Online FDR control for probabilistic time series anomaly detection

Authors

May 20, 2021

Abstract

This article proposes online false discovery rate control (FDRC) rules adapted for probabilistic anomaly detection tasks. Online FDRC rules allows researchers to control the properties of a sequence of statistical tests. In the context of anomaly detection, the null hypothesis is that an observation is non-anomalous and the alternative is that it is anomalous. We propose methods to ensure that power remains high even when the alternative is extremely rare and the test statistics serially dependent. We document the ability of our method to control the false discovery rate on artificial and real-world data.

1 Introduction

We consider the problem of false discovery rate control for anomaly detection in time-series data. Let $\{z_t\}_{t=1}^{\infty}$ be a stream of values that the scientist is able to observe in an online fashion. At each time step t, it must be decided whether z_t is anomalous or not, based solely on past observations.

Wang [2019] propose to perform a STL decomposition of the observations, resulting in a trend, a seasonality, and a residual components. From a model of the residual component as a mixture model of noise and anomalies, they control the FDR with z-values [Sun and Cai, 2007]. Due to the temporal nature of the data, p-values come one at a time with some granularity, which makes the problem inherently online.

Shipmon et al. [2017] Chandola et al. [2009] Braei and Wagner [2020] Siffer et al. [2017] iid assumption not realistic. No FDR control guarantee (only heuristic observations) Sadik and Gruenwald [2014]

From the p-values, hypotheses are rejected based on decision rules that should fit at best the location on the precision-recall curve required by the customer. A natural way to do so is to estimate the desired precision $1 - \alpha$, and take decisions controlling the *false discovery rate* (expected ratio of false positives to total discoveries) at level α. For anomaly detection purposes, α could be high in order to maximize the recall in settings where it is critical to avoid missing any anomaly, or lower in settings where avoid false alarms, even at the price of missed anomalies, is important.

2 Multiple hypothesis testing for anomaly detection

The problem of anomaly detection in time series can be posed as an online multiple hypotheses testing problem. Specifically, at each time step, we observe a new data point z_t and emit a null hypothesis (z_t is not anomalous) H_t. This hypothesis is tested with a p-value computed from past observations, and may be rejected (detection of an anomaly) if there is enough evidence against the null. We note $\{H_t\}_{t=1}^{\infty}$ the stream of hypotheses and $\mathcal{H}^0$ the unknown set of truly null hypotheses (no anomaly).

Deriving p-values. At each time t, hypothesis H_t is tested based on a p-value which must be derived from past observations only. The concept of anomaly is ambiguous and suffers from a subjective and imprecise definition. Nonetheless, we assume that anomalies are always characterized by rare events that deviate significantly from expectations. In other words, they are very unlikely and can hardly be anticipated. For this reason, the problem of anomaly detection is closely related to *probabilistic* forecasting. A probabilistic forecast is an estimation of the probability distribution of z_T given the past, that is

$$\mathbf{Prob}\{z_T \mid z_t, t < T\},$$

or equivalently its c.d.f. F_T. There is a rich literature describing probabilistic forecasts and how to generate them [Box et al., 1974, Hyndman et al., 2008, Seeger et al., 2016, Salinas et al., 2017]. Note that not all models require to assume a probability distribution, and there exists alternative paradigms for anomaly detection. We do not aim to contribute to the literature on probabilistic anomaly detection models, but rather show that they naturally lend themselves to a decision approach based on false discovery rate control. Under the null hypothesis, such a forecast should be a decent estimate of z_T. We may derive a sequence of p-values $\{p_t\}_{t=1}^{\infty}$, such that p_t is the probability to observe an event more extreme than z_T, hence low p_t is an indication that H_t must be rejected. From the predicted c.d.f. F_t, two-sided p-values may be derived as

$$p_t = 2\min\left(F_t(z_t), 1 - F_t(z_t)\right).$$

From this, a sequence of p-values $\{p_t\}_{t=1}^{\infty}$ can be derived iteratively as observations come. In this setting, the problem is unsupervised, in the sense that we don't need the time series to come with labels to indicate anomalous points. For certain usages, it makes sense to define anomalies as values either extremely high or low compared to the average. For example, if an application is designed to monitor the resources usage of a service, only large values should send an alert. Accordingly, one-sided p-values can be derived to account for only one end of the distribution. More generally, p-values need only to be stochastically larger than the uniform distribution under the null hypothesis, that is:

$$\text{If } H_t \text{ is truly null then } \mathbf{Prob}\{p_t \leq u\} \leq u \text{ for all } u \in [0,1]. \tag{1}$$

With this definition of anomalies, the forecasting model plays a major role to feed p-values in line with the application. The question of which criterion on the p-values should be used to reject hypotheses remains.

False discovery rate control. In the case of a single point to test, the most common convention is to report a discovery (an anomaly here) if the associated p-value is smaller than some preselected threshold α. The effect is said to be *statistically significant* and the probability to falsely declare the point anomalous is less than α. In other words, this method controls the type I error at level α, which is an immediate consequence of the definition of the p-value (1). In the case of multiple hypotheses, making rejections based on a sole threshold α doesn't provide similar guarantees and the fraction of false positives may be arbitrarily close to 1. This particularly applies to anomaly detection because the proportion of alternative hypotheses tends to be very low. For this reason, rejections are typically done with procedures controlling other sort of metrics in multiple hypotheses settings. One of those is the family-wise error rate (FWER) [Hochberg and Tamhane, 1987], which is the probability to make at least one false discovery. As the probability to reject a truly null hypothesis approaches 1 when the number of tests grows, controlling such a metric is too conservative and leads to very few discoveries for very long (possibly infinite) streams of data.

In the setting of anomaly detection, we claim that a natural quantity to control is the false discovery rate (FDR), as introduced by Benjamini and Hochberg [1995] along with a

selection procedure to control it. It is formally defined as the expected ratio of false positives to total discoveries:

$$\mathrm{FDR} \equiv \mathbf{E}\left[\mathrm{FDP}\right] \equiv \mathbf{E}\left[\frac{|\mathcal{H}^0 \cap \mathcal{R}|}{|\mathcal{R}| \vee 1}\right],$$

where $\mathcal{H}^0$ and $\mathcal{R}$ are the sets of truly null hypotheses and rejections respectively, and FDP is the false discovery proportion of the samples. Most offline procedures that control the FDR rely on an ordering of the p-values; they all need to be known in advance.

Batching. Rather than making a decision at every time step, hybrid schemes can be considered [Zrnic et al., 2019]. They consist in testing sequences of hypotheses by batches whose length is variable, which can take advantage of a partial ordering of the p-values. However, gains in terms of power are not significant and this method introduces a delay in the decision which not desirable in online anomaly detection.

Anomalies over multiple time steps. It is sometimes desirable to characterize anomalies as unlikely events over several time steps. We believe that the online multiple hypotheses testing approach is adaptable and remains appealing. To this end, forecasts can be multivariate [Salinas et al., 2019] and p-values derived over a range of steps. Another option is to aggregate a series of p-values [Meinshausen et al., 2008, Heard and Rubin-Delanchy, 2017] and perform FDR control consequently.

3 Online FDR control

At every time step t, a decision threshold α_t is chosen to test hypothesis H_t; it is rejected if the associated p-value is less than α_t, and the rejection indicator is noted $R_t = \mathbf{1}\{p_t \leq \alpha_t\}$. Decision thresholds must depend solely on past observations, so that for any time t there is a function f_t such that

$$\alpha_t = f_t(R_1, \ldots, R_{t-1}) \in \mathcal{F}^{t-1}, \tag{2}$$

where $\mathcal{F}^t \equiv \sigma(R_1, \ldots, R_t)$ is the σ-field (information known by the scientist) at time t.

In the online FDR control problem, we define a time-dependent notion of the number of rejections $R(T)$ and of the number of false positives $V(T)$ up to time T:

$$R(T) = \sum_{t=1}^{T} R_t \quad \text{and} \quad V(T) = \sum_{t=1}^{T} R_t \mathbf{1}\{t \in \mathcal{H}^0\}.$$

Naturally, the FDP and FDR at time T are defined as $\mathrm{FDP}(T) = \frac{V(T)}{R(T)\vee 1}$ and $\mathrm{FDR}(T) = \mathbf{E}\left[\mathrm{FDP}(T)\right]$ respectively. We wish to pick the rejection thresholds $\{\alpha_t\}_{t=1}^{\infty}$ in such a way that $\mathrm{FDR}(T) \leq \alpha$ at any time T for some FDR target $\alpha \in (0,1)$. In this section, we make the standard conditional super-uniformity assumption:

$$\text{If } H_t \text{ is truly null then } \mathbf{Prob}\{p_t \leq u \mid \mathcal{F}^{t-1}\} \leq u \text{ for all } u \in [0,1], \tag{3}$$

which is satisfied if $\{p_t\}$ are valid p-values and null p-values are independent of all others. The independence assumption can be problematic for time series data in certain settings, we propose a method to relax it in Section 5 but will maintain it until then to simplify the exposition.

Generalized α-investing. Foster and Stine [2008] introduced the idea of α-investing in order to control the marginal FDR, which is defined as the ratio of the expectations, in contrast to the expectation of the ratio:

$$\mathrm{mFDR} = \frac{\mathbf{E}\left[V(T)\right]}{\mathbf{E}\left[R(T)\right] \vee 1}.$$

[Aharoni and Rosset, 2014] showed α-investing to be part of a broader class of algorithms, known as generalized α-investing (GAI) rules. Algorithms of the GAI class record a quantity w_t called the wealth, which represents the allocation potential for future decision thresholds. They start with an initial wealth $w_0 \in (0, \alpha)$ and at every time step t, an amount ϕ_t is spent from the remaining wealth in order to test the hypothesis H_t at level $\alpha_t \in \mathcal{F}^{t-1}$. When a rejection occurs, an amount of ψ_t is earned back, which induces the update $w_t = w_{t-1} - \phi_t + R_t \psi_t$. The wealth must always be non-negative, which imposes $\phi_t \leq w_{t-1}$, and the mFDR is controlled only if rewards satisfy the inequality

$$\psi_t \leq \min\left(\phi_t + b_t, \frac{\phi_t}{\alpha_t} + b_t - 1\right).$$

The first work proposing algorithms controlling the exact FDR is done by Javanmard and Montanari [2015], introducing LORD and LOND. These algorithms are part of the GAI framework, and satisfy a *monotonic* property, meaning that f_t is a coordinatewise nondecreasing function in Equation (2). The same authors later show that any monotonic GAI rule controls the FDR, under the assumption that p-values are independent [Javanmard and Montanari, 2016].

Oracle rules. Monotonic GAI rules can be seen as a special case of decision rules relying oracle estimates of the FDP, as shown by Ramdas et al. [2017, 2018], Tian and Ramdas [2019]. Maintaining the quantity

$$\mathrm{FDP}^\star(T) = \frac{\sum_{t \leq T, t \in \mathcal{H}^0} \alpha_t}{R(T) \vee 1}$$

below α for all T guarantees to control the FDR at every time step. This oracle cannot be evaluated as the set of null hypotheses $\mathcal{H}^0$ is not known. A naive way to upper-bound $\mathrm{FDP}^\star$ is to suppress the condition $t \in \mathcal{H}^0$ in the summation, which leads to algorithms of the class of LORD. More refined upper-bounds give rise to adaptive algorithms known as SAFFRON and ADDIS [Ramdas et al., 2018, Tian and Ramdas, 2019]. In short, LORD can be seen as an online version of the BH procedure [Benjamini and Hochberg, 1995], SAFFRON as an online version of Storey-BH [Storey, 2002], and ADDIS is able to slightly improve on SAFFRON's performances when null p-values are not exactly uniformly distributed, but stochastically larger than uniform. The oracle estimates are given by

$$\widehat{\mathrm{FDP}}_{\text{LORD}}(T) = \frac{\sum_{t \leq T} \alpha_t}{R(T) \vee 1} \quad \text{and} \quad \widehat{\mathrm{FDP}}_{\text{ADDIS}}(T) = \frac{\sum_{t \leq T} \alpha_t \frac{\mathbf{1}\{\lambda_t < p_t \leq \tau_t\}}{\tau_t - \lambda_t}}{R(T) \vee 1},$$

and SAFFRON can be seen as a special case of ADDIS where $\tau_t = 1$ for all t, that is

$$\widehat{\mathrm{FDP}}_{\text{SAFFRON}}(T) = \frac{\sum_{t \leq T} \alpha_t \frac{\mathbf{1}\{\lambda_t < p_t\}}{1 - \lambda_t}}{R(T) \vee 1}.$$

Oracle rules offer a lot of flexibility: FDR control holds for any sequence $\{\alpha_t\}_{t=1}^\infty$ allocated such that $\widehat{\mathrm{FDP}} \leq \alpha$, provided that the functions f_t (in Equation 2) are non-decreasing. Table 1 presents special instances of the three state of the art algorithms described above. All of them make use of a non-increasing sequence $\{\gamma_t\}_{t=1}^\infty$ summing to 1, where we set $\gamma_t = 0$ for all $t \leq 0$ for convenience. We note

$$S_{j+}^{\lambda,\tau}(t) = \mathbf{1}\{t > \rho_j\} + \sum_{i=\rho_j+1}^{t-1} \mathbf{1}\{\lambda_i < p_i \leq \tau_i\},$$

where $\rho_j = \min\{t \geq 0 \mid \sum_{i=1}^t R_i \geq j\}$ is the time of jth rejection (∞ if it doesn't exist). The respective power of these algorithms highly depends on $\{\gamma_t\}, \{\lambda_t\}, \{\tau_t\}$ and on the nature

Algorithm	Threshold α_t
LORD	$w_0(\gamma_t - \gamma_{t-\rho_1}) + \alpha \sum_j \gamma_{t-\rho_j}$
LOND	$\alpha\gamma_t \left(R\left(t-1\right)+1\right)$
SAFFRON	$(1-\lambda_t)\left(w_0\left(\gamma_{S^{\lambda,1}_{0+}(t)} - \gamma_{S^{\lambda,1}_{1+}(t)}\right) + \alpha \sum_j \gamma_{S^{\lambda,1}_{j+}(t)}\right) \wedge \lambda_t$
ADDIS	$(\tau_t-\lambda_t)\left(w_0\left(\gamma_{S^{\lambda,\tau}_{0+}(t)} - \gamma_{S^{\lambda,\tau}_{1+}(t)}\right) + \alpha \sum_j \gamma_{S^{\lambda,\tau}_{j+}(t)}\right) \wedge \lambda_t$

Table 1: Examples of decision thresholds for LORD, SAFFRON and ADDIS controllers. Note that setting $\lambda_t = 0$ recovers LORD but ADDIS and SAFFRON algorithms impose $\alpha_t \leq \lambda_t$; for this reason they cannot be seen as a strict generalization of LORD.

of the data since they are general and make minimal assumptions. In the remaining, we use the default and robust sequences derived by Ramdas et al. [2017, 2018], Tian and Ramdas [2019], that is, $\gamma_t \propto \frac{\log(t \wedge 2)}{t \exp\sqrt{\log(t)}}$ for LORD, $\lambda_t = 1/2$ for SAFFRON, $\lambda_t = 1/4$, $\tau_t = 1/2$ for ADDIS, and $\gamma_t \propto t^{-s}$ for both SAFFRON and ADDIS, where $s = 1.6$ unless otherwise specified.

Limitations of FDR rules. Anomalies are by definition rare events and the proportion of alternative hypotheses in the stream is expected to be very low. This induces two distinct problems for the discovery of anomalies:

P.1 The lower the proportion of alternatives, the harder it is to distinguish them from nulls with extreme values. This difficulty is inherent to FDR control, be it offline or online, and makes anomaly detection challenging.

P.2 In the online setting, decision thresholds α_t decay to zero except at rejection times. Hence, if no rejection is made for a long time, the thresholds will be essentially null and no additional discovery is possible. This phenomenon is known as α-*death* and causes a loss of power in anomaly detection regimes.

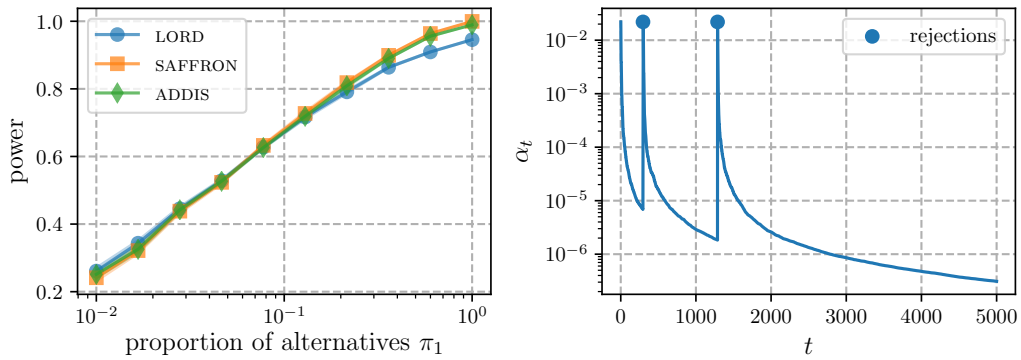

Figure 1: Illustration of the difficulties encountered in FDR control for anomaly detection. *(Left)* The power decreases to zero when the proportion of alternative hypotheses π_1 is low for all FDR controllers. Stream of artificially generated p-values from observations either from $\mathcal{N}(0,1)$ (null) or $\mathcal{N}(3,1)$ (anomaly) with proportion π_1. *(Right)* SAFFRON's decision thresholds α_t as a function of time for a target $\alpha = 0.1$ and the canonical sequence $\{\gamma_t\}$.

Figure 1 illustrates both of these problems. In the regime where the proportion of alternatives is low, LORD, SAFFRON and ADDIS all exhibit the same (poor) performances. Selecting a sequence $\{\gamma_t\}$ going to zero as slowly as possible helps to overcome P.2 to some

extent but worsens P.1 in contrast. For anomaly detection purposes, α should arguably be high in order to maximize the recall and avoid missing any anomaly.

4 Circumventing α-death

In this section, we tackle the issue of α-death described in Section 3, which is the biggest practical problem of FDR control for anomaly detection. This suggests a change in the FDR metric, as indicated by Ramdas et al. [2017], who introduced a memory-decay version of the FDR, meaning that the past is forgotten with a discount factor $\delta \in (0, 1)$.

4.1 Memory decay

Decaying rejections count $R_\delta(T)$ and false positives count $V_\delta(T)$ are defined as follows:

$$R_\delta(T) = \sum_{t=1}^{T} \delta^{T-t} R_t \quad \text{and} \quad V_\delta(T) = \sum_{t=1}^{T} \delta^{T-t} R_t \mathbf{1}\{t \in \mathcal{H}^0\}\,.$$

In words, more weight is given to recent rejections, and the past shrinks exponentially. Consistently, the memory FDR at level δ is defined as

$$\mathrm{FDR}_\delta(T) = \mathbf{E}\left[\mathrm{FDP}_\delta(T)\right] = \mathbf{E}\left[\frac{V_\delta(T)}{R_\delta(T) \vee 1}\right]\,. \tag{4}$$

Ramdas et al. [2017] show that adapting the algorithm LORD with the rejection thresholds

$$\alpha_t = w_0 \left(\delta^{t-\min(\rho_1,t)}\gamma_t - \delta^{t-\rho_1}\gamma_{t-\rho_1}\right) + \alpha \sum_{j\geq 1} \delta^{t-\rho_j}\gamma_{t-\rho_j} \tag{5}$$

allows to control the memory FDR at the desired level α. It prevents the phenomenon known as piggybacking to happen. However, after the first rejection, the decision threshold 5 suffers from an exponential decay and quickly goes to zero if no rejection is done. The authors suggest that when the wealth is too low, the rejection process may paused until the wealth has recovered to some pre-specified threshold and then resumed. This scheme poses two problems: *(i)* during an abstention phase we are not able to reject any hypothesis, even those with very low p-values that might be critical and *(ii)* there is no theoretical guarantee that the FDR is still controlled after resetting the process with initial values.

In this section we establish one of the key contributions of this paper, showing that the idea of memory decay, *i.e.*, forgetting the past, allows to avoid α-death. The standard expression of the FDR contains an asymmetry before and after the first rejection time ρ_1. For this reason, our results are more naturally expressed in terms of *smoothed* FDR, that is

$$\mathrm{sFDR}_\delta(T) = \mathbf{E}\left[\mathrm{sFDP}_\delta(T)\right] = \mathbf{E}\left[\frac{V_\delta(T)}{R_\delta(T) + \eta}\right]\,,$$

for some smoothing parameter $\eta > 0$. We show in Appendix A.3 how to deal with the case where the quantity to control is the memory FDR at level δ as defined in (4). Both V_δ and R_δ decay to zero, and so does the FDP estimate, as its denominator is lower bounded (by either 1 or η). Intuitively, this translates into a minimal rejection threshold that doesn't depend on the last rejection time.

Choice of δ.

4.2 LORD memory decay

As in Section 3, we express algorithms in terms of oracle rules.

$$\widehat{\mathrm{FDP}}^{\delta}_{\mathrm{LORD}}(T) = \frac{\sum_{t \le T} \delta^{T-t} \alpha_t}{R(T) + \eta}$$

Proposition 4.1. *Under independence of the p-values, picking decision thresholds α_t such that $\widehat{FDP}^{\delta}_{\mathrm{LORD}}(T) \le \alpha$ ensures that $sFDR_\delta(T) \le \alpha$ at any time T.*

Proof given in Appendix A.1.

Special instance. As for all oracle rules presented above, Proposition 4.1 gives a lot of freedom regarding the choice of the sequence $\{\alpha_t\}$. We show here a special instance that naturally generalizes the standard LORD algorithm. Let $\{\tilde{\gamma}_t\}_{t=1}^{\infty}$ be a positive non-increasing sequence such that

$$\sum_{t=1}^{T} \delta^{T-t} \tilde{\gamma}_t \le 1 \quad \forall\, T. \tag{6}$$

Remark. *Such a sequence admits a limit no larger than $1-\delta$, and may be chosen lower-bounded by $1-\delta$. A natural choice is for example $\tilde{\gamma}_t = \max(\gamma_t, 1-\delta)$.*

At time step t, the hypothesis H_t is rejected if $p_t \le \alpha_t$ where

$$\alpha_t = \alpha\eta\tilde{\gamma}_t + \alpha \sum_j \delta^{t-\rho_j} \gamma_{t-\rho_j} \tag{7}$$

In the case where $\tilde{\gamma}_t = \max(\gamma_t, 1-\delta)$, this means that there is a minimal rejection threshold, namely $\alpha\eta(1-\delta)$, no matter how long ago the last rejection occurred.

4.3 SAFFRON and ADDIS memory decay

Memory decay for ADDIS at level δ.

$$\widehat{\mathrm{FDP}}^{\delta}_{\mathrm{ADDIS}}(T) = \frac{\sum_{t \le T} \delta^{T-t} \alpha_t \frac{\mathbf{1}\{\lambda_t < p_t \le \tau_t\}}{\tau_t - \lambda_t}}{R(T) + \eta},$$

As shown in Section 3, SAFFRON can be recovered by setting $\tau_t = 1$ for all t.

Proposition 4.2. *Under independence of the p-values, picking decision thresholds α_t such that $\widehat{FDP}^{\delta}_{\mathrm{ADDIS}}(T) \le \alpha$ and $\alpha_t \le \lambda_t$ ensures that $sFDR_\delta(T) \le \alpha$ at any time T.*

Proof given in Appendix A.2.

Special instance. Let $\{\tilde{\gamma}_t\}_{t=1}^{\infty}$ be a sequence satisfying Property 6. Then the sequence of thresholds

$$\alpha_t = \alpha(\tau_t - \lambda_t)\Big(\eta\tilde{\gamma}_{S^{\lambda,\tau}_{0+}} + \sum_j \delta^{t-\tau_j} \gamma_{S^{\lambda,\tau}_{j+}}\Big) \wedge \lambda_t\,. \tag{8}$$

satisfies the inequality required by Proposition 4.2. Here again, in the case where $\tilde{\gamma}_t = \max(\gamma_t, 1-\delta)$ and λ_t, τ_t are constant, this means that there is a minimal rejection threshold, namely $\alpha\eta(\tau-\lambda)(1-\delta)$.

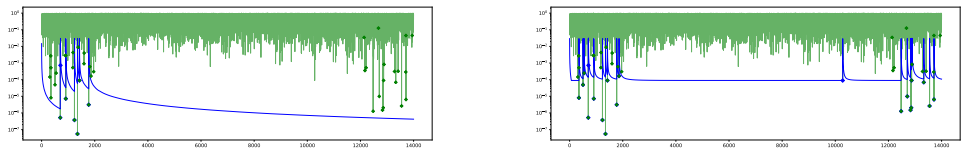

Figure 2: SAFFRON rejection threshold with memory decay (right) and without (left). Without memory decay, the rejection threshold decreases so much that further discoveries are impossible.

4.4 Practical implementation

Online FDR controllers based on LORD, SAFFRON or ADDIS all rely on the non-negative and non-increasing sequence $\{\gamma_t\}_{t=1}^{\infty}$ summing to 1. In practice, an infinite sequence is unrealistic, and a reasonable implementation of these algorithms would use a finite support sequence summing to 1, that is $\{\gamma_t\}_{t=1}^{B}$ for some large constant B. This is primarily because the buffer of rejection times $\{\rho_t\}$ cannot be infinite, for memory purposes, but also because computing the exact normalization constant may often be impossible. In this context, if no rejection is done for B time steps, the thresholds α_t become exactly zero. No further rejection is possible even with extremely low p-values, and the α-death cannot be recovered.

In contrast,

5 Local dependency

The conditional super-uniformity assumption (3) is overly optimistic. Time-series data contains dependencies that may leak to the p-values. Javanmard and Montanari [2015, 2016] prove that thresholds of both LOND and LORD can be adapted to control FDR under arbitrary dependence of p-values, in the same way as Benjamini and Yekutieli [2001] in the offline setting. It consists essentially in dividing the thresholds α_T by the quantity $q(T) \triangleq \sum_{t=1}^{T} \frac{1}{t}$ at all steps. By extension, SAFFRON and ADDIS could be corrected the same way. However, this calibration makes thresholds essentially null in the long term as $q(T)$ goes to infinity. The reason is that arbitrary dependency is an overly restrictive assumption. As shown in Section 2, the most natural way to derive p-values is to generate probabilistic forecast. At time step t such a forecast is typically based on a *context*, that is, a sequence of past observations of finite length L_t. This means that the p-value p_t depends on the last L_t p-values at most, or, in other words,

$$p_t \perp\!\!\!\perp p_1, \ldots, p_{t-L_t-1}. \tag{9}$$

This fits into the *local dependency* framework of Zrnic et al. [2021].

The sole requirement regarding dependency ranges is that $L_{t+1} \leq L_t + 1$, which means that p_{t+1} cannot depend on p-values on which p_t did not depend. Most commonly, the dependency range is constant over time, that is, $L_t = L$ for all t. Notice that $L = 0$ and $L = \infty$ recover the independence and arbitrary dependence settings respectively. We note $\mathcal{G}^t = \sigma(R_1, \ldots, R_{t-L_t-1})$ the non-conflicting filtration, *i.e.*, the information known at time t which is not in conflict with the current p-value (it is indeed a filtration because $t - L_t - 1$ is non-decreasing). If we impose that $\alpha_t \in \mathcal{G}^{t-1}$ then the mFDRis controlled at all times, as shown in Zrnic et al. [2021]. We propose to pick these thresholds in such a way that prevents α-death as in Section 4.

$$\alpha_t = \alpha\eta\tilde{\gamma}_t + \alpha \sum_j \delta^{t-\rho_j} \gamma_{t-\rho_j}$$

6 Experiments

We perform a series of experiments on artificial and real-world data to support our claims.

6.1 Experiment on importance of correct p-value

Not an essential experiment: we could try to tease apart how robust different FDRC methods are to the same, miscallibrated model. This could be on synthetic data where we have true generating model with which we generated the original data, injected synthetic anomalies and then the true model with its p-values versus a, say, GP or RNN model. Feedback from others on this experiment?

6.2 FDRC for anomaly detection experiment (from Tim)

Would we need to compare against this approach in the experiments: `https://milets19.github.io/papers/milets19_paper_6.pdf` They seem to have R code and it would be highly related. But, we would have to tease apart where an improvement comes from: better FDRC or a better model compared to their STL. Thoughts on this, Laurent and Baris?

The experiment I have in mind is: given a realistic data set (maybe from the NCAD paper or the GP paper), how much better is an FDRC controlled anomaly detection versus a statically defined threshold? Here, "better" would mean, more intuitive. So basically, if we set alpha to some number 0.05 in the static thresholding approach, how many anomalies do we actually get versus what's the ration of anomalies do we get with FDRC?

Another experiment similar to what Laurent suggested: have a time series that gets more and more noisy over time (so the tails of the distribution get fatter), but the ratio of anomalies stays constant – I would expect FDRC to help detect only the anomalies whereas a constant alpha would start capturing noise as anomalies. Can we show this? This would be a synthetic data experiment. Does this make sense Laurent?

The above is similar to what Laurent suggested, but a bit different from the original idea. The original idea was to have the rate of anomalies increase over time and for FDRC then ensure that we only select the most extreme anomalies. At least on second thought I can't really think of an application where this is useful – e.g., if an app running on AWS starts to get buggy, we want to be alerted at every bug and not just the most extreme bugs. Maybe, Laurent, you can think of another example from the Poirot context?

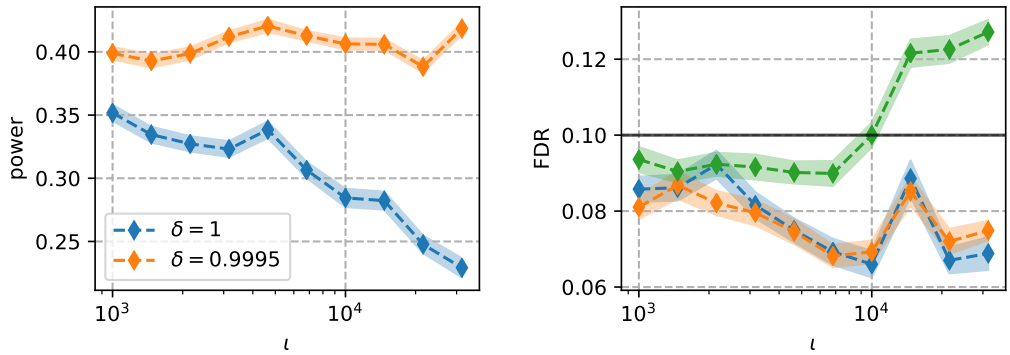

Figure 3:

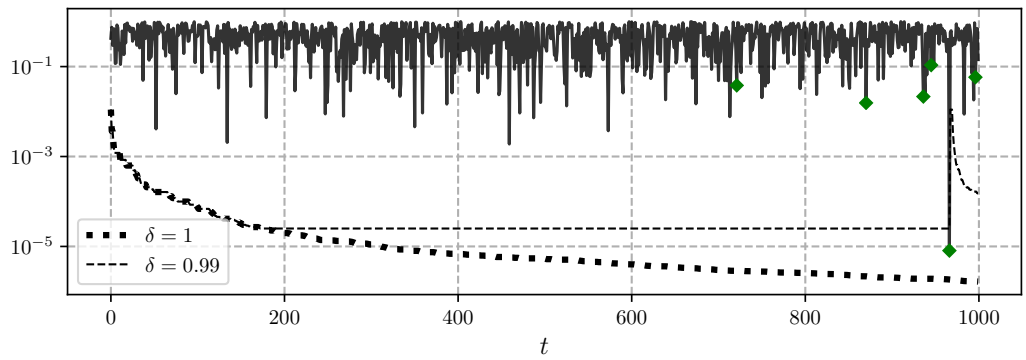

Figure 4:

7 Conclusion

References

Ehud Aharoni and Saharon Rosset. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 76:771–794, 2014.

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

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29:24, 08 2001. doi: 10.1214/aos/1013699998.

G. E. P. Box, G. M. Jenkins, and J. F. MacGregor. Some recent advances in forecasting and control. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(2): 158–179, 1974. ISSN 00359254, 14679876.

Mohammad Braei and Sebastian Wagner. Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art. *arXiv e-prints*, art. arXiv:2004.00433, April 2020.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41:72, 07 2009. doi: 10.1145/1541880.1541882.

Dean Foster and Robert Stine. Alpha-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B*, 70:429–444, 04 2008. doi: 10.1111/j.1467-9868.2007.00643.x.

Nicholas Heard and Patrick Rubin-Delanchy. Choosing Between Methods of Combining p-values. *arXiv e-prints*, art. arXiv:1707.06897, July 2017.

Y. Hochberg and A. Tamhane. *Multiple Comparison Procedures*. John Wiley, 1987.

Rob Hyndman, Anne Koehler, Keith Ord, and Ralph Snyder. *Forecasting with exponential smoothing. The state space approach*. Springer, 01 2008. doi: 10.1007/978-3-540-71918-2.

Adel Javanmard and Andrea Montanari. On Online Control of False Discovery Rate. *arXiv e-prints*, art. arXiv:1502.06197, February 2015.

Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, 46, 03 2016. doi: 10.1214/17-AOS1559.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *arXiv e-prints*, art. arXiv:0811.2177, November 2008.

Aaditya Ramdas, Fanny Yang, Martin J. Wainwright, and Michael I. Jordan. Online control of the false discovery rate with decaying memory. *arXiv e-prints*, art. arXiv:1710.00499, October 2017.

Aaditya Ramdas, Tijana Zrnic, Martin Wainwright, and Michael Jordan. SAFFRON: an adaptive algorithm for online control of the false discovery rate. *arXiv e-prints*, art. arXiv:1802.09098, February 2018.

Shiblee Sadik and Le Gruenwald. Research issues in outlier detection for data streams. *SIGKDD Explor. Newsl.*, 15(1):33–40, March 2014. ISSN 1931-0145. doi: 10.1145/2594473.2594479. URL `https://doi.org/10.1145/2594473.2594479`.

David Salinas, Valentin Flunkert, and Jan Gasthaus. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *arXiv e-prints*, art. arXiv:1704.04110, April 2017.

David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-Dimensional Multivariate Forecasting with Low-Rank Gaussian Copula Processes. *arXiv e-prints*, art. arXiv:1910.03002, October 2019.

Matthias W Seeger, David Salinas, and Valentin Flunkert. Bayesian intermittent demand forecasting for large inventories. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4646–4654. Curran Associates, Inc., 2016.

Dominique T. Shipmon, Jason M. Gurevitch, Paolo M. Piselli, and Stephen T. Edwards. Time Series Anomaly Detection; Detection of anomalous drops with limited features and sparse examples in noisy highly periodic data. *arXiv e-prints*, art. arXiv:1708.03665, August 2017.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1067–1075, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098144. URL `https://doi.org/10.1145/3097983.3098144`.

John Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, 64:479–498, 08 2002. doi: 10.1111/1467-9868.00346.

Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102:901 – 912, 2007.

Jinjin Tian and Aaditya Ramdas. Addis: an adaptive discarding algorithm for online fdr control with conservative nulls. In *NeurIPS*, 2019.

Weinan Wang. Online fdr controlled anomaly detection for streaming time series. 2019.

Tijana Zrnic, Daniel L. Jiang, Aaditya Ramdas, and Michael I. Jordan. The Power of Batching in Multiple Hypothesis Testing. *arXiv e-prints*, art. arXiv:1910.04968, October 2019.

Tijana Zrnic, Aaditya Ramdas, and Michael I Jordan. Asynchronous online testing of multiple hypotheses. *Journal of Machine Learning Research*, 22(33):1–39, 2021.

A Proofs

Our proofs are based on the following three lemmas, proposed by Ramdas et al. [2017, 2018] and Tian and Ramdas [2019].

Lemma A.1. *Suppose that the sequence $\{p_t\}_{t=1}^{\infty}$ is composed of independent p-values. Let f_t be a sequence of non-decreasing functions such that $\alpha_t = f_t(R_1, \ldots, R_{t-1})$. Then for any non-decreasing function h,*

$$\mathbf{E}\left[\frac{\alpha_t}{h(R_{1:T})} \mid \mathcal{F}^{t-1}\right] \geq \mathbf{E}\left[\frac{\mathbf{1}\{p_t \leq \alpha_t\}}{h(R_{1:T})} \mid \mathcal{F}^{t-1}\right]$$

Lemma A.2. *Suppose that the sequence $\{p_t\}_{t=1}^{\infty}$ is composed of independent p-values. let f_t, g_t be two sequences of non-decreasing functions such that $\alpha_t = f_t(R_1, \ldots, R_{t-1}, C_1, \ldots, C_{t-1})$, $\lambda_t = g_t(R_1, \ldots, R_{t-1}, C_1, \ldots, C_{t-1})$, and $\alpha_t \leq \lambda_t$. Then for any non-decreasing function h,*

$$\begin{aligned}\mathbf{E}\left[\frac{\alpha_t \mathbf{1}\{p_t > \lambda_t\}}{(1-\lambda_t)h(R_{1:T})} \mid \mathcal{F}^{t-1}\right] &\geq \mathbf{E}\left[\frac{\alpha_t}{h(R_{1:T})} \mid \mathcal{F}^{t-1}\right] \\ &\geq \mathbf{E}\left[\frac{\mathbf{1}\{p_t \leq \alpha_t\}}{h(R_{1:T})} \mid \mathcal{F}^{t-1}\right]\end{aligned}$$

Note that the second inequality is a direct consequence of Lemma A.1.

Lemma A.3. *Suppose that the sequence $\{p_t\}_{t=1}^{\infty}$ is composed of independent p-values. let $f_t, g_t, \tilde{g}_t$ be three sequences of non-decreasing functions such that $\alpha_t = f_t(R_1, \ldots, R_{t-1}, C_1, \ldots, C_{t-1})$, $\lambda_t = g_t(R_1, \ldots, R_{t-1}, C_1, \ldots, C_{t-1})$, and $\alpha_t \leq \lambda_t < \tau_t$. Then for any non-decreasing function h,*

$$\begin{aligned}\mathbf{E}\left[\frac{\alpha_t \mathbf{1}\{\lambda_t p_t \leq \tau_t\}}{(\tau_t - \lambda_t)(h(R_{1:T})} \mid \mathcal{F}^{t-1}, S_t = 1\right] &\geq \mathbf{E}\left[\frac{\alpha_t}{\tau_t(h(R_{1:T})} \mid \mathcal{F}^{t-1}, S_t = 1\right] \\ &\geq \mathbf{E}\left[\frac{\mathbf{1}\{p_t \leq \alpha_t\}}{h(R_{1:T})}\right]\end{aligned}$$

Detailed proofs of these lemma may be found in Ramdas et al. [2017, 2018] and Tian and Ramdas [2019].

We let $\mathcal{R}(T) = \{t \in [T] \mid R_t = 1\}$ denote the set containing the times of rejections until step T. At any time T we have

$$\begin{aligned}\mathrm{sFDR}_\delta(T) &= \mathbf{E}\left[\frac{V_\delta(T)}{R_\delta(T) + \eta}\right] \\ &= \mathbf{E}\left[\sum_{t=1}^{T} \frac{\delta^{T-t} R_t \mathbf{1}\{t \in \mathcal{H}^0\}}{R_\delta(T) + \eta}\right] \\ &\leq \mathbf{E}\left[\sum_{t=1}^{T} \frac{\delta^{T-t} \mathbf{1}\{p_t \leq \alpha_t\}}{R(T) + \eta}\right].\end{aligned}$$

So if we prove that the sum is less than α at any T then the FDR is always controlled. We show below that this is the case for LORD, SAFFRON and ADDIS.

A.1 Proof of Proposition 4.1

The mapping $h(R_{1:T}) = \sum_{t=1}^{T} \delta^{T-t} \mathbf{1}\{p_t \leq \alpha_t\} + \eta$ is coordinate-wise non-decreasing. If we apply Lemma A.1 with it we find that

$$\begin{aligned} \mathrm{sFDR}_\delta(T) &\leq \mathbf{E}\left[\sum_{t=1}^{T} \frac{\delta^{T-t}\alpha_t}{R(T)+\eta}\right] \\ &= \mathbf{E}\left[\widehat{\mathrm{FDR}}^{\delta}_{\mathrm{LORD}}(T)\right]. \end{aligned}$$

This shows that if the thresholds are chosen such that $\widehat{\mathrm{FDR}}^{\delta}_{\mathrm{LORD}}(T) \leq \alpha$ for all T then the FDR is controlled.

Let us now show that the sequence

$$\alpha_t = \alpha\eta\tilde{\gamma}_t + \alpha \sum_j \delta^{t-\rho_j}\gamma_{t-\rho_j}$$

satisfies this property. This is equivalent to show that the quantity

$$P(T) = \alpha(R_\delta(T)+\eta) - \sum_{t=1}^{T} \delta^{T-t}\alpha_t \geq 0$$

is non-negative for all T. We find that

$$\begin{aligned} P(T) &= \alpha\left(\sum_{t=1}^{T} \delta^{T-t} R_t + \eta\right) - \alpha \sum_{t=1}^{T}\left(\delta^{T-t}\eta\tilde{\gamma}_t + \sum_j \delta^{t-\rho_j}\gamma_{t-\rho_j}\right) \\ &= \alpha\eta\left(1 - \sum_{t=1}^{T}\delta^{T-t}\tilde{\gamma}_t\right) + \alpha \sum_{t\in\mathcal{R}(T)} \delta^{T-t}\left(1 - \sum_{j=1}^{T-t}\gamma_j\right) \\ &\geq 0, \end{aligned}$$

where the inequality follows from the definitions of the sequences $\{\gamma_t\}$ and $\{\tilde{\gamma}_t\}$.

A.2 Proof of Proposition 4.2

The mapping $h(R_{1:T}) = \sum_{t=1}^{T} \delta^{T-t} \mathbf{1}\{p_t \leq \alpha_t\} + \eta$ is coordinate-wise non-decreasing. If we apply Lemma A.2 with it we find that

$$\begin{aligned} \mathrm{sFDR}_\delta(T) &\leq \mathbf{E}\left[\sum_{t=1}^{T} \frac{\delta^{T-t}\alpha_t \mathbf{1}\{p_t > \lambda_t\}}{(1-\lambda_t)(R_\delta(T)+\eta)}\right] \\ &= \mathbf{E}\left[\widehat{\mathrm{FDR}}^{\delta}_{\mathrm{SAFFRON}}(T)\right]. \end{aligned}$$

This shows that if the thresholds are chosen such that $\widehat{\mathrm{FDR}}^{\delta}_{\mathrm{SAFFRON}}(T) \leq \alpha$ for all T then the FDR is controlled.

Let us now show that the sequence

$$\alpha_t = \alpha(\tau_t - \lambda_t)\left(\eta\tilde{\gamma}_{S^{\lambda,\tau}_{0+}} + \sum_j \delta^{t-\tau_j}\gamma_{S^{\lambda,\tau}_{j+}}\right) \wedge \lambda_t$$

satisfies this property. This is equivalent to show that the quantity $P(T) = \alpha(R_\delta(T)+\eta) - \sum_{t=1}^{T} \frac{\delta^{T-t}\alpha_t \mathbf{1}\{p_t > \lambda_t\}}{(1-\lambda_t)}$ is always non-negative. If the decision thresholds $\{\alpha_t\}_{t=1}^{\infty}$ are chosen

such that $P(T) \geq 0$ for all T, it ensures that the FDR is controlled at level α.

$$\begin{aligned}
P(T) &= \alpha\Big(\sum_{t=1}^{T} \delta^{T-t} R_t + \eta\Big) - \alpha \sum_{t=1}^{T} \left(\delta^{T-t}\left(1 - C_t\right)\left(\eta\tilde{\gamma}_{t-C_{0+}(t)} + \sum_j \delta^{t-\rho_j}\gamma_{t-\rho_j-C_{j+}(t)}\right)\right) \\
&= \alpha\eta\Big(1 - \sum_{t=1}^{T} \delta^{T-t}(1-C_t)\tilde{\gamma}_{t-C_{0+}(t)}\Big) + \alpha \sum_{t\in\mathcal{R}(T)} \delta^{T-t}\Big(1 - \sum_{j=1}^{T} \gamma_{j-t-C_{t+}(j)}(1-C_j)\Big) \\
&\geq 0
\end{aligned}$$

A.3 Adapting algorithms to standard FDR

We show here how to adapt the aforementioned algorithms to control the FDR without smoothing, that is,

$$\mathrm{FDR}_\delta(T) = \mathbf{E}\left[\frac{V_\delta(T)}{R_\delta(T) \vee 1}\right].$$

To do so we pick $w_0 \in (0, \alpha)$.

LORD In the case of LORD we can choose the thresholds as follows:

$$\alpha_t = w_0\tilde{\gamma}_t + (\alpha - w_0)\sum_j \delta^{t-\rho_j}\gamma_{t-\rho_j} \tag{10}$$

We can check that at any time T,

$$\widehat{\mathrm{FDP}}^{\delta}_{\mathrm{LORD}}(T) \leq \alpha$$

which shows that FDR_δ is controlled.

SAFFRON For SAFFRON we can choose thresholds as follows:

$$(1-\lambda_t)\Big(w_0\tilde{\gamma}_{S^{\lambda,1}_{0+}(t)} + (\alpha - w_0)\sum_j \delta^{t-\rho_j}\gamma_{S^{\lambda,1}_{j+}(t)}\Big) \wedge \lambda_t$$

We can check that at any time T,

$$\widehat{\mathrm{FDP}}^{\delta}_{\mathrm{ADDIS}}(T) \leq \alpha$$

which shows that FDR_δ is controlled.