Supplementary Material for: Breakpoint based online anomaly detection

Etienne Krönert[1], Dalila Hattab[1] and Alain Celisse[2]

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

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

1. Experiment about segment assignment change probability

In Appendix B.1 of the main article, two algorithms for estimating the probability of a segment assignment change were described: the exact estimation using Algorithm 3 and an efficient estimation using Algorithm 5. In this section, these different methods are evaluated by experiments on simulated data.

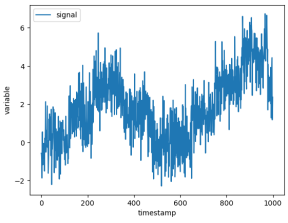
The following notations are used: Let T be the length of the time series, θ the average segment length, Δ the size jumps to generate a breakpoint and σ the standard deviation of the data point within a segment.

Time series are generated according to the following rules:

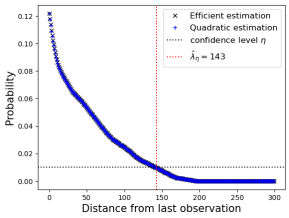
- The number of breakpoints follows the exponential distribution $D \sim Exp(T/\theta)$.
- Each breakpoint position is generated according to uniform distribution $\forall i \in [1, D], \tau_i \sim U(1, T)$
- The mean of the time series μ_i is piecewise constant with respect to the segmentation τ_i, with $\mu_{\tau_i} - \mu_{\tau_i+1} = \xi\Delta\sigma$
- The time series is generated according to the following rule $X_t \sim \mathcal{N}(\mu_t, \sigma)$

Then $\hat{f}_\tau(\lambda)$ is estimated using the two different methods: Algorithm 3 and Algorithm 5.

An example of generated time series is illustrated in Figure 1a. Figure 1b gives the estimated probability of segment assignment change according to the two estimation Algorithms 3 and 5. The two algorithms give results that are almost the same, as shown in Figure 1b. The selected $\lambda^\star_\eta$ is equal to 143, in the two cases. This supports assumption that (**Last**) is verified. In practice, we recommend to use the Algorithm 5 since it is more computationally efficient. To compute the probability $\hat{f}_\tau(\lambda)$ on a PC (4 CPU, 16G), the Algorithm 5 gives results within 30 seconds compared to the exact computation which gives the results within 15mn, for a time series of length 10^4.



(a) Generated time series



(b) Probability of assignment change as a function of distance to time series end.

Fig 1: Results for estimation of the "segment assignment change" probability

2. Experiment about the status change

The training procedure in Appendix B.2 of the main article is applied for different scoring functions adapted to different types of time series considered in Appendix B.4. The goal is to check if the estimation approach of $\hat{f}_d(\ell)$ can be applied to different scoring functions.

2.1. Description of the experiment

Different series that require different scoring functions are considered: Gaussian and Mixture of Gaussian.

- Figure 2a shows a Gaussian white noise with anomalies in distribution tail.

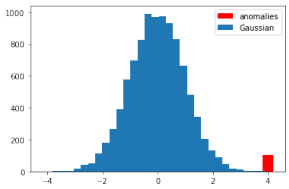
$$\begin{aligned} \forall t \in [\![1, T]\!], \qquad & A_t \sim Ber(\pi), \\ & \text{if } A_t = 0, X_t \sim \mathcal{N}(0, 1) \\ & \text{else } X_t = \Delta \end{aligned}$$

  The z-score applied on X_t to detect anomalies that are in the tail of the distribution, is computed by,

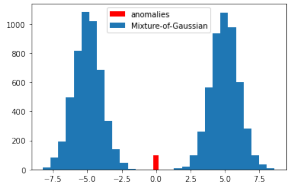
$$\overline{a}(X_t, S) = |X_t - \hat{\mu}_S|/\hat{\sigma}_S \tag{1}$$

  where S is a segment of data, $\hat{\mu}_S$ the mean estimator on S and , $\hat{\mu}_S$ the standard deviation on S
- Figure 2b shows a Mixture of Gaussians with anomalies between the distribution modes.

$$\begin{aligned} \forall t \in [\![1, T]\!], \qquad & A_t \sim Ber(\pi), \\ & \text{if } A_t = 0, X_t \sim 0.5\mathcal{N}(\Delta, 1) + 0.5\mathcal{N}(-\Delta, 1) \\ & \text{else } X_t = 0 \end{aligned}$$



(a) Gaussian with anomalies in tail



(b) Gaussian mixture with anomalies in center

Fig 2: Different time series distributions and anomalies.

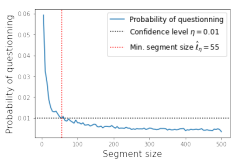
The kernel based score, inspired from other works on kernel based anomaly detection [1, 2], applied to detect anomalies having large distance from the normal data, is computed by,

$$\overline{a}(X_t, S) = \frac{1}{|S|^2} \sum_{s,s' \in S^2} K(s, s') - \frac{2}{|S|} \sum_{s \in S} K(X_t, s) + K(X_t, X_t) \qquad (2)$$

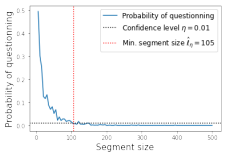
2.2. Results and analysis

As stated previously, two types of time series are considered in the experiments: results of Gaussian data shown in Figure 3 and results of Gaussian mixture data shown in Figure 4. For both, three line charts representing the probability of status change as a function of the current segment length in relation to the initial status: (a) the status is normal, (b) the status is abnormal and (c) unknown status.

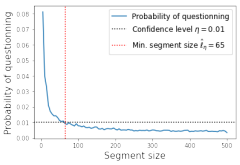
For Gaussian data and in the unknown status, Figure 3c shows clearly that the probability of status change decreases with the length of the current segment. This probability is higher when the status is abnormal, as shown in Figure 3b. Nevertheless, with a segment length of 100, the probability is less than 1%. For Gaussian mixture data and in the abnormal status scenario shown in Figure 4b, the length of the current segment needs to be at least equal to 500 to get a



(a) normal status

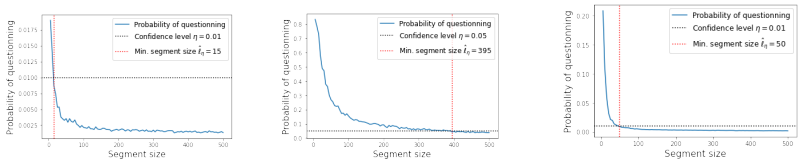


(b) abnormal status



(c) unknown status

Fig 3: Probability that status changes under stable breakpoints as a function of segment length, for Gaussian data.



(a) normal status (b) abnormal status (c) unknown status

Fig 4: Probability that status change under stable breakpoint as a function of segment length, for Gaussian mixture data.

probability of changing status around 5%. For the normal status scenario in Figure 4a, the probability of changing quickly decreases to 0. The results are also promising in the unknown status scenario in Figure 4c, where the change probability is low.

2.3. Conclusion

Empirical results show that the choice of an optimal $\hat{\ell}_\eta$ which reduces the uncertainty of a data point status depends on the type of data and the scoring function that is used. The method can help to select an atypicity score. A good atypicity score, satisfying requirements discussed in Appendix B.4 of the main article (been robust and efficient) should have low $\hat{\ell}_\eta$ value.

3. Experiments about the atypicity score

In Appendix B.4 of the main article, it has been affirmed that to build a good score function, the estimators used must verify the robustness and efficiency properties. To assess the robustness and the efficiency of the atypicity score, synthetic data are used for experimentation and analysis. The robustness of an estimator is its ability to be unbiased in the presence of anomalies. An estimator is said efficient when it is close to the parameter value with a limited number of data points. In this analysis, three categories of estimators are tested: one "efficient and not robust", a second "not efficient and robust" and a third "robust and efficient". These three estimators are analyzed considering the absence or presence of anomalies. The assessment is based on the parameter estimation error and on the anomaly detection performances using FDR and FNR.

3.1. Description

In this experiment, the focus is on the z-score. The atypicity of a data point x is calculated from the mean μ and standard deviation σ as follows $a_z(x, \mu, \sigma) = (x - \mu)/\sigma$. In an anomaly detection context, the mean and standard deviation are unknown and need to be estimated. There are many estimators of the mean

and standard deviation. These estimators have different properties in terms of robustness and efficiency. In order to study the relationship between these properties and the performance of the anomaly detector, three estimators are chosen for each of these two values.
For the mean value the three estimators are defined as the following:

- Maximum Likelihood Estimator: $\mu_{mle} = \frac{1}{n}\sum_i x_i$. This estimator is efficient but not robust against anomaly contamination.
- Median: $\mu_r = \text{median}(x_1, ..., x_n)$. This estimator is robust but less efficient than the MLE estimator.
- Biweight location, introduced in [3]. This estimator is robust and efficient.

$$\mu_{bw} = \frac{\sum_{i=1}^{\ell}(1-u_i^2)x_i\mathbb{1}[|u_i|<1]}{\sum_{i=1}^{\ell}(1-u_i^2)}$$
$$u_i = \frac{x_i - \overline{x}}{9MAD}$$

  Where $\overline{x}$ is median of the x_i and MAD is the median absolute deviation.

For the standard deviation, the three estimators are defined as the following:

- Maximum Likelihood Estimator: $\sigma_{mle} = \sqrt{\frac{1}{n}\sum_i (x_i - \mu)^2}$. This estimator is efficient but not robust against anomaly contamination.
- Median: $\sigma_{mad} = \text{median}(||x_i - \mu|)$. This estimator is robust but less efficient than the MLE estimator.
- Biweight Midvariance estimator: introduced in [4]. This estimator is robust and efficient.

$$\sigma_{bw}^2 = \frac{\ell\sum_{i=1}^{\ell}(x_i-\overline{x})^2(1-u_i^2)^4\mathbb{1}[|u_i|<1]}{(\sum_{i=1}^{\ell}(1-u_i^2)(1-5u_i^2)\mathbb{1}[|u_i|<1])^2}$$
$$u_i = \frac{x_i - \overline{x}}{9MAD}$$

  Where $\overline{x}$ is the median of the x_i and MAD is the median absolute deviation.

All the six estimators are evaluated according to two measures:

1. First, the precision and the robustness of the estimator is evaluated using the Mean Squared Error (MSE), applying the following procedure: Let θ be either the mean or the standard deviation parameter, and $\hat{\theta}$ be an one estimator of the parameter θ. Let ℓ be the cardinality of the segment used to estimate θ. Let B be the number of repetitions for the experiments.
   (a) Generate the segment data: For b in $[1, B]$ and for i in $[1, \ell]$, $X_{b,i} \sim \mathcal{N}(0, 1)$, if the segment contains only normal data. For b in $[1, B]$ and for i in $[1, \ell_0]$, $X_{b,i} \sim \mathcal{N}(0, 1)$ and for i in $[\ell_1, \ell]$, $X_{b,i} = 4$, if the segment is contaminated by anomalies.

(b) Estimate the parameter using the estimator: For b in $[1, B]$, $\hat{\theta}_b = \hat{\theta}(X_{b,1}, ..., X_{b,\ell})$.

(c) Compute the MSE, $MSE = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \theta)^2$

Different values of the segment length ℓ are tested, from 10 to 1000. For each value of ℓ, two values of ℓ_1 are tested. One with $\ell_1 = 0$, for the case where there are no anomaly in the training set. The other with $\ell_1 = \lfloor 0.02\ell \rfloor$ for the case of contamination with anomalies. For each set of parameter values, the experiment is repeated $B = 1000$ times.

2. Then, the Anomaly Detection capacity is evaluated using the FDR and FNR criteria. This is done by simulating multiple detections inside a segment applying the following procedure: using n the calibration set cardinality, ℓ the length of the segment, ℓ_1 the number of anomalies in the training set, m the test set cardinality and m_1 the number of anomalies in the test set:

(a) Generate training segment data with ℓ_1 anomalies.

$$\forall i \in [\![1, \ell_1]\!], \quad X_i \sim \mathcal{N}(4, 0.1), \text{ and } \forall i \in [\![\ell_1, \ell]\!], \quad X_i \sim \mathcal{N}(0, 1)$$

And estimate the segment mean and standard deviation

$$\hat{\mu} = \hat{\mu}(X_1^\ell), \quad \hat{\sigma} = \hat{\sigma}(X_1^\ell)$$

(b) Generate the calibration set

$$\forall j \in [\![1, n]\!], \quad Y_j \sim \mathcal{N}(0, 1)$$

(c) Generate the test segment data

$$\forall i \in [\![1, m_1]\!], \quad Z_i \sim \mathcal{N}(4, 0.1), \text{ and } \forall i \in [\![m_1, m]\!], \quad Z_i \sim \mathcal{N}(0, 1)$$

(d) Compute the p-values of the test set, using calibration set and affected by the parameter estimations

$$\forall i \in [\![1, m]\!], \quad \hat{p}_i = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X_j > (Z_i - \hat{\mu})/\hat{\sigma}]$$

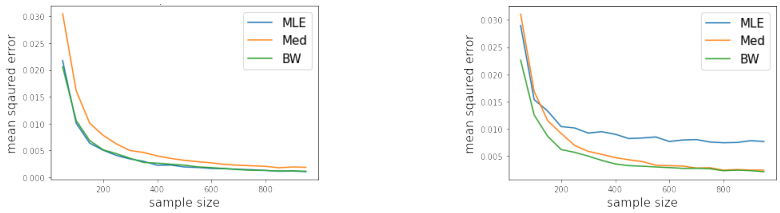
(e) Anomalies are detected using the Benjamini-Hochberg procedure on the p-values. The threshold of the BH procedure is noted $\hat{\varepsilon}_{BH_\alpha}$ as defined in our previous work [5]:

$$\hat{\varepsilon} = \hat{\varepsilon}_{BH_\alpha}(\hat{p}_1, \ldots, \hat{p}_m)$$

(f) Compute FDP and FNP. Remembering that anomalies are generated in the first m_1 values of the test set:

$$FDP = \frac{\sum_{j=m_1}^{m} \mathbb{1}[\hat{p}_j \leq \hat{\varepsilon}]}{\sum_{j=1}^{m} \mathbb{1}[\hat{p}_j \leq \hat{\varepsilon}]}$$

$$FNP = \frac{\sum_{j=1}^{m_1} \mathbb{1}[\hat{p}_j > \hat{\varepsilon}]}{m_1}$$



(a) Without anomaly contamination. (b) With anomaly contamination.

Fig 5: Estimation error of the mean as a function of segment length and mean estimator used.

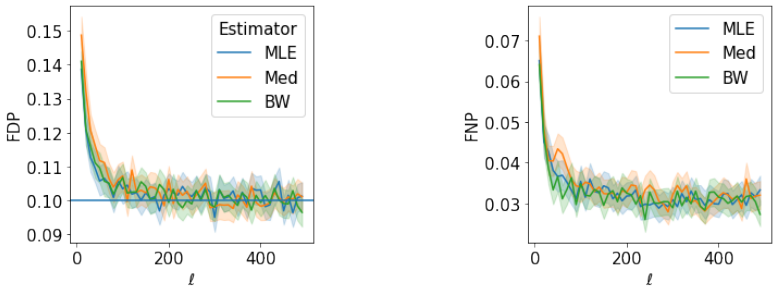
Different values of segment length ℓ are tested, from 10 to 500. For each value of ℓ, two values of ℓ_1 are tested. One with $\ell_1 = 0$, for the case where there are no anomaly in the training set. The other with $\ell_1 = \lfloor 0.02\ell \rfloor$ for the case of contamination with anomalies. The test set contain $m = 100$ data points with $m_1 = 1$ anomaly and the calibration set contains $n = 999$ data points. For each set of parameter values, the experiment is repeated $B = 10^4$ times.

3.2. Results

Figures 5 and 6 illustrate the estimators performances of the mean estimators. Figure 5 compares different mean estimators according to the segment length. The MSE decreases rapidly with the sample size for all estimators in Figure 5a. However the MLE and BW estimators have very close and slightly better performances compared to the median estimator. This illustrates the efficiency of the MLE and BW estimators. But in the presence of anomalies, the performance of the MLE estimator is severely degraded as shown in Figure 5b compared to the median and BW estimators showing more robustness in the presence of outliers.

Figure 6 illustrates the FDR and FNR of the anomaly detector according to the mean estimator used. As shown in Figure 6a and 6b, in the case of non contamination by anomalies, the FDR and FNR results are very close to the target for all the estimators. However, in presence of anomalies, the MLE performance is degraded. In Figure 6c, the FDR is below the targeted level and in Figure 6d, the FNR is higher than other estimators. Either Med or BW can be used to do anomaly detection.

Figures 7 and 8 illustrate the performances of the standard deviation estimators. Figure 7 compares the precision using the MSE of the different standard deviation estimators according to the segment length. The MSE decreases rapidly with the sample size for all estimators in Figure 7a. However the MLE and BW estimators have very close and better performances when compared with the MAD estimator. This illustrates the efficiency of the MLE



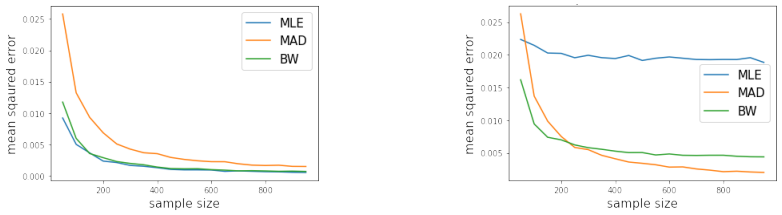
(a) FDR, without anomaly

(b) FNR, without anomaly

(c) FDR, with anomalies

(d) FNR, with anomalies

Fig 6: Anomaly detector performances as a function of the segment length and the mean estimator used.



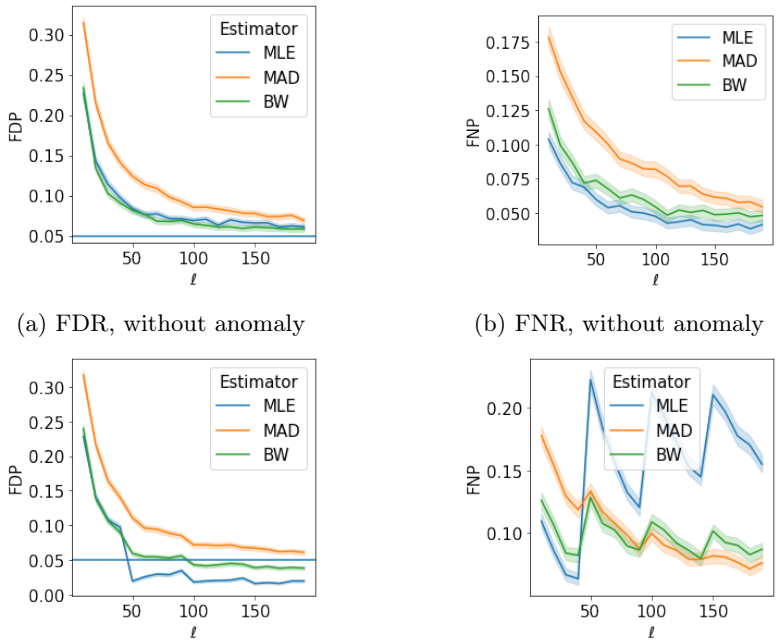
(a) Without anomaly contamination.

(b) With anomaly contamination.

Fig 7: Estimation error of standard deviation as a function of the segment length and the standard deviation estimator used.

and BW estimators. But in the presence of anomalies, the performance of the MLE estimator is severely degraded as shown in Figure 7b. On the contrary, the MAD and BW estimators are less degraded and show more robustness in presence of outliers.

Figure 8 shows the performances measured by FDR and FNR once the anomaly detection is applied. As illustrated in Figures 8a and 8b, FDR and FNR



(a) FDR, without anomaly

(b) FNR, without anomaly

(c) FDR, with anomalies

(d) FNR, with anomalies

Fig 8: Anomaly detector performances as a function of segment length and standard deviation estimator used.

for MAD are higher compared to MLE and BW. But in presence of anomalies, the MLE performance is degraded. The FDR is below the targeted level, as shown Figure 8c, and the FNR is higher than other estimators, as shown in Figure 8d. The strange behavior of the MLE curve in Figure 8d with spikes in the FNR is due to the number of anomalies increasing with every 50 data points because $\ell_1 = \lceil 0.02\ell \rceil$. The best estimator for standard deviation in case of anomaly detection is BW.

3.3. Conclusion

The experiments show the importance of the robustness and efficiency to build a good atypicity score. High MSE implies lower performance in terms of FDR and FNR control. The classical standard deviation estimators, MLE and MAD, are underperforming. For the following sections of this paper, the BW (Biweight midvariance) estimator is used to implement the scoring function.

4. More details on experiments about FDR control assessment on different scenarios

This section collects the detailed results of the experiment presented in Section 4.2 of the main article. Each subsection corresponds to a different scenario.

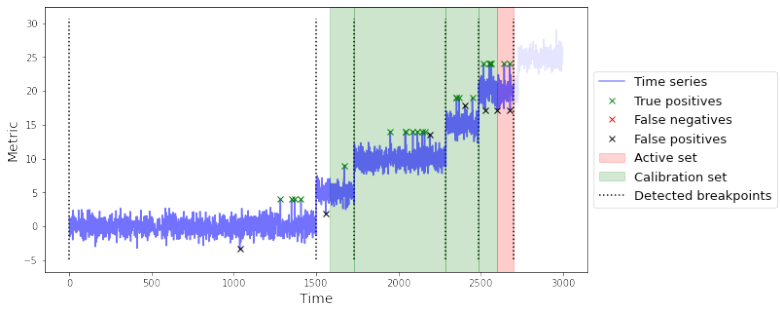
4.1. Gaussian time series with breakpoints in the mean

This scenario considers Gaussian data with breakpoints in the mean. The z-score is used to capture anomalies. The ability of the detector to control the FDR with a low FNR on different difficulties is assessed by varying the desired level of FDR control α and the size of the shift between two segments Δ.

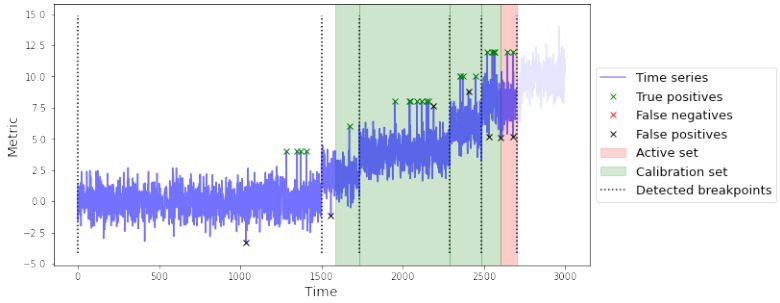
By applying the framework presented in Section 4.1 of the main article, multiple choices have been made:

- The Gaussian distribution is considered as the reference $\mathcal{P}_{0,1}$ and the proportion of anomalies is equal to $\pi = 0.01$. These anomalies are generated in the tail of the reference distribution and follow $\Delta'\zeta$, where ζ is the Rademacher distribution and $\Delta' = 4$ is the spike size of the anomalies.
- The transition rule between two breakpoints is a jump in the mean of size Δ taking values in $\{2, 3, 5\}$.
- For the breakpoint detector, the Gaussian kernel with bandwidth estimated using the median heuristic is considered, as presented in Appendix B.3 of the main article. The z-score is used as the scoring function with the mean estimated using the median estimator and the standard deviation estimated using the biweight midvariance estimator, as defined in Section 3.
- According to preliminary experiments in Section 1 and Section 2, the active set is built using $\hat{\lambda} = \hat{\ell} = 100$. Based on the rules defined in Appendix B.6 of the main article, Benjamini-Hochberg is applied on the active set with the modified parameter $\alpha' = \frac{\alpha}{1+\frac{1-\alpha}{m\pi}}$. The calibration set is built according to the rules in Appendix B.6 of the main article, where the value n is chosen equal to $m/\alpha' - 1$. Two cases are considered $\alpha = 0.2$ and $\alpha = 0.1$. In the case $\alpha = 0.2$, then the following values are chosen $\alpha' = 0.1$ and $n = 999$. In the case $\alpha = 0.1$, then $\alpha' = 0.05$ and $n = 1999$.

Figure 9 shows an example of anomaly detection for one time series. The x-axis is the timestamp and the y-axis the value of the generated time series, shown in blue. The light blue data points are those that are not observed at the time the results are presented. The vertical black lines are the detected breakpoints, the red band is the subseries defined as the active set, the green band is the subseries used to build the calibration set. Detected anomalies are the green crosses, false positives are the black crosses and red crosses are the false negatives. As shown in Figure 9a, there are no false negative and the false



(a) $\Delta = 5$



(b) $\Delta = 2$

Fig 9: Application of our anomaly detector on Gaussian time series having breakpoints in the mean, for different shift size values Δ.

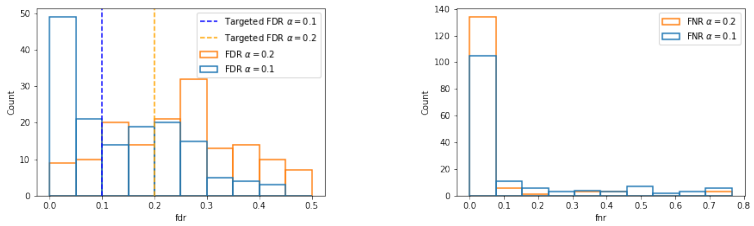
positives seem to be a small fraction of the true detected anomalies. As expected, the breakpoints are positioned exactly where the means of the series change. The active set contains the most recent observations. And the calibration set gathers data from several segments since the current segment does not contain enough data.

Table 1 gives the FDR and the FNR after having applied the anomaly detector to a collection of $B = 50$ Gaussian time series with breakpoint in the

α	Δ	FDR	FNR
0.10	2	0.133	0.123
	3	0.134	0.111
	5	0.129	0.106
0.20	2	0.242	0.039
	3	0.242	0.042
	5	0.236	0.037

TABLE 1

FDR and FNR for anomaly detection on Gaussian time series having breakpoints in the mean according to the α level and the shift size Δ.



(a) FDR (b) FNR

Fig 10: Histograms of the FDR and FNR for different targeted FDR α levels.

mean for different shift sizes Δ. The FNR is always close to 0. This is necessary to ensure the FDR control with the modified BH procedure. For all the cases, the FDR remains close to the desired α level. The FDR is well influenced by the choice of α level but less by the value of Δ. However, it is always slightly higher than alpha. Indeed, for $\Delta = 5$, it is equal to 0.23 instead of $\alpha = 0.20$, as shown in Table 1.

The histogram in Figure 10 shows more detailed results applied to the collection of time series for different values of α parameter. Figure 10a shows the distribution of the FDR values compared to the target FDR in vertical lines. Figure 10b shows the distribution of the FNR values. As shown in Figure 10a,the performance of the anomaly detector is poor for some time series since the FDR values are higher and far from the target FDR. This explains why the measured FDR is slightly higher than the targeted FDR in Table 1. The diagnosis of this inefficiency will be examined in Section 5. In the next Sections 4.2, 4.3, 4.3, 4.4 and 4.5 the anomaly detector is applied and checked to more complex time series.

4.2. Gaussian mixture time series with breakpoints in the mean

In this section, the aim is to show how to handle anomalies that occur between two modes of a Gaussian mixture. These anomalies, which do not occur in the tail of a distribution, cannot be detected by z-scores because they are close to the mean. Therefore, it is necessary to adapt to this new situation by using another atypicity score, such as the kNN score introduced in [6]. Indeed, in this case, anomalies can be characterized by their distance from other segment data.

The anomaly detector applied to Gaussian mixture data considers the reference distribution $\mathcal{P}_{0,1} = 0.5\mathcal{N}(\Delta', 1) + 0.5\mathcal{N}(-\Delta', 1)$, with an anomaly spike size of $\Delta' = 6$. The anomalies are chosen to be equal to 0 to ensure they lie in the middle between the two Gaussian distributions.

As explained previously, to adapt to this new difficulty of time series data with Gaussian mixture, the atypicity score needs to be chosen accordingly. The kNN score introduced in [6] is applied. To ensure that the distribution of the

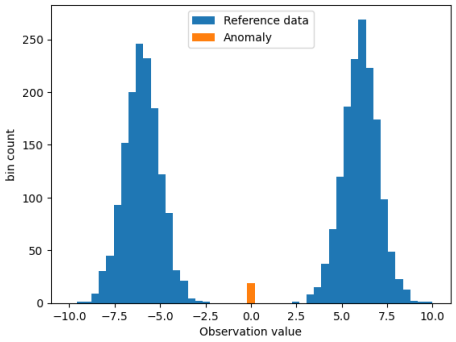


Fig 11: Histogram that represent the Gaussian mixture reference distribution with anomalies in the center.

score is the same between two segments and not affected by segment cardinality, the kNN distance is computed after having resampled $B_s = 100$ points from the segment. To obtain a robust score, the number k of nearest neighbors should be chosen carefully because the kNN distance should not be affected by the presence of anomalies in the segment. In particular, the k nearest neighbors of an anomaly should not be an anomaly, otherwise the distance will be close to 0, which leads to a false positive. By choosing $k = 10$ and ensuring $k/B = 0.1 >> 0.01 = \pi$, this issue is avoided with high probability. Experimental parameters not specified in this section have the same values as in Section 4.1.

The result in Figure 12 clearly shows that for this example, the anomaly detector is able to detect the breakpoints, in the dashed black lines, and the anomalies, represented by the green crosses, with few false positives. The anomaly detector has been applied to 50 time series and the results are summarized in Table 2. The FDR is controlled at the desired level of 0.1 or

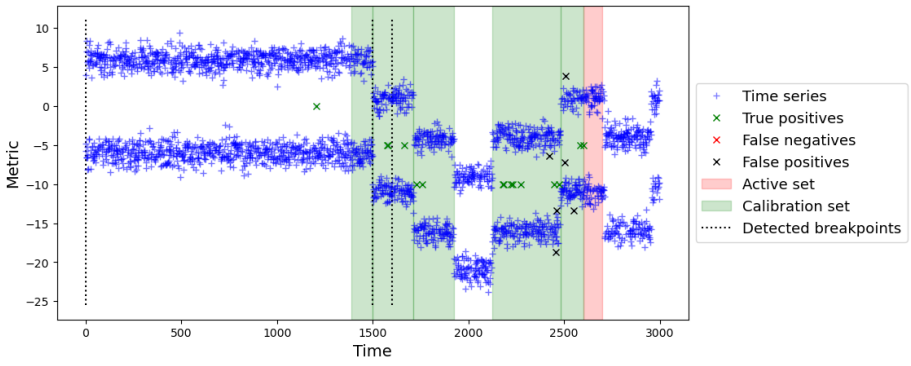


Fig 12: Application of our anomaly detector on Gaussian mixture time series having breakpoints in the mean.

α	FDR	FNR
0.1	0.118	0.246
0.2	0.202	0.137

TABLE 2
FDR and FNR for anomaly detection on Gaussian mixture time series with breakpoints in the mean according to α level.

0.2 while the FNR is slightly higher compared to the Gaussian case in Table 1. This is probably due to the kNN score, which is less efficient than the z-score.

4.3. 2D Gaussian time series with breakpoint in the covariance

In this section, the aim is to show how to handle anomalies that occur on multidimensional data. Previously, the kernel method in KCP demonstrated high accuracy in detecting breakpoints for univariate time series data. Hopefully, the paper [7] shows that this kernel method is also applicable to multivariate time series. Once the time series is segmented, a scalar atypicity score is computed for the multidimensional data points. An alternative would be to apply univariate anomaly detection to each univariate time series. However, some breakpoints, such as those occurring in the covariance, cannot be detected by this alternative method.

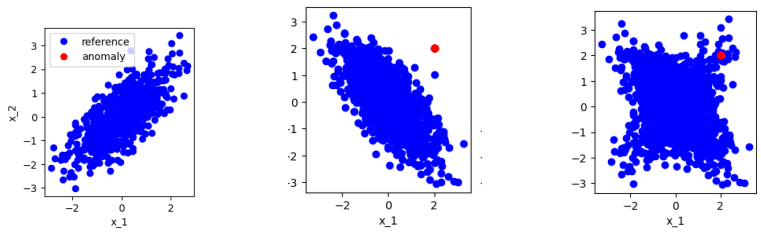
Data are generated according the following rule:

$$\forall t \in [\![1, T]\!], \quad X_t = \begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} \sim \mathcal{N}(0, \Sigma_t) \tag{3}$$

With the covariant matrix equal to:

$$\Sigma_t = \begin{cases} \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} & \text{if } t \leq \tau_1 \\ \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix} & \text{else} \end{cases} \tag{4}$$

The reference distribution generates two-dimensional Gaussian data X_t. For each component the mean is 0 and the standard deviation is 1. To simplify the generation, one breakpoint τ_1 is considered linked to the change of the covariance from 0.7 to -0.7. Figure 13a shows that the covariance is positive before the breakpoint and negative after the breakpoint in Figure 13a. Anomalies are considered in the second segment, and are set to the value $(1, 1)$. This value has interesting properties to evaluate the capacity of the anomaly detector. First, "1" appears as a typical value at each one dimensional component of the time series. This implies that the anomalies cannot be detected by working on each component independently. Second, the value $(1, 1)$ is fairly typical before the breakpoint τ_1, as shown Figure 13a. Consequently, the breakpoint detector enables the detection of anomalies while they are hidden in the data mixture as shown in Figure 13c.



(a) Before the breakpoint. (b) After the breakpoint. (c) Data mixture.

Fig 13: 2D Gaussian data with breakpoint in covariance matrix

For this scenario, the Gaussian kernel is used to detect the breakpoint in the covariance. As a characteristic kernel, it should detect the change in the covariance, which is the change at the second moment order. The median heuristic is used to select the bandwidth. Since each component cannot be treated independently to detect anomalies, the Mahalanobis distance [8] is preferred over the Euclidean distance. The Mahalanobis distance is defined as the following, where $\hat{\mu}$ is the estimated mean vector and $\hat{\Sigma}$ is the estimated covariance matrix.

$$s_t = \sqrt{(X_t - \hat{\mu})^T \hat{\Sigma}^{-1} (X_t - \hat{\mu})}$$

To ensure a good atypicity score, the estimator of the covariance has to be robust and efficient, as shown in Appendix B.4 of the main article. Inspired by the results of Section 3, the biweight-midcovariance [9] is used to estimate each coefficient of the matrix $\hat{\Sigma}$.

The result is represented for one example in Figure 14. The multidimensional time series is represented using one plot for each dimension. The anomaly detector successfully detects the breakpoints in the dashed black lines, and the anomalies that are represented by green dots with few false positives.

The anomaly detector has been applied to 50 time series and the results are

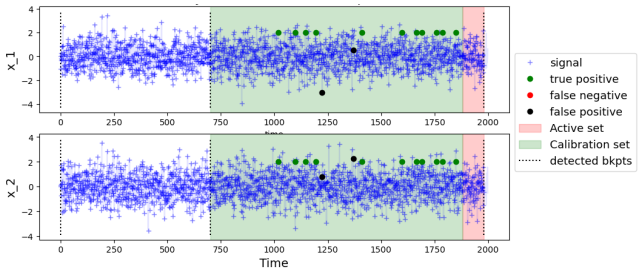


Fig 14: Application of our anomaly detector on 2D Gaussian time series having breakpoints in the covariance.

α	FDR	FNR
0.2	0.126	0.054

TABLE 3
FNR and FDR for anomaly detection on 2D Gaussian time series with breakpoint in the covariance.

summarized in Table 3. The FNR is close to 0 and the FDR is smaller than expected, 0.12 instead of 0.2. This confirms that the detector can be applied to multidimensional data with minor adaptation.

4.4. Gaussian data with breakpoints in the mean and in the variance

In the experiments conducted so far, all scenarios considered homoscedastic time series where the change is in the mean while the variance is constant between two segments. In this section, the case of heteroscedasticity in time series is studied where the variance changes between two segments. Therefore, time series will have parts where the variance is very low and parts where it is very high. The struggle is that a kernel may be good at detecting breakpoints in a low variance context, but have difficulty when the variance is high, and vice versa. Therefore, several kernels are tested by varying the bandwidth size.

Let's consider a time series generation process and an anomaly detector described in Section 4.1 of the main article. To adapt to the heteroscedasticity hypothesis, the transition rule is modified so that at each breakpoint the variance changes as follows, where Δ_σ is the variance shift size equal to 2:

$$\sigma_{i+1} = \exp(\zeta_{\sigma,i} \ln \Delta_\sigma) * \sigma_i$$

To ensure that the variance covers a wide range of values, the variable ζ_i is chosen asymmetric. In the case of this experiment, ζ_i has a probability of 0.9 of being $+1$. Thus, the variance is more likely to increase than to decrease at each breakpoint. To ensure the visibility of the breakpoint in the mean to any variance, the size of the shift in the mean needs to be proportional to the maximum of the variance of the segment before and after the breakpoint, as described in the following:, where Δ_μ is the mean shift size equal to 2:

$$\mu_{i+1} = \zeta_{\mu,i+1} \Delta_\mu \max(\sigma_i, \sigma_{i+1}) + \mu_i$$

According to the median heuristic, breakpoints are easily detected by a Gaussian kernel when the standard deviation of the data is of the same order as the bandwidth h. Several kernels are tested:

- Gaussian kernel with bandwidth $h = 1$. This kernel with a small bandwidth is relevant to detect breakpoints when the variance of the time series is small, but may fail when the variance is high.
- Gaussian kernel with bandwidth $h = 100$. In this situation, the kernel is more relevant to detect breakpoints when the variance is high.

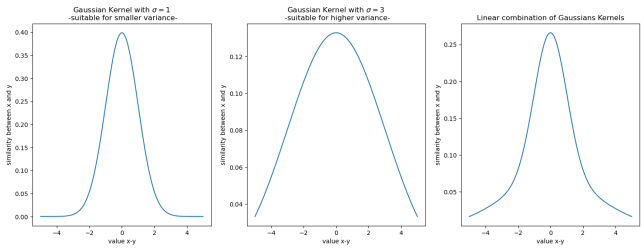


Fig 15: Illustration of the different kernels

- To consider both scenarios, where breakpoints appear in some parts of the series with high variance and in parts of the series with low variance, a linear combination of the two Gaussian kernels may be a good response. This kernel is characteristic as a sum of two characteristic kernels and is defined by:

$$K(x, y) = 0.5K_{h_1}(x, y) + 0.5K_{h_2}(x, y) \tag{5}$$

The anomaly detector is applied three times to the same time series, changing only the kernel used in Figures 16, 17 and 18:

- Figure 16 illustrates the result using the Gaussian kernel with small bandwidth, $h = 1$. The breakpoint was not detected at ①, which leads to a false negative ② and a large number of false positives at ③.
- Figure 17 illustrates the result using the Gaussian kernel with large bandwidth, $h = 100$. At the position ①, the breakpoint with low variance is not detected. It leads to false positives at ② because data with different variances belong to the same calibration set.
- Figure 18 illustrates the result when using the linear combination of the two Gaussian kernels. All breakpoints are detected, reducing the number of false positives and false negatives.

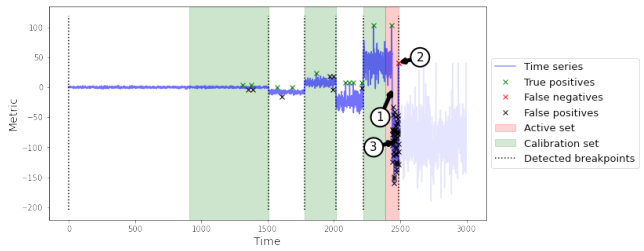


Fig 16: Application of our anomaly detector on Gaussian time series having breakpoints in the mean and in the variance, using a Gaussian kernel having a small bandwidth.

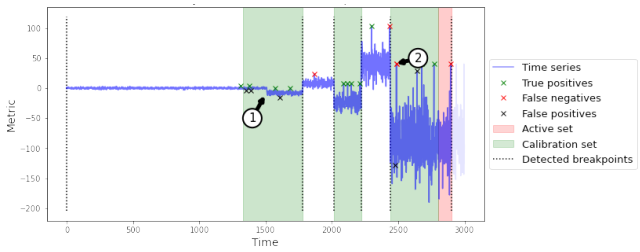


Fig 17: Application of our anomaly detector on Gaussian time series having breakpoints in the mean and in the variance, using a Gaussian kernel having a large bandwidth.

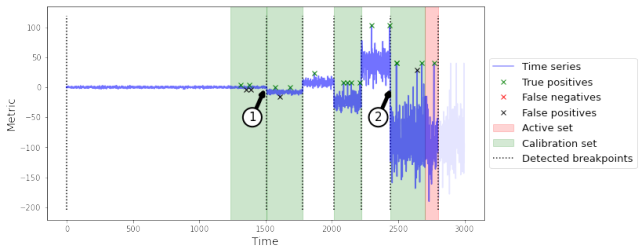


Fig 18: Application of our anomaly detector on Gaussian time series having breakpoints in the mean and in the variance, using a linear combination of two Gaussian kernels.

The anomaly detector has been applied to 50 time series and the FDR and FNR results are summarized in Table 4. Different kernels, bandwidth h, are considered in combination with α levels in $\{0.1, 0.2\}$:

- Gaussian kernel (labeled Gaussianh) with bandwidth h in $\{1, 10, 100\}$
- Linear combination of two Gaussians (labeled CombG1G100)

The performances are strongly influenced by the kernel bandwidth: The FNR is lower when using the Gaussian kernel with bandwidth $h = 1$ or using the combination of Gaussian kernels while it is high when using kernels with larger bandwidth. The FDR is slightly higher than expected α for all tested kernels. However, the FDR is smaller when using the combination of Gaussians compared to the Gaussian kernel with $h = 1$. Thus, anomaly detection remains possible when the variance of the time series changes under heteroscedasticity. However, there is no general way to build a dedicated kernel that responds to this scenario, but combining specialized kernels to adapt to the different regimes of the time series seems to be a promising approach.

α	Kernel	FDR	FNR
0.10	Gaussian1	0.188	0.054
	Gaussian10	0.127	0.136
	Gaussian100	0.148	0.456
	CombG1G100	0.134	0.057
0.20	Gaussian1	0.323	0.017
	Gaussian10	0.232	0.102
	Gaussian100	0.232	0.397
	CombG1G100	0.253	0.018

TABLE 4

FDR and FNR for anomaly detection on Gaussian time series with breakpoints in the mean and in the variance according to the α level and the chosen kernel

4.5. Gaussian data with breakpoints in the variance

In this section, the more challenging scenario of time series with changes in variance without a shift in mean is addressed.

To generate the data, a Gaussian distribution is used as the reference one. The breakpoints in the variance are generated according to the rule described in Eq. 6.

$$\forall i \in [\![1, D-1]\!], \quad \sigma_{i+1} = \exp(\zeta_i \ln \Delta/2) * \sigma_i \tag{6}$$

Since the variance of the time series changes along the time series, it may be difficult to detect all the breakpoints with the same kernel. To evaluate the detector in this scenario, it is based on the same kernels defined in Section 4.4 and on the z-score atypicity function.

Figures 19 and 20 show two examples of anomaly detection. In Figure 19, all the breakpoints are successfully detected, allowing correct anomaly detection with few false positives. In Figure 20, the procedure fails and no breakpoint is detected in ①. After the change with higher variance, all data are considered as anomalies. It is an evidence that the efficiency of the anomaly detector is strongly influenced by its ability to detect the true breakpoints.

Table 5 summarizes the FDR and FNR results obtained for 50 time series using the same kernels and α levels as in Table 4. In all cases, the anomaly

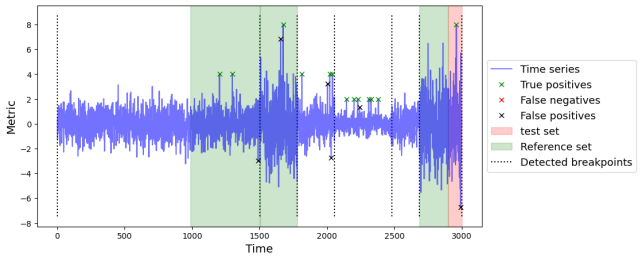


Fig 19: Example of successful anomaly detection on time series with breakpoints in the variance.

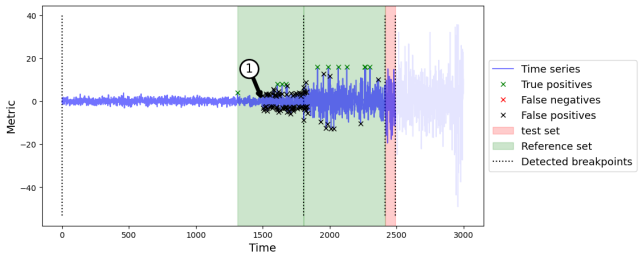


Fig 20: Examples of failure of anomaly detection on time series with change in the variance.

α	Kernel	FDR	FNR
0.10	Gaussian1	0.272	0.321
	Gaussian10	0.806	0.712
	Gaussian100	0.835	0.599
	CombG1G100	0.229	0.298
0.20	Gaussian1	0.313	0.241
	Gaussian10	0.649	0.511
	Gaussian100	0.685	0.396
	CombG1G100	0.282	0.225

TABLE 5
FDR and FNR for anomaly detection on Gaussian time series with breakpoint in the variance according α level and chosen kernel.

detection shows a poor accuracy, since on one side there is a lack of control of the FDR with respect to the target value alpha, and on the other side the FNR is very high. However, the best FNR and FDR values are obtained for the combination of Gaussian kernels, which allows better detection of breakpoints in the variance.

These results show how challenging the case of time series with breakpoints in the variance is. Indeed, the change in the variance is much harder to detect than the shift in the mean presented in Section 4.1. One approach is to carefully tune the kernel by choosing the right combination of kernels to enable the detection of specific types of breakpoints.

5. More details on experiments about underperformance diagnosis

This section collects the detailed results of the experiment presented in Section 4.3 of the main article.

5.1. Description of the experiment

The BKAD is applied to the synthetic time series, where some estimators are replaced by true knowledge, called oracle version. Three estimators are chosen to be replaced by their oracle versions:

Detector	Breakpoint	Mean and variance	Anomaly Removing
Detector 1	O	O	O
Detector 2	O	O	E
Detector 3	O	E	O
Detector 4	E	E	O
Detector 5	E	E	E

TABLE 6
Description of the different detectors.

- The breakpoint estimator: can be replaced by the true breakpoint position,
- The mean and standard deviation estimators: can be replaced by their true values,
- The anomaly removed: As described in Appendix B.5 of the main article when building the calibration set, estimated anomalies are removed to avoid biasing the estimation of the p-values. The oracle version of this is to remove the true anomalies.

Using the framework from Section 4.1 of the main article, five anomaly detectors are applied to each time series. Multiple combinations of the true knowledge (marked "O") versus estimated values (marked "E") are used to produce different versions of anomaly detectors described in Table 6. As an example, for detector 3, the breakpoints and anomalies in the calibration set are detected using their true values, but the segment mean and variance parameters are estimated.

The significance of the results is checked using permutation tests. It is possible that the cause of this underperformance depends on the data distribution or on breakpoint types. Different probability distributions are tested with different kinds of shifts.

5.2. Results and analysis

The complete empirical results can be found in Section 6. The performances of the different detectors are evaluated on a different laws generating the time series (Student, Gaussian, Mixture of Gaussians noted MoG). The FDR and FNR distributions are represented by a boxplot with the significance differentiating two detectors ("ns" the difference is not significant, "*" significance at 5%, "**" significance at 1%, "***" significance at 0.1%).

In the following paragraphs, the effects of the various core components are studied: breakpoint detector, mean and variance segment estimator and anomalies removed from the calibration set.

Breakpoint Estimation Table 7 shows the performance of anomaly detectors 3 and 4 (see Table 6) for different types of data and shifts. The only difference between the two detectors is that Detector 3 uses a breakpoint detector while Detector 4 has knowledge of true breakpoints. The bold values highlight the cases where the difference between the two estimators is significant. Table 7 illustrates that the breakpoint estimation does not strongly

Type of shift	law	α	Breakpoints	FDR	FNR
Mean	Gaussian	0.10	E	0.104	0.105
			O	0.100	0.091
		0.20	E	0.182	0.054
			O	0.176	0.054
	Student	0.10	E	0.119	0.066
			O	0.117	0.065
		0.20	E	0.199	0.033
			O	0.198	0.032
	MoG	0.10	E	0.113	0.131
			O	0.108	0.124
		0.20	E	0.186	0.072
			O	0.190	0.071
Mean and var.	Gaussian	0.10	E	**0.114**	0.090
			O	0.099	0.078
		0.20	E	**0.188**	0.051
			O	0.167	0.040
Variance	Gaussian	0.10	E	**0.200**	**0.214**
			O	0.110	0.109
		0.20	E	**0.257**	**0.128**
			O	0.174	0.062

TABLE 7
Anomaly detector performances with and without knowledge of true breakpoint positions, according different time series.

affect the FDR performance except in the case where breakpoints occur in the variance. This is expected since breakpoints in the variance are more difficult to to detect, as discussed earlier in Section 4.5. FNR increases in few cases where the breakpoint positions are estimated.

Segment mean and variance parameters Table 8 shows the performance of anomaly detectors 1 and 3 (see Table 6) for different types of data and shifts. The only difference between the two detectors is that Detector 3 estimates the mean and the variance parameters of the segments while Detector 1 has knowledge of the true parameters. According to Table 8, the estimators do not strongly affect the FDR of the anomaly detector. There are few significant differences, displayed in bold, which are smaller than in Table 7.

Anomalies Removing Table 9 represents the results for different detectors considering different laws, alpha levels and kind of shift. The four detectors are chosen to identify the effect of removing detected anomalies from the calibration set instead of removing the true anomalies, in case other components are estimators and in case other components are oracles. Note that for Gaussian Mixture (MoG), the "Mean and Variance" component is marked with a "X", since the kNN atypicity score does not use mean and variance parameters. It is clear that the control of the FDR is worse when the calibration set is built based on detected anomalies. Indeed, the false positives and false negatives detected at time t will badly affect the detection at time $t+1$. Despite the fact that a robust score is chosen, these observations lead to a conclusion that the p-value estimator is sensitive to:

type of shift	law	α	Mean and variance	FDR	FNR
Mean	Gaussian	0.10	E	0.082	0.043
			O	0.105	0.000
		0.20	E	0.167	0.000
			O	0.188	0.000
	Student	0.10	E	0.117	0.065
			O	0.113	0.056
		0.20	E	0.198	0.032
			O	0.196	0.026
Mean and var.	Gaussian	0.10	E	**0.091**	0.000
			O	0.107	0.000
		0.20	E	**0.167**	0.000
			O	0.200	0.000

TABLE 8

Anomaly detector performances with knowledge of the true segment mean and standard deviation values and with estimation of these parameters, according different time series.

- False negatives: If there is a missed anomaly in the calibration set, the p-values of all data points in the active set will be underestimated. This situation leads to generate more false negatives, which will confound the calibration sets of subsequent instants.
- False positives: The p-value estimator is also sensitive to false positives due to the way the calibration set is constructed. As a reminder, detected anomalies are replaced by a random points belonging to a segment similar to the current segment. The problem arises when an anomaly is falsely detected. Generally speaking a false positive is a point with a high score. When a false positive is replaced with a random point, its score will be statistically lower. Thus, removing the false positives from the calibration set reduces the average score in the calibration set and consequently reduces the p-values of the data points in the active set. This leads to more false positives, which will affect the construction of calibration sets at later times.

5.3. Conclusion

The conclusion of this analysis is that most of the underperformance relative to the ideal case, such as higher than expected FDR, is explained by the non-robustness of the empirical p-value estimator and the contamination of the calibration set by false negatives and false positives.

Type of shift	Law	α	Breakpoint	Mean and variance	**Anomaly removing**	FDR	FNR
Mean	Gaussian	0.1	E	E	E	**0.134**	0.123
			E	E	O	0.104	0.105
			O	O	E	**0.165**	0.041
			O	O	O	0.121	0.048
		0.2	E	E	E	**0.242**	0.039
			E	E	O	0.182	0.054
			O	O	E	**0.301**	0.018
			O	O	O	0.197	0.018
	Student	0.1	E	E	E	**0.158**	0.059
			E	E	O	0.119	0.066
			O	O	E	**0.154**	0.035
			O	O	O	0.113	0.056
		0.2	E	E	E	**0.289**	0.026
			E	E	O	0.199	0.033
			O	O	E	**0.301**	0.013
			O	O	O	0.196	0.026
	MoG	0.1	E	X	E	0.118	0.246
			E	X	O	0.113	0.131
			O	X	E	0.103	0.294
			O	X	O	0.108	0.124
		0.2	E	X	E	0.202	0.137
			E	X	O	0.186	0.072
			O	X	E	0.221	0.111
			O	X	O	0.190	0.071
Mean and var.	Gaussian	0.1	E	E	E	0.134	0.057
			E	E	O	0.114	0.090
			O	O	E	**0.955**	0.022
			O	O	O	0.119	0.054
		0.2	E	E	E	**0.253**	0.018
			E	E	O	0.188	0.051
			O	O	E	**0.961**	0.021
			O	O	O	0.205	0.029

TABLE 9

Anomaly detector performances with and without knowledge of true anomalies for removing anomalies, according different time series.

6. Figures related to experiment of Section 5

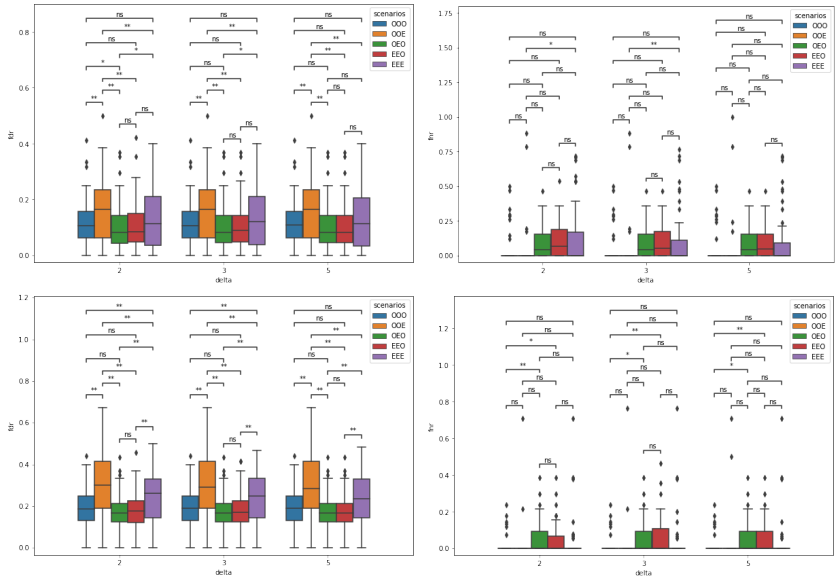


Fig 21: Boxplots of FDR and FNR for anomaly detection on Gaussian time series having breakpoint in the mean according to the different Detectors described in Table 6 and shift size Δ. Top-left: FDR while $\alpha = 0.1$, Top-right: FNR while $\alpha = 0.1$, Bottom-left: FDR while $\alpha = 0.2$, Top-right: FNR while $\alpha = 0.2$.

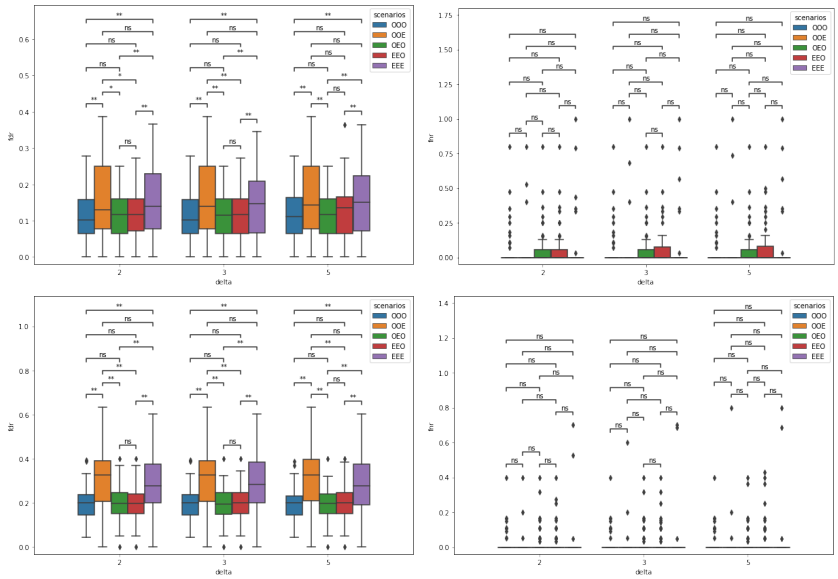


Fig 22: Boxplots of FDR and FNR for anomaly detection on Student time series having breakpoint in the mean according to the different Detectors described in Table 6 and shift size Δ. Top-left: FDR while $\alpha = 0.1$, Top-right: FNR while $\alpha = 0.1$, Bottom-left: FDR while $\alpha = 0.2$, Top-right: FNR while $\alpha = 0.2$

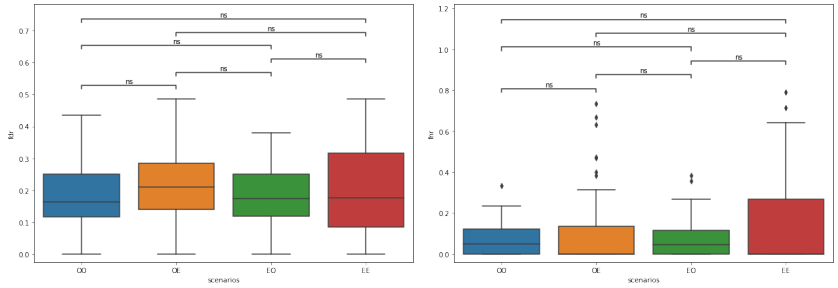


Fig 23: Boxplots of FDR and FNR for anomaly detection on Gaussian Mixture time series having breakpoint in the mean according to the different Detectors described in Table 6. Left: FDR while $\alpha = 0.2$, right: FNR while $\alpha = 0.2$.

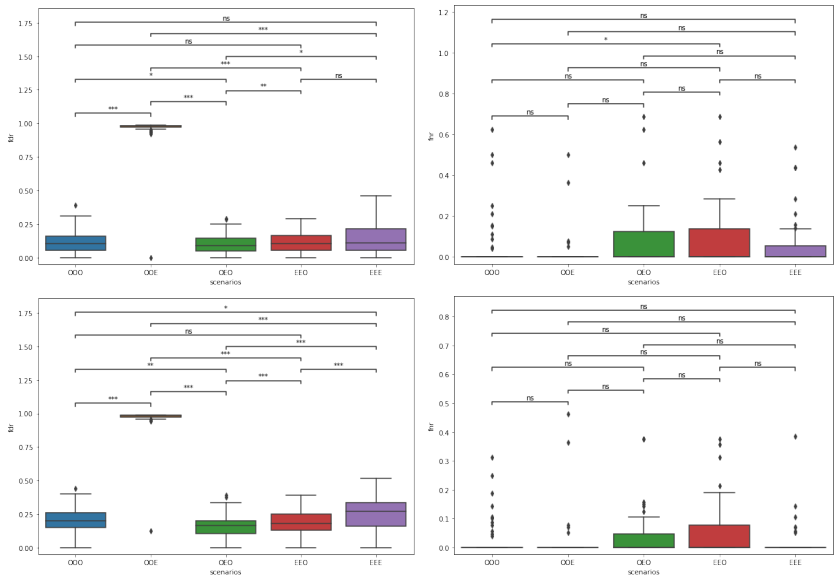


Fig 24: Boxplots of FDR and FNR for anomaly detection on Gaussian time series having breakpoint in the mean and variance according to the different Detectors described in Table 6. Top-left: FDR while $\alpha = 0.1$, Top-right: FNR while $\alpha = 0.1$, Bottom-left: FDR while $\alpha = 0.2$, Top-right: FNR while $\alpha = 0.2$

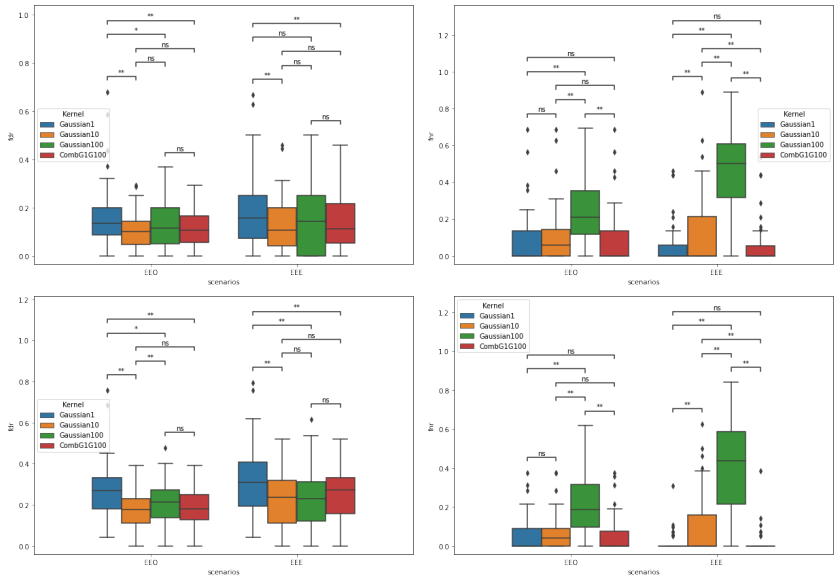


Fig 25: Boxplots of FDR and FNR for anomaly detection on Gaussian time series hawing breakpoint in the mean and and in the variance according to the chosen Kernel. Top-left: FDR while $\alpha = 0.1$, Top-right: FNR while $\alpha = 0.1$, Bottom-left: FDR while $\alpha = 0.2$, Top-right: FNR while $\alpha = 0.2$

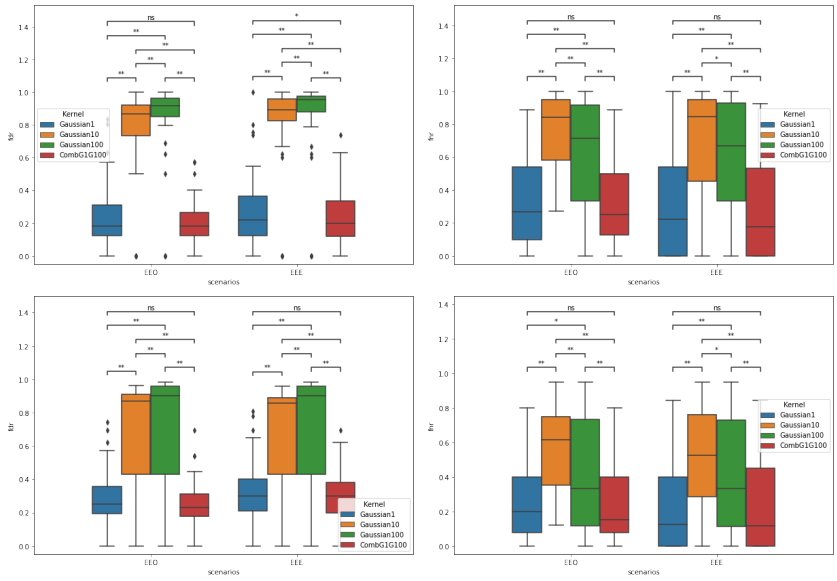


Fig 26: Boxplots of FDR and FNR for anomaly detection on Gaussian time series having breakpoints in the variance according to different the chosen Kernel. Top-left: FDR while $\alpha = 0.1$, Top-right: FNR while $\alpha = 0.1$, Bottom-left: FDR while $\alpha = 0.2$, Top-right: FNR while $\alpha = 0.2$

7. Supplementary experiment: empirical study on the effect of hyperparameter choice

The goal of this section is to show how incorrect hyperparameter values of the anomaly detector lead to a degradation of the anomaly detector's performances. This evaluation is done for three core components: the variance estimator, the cardinality of the calibration set, and the cardinality of the active set. The hyperparameters of these components are intentionally set very far from the recommendations given in Appendixes B.1, B.2 and B.4 of the main article and the consequences are observed and discussed to confirm the recommendations, the rules and the analyses stated in the paper.

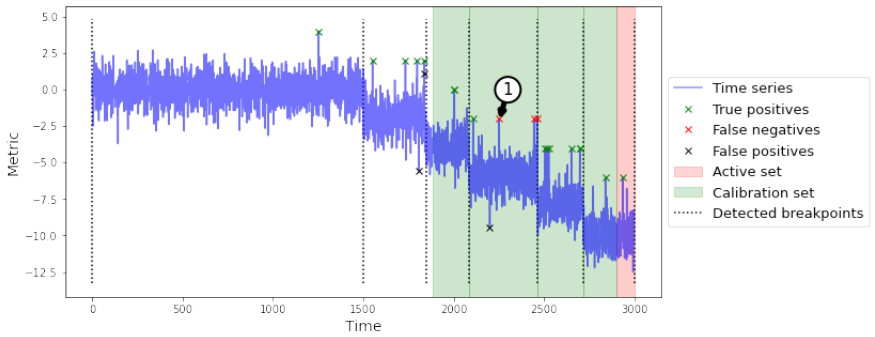
7.1. Bad choice of variance of segments estimator

In Appendix B.4 of the main article , it was established that a good atypicity score should respect two properties: robustness to the presence of anomalies in the training set and efficiency. In this scenario, a bad choice is made for an atypicity score that does not respect the requirements of being robust and efficient. To simulate this case and evaluate the effects, the experiment with Gaussian time series having breakpoint in the mean introduced in Section 4.1 is reused by replacing the biweight estimator of the variance in the z-score function by the MLE estimator or the MAD estimator. Indeed, MLE estimator is efficient but not robust, and the MAD estimator is robust but not efficient while the biweight midvariance is robust and efficient.

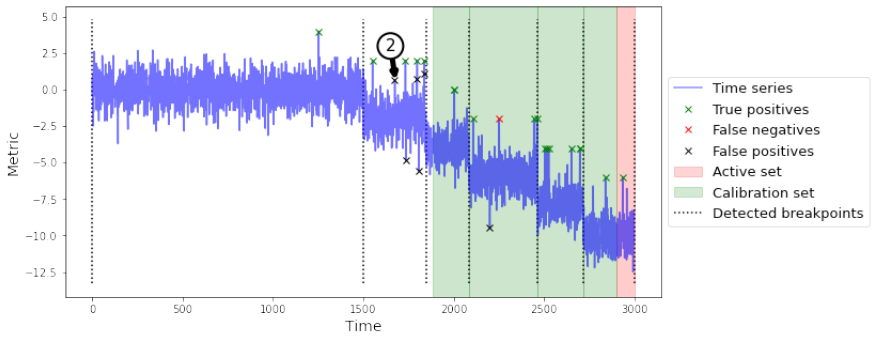
To analyze and compare the effect of the different estimators, the same example is considered in Figure 27, for different variance estimators. Since the MLE estimator is not robust, Figure 27a at ① shows false negatives due to variance overestimation caused by the presence of anomalies in the current segment. The choice of the robust MAD estimator reduces the false negatives while it generates a higher number of false positives as shown in Figure 27b at ②. The variance is underestimated due to the lack of data points and the lower efficiency of MAD. The Biweight estimator is advantageous as it is both robust and efficient and is able to reduce false positives and false negatives, as shown in Figure 27c.

In Figure 28, the boxplots represent the distribution of FNR and the FDR over a set of 50 time series based on the standard deviation estimator (MLE, MAD or BW). Paired permutation tests [10] are used to compare the performances of two estimators. For each pair of estimators, the hypothesis tested is: "The mean FDR (or FNR) is the same using these two variance estimators". The results are represented by adding a symbol ("ns" the difference is not significant, "*" significance at 5%, "**" significance at 1%, "***" significance at 0.1%) between the two tested estimators.

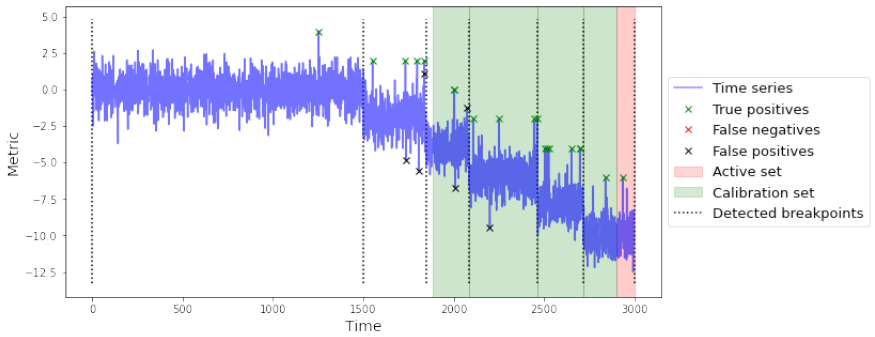
The FDR and FNR results are summarized in Table 10. The FNR is significantly higher when the MLE variance estimator is used compared to the more robust MAD and biweight midvariance estimators, which have close



(a) MLE



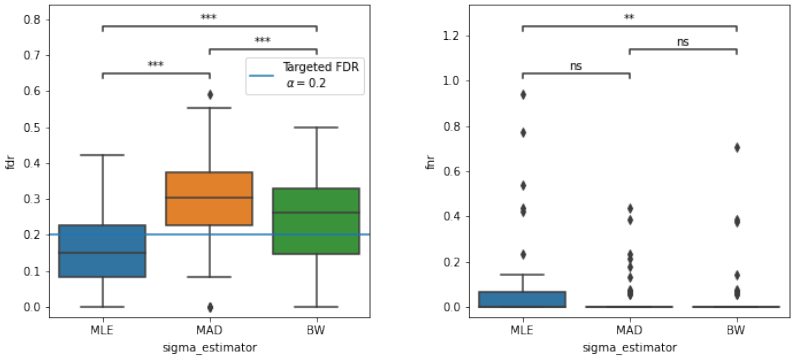
(b) MAD



(c) BW

Fig 27: Application of our anomaly detector on Gaussian time series having breakpoints in the mean, using different variance estimators.

performances. However, the FDR is significantly higher when the MAD is used compared to the biweight midvariance estimator, for which the FDR is better controlled.



(a) FDR (b) FNR

Fig 28: Boxplots of the FNR and FDR according to the choice of the variance estimator.

sigma_estimator	FDR	FNR
MLE	0.16	0.08
MAD	0.29	0.04
BW	0.24	0.04

TABLE 10
FNR and FDR according to the choice of the variance estimator.

7.2. Bad choice for active set cardinality

Section 3.2 of the main article introduced the notion of an active set to deal with the uncertainty of status. It also provides rules to compute the cardinality of the active test, described more precisely in Appendix B.1 and Appendix B.2. In this section the relevance of this rule is evaluated.

To evaluate the performance degradation due to a bad choice of the active set cardinality, the experiment framework introduced in Section 4.1 is reused. According to the results of the experiments in Sections 1 and 2, status can be ensured with strong confidence with an active set cardinality equal to $m = 100$. For each time series generated, two anomaly detectors are applied, one with an active set cardinality equal to 100 and the second with an active set cardinality equal to 10.

Figure 29 illustrates the boxplots of the FDR distribution according to the active set cardinality. The results, summarized in Table 11, show that the FDR is significantly higher when the active set has a cardinality of $m = 10$. On the contrary, using a cardinality of $m = 100$ allows to control the FDR at the desired level $\alpha = 0.2$. This experiment illustrates the benefits of following the recommendations defined in Appendix B.1 and Appendix B.2 of the main article

to improve anomaly detection performances.

α	m	FDR	FNR
0.2	10	0.529	0.006
	100	0.186	0.053

TABLE 11
FDR and FNR mean according to the active set cardinality.

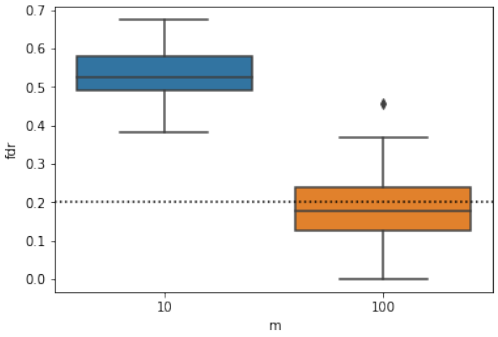
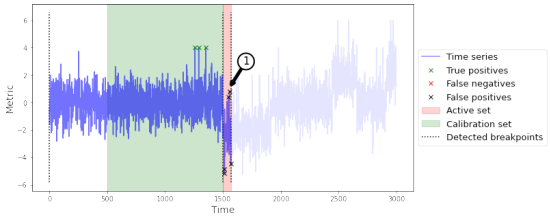
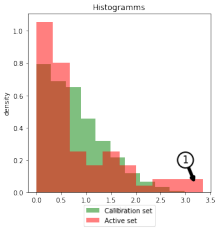


Fig 29: FDR boxplots according to the active set cardinality

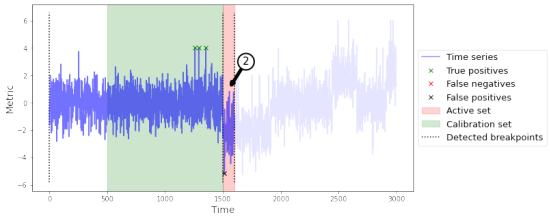
In order to understand how the active set improves the anomaly detector, the results are observed, for one time series, at two different instants: at time $t = 1570$ in Figures 30a and 30b, and at time $t = 1600$ in Figures 30c and 30d. The histograms of the z-scores of the calibration set in green and the active set in red are shown in Figures 30b and 30d. At time $t = 1570$, the new current segment contains few points, resulting in a variance estimation error and an overestimation of the z-score of the active set in ① Figure 30b and false positives in ① Figure 30a. At time $t = 1600$, the segment has acquired new data points, the variance estimate is improved and the z-score is not overestimated in ② Figure 30d. The number of false positives is reduced in ② Figure 30c. The status of the data point at $t = 1570$ is corrected at time $t = 1600$ because the active set is large enough, otherwise its status would be fixed to the wrong one.



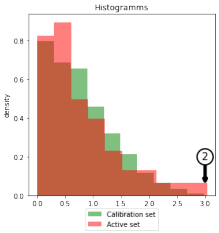
(a) Example of false positive, $t = 1570$



(b) Example of false positive



(c) Example of accurate detection, $t = 1600$



(d) Example of false negative

Fig 30: Abnormality status update after new data points acquisition in the current segment.

7.3. Bad choice of cardinality for the calibration set

It is established in Appendix B.6 of the main article that the FDR can only be controlled if the cardinality of the calibration set takes specific values. This section verifies this claim in the case of breakpoint based anomaly detection.

To evaluate the degradation of the FDR control due to a bad choice of the calibration set cardinality, time series are generated according to the framework design introduced in Section 4.1. For each generated time series, with the target FDR $\alpha = 0.2$ (resp. $\alpha = 0.1$) and the active set cardinality $m = 100$, two anomaly detectors are applied: one with a calibration set cardinality equal to 999 (resp. 1999) respecting the recommendation. The second with a calibration set cardinality equal to 1000 (resp. 2000), not respecting the recommendation. Indeed, since the proportion of anomalies is equal to $\pi = 0.01$, the goal of a FDR equal to $\alpha = 0.2$ (resp. $\alpha = 0.1$) can be achieved using Benjamini-Hochberg with $\alpha' = 0.1$ (resp. 0.05) according to Appendix B.6. The calibration set cardinality should then be equal to 999 (resp. 1999) according to Eq. B.15.

The results in Table 12 show that the FDR is controlled at the desired level for $n = 999$ and $n = 1999$, while the FNR is higher. This confirms that the FDR can only be controlled by selecting the parameter n using the rule in Appendix B.6. To reduce the FNR while maintaining control of the FDR, the values n must be chosen among the values $\{1999, 2999, 3999, \ldots\}$ as discussed

α	n	FDR	FNR
0.2	999	0.21	0.030
	1000	0.30	0.0
0.1	1999	0.1	0.1
	2000	0.16	0.03

TABLE 12
FDR and FNR according to the calibration set cardinality.

in the paper [5].

References

[1] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, The Journal of Machine Learning Research 13 (1) (2012) 723–773.

[2] A. M. Safin, E. Burnaev, Conformal kernel expected similarity for anomaly detection in time-series data, Advances in Systems Science and Applications 17 (3) (2017) 22–33.

[3] A. Gross, J. W. Tukey, The estimators of the Princeton robustness study, Department of Statistics, Univ., 1973.

[4] L. H. Shoemaker, T. P. Hettmansperger, Robust estimates and tests for the one-and two-sample scale models, Biometrika 69 (1) (1982) 47–53.

[5] E. Krönert, A. Célisse, D. Hattab, Fdr control for online anomaly detection, arXiv preprint arXiv:2312.01969 (2023).

[6] V. Ishimtsev, A. Bernstein, E. Burnaev, I. Nazarov, Conformal k-nn anomaly detector for univariate data streams, in: Conformal and Probabilistic Prediction and Applications, PMLR, 2017, pp. 213–227.

[7] D. S. Matteson, N. A. James, A nonparametric approach for multiple change point analysis of multivariate data, Journal of the American Statistical Association 109 (505) (2014) 334–345.

[8] P. C. Mahalanobis, On the generalized distance in statistics, Sankhyā: The Indian Journal of Statistics, Series A (2008-) 80 (2018) S1–S7.

[9] F. Mosteller, J. W. Tukey, Data analysis and regression. a second course in statistics, Addison-Wesley series in behavioral science: quantitative methods (1977).

[10] R. A. Fisher, et al., The design of experiments., no. 7th Ed, Oliver and Boyd. London and Edinburgh, 1960.