$$

It implies that $\lfloor \frac{\alpha k(n_\ell - 1)}{m} \rfloor < \lfloor \frac{\alpha k(n_\ell)}{m} \rfloor$. Also, Eq. A.8 is expressed as $q_{n_\ell,k} - q_{n_\ell - 1,k} \equiv \frac{\alpha k}{m} \pmod 1$:

$$1 - \frac{\alpha k}{m} \leq q_{n_\ell - 1,k} < 1 \tag{A.9}$$

Indeed, the fractional part of a number as to be larger than $1 - \alpha k/m$ so that adding $\alpha k/m$ increase the integer part.

By plugin the bounds of $q_{n_\ell - 1,k}$ into Eq. A.7, it can gives the bounds of the FDR. At first, to compute the upper bound of the FDR the lower bound of $q_{n_\ell - 1,k}$ is used:

$$FDR \leq \frac{m_0(n_\ell - 1)\alpha}{mn_\ell} + \frac{m_0}{n+1} \sum_{k=1}^{m} \frac{\alpha k}{km} \mathbb{P}(R(1) = k)$$

With the same calculations as for the "Particular case", it gives:

$$FDR \leq \frac{m_0 \alpha}{m}$$

Similarly, the lower bound of the FDR can be obtained using the $q_{n,k}$ upper bound from Eq. A.9 plugged into Eq. A.7:

$$\frac{m_0(n_\ell - 1)\alpha}{mn_\ell} + \frac{m_0}{n+1} \sum_{k=1}^{m} \frac{(1-1)}{k} \mathbb{P}(R_1 = k) < FDR$$

$$\frac{m_0(n_\ell - 1)\alpha}{mn_\ell} < FDR$$

□

A.5. Proof of Proposition 4

Proof of Proposition 4. By definition $mFDR_1^m = \frac{\mathbb{E}[FP_1^m]}{\mathbb{E}R_1^m}$, and $R_1^m = FP_1^m + TP_1^m$. With hypothesis the $mFDR$ is equal to α, this gives:

$$\begin{aligned}
\alpha &= \frac{\mathbb{E}[FP_1^m]}{\mathbb{E}R_1^m} \\
\alpha &= \frac{\mathbb{E}[FP_1^m]}{\mathbb{E}[FP_1^m + TP_1^m]} \\
\alpha(\mathbb{E}[FP_1^m] + \mathbb{E}[TP_1^m]) &= \mathbb{E}[FP_1^m] \\
(\alpha - 1)\mathbb{E}[FP_1^m] &= -\alpha\mathbb{E}[TP_1^m] \\
\mathbb{E}[FP_1^m] &= \frac{\alpha}{1-\alpha}\mathbb{E}[TP_1^m]
\end{aligned}$$

Then, the expectation of true positives is expressed using the proportion of false negatives β, the proportion of anomaly π in the m observations, A_i the random variable equal to 1 if the observation X_i is an anomaly and D_i the random variable equal to 1 if the observation X_i is detected as anomaly :

$$\begin{aligned}
\mathbb{E}[TP_1^m] &= \sum_{i=1}^{m} \mathbb{P}[A_i = 1 \text{ and } D_i = 1] \\
&= \sum_{i=1}^{m} \mathbb{P}[A_i = 1]\mathbb{P}[D_i = 1 | A_i = 1] \\
&= m\pi(1-\beta)
\end{aligned}$$

Therefore, the $\mathbb{E}[FP_1^m]$ can be expressed as:

$$\mathbb{E}[FP_1^m] = \frac{\alpha m\pi(1-\beta)}{1-\alpha}$$

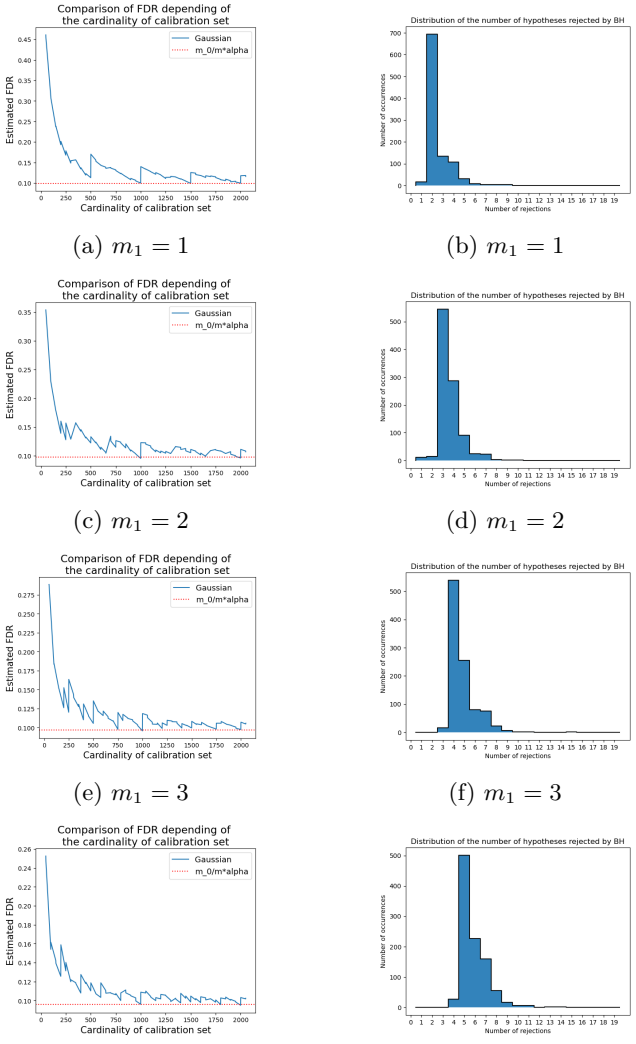
So the $\mathbb{E}[R_1^m]$ is expressed as follows:

$$\begin{aligned}
\mathbb{E}[R_1^m] &= \frac{\alpha m\pi(1-\beta)}{1-\alpha} + m\pi(1-\beta) \\
&= \frac{m\pi(1-\beta)}{1-\alpha}
\end{aligned}$$

□

Appendix B: Figures

B.1. Effect of the number detections by BH on the intermediate drops for the FDR control in Section 3.4



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

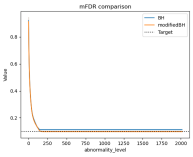
(c) $m_1 = 2$ (d) $m_1 = 2$

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

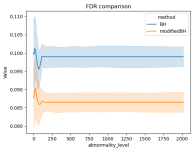
(g) $m_1 = 4$ (h) $m_1 = 4$

Figure 12: Effect of the number detections by BH on the intermediate drops for the FDR control

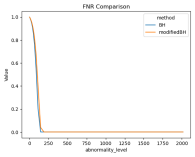
B.2. Figures related to experiment of Section 4.4.1



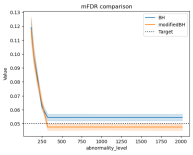
(a) mFDR, $\alpha = 0.1$, $\pi = 0.07$



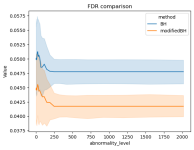
(b) FDR, $\alpha = 0.1$, $\pi = 0.07$



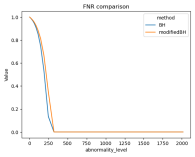
(c) FNR, $\alpha = 0.1$, $\pi = 0.07$



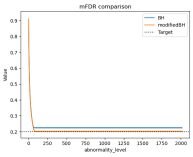
(d) mFDR, $\alpha = 0.05$, $\pi = 0.07$



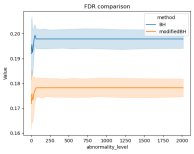
(e) FDR, $\alpha = 0.05$, $\pi = 0.07$



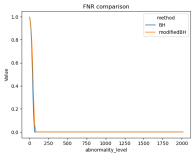
(f) FNR, $\alpha = 0.05$, $\pi = 0.07$



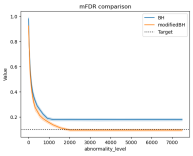
(g) mFDR, $\alpha = 0.2$, $\pi = 0.07$



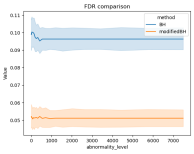
(h) FDR, $\alpha = 0.2$, $\pi = 0.07$



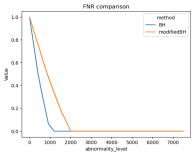
(i) FNR, $\alpha = 0.2$, $\pi = 0.07$



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



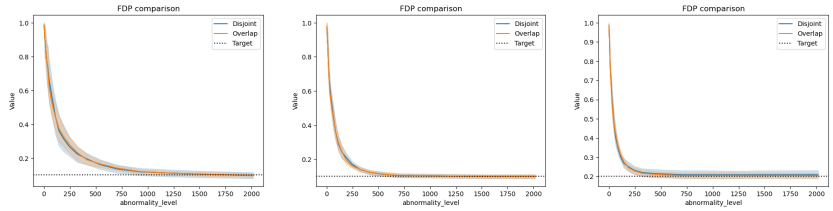
(k) FDR, $\alpha = 0.1$, $\pi = 0.01$



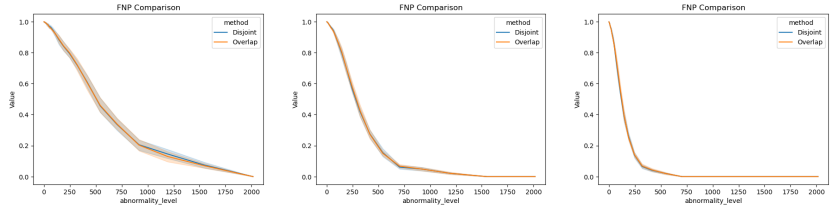
(l) FNR, $\alpha = 0.1$, $\pi = 0.01$

Figure 13: Effect of the atypicity level on the mFDR, FDR and FNR, according to different multiple testing procedures.

B.3. Figures related to experiment of Section 4.4.1



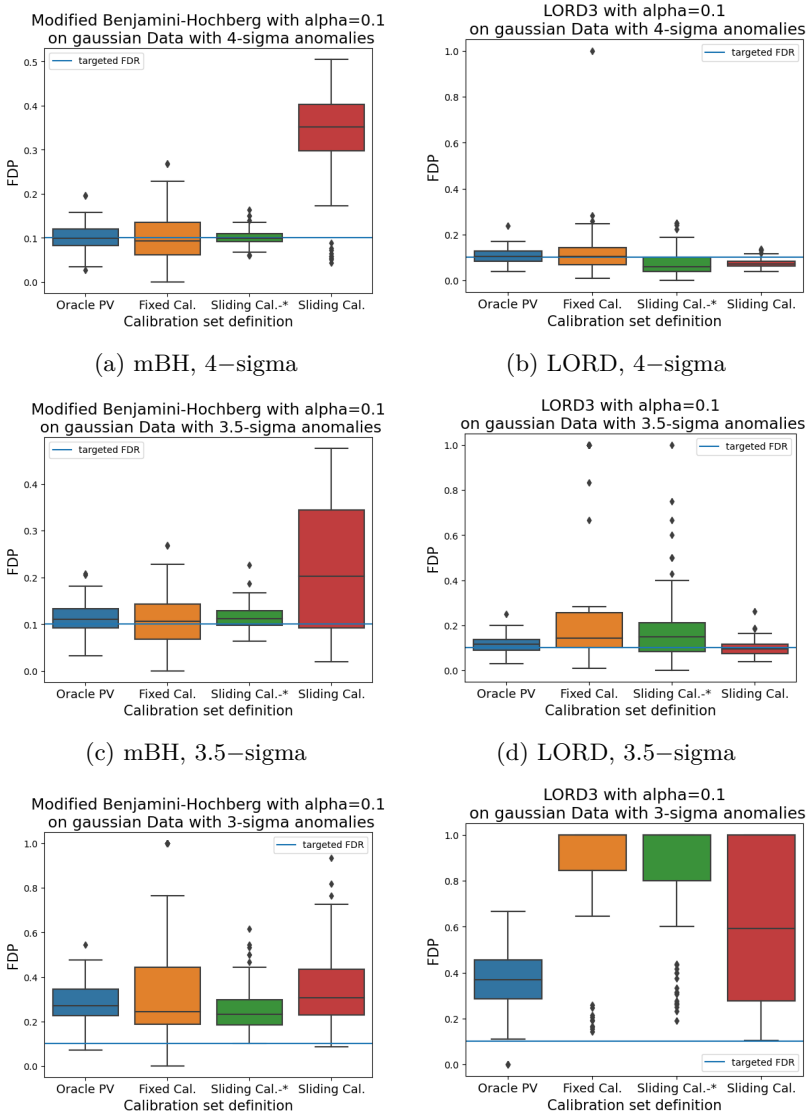
(a) mFDR, $\alpha = 0.1$, $\pi = 0.01$ (b) mFDR, $\alpha = 0.1$, $\pi = 0.02$ (c) mFDR, $\alpha = 0.2$, $\pi = 0.02$



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

B.4. Figures related to the experiment of Section 5.4



(a) mBH, 4−sigma

(b) LORD, 4−sigma

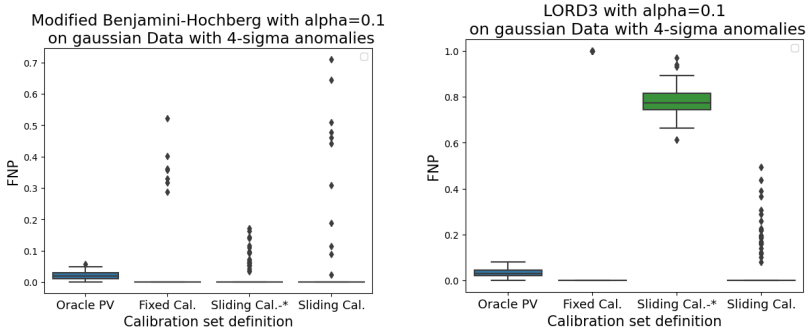
(c) mBH, 3.5−sigma

(d) LORD, 3.5−sigma

(e) mBH, 3−sigma

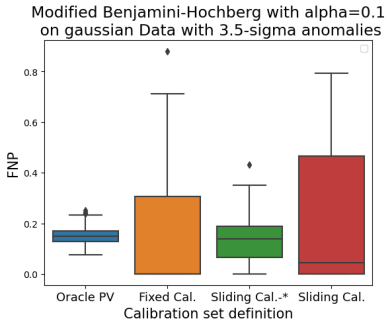
(f) LORD, 3−sigma

Figure 15: Comparison of the FDPs acquired from using different multiple testing procedures, mBH or LORD, and from the way the p-values are calculated, in the case $\alpha = 0.1$.

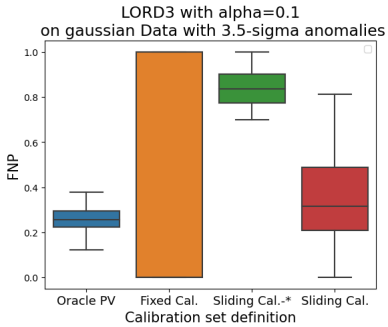


(a) mBH, 4−sigma

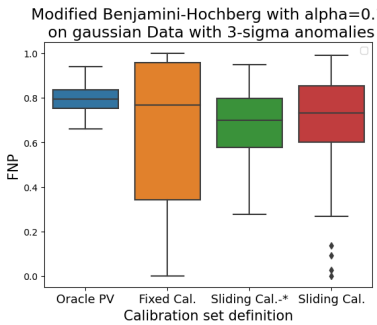
(b) LORD, 4−sigma



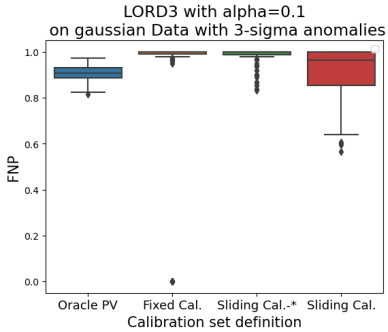
(c) mBH, 3.5−sigma



(d) LORD, 3.5−sigma

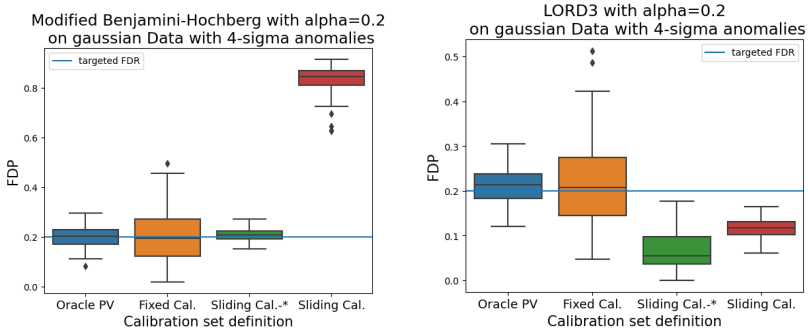


(e) mBH, 3−sigma



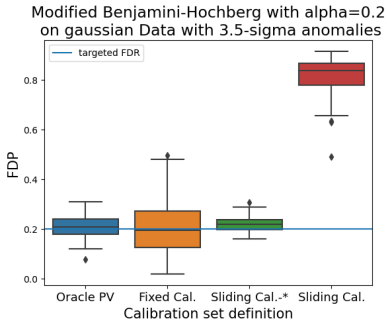
(f) LORD, 3−sigma

Figure 16: Comparison of the FNPs acquired from using different multiple testing procedures, mBH or LORD, and from the way the p-values are calculated, in the case $\alpha = 0.1$.

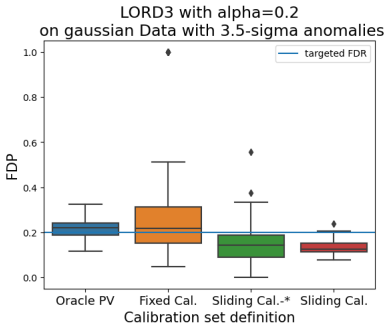


(a) mBH, 4−sigma

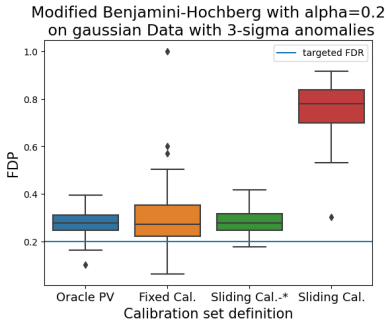
(b) LORD, 4−sigma



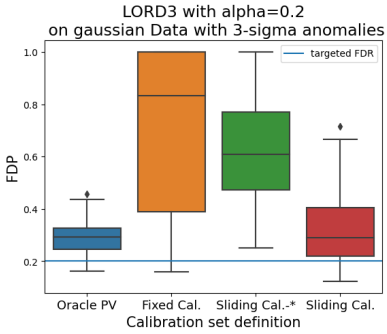
(c) mBH, 3.5−sigma



(d) LORD, 3.5−sigma

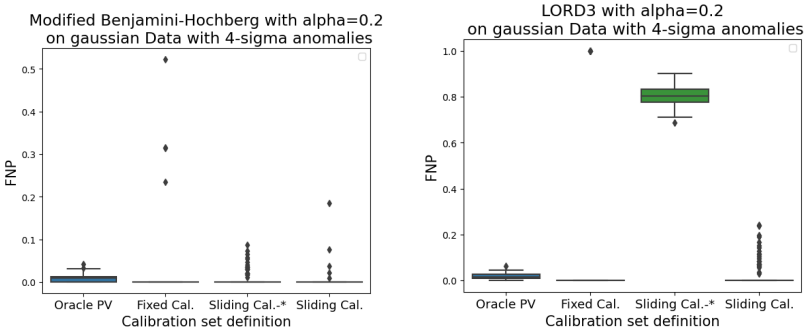


(e) mBH, 3−sigma



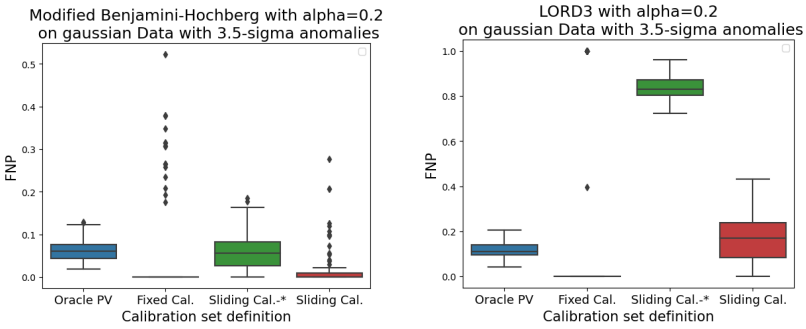
(f) LORD, 3−sigma

Figure 17: Comparison of the FDPs acquired from using different multiple testing procedures, mBH or LORD, and from the way the p-values are calculated, in the case $\alpha = 0.2$



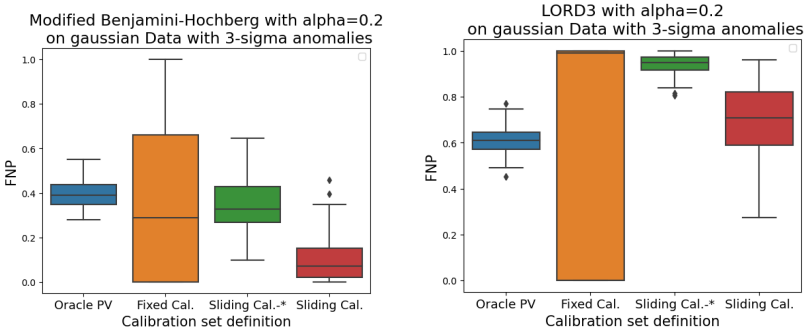
(a) mBH, 4−sigma

(b) LORD, 4−sigma



(c) mBH, 3.5−sigma

(d) LORD, 3.5−sigma



(e) mBH, 3−sigma

(f) LORD, 3−sigma

Figure 18: Comparison of the FNPs acquired from using different multiple testing procedures, mBH or LORD, and from the way the p-values are calculated, in the case $\alpha = 0.2$.