Journal of *Marine Science and Engineering*

MDPI

Article

Enhancing Online Statistical Decision-Making in Maritime C2 Systems: A Resilience Analysis of the LORD Procedure Under Adversarial Data Perturbations

Victor Benicio Ardilha da Allen Alves [1], Gabriel Custódio Rangel [1], Miguel Ângelo Lellis Moreira [1,2], Igor Pinheiro de Araújo Costa [1,2,*], Carlos Francisco Simões Gomes [2] and Marcos dos Santos [3]

[1] Naval Systems Analysis Center (CASNAV), Brazilian Navy, Rio de Janeiro 20091-000, Brazil; benicio@marinha.mil.br (V.B.A.d.A.A.); crkster@gmail.com (G.C.R.); miguellellis@hotmail.com (M.Â.L.M.)
[2] Operations Research Department, Fluminense Federal University (UFF), Niterói 24020-007, Brazil; cfsg1@bol.com.br
[3] Department of Systems and Computing, Military Institute of Engineering (IME), Rio de Janeiro 22290-270, Brazil; marcosdossantos_doutorado_uff@yahoo.com.br
* Correspondence: costa_igor@id.uff.br

Abstract

Real-time statistical inference plays a pivotal role in maritime Command and Control (C2) environments, particularly for applications such as satellite-based object detection and underwater signal interpretation. These contexts often require online multiple hypothesis testing mechanisms capable of sequential decision-making while preserving statistical rigor. A primary concern is the control of the False Discovery Rate (FDR), as erroneous detections can impair operational effectiveness. In this study, we investigate the robustness of the Levels based On Recent Discovery (LORD) algorithm under adversarial conditions by introducing controlled perturbations to the data stream—specifically, missing or corrupted *p*-values derived from simulated Gaussian distributions. Inspired by developments in corruption-aware multi-armed bandit models, we formulate adversarial scenarios and propose defense strategies that modify the LORD algorithm's threshold sequence and integrate an online Benjamini–Hochberg procedure. The results, based on extensive Monte Carlo simulations, demonstrate that even a single missing *p*-value can trigger a cascading effect that reduces statistical power, and that our proposed mitigation strategies significantly improve algorithmic resilience while maintaining FDR control. These contributions advance the development of robust online statistical decision-making tools for real-time maritime surveillance systems operating under uncertain and error-prone conditions.

Keywords: false discovery rate; online hypothesis testing; adversarial scenarios; cascading effect; LORD algorithm

check for updates

Academic Editor: Tao Wang

Received: 2 July 2025
Revised: 29 July 2025
Accepted: 5 August 2025
Published: 12 August 2025

Citation: Alves, V.B.A.d.A.; Rangel, G.C.; Moreira, M.Â.L.; Costa, I.P.d.A.; Gomes, C.F.S.; Santos, M.d. Enhancing Online Statistical Decision-Making in Maritime C2 Systems: A Resilience Analysis of the LORD Procedure Under Adversarial Data Perturbations. *J. Mar. Sci. Eng.* **2025**, *13*, 1547. https://doi.org/10.3390/jmse13081547

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In today's globalized world, the volume of maritime transportation and the scale of international trade continue to expand, resulting in a rapid increase in maritime traffic and a progressively complex traffic environment [1]. In such a context, anomaly detection in maritime issues becomes integral to identifying deviations from normative patterns or expected behavior [2–6]. These deviations often signify unusual events, systemic errors, or potential critical threats, thereby playing a crucial role in enhancing the robustness of maritime monitoring systems and ensuring stringent quality control in increasingly intricate operational settings.

In modern maritime command and control (C2) systems, such as the Brazilian Navy (BN)'s Blue Amazon Management System (SisGAAz), real-time statistical decision-making plays a vital role in anomaly detection, target classification, and the prioritization of operational responses. These systems must process streaming data from heterogeneous sensors—such as satellite imagery, over-the-horizon radar, and geolocation—requiring robust statistical mechanisms that operate sequentially and adaptively. Each decision corresponds to a hypothesis test, and classification errors can lead to critical operational consequences: false positives may waste naval resources, while false negatives may delay or prevent responses to real threats. In such high-stakes environments, false discovery rate (FDR) control is not merely a statistical consideration but a strategic imperative. Algorithms like Levels Based On Recent Discovery (LORD) enable this kind of online FDR control, but their reliance on uncorrupted data makes them vulnerable in adversarial or degraded environments—a gap this study seeks to address.

Mathematical and statistical methods have increasingly become essential tools in addressing real-world problems across tactical, operational, and strategic domains [7–10], particularly in complex and dynamic environments such as maritime security [11]. Complementing this, hypothesis testing is a foundational tool in empirical research, offering a systematic approach for validating theories and assumptions across diverse fields [12,13]. Its broad applicability ranges from determining the efficacy of new medical treatments [14] to evaluating the environmental impacts of human activities [15] and assessing investment strategies and market trends in finance [16]. Additionally, A/B testing is widely used to compare the performance of two versions of a product or web feature, further demonstrating its utility in both scientific and commercial applications [17]. This versatility underscores its critical contribution to evidence-based decision-making across scientific and practical applications.

Within the military context, particularly for the Brazilian Navy (BN), hypothesis testing takes on an even more strategic dimension [18]. As a key player in safeguarding Brazil's sovereignty and maritime interests, the BN increasingly relies on data-driven decision-making processes [19]. In this context, SisGAAz epitomizes this shift, employing large-scale datasets and sophisticated statistical inference mechanisms to monitor and classify targets in real-time [20]. In such systems, maintaining control over the FDR is paramount, as classification errors can have severe consequences for downstream operational decisions [20]. Algorithms like LORD, proposed by [21] have been pivotal in managing FDR in online hypothesis testing environments, where hypotheses are tested sequentially as new data becomes available [22]. However, these algorithms typically assume the integrity of the incoming data. In military operations, this assumption often does not hold, as data corruption due to technical failures or adversarial interference is a persistent risk [23].

Data corruption greatly affects the reliability of real-time analysis [24–26]. Despite this, most existing FDR studies assume that incoming data streams are intact and free from adversarial interference. This assumption may not apply in critical real-time environments, such as maritime C2 systems, where sensor failures, communication errors, or intentional manipulation can degrade data integrity. Therefore, no previous work has systematically investigated how LORD-type algorithms behave under such conditions. This research aims to fill this gap by modeling adversarial perturbations, quantifying their effects (including the emergence of cascading failures), and proposing algorithmic defenses to preserve both the power and control of the FDR in the face of corrupted inputs.

Our proposed adaptation introduces dynamic threshold adjustments and corruption-resistant mechanisms, ensuring that the algorithm can maintain its performance even in the presence of noisy or manipulated data. This additional resilience is essential to maintaining operational efficiency and safety in maritime environments, where delays or false discoveries can have serious strategic consequences. In short, the improved LORD algorithm can support a wide range of maritime applications, from surveillance and navigation systems to environmental monitoring and resource management, strengthening overall maritime situational awareness.

The remainder of this paper is structured as follows. Section 2 introduces the background concepts of single and multiple hypothesis testing, including the theoretical foundation of the LORD algorithm and its application to maritime Command and Control systems. Section 3 presents the methodology and describes the simulation design, including adversarial models and defensive strategies. Section 4 reports and discusses the main results, comparing the resilience of the LORD algorithm and the proposed hybrid strategies under varying attack scenarios. Finally, Section 5 summarizes the key findings, outlines the study's limitations, and proposes directions for future research.

2. Materials and Methods

In this section, we provide a background on potential applications within maritime command and control systems, alongside a comprehensive review of the literature on hypothesis testing. We begin by discussing the fundamental concepts of single hypothesis testing, gradually transitioning to the more complex and evolving field of online multiple hypothesis testing. Furthermore, we detail the methodology employed in this work, offering a clear and cohesive understanding of both the theoretical framework and the practical approach adopted.

2.1. Background

Brazil's sea and inland waterways, crucial for the nation's well-being, necessitate effective protection. In this regard, the Strategic Plan of the Brazilian Navy (PEM) is instrumental, steering both medium- and long-term strategic planning via Naval Objectives. Then, the Naval Strategic Actions are meticulously crafted by dissecting these objectives, delineating a clear execution strategy to fulfill the overarching mission of the BN.

According to the Third United Nations Convention on the Law of the Sea (UNCLOS III), Brazil has property rights and sovereignty in the AJB up to 200 nautical miles. Beyond that, the nation also has the extension of the soil and subsoil of the submarine areas, defined by the limits of the continental Shelf. This area encompasses about 5.7 million square kilometers rich in natural resources, accounting for approximately 95% of Brazil's oil and 83% of the country's natural gas [27]. Andrade et al. [28], citing a 2002 report from the ANP, asserted that Brazil's reserves of these resources amounted to 9.81 billion barrels, and most of this, 8.87 billion barrels, originated from the AJB. In a related discussion, ref. [29] hussein (2018) mentioned the discovery in 2006 of substantial oil deposits located beneath a salt layer approximately 2000 meters thick, under a layer of sediment of similar thickness in the Santos Basin, about 300 kilometers southeast of Brazil's coast [30]. More than two decades later, based on a new report released by the ANP, ref. [31] highlighted that July's production from the pre-salt layers constituted 75% of Brazil's total oil output for that month. This substantial share underscores Brazil's capacity to emerge as the fourth-biggest oil producer worldwide .

Additionally, ref. [28] claimed that data from the Brazilian Institute of Geography and Statistics reveals that a significant portion of Brazil's population, approximately 80%, resides within 200 km of the coast. Consequently, according to these authors, this coastal proximity is a hub for economic activity, containing about 90% of the country's infrastructure and industrial production and roughly 80% of overall production. Furthermore, the oceans and river basins play a vital role as an intercommunicating element: 90% of the volume of this trade is made by sea [32].

The AJB are not merely conduits for transporting commodities; they represent an extensive reservoir of biodiversity and natural resources that are pivotal for the nation's advancement. The exigency for safeguarding and conserving these waters as a legacy for succeeding generations is of key importance [32].

2.2. Blue Amazon

The PEM explains the Blue Amazon concept, a term that the BN has spread to raise awareness among society and national institutions about the importance of the AJB's protection, a domain as vast as the Amazon rainforest. Figure 1 should not be perceived merely as an area encompassing the sea surface, waters overlying the seabed, and marine soil and subsoil within the Atlantic extension from the coast to the outer limit of the Brazilian continental Shelf. Rather, according to the Navy's plan document, it should be understood as a multifaceted concept embodying four distinct aspects:

1. Sovereignty—linked to the roles of the BN, which represents the authority of the state and oversees the use of force at sea.
2. Scientific—addresses the opportunities for research and technological advancement, the economic impact of using marine biodiversity, and the importance of maintaining knowledge about the maritime environment. Naval forces can use this knowledge to protect the interests of their respective nations.
3. Environmental—adopts a stance that goes beyond mere regulatory matters, considering that the unbroken expanse of oceanic areas and the movement of ocean currents enhance the risk of introducing and spreading non-native species and activities that endanger the marine ecosystem. This includes the need for mechanisms to monitor and tackle pollution, whether by accident or deliberate.
4. Economic—related to national development, based on the wealth of living and non-living resources in the AJB and the importance of maritime transportation for foreign trade.

Regarding the aspect of sovereignty, the BN undertakes strategic programs aligned with its institutional mission—the preparation and deployment of naval power as a component of national defense. Andrade et al. [28] articulated that these initiatives are instrumental in overseeing and administrating the Blue Amazon.

Central to these projects is the Blue Amazon Management System (SisGAAz), a system primarily aimed at extensively monitoring and managing the BN's area of responsibility, enhancing the Navy's capability to respond to challenges, including threats, hostilities, illicit activities, emergencies, and ecological crises. Consequently, ref. [28] concluded that this system will bolster the situational awareness of national authorities in these zones, elevating their monitoring and regulatory capabilities and strengthening their surveillance and protection of these maritime areas.



Figure 1. Brazilian "Blue Amazon". Adapted from [33]. The dashed line indicates the maritime boundary.

2.3. Blue Amazon Management System

The SisGAAz project, initiated in 2009, as explained by [28], was designed to fulfill the need for effective monitoring, surveillance, and defense within the Blue Amazon. The study claimed that its objective is to establish a unified and cohesive monitoring system that leverages and integrates others, improving the application of resources that are already in place instead of creating something entirely new. The research also noted that it will enable data gathering, analysis, and generating supportive information for decision-making, ultimately facilitating informed decisions that will guide the deployment of available resources (protection). Figure 2 illustrates SisGAAz's conception.

In Figure 2, solid yellow arrows represent the communication between platforms and satellites, while white dashed arrows indicate the communication between satellites, coastal stations, and the system's control center. According to [28], in terms of its integration with other platforms, SisGAAz will be interconnected with various systems both within and outside the BN, including the Military Command and Control System of the Ministry of Defense, which encompasses the Brazilian Army Integrated Border Monitoring System and the Brazilian Aerospace Defense System of the BAF. Additionally, the authors pointed out that SisGAAz will integrate with institutions outside the national defense realm, including those affiliated with ministries such as Finance, Transportation, Mines and Energy, Science and Technology, and Justice, as well as regulatory bodies and corporations.

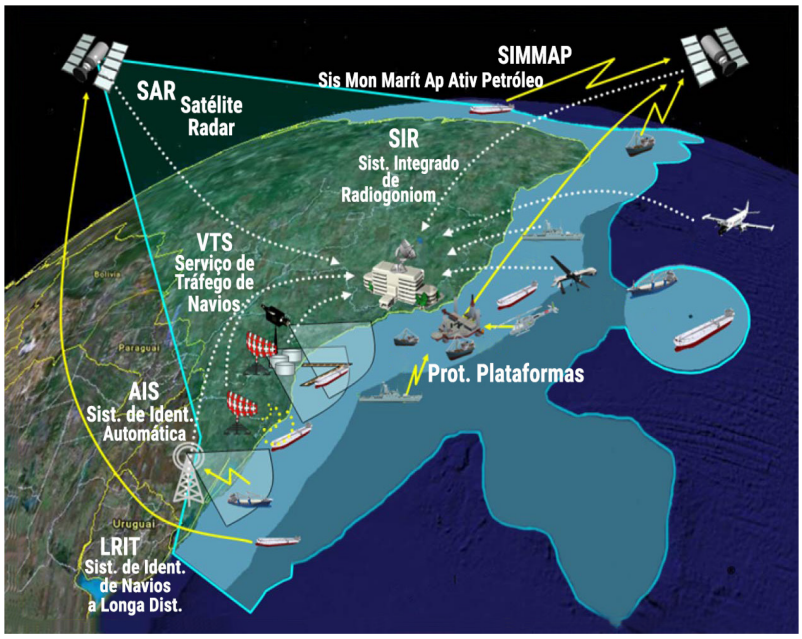


Figure 2. Operation and functioning of SisGAAz. Source: Adapted from [28].

Moreover, the paper stated further that the system will also receive data from various external sources, such as over-the-horizon radar, maritime patrol aircraft from the Brazilian Air Force (BAF), and unmanned aerial vehicles, and with systems from other nations and global entities, like the International Maritime Organization's Long Range Identification System (LRIT) and the Trans-Regional Maritime Network (T-RMN).

SisGAAz, tasked with the vigilant surveillance of the Blue Amazon, is instrumental in the continuous collection and analysis of comprehensive data to safeguard Brazil's maritime domain. It operates as an unceasing sentinel, meticulously sifting through current and historical data—unperturbed by the potential for data corruption—to categorize each vessel as either suspect or non-suspect. This ongoing vigilance mirrors the principles of online hypothesis testing, whereby the system dynamically assesses new (online) information to make informed decisions.

2.4. Single Hypothesis Testing

Devore [34] claimed that a statistical hypothesis, often referred to simply as a hypothesis, represents a statement or assertion regarding the value of a single parameter, of multiple parameters, or of the shape of an entire probability distribution. The book emphasized that, in hypothesis testing, there are typically two conflicting hypotheses to examine: the null hypothesis (H_0), the initially assumed claim, and the alternative hypothesis (H_a), which contradicts H_0. Using data from a sample, the author remarked that the prime objective is to determine which of these two hypotheses is true, pointing out that the null hypothesis will only be discarded in favor of the alternative hypothesis if the evidence from the sample strongly suggests that H_0 is incorrect. Consequently, hypothesis testing has two potential outcomes: rejecting H_0 or failing to reject H_0.

The statistician Fisher [35] defined p-value as "the probability of the observed result, plus more extreme results if the null hypothesis were true". This means that the p-value serves as a critical piece of information in hypothesis testing, quantifying the strength of evidence against the null hypothesis. Then, a p-value smaller than α (the test's significance level chosen by the analyst) suggests strong evidence against H_0, while a p-value greater than α suggests weaker evidence and the inability to reject H_0.

Mathematically, Pr is the probability distribution of the observed data x under H_0, for any value of α between 0 and 1. Efron [36] defined a rejection region R_α as:

$$\Pr\{x \in R_\alpha\} = \alpha. \tag{1}$$

Moreover, the p-value $p(x)$ is also defined as the smallest α such that $x \in R_\alpha$:

$$p(x) = \inf_{\alpha}\{x \in R_\alpha\}. \tag{2}$$

Since the area under the curve for any Probability Density Function (PDF) always equals 1, $p(x)$ conforms to a uniform distribution across the interval (0, 1):

$$p(x) \sim U(0,1). \tag{3}$$

In this context, Austin et al. [37] characterized the two potential mistakes that can be made: type I and type II errors. A type I error, also known as a false positive, happens when we reject the null hypothesis even though it is true. On the other hand, a type II error, or false negative, occurs when we fail to reject the null hypothesis when the alternative hypothesis is true. Table 1 summarizes the potential results.

Table 1. Possible outcomes for single hypothesis testing.

Reality	Stated as True	Stated as False
True null hypothesis	Correct $(1-\alpha)$	Type I error (α)
False null hypothesis	Type II error (β)	Correct $(1-\beta)$

Moreover, the authors emphasized that the test's significance level α is typically chosen to limit the probability of a type I error to a predetermined level, and the main objective is to maximize power (i.e., $1-\beta$), while ensuring the probability of a type I error remains at the intended level.

2.5. Multiple Hypothesis Testing

In various fields where statistics are employed, military applications included, decisions are made by assessing many hypotheses. In these scenarios, as outlined by [37], single hypothesis testing procedures are ineffective because the probability of committing at least one type I error significantly exceeds the nominal significance level employed for each test. The authors demonstrated that for N number of independent tests, with α as the threshold for each p-value, the probability of not committing any type I errors is $(1-\alpha)^N$. Given that α falls between 0 and 1:

$$(1-\alpha)^N < (1-\alpha). \tag{4}$$

Hence, they deduced that when conducting multiple tests, the likelihood of avoiding any type I errors becomes significantly reduced compared to when only one test is performed. As a result, the chances of committing at least one type I error increase with the number of tests conducted. This situation highlights the increased complexity of controlling the rate of false positives while effectively managing the type I error rate in multiple testing scenarios.

In the literature, the most relevant type I error rates are the Family-Wise Error Rate (FWER) and the False Discovery Rate (FDR). While FWER controls the probability of making even a single false discovery, it is known to be overly conservative in high-dimensional or streaming settings, where its strictness may lead to low statistical power. In contrast, FDR

offers a more flexible error control criterion, bounding the expected proportion of false positives among the rejected hypotheses. This balance between error control and discovery potential makes FDR particularly well-suited for online, sequential testing frameworks such as those needed in real-time maritime surveillance and decision-making environments [38]. Therefore, this article will focus on the FDR.

2.6. False Discovery Rate

Since it was introduced by [39], the concept of FDR has become a prominent focus in statistical research and remains the prevailing method applied, apparently attaining the "accepted methodology" status in scientific subject-matter journals [36].

Following the explanation provided by [37], the FDR is defined as "the expected proportion of rejected hypotheses that have been wrongly rejected". Robertson et al. [40] defined the False Discovery Proportion (FDP) up to time t, considering $R(t)$ as the number of rejected tests, and $V(t)$ as the number of falsely rejected hypotheses:

$$\mathrm{FDP}(t) = \frac{V(t)}{R(t) \vee 1}, \tag{5}$$

where $R(t) \vee 1 = \max(R(t), 1)$. This notation ensures that the denominator is always at least 1, even if there are no rejections ($R(t) = 0$). This avoids division by zero and maintains a valid proportion when calculating the FDP.

The FDR is the expectation of the FDP:

$$\mathrm{FDR}(t) = \mathbb{E}\{\mathrm{FDP}(t)\}. \tag{6}$$

Benjamini and Hochberg (BH) [39] developed a method to maintain the FDR under a predetermined threshold, and, in line with [41], the BH procedure is effective not just with independent tests but also with positive regression dependence on those test statistics associated with the true null hypotheses. This approach is advantageous in situations with a high number of true discoveries, particularly when numerous non-null hypotheses exist [21]. The algorithm, instead of controlling the probability of a type I error at a set level for each test, controls the overall FDR at level α in (0, 1):

$$\begin{aligned} &i_{\max} \text{is the greatest index for which } p_{(i)} \leq \frac{i}{N}\alpha. \\ &\text{Reject all } H_{(i)} \text{where: } i \leq i_{\max}. \end{aligned} \tag{7}$$

The BH procedure, enhanced with certain improvements, continues to be the leading approach in the field of multiple hypothesis testing [21].

2.7. Online Multiple Hypothesis Testing

Javanmard and Montanari [21] asserted that standard FDR control methods, like the BH procedure, require the presence of all p-values under consideration before any discoveries are made. For them, this implies that decisions are made only after all the necessary data has been gathered. However, they argued that this approach is unfeasible in several applications better suited to an online hypothesis testing framework. The study defined online hypothesis testing as follows: "Hypotheses arrive sequentially in a stream. At each step, the analyst must decide whether to reject the current null hypothesis without having access to the number of hypotheses (potentially infinite) or the future p-values but solely based on the previous decisions".

More formally, the authors considered a sequence of hypotheses $H_1, \ldots, H_N$ arriving sequentially in a stream, as depicted in Figure 3, with p-values $p_1, \ldots, p_N$. The primary

objective remains to keep the FDR under a predefined threshold α. A desired testing procedure offers, they proclaimed, a series of significance levels α_i with the following decision rule:

$$R_i = \begin{cases} 1, & \text{if } p_i \leq \alpha_i \text{ (reject } H_i\text{)}, \\ 0, & \text{otherwise (accept } H_i\text{)}. \end{cases} \tag{8}$$

Furthermore, each α_i depends on prior outcomes:

$$\alpha_i = \alpha_i(R_1, R_2, \ldots, R_{i-1}). \tag{9}$$

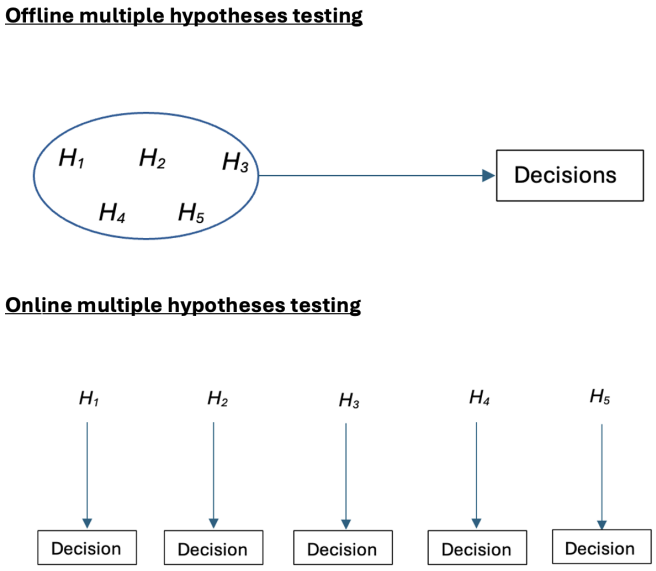


Figure 3. Offline and online FDR control. Decisions are made after all hypotheses have been available versus conclusions made sequentially for each incoming hypothesis online. Source: Adapted from [42].

The alpha-investing algorithm, first presented by [43], marked the beginning of online rate management techniques. According to [44], the alpha-investing method focuses on controlling the marginal false discovery rate mFDR_η at level α for any given choice of η and α, a variant of the FDR. The mFDR_η is defined as

$$\text{mFDR}_\eta = \frac{\mathbb{E}\{\text{V}(t)\}}{\mathbb{E}\{\text{R}(t)\} + \eta}. \tag{10}$$

Aharoni and Rosset [44] defined the wealth at any time point t as

$$W(t) = W(t-1) - (1 - R_t)\frac{\alpha_t}{1 - \alpha_t} + R_t\omega, \tag{11}$$

where $W(0) = \alpha\eta$.

If H_t is true, then $\frac{\alpha_t}{1-\alpha_t}$ is reduced from the wealth. If H_t is rejected, a reward ω is gained. By convention, ω is usually set to the maximal allowed value $\omega = \alpha$.

The research also extended the alpha-investing method to Generalized Alpha-Investing (GAI) algorithms. The potential function, previously known as alpha-wealth, operates as follows:

$$W(t) = W(t-1) - \phi_t + R_t\psi_t, \tag{12}$$

where $W(0) = \alpha\eta$.

Moreover, ref. [44] emphasized an important distinction: in the original alpha-investing, the quantity $\frac{\alpha_t}{1-\alpha_t}$ is deducted from the wealth only if the hypothesis H_t is not rejected. In contrast, in the GAI approach, ϕ_t is subtracted regardless of the test outcome.

Figure 4 summarizes this alpha-investing concept: when setting the initial FDR level, the algorithm is allocated a certain "initial wealth" W_0. At each time point t, the alpha-wealth $W(t)$ decreases by ϕ_t. If the hypothesis H_t is rejected ($R_t = 1$), then $W(t)$ is increased by ψ_t.

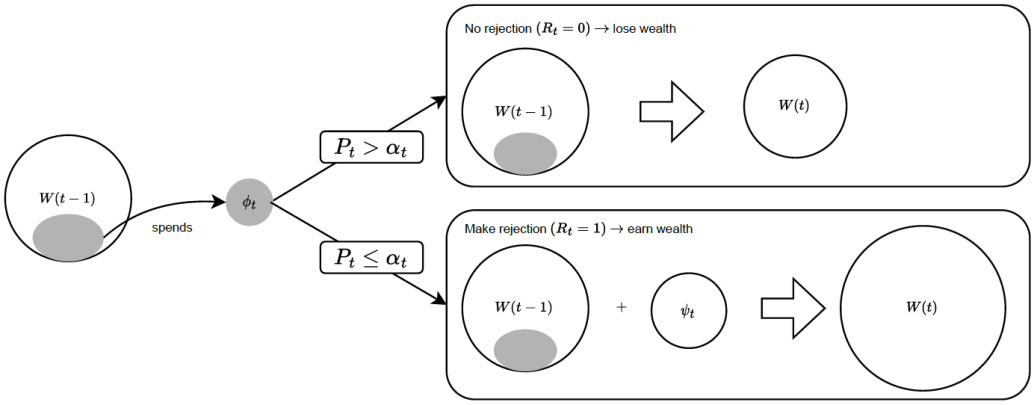


Figure 4. GAI representation showing how the wealth $W(t)$ changes depending on whether the hypothesis H_t is rejected. Source: Adapted from [40].

Suppose that θ_t is the actual parameter value for test t, and $\mathcal{H}^0$ is the null hypothesis space, which includes all parameter values that would lead to the null hypothesis not being rejected. If $\theta_t \notin \mathcal{H}^0$, then it is in the alternative hypothesis space, and, according to [44], the best power of the t-th test is defined as

$$\rho_t = \sup \text{Prob}_{\theta_t}(R_t = 1). \tag{13}$$

In simpler terms, this function finds the maximum probability that a test can achieve when it correctly rejects a null hypothesis across all possible alternative parameter values.

For the GAI method, as explained by the academics, any choice for the parameters α_t, ϕ_t, ψ_t is valid, as long as $W(t)$ does not become negative, meaning $\phi_t \leq W(t-1)$, and

$$\forall t : 0 \leq \psi_t \leq \min\left(\frac{\phi_t}{\rho_t} + \alpha, \frac{\phi_t}{\alpha_t} + \alpha - 1\right) \tag{14}$$

where ρ_t is the best power of the t-th test.

Javanmard and Montanari [21] presented alternative versions of the GAI algorithms, which are designed to control the FDR, in contrast to the mFDR proposed by [43] As described in [38], the parameter B_0 and proved for monotone GAI rules and under independence, with $B_0 = \alpha - W_0$, the FDR is controlled. Now, for some user-defined B_0

$$\psi_t \leq \min\left\{\phi_t + B_0, \frac{\phi_t}{\alpha_t} + B_0 - 1\right\}. \tag{15}$$

Ramdas et al. [38] also defined a class of improved GAI algorithms called GAI ++: the initial wealth W_0 is set to be $0 \leq w_0 \leq \alpha$ and the payout satisfies $\psi_t \leq \min\left\{\phi_t + b_t, \frac{\phi_t}{\alpha_t} + b_t - 1\right\}$, a modified version of Equation (15), where $b_t = \alpha - w_0 \mathbb{1}\{R(t-1) = 0\}$. As they demonstrated in [38], any monotone GAI++ rule comes with the following guarantee:

Theorem 1. *If the null p-values (i.e., the p-values corresponding to the true null hypotheses) are independent of all other p-values, any monotone GAI++ rule satisfies the bound* $\mathbb{E}\left[\frac{V(t)+W(t)}{R(t)\vee 1}\right] \leq \alpha$ *for all* $t \geq 1$. *Since* $W(t) \geq 0$, *the FDR is controlled at level* α.

Finally, ref. [21] conceptualized the Levels Based On Recent Discovery (LORD) algorithm, an instance of GAI algorithms. Later enhanced by [45], the so-called LORD++ (henceforth LORD) is widely considered one of the most advanced techniques in online multiple hypothesis testing.

2.8. Levels Based on Recent Discovery (LORD) Algorithm

We follow [38] to explain the LORD algorithm. Given a sequence of p-values, the decisions (rejections or non-rejections) $R_1, \ldots, R_t$, where each R_i is an indicator of whether the i-th hypothesis is rejected. The decision at time t is adapted to the sequence of decisions until $t-1$ (meaning that it can depend on them); we store this information via $\mathcal{F}^t = \sigma(R_1, \ldots, R_{t-1})$. The same applies to the rejection thresholds $\alpha_t \in [0,1]$; they are adapted to the history up to $t-1$, which means $\alpha_t = f_t(R_1, \ldots, R_{t-1})$, where f_t is an arbitrary [0,1]-valued function of the first $t-1$ decisions.

If the hypothesis H_i is truly null, its corresponding p-value has a $U(0,1)$ distribution, so the p-value is unlikely to take on very small values. By definition, these p-values are super-uniformly distributed, meaning that

$$\text{Prob}\left\{p_t \leq \alpha_t \mid \mathcal{F}^{t-1}\right\} \leq \alpha_t, \text{ or equivalently, } \mathbb{E}\left[\frac{\mathbb{1}\{p_t \leq \alpha_t\}}{\alpha_t}\middle| \mathcal{F}^{t-1}\right] \leq 1. \tag{16}$$

The interpretation is that the probability of the t-th null p-value p_t being less than or equal to its corresponding threshold α_t is at most α_t, given the past information $\mathcal{F}^{t-1}$.

Ramdas et al. [38] also defined, given any non-negative predictable sequence $\{\alpha_t\}$, the oracle FDP:

$$\text{FDP}^*(t) = \frac{\sum_{j \leq t, j \in \mathcal{H}^0} \alpha_j}{R(t)}. \tag{17}$$

The interpretation is that the expected number of null (i.e., false) rejections up to time t is approximately the sum of all α_j, for $j \leq t$ and H_j be a null hypothesis.

Since $\text{FDP}^*(t)$ cannot be calculated because the contents of $\mathcal{H}^0$ are unknown, a conservative estimate of the oracle FDP is

$$\widehat{\text{FDP}}_{\text{LORD}}(t) = \frac{\sum_{j=1}^{t} \alpha_j}{R(t)}. \tag{18}$$

The implication is that $\widehat{\text{FDP}}_{\text{LORD}}(t)$ overestimates the unknown FDP(t):

$$\widehat{\text{FDP}}_{\text{LORD}}(t) \geq \frac{\sum_{j \leq t, j \in \mathcal{H}^0} \alpha_j}{R(t)} \approx \frac{\sum_{j \leq t, j \in \mathcal{H}^0} \mathbb{1}\{p_j \leq \alpha_j\}}{R(t)} = \text{FDP}(t). \tag{19}$$

The authors declared that a more straightforward approach to developing online FDR methods is to guarantee that $\sup_{t \in \mathbb{N}} \widehat{\text{FDP}}_{\text{LORD}}(t) \leq \alpha$, eliminating the need for wealth, penalties, and rewards, as seen in the GAI procedure. Based on these definitions, the following theorem is also proved in [38]:

Theorem 2. *(a) If the null p-values are conditionally super-uniform, then the condition* $\widehat{FDP}_{LORD}(t) \leq \alpha$, $\forall t \in \mathbb{N}$, *implies that* $mFDR(t) \leq \alpha$, $\forall t \in \mathbb{N}$. *(b) If the null p-values are independent of each other and of the p-values corresponding to the non-null hypotheses, and*

$\{\alpha_t\}$ is chosen to be a monotone function of past rejections, then the condition $\widehat{FDP}_{LORD}(t) \leq \alpha$, $\forall t \in \mathbb{N}$, implies that $FDR(t) \leq \alpha$, $\forall t \in \mathbb{N}$.

Leveraging this theorem, ref. [38] presented the LORD algorithm: given an infinite, non-increasing sequence of positive constants $\{\gamma_t\}_{t=1}^{\infty}$ that sums to one, and τ_j as the time of j-th rejection, the test level α_t is

$$\alpha_t = w_0\gamma_t + (\alpha - w_0)\gamma_{t-\tau_1}\mathbb{1}\{\tau_1 < t\} + \alpha \sum_{j:\tau_j<t,\tau_j\neq\tau_1} \gamma_{t-\tau_j}. \tag{20}$$

As explained by [40], the initial term $w_0\gamma_t$ represents the portion of the starting wealth w_0 allocated to the t-th test, while the subsequent terms are the gains from previous rejections before t that are used in round t: the reward for the first rejection is $(\alpha - w_0)$, and for subsequent rejections is α. Once these earnings are received, they are allocated to future rounds according to the same constants $\{\gamma_t\}$, shifted to start at the next instant. Ramdas et al. [38] showed this rule ensures LORD always operates within its earned resources and maintains $\widehat{\text{FDP}}_{\text{LORD}}(t) \leq \alpha$. For them, default values are $w_0 = \frac{\alpha}{10}$ and $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, the former calculated in a Gaussian setting to maximize power.

3. Methodology

This article employs the open-source R software (version 4.4.1) "online FDR," encompassing implementations of LORD and nearly all subsequent advancements in online error rate control methods. Additionally, modifications to the source code of the LORD algorithm enabled further analysis under various data corruption scenarios.

Our study implements a straightforward experimental framework that tests Gaussian means across N hypotheses to evaluate the comparative efficacy among the algorithms, utilizing the default configurations recommended in the existing literature.

For all simulations conducted, null hypotheses $H_t : \mu_t = 0$ are tested against the alternative: $\mu_t > 0$, for t = 1, ..., N. Consequently, we observe independent $Z_t \sim \mathcal{N}(\mu_t, 1)$ transformed into one-sided p-values $p_t = \Phi(-Z_t)$, where Φ denotes the standard Gaussian Cumulative Density Function (CDF). The values of μ_t are determined based on the mixture distribution:

$$\mu_t = \begin{cases} \mathcal{N}(0,1) & \text{with probability } \pi_0 = 1 - \pi_1 \\ \mathcal{N}(3,1) & \text{with probability } \pi_1. \end{cases} \tag{21}$$

In the composite model designated as G, the null hypothesis H_t stipulates that p_t is uniformly distributed within the interval [0,1]. Contrarily, the alternative hypothesis posits that p-values are derived from a distribution with the CDF represented by F. Consequently, the marginal distribution of these simulated p-values is expressed as $G(x) = \pi_0 x + \pi_1 F(x)$. As depicted in Figure 5, the histogram illustrates why online methodologies have more likelihood of rejecting non-null hypotheses, as they are characterized by lower p-values.

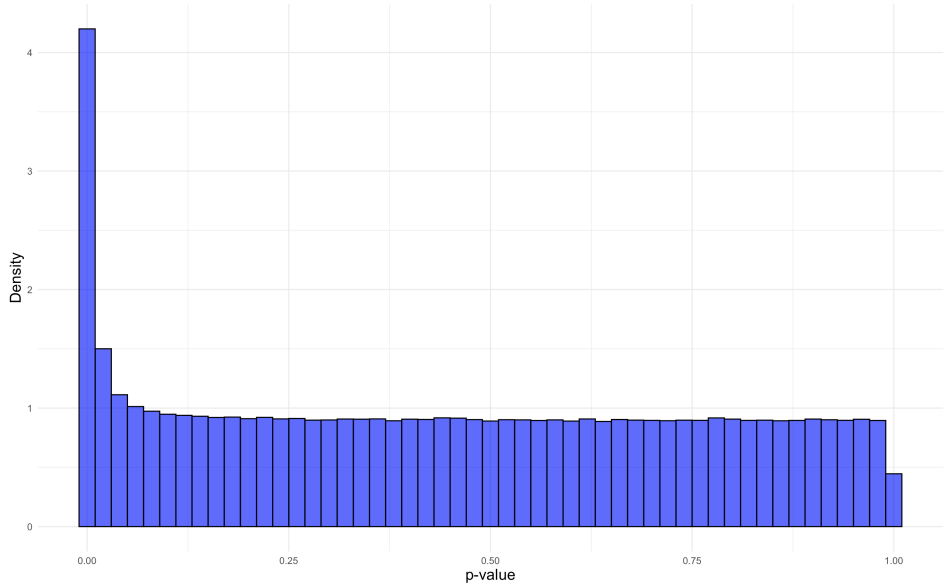


Figure 5. Histogram of the mixed model. Results are based on N = 100,000.

4. Results

To rigorously evaluate data corruption in online multiple hypothesis testing, we propose a controlled adversarial setup featuring two entities: Blue, representing the side that performs the tests attempting to make true discoveries, and Red, which acts as the offensive agent by stealing discoveries. The model operates as follows:

1. In period t, Blue receives a single p-value, p_t, and must decide whether to accept or reject the hypothesis H_t, using only the information collected in rounds $1, \ldots, t-1$.
2. Red knows if H_t is true but does not know the p-values past p_t.
3. In case p_t is stolen, it is removed from the data stream without Blue noticing. This way, we simplify potentially more complicated scenarios, such as setting a new (corrupted) value of p_t.

4.1. Problem Formulation

The primary objective of Blue is to maximize power, ensuring that the FDR does not exceed the chosen threshold α over a time horizon of interest. Conversely, Red aims to min–max Blue's power, subject to an effort constraint.

In this setting, there are T periods. In periods $t = 1, \ldots, T$, Blue receives a single p-value and has to reject or fail to reject the hypothesis associated with p_t. This is done by comparing p_t with α_t, where α_t is determined by the LORD algorithm. If the p-value associated with the hypothesis H_t is greater than α_t, the hypothesis will not be rejected, but if $p_t \leq \alpha_t$ it will be rejected in favor of the alternative, and we get a so-called discovery (which could be true or false).

Before examining the impact of Red's attack, it is pertinent to revisit the LORD α_t equation, as delineated in Equation (20):

$$\alpha_t = w_0\gamma_t + (\alpha - w_0)\gamma_{t-\tau_1}\mathbb{1}\{\tau_1 < t\} + \alpha \sum_{j:\tau_j<t,\tau_j\neq\tau_1} \gamma_{t-\tau_j}.$$

Suppose the p-values are generated as those in Section 2, $w_0 = \frac{\alpha}{10}$ and $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$:

1. For time $t = 1$, $\alpha_1 = w_0\gamma_1$. If $p_1 > \alpha_1$, there is no discovery, and no wealth is added to the initial budget.

2. For time $t = 2$, $\alpha_2 = w_0\gamma_2$. However, if $p_2 \leq \alpha_2$, a discovery is made and $(\alpha - w_0)\gamma_1$ is added to the budget. Thus, $\tau_1 = 2$.
3. For time $t = 3$, $\alpha_3 = w_0\gamma_3 + (\alpha - w_0)\gamma_1$. If $p_3 \leq \alpha_3$, then $\alpha\gamma_1$ is added to the budget.
4. For time $t = 4$, $\alpha_4 = w_0\gamma_4 + (\alpha - w_0)\gamma_2 + \alpha\gamma_1$. If $p_4 \leq \alpha_4$, then $\alpha\gamma_2$ is added to the budget.
5. For time $t = 5$, $\alpha_5 = w_0\gamma_5 + (\alpha - w_0)\gamma_3 + \alpha(\gamma_1 + \gamma_2)$. If $p_5 \leq \alpha_5$, then $\alpha\gamma_3$ is added to the budget.
6. This process continues until the last unit of time.

In this case, let us consider that the p-value p_4 at time $t = 4$ came from the alternative hypothesis and is stolen by Red. This results in a wealth amount equal to $\alpha\gamma_1$ being removed from α_5, $\alpha\gamma_2$ removed from α_6, and so on, for a total α removed from the subsequence α_{5+t}, for $t \geq 0$. This may induce a "cascade effect" in future values of α_t.

This "cascade effect" is significant because it influences the likelihood of future discoveries. Normally, if the conditions for discovery ($p_t \leq \alpha_t$) were met, additional wealth would be added to the budget. However, due to Red's manipulation at $t = 4$, the subsequent α_t values are impacted. This means that potential discoveries that might have occurred under normal circumstances may no longer happen, as the altered α_t levels are now lower, making it harder to meet the discovery criteria. This illustrates how a single attack at a point in time can have lasting effects on the entire process, altering the trajectory outcomes.

4.2. Cascade Effect Formulation

This section estimates the expected number of lost discoveries until the next discovery, which is a lower bound for the expected number of discoveries lost.

Let $p_k^{(1)}$ be a random p-value from $\mathcal{H}_1$ at time k. As an example, in case the alternative distribution is $N(\mu_1, 1)$, we know that $p_k^{(1)} \sim 1 - \Phi(\mu_1 + Z)$, for $Z \sim N(0,1)$. Likewise, $p_k^{(1)} \sim U(0,1)$, if the alternative and null distributions coincide.

The starting wealth at time t, α_t, depends on rejections up to $t-1$. We consider an attack taking place at time t. That is, $p_t \leq \alpha_t$, but the attacker prevents a rejection from taking place—it does not matter how this is done, whether by stealing the p-value or by corrupting it, the end effect is that there is no rejection in period t when there should have been one. Importantly, the decision-maker is unaware of this fact.

Let $\tilde{\alpha}_t = \alpha_t$ and $\tilde{\alpha}_{t+k}$ as the value of α_{t+k} if there were no rejections at $t, t+1, \ldots, t+k$. That is, the sequence of thresholds $\tilde{\alpha}_t$ is deterministic, conditional on the discoveries until $t-1$.

Hence, the expected number of lost discoveries until the next discovery is

$$\underbrace{\pi_1 P(p_{t+1}^{(1)} \in (\tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1} + \alpha\gamma_1))}_{\text{P(missing an } \mathcal{H}_1 \text{ rejection in period } t+1 \text{ due to stolen discovery at } t)} + \sum_{k=2}^{\infty} \underbrace{\left(\prod_{j=1}^{k-1} \left(\pi_1 P(p_{t+j}^{(1)} > \tilde{\alpha}_{t+j}) + \pi_0 P(p_{t+j}^{(0)} > \tilde{\alpha}_{t+j}) \right) \right)}_{\text{prob of no rejections up to period } t+k-1} \underbrace{\pi_1 P(p_{t+k}^{(1)} \in (\tilde{\alpha}_{t+k}, \tilde{\alpha}_{t+k} + \alpha\gamma_k))}_{\text{P(missing an } \mathcal{H}_1 \text{ rejection in period } t+k \text{ due to stolen discovery at } t)}. \quad (22)$$

The reasoning of Expression (22) is that the expected number of true discoveries lost in period $t+k$ is the sum of (i) the probability of the p-value in $t+1$ being from $\mathcal{H}_1$ and falling in the range of values that would have triggered a rejection had an attack in period t not taken place, with the product of two terms, (ii) the probability of the p-values in $t+2, \ldots, t+k-1$ being below the rejection threshold, and (iii) a p_{t+k} being from $\mathcal{H}_1$ and falling in the range of values that would have triggered a rejection had an attack in period t not taken place.

Since α_k is decreasing as long as there are no discoveries, and the PDF of $p_t^{(1)}$ is non-increasing, the above expression can be lower bounded by

$$\sum_{k=1}^{\infty}(\pi_1 P(p^{(1)} > \tilde{\alpha}_{t+1}) + \pi_0 P(p^{(0)} > \tilde{\alpha}_{t+1}))^{k-1}\pi_1 P(p_{t+k}^{(1)} \in (\tilde{\alpha}_{t+k}, \tilde{\alpha}_{t+k} + \alpha\gamma_k)). \quad (23)$$

An even weaker lower bound is obtained by replacing $\tilde{\alpha}_{t+k}$ with $\tilde{\alpha}_{t+1}$ in Expression (23),

$$\sum_{k=1}^{\infty}(\pi_1 P(p^{(1)} > \tilde{\alpha}_{t+1}) + \pi_0 P(p^{(0)} > \tilde{\alpha}_{t+1}))^{k-1}\pi_1 P(p_{t+k}^{(1)} \in (\tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1} + \alpha\gamma_k)). \quad (24)$$

In case the alternative distribution is $N(\mu_1, 1)$, we get

$$P(p^{(1)} > \tilde{\alpha}_{t+1}) = P(1 - \Phi(\mu_1 + Z) > \tilde{\alpha}_{t+1}) = P(\Phi(\mu_1 + Z) < 1 - \tilde{\alpha}_{t+1})$$

$$= P(\mu_1 + Z < \Phi^{-1}(1 - \tilde{\alpha}_{t+1})) = \Phi(-\mu_1 + \Phi^{-1}(1 - \tilde{\alpha}_{t+1})).$$

Likewise,

$$P(p_{t+k}^{(1)} \in (\tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1} + \alpha\gamma_k)) = \Phi(\mu_1 - \Phi^{-1}(1 - \tilde{\alpha}_{t+1} - \alpha\gamma_k)) - \Phi(\mu_1 - \Phi^{-1}(1 - \tilde{\alpha}_{t+1})).$$

From here, we can compute Expression (23) numerically.

In contrast, when $p^{(1)} \sim U(0,1)$ (meaning that the null and alternative distributions coincide), we get in Expression (23),

$$\alpha\pi_1 \sum_{k=1}^{\infty}(1 - \tilde{\alpha}_{t+1})^{k-1}\gamma_k. \quad (25)$$

It follows that the expected number of true discoveries lost approaches $\alpha\pi_1$ in (25), as $\tilde{\alpha}_{t+1} \to 0$.

Thusly motivated, we investigate two distinct scenarios of online hypothesis testing with corrupted data:

1. Single attack: Red's capacity to attack is limited by a single attack.
2. Stochastic attacks: Red attacks each alternative p-value with probability ζ.

The simulation of each scenario uses altered forms of the LORD algorithm, derived from the "onlineFDR" R package, to effectively incorporate Red's and Blue's strategies. The generation of p-values adhered to the process detailed in Section 2.

4.3. Single Attack

As previously discussed, the corruption of a single alternative p-value may initiate a "cascade effect," where the stolen wealth imposes future reduced α_t values, leading consequently to fewer discoveries.

From the attacker's perspective, the earlier Red intervenes in the data stream directed towards Blue, the more promptly α_t will decrease, thereby suppressing a greater number of potential discoveries. Accordingly, this scenario examines the dynamics of power and FDR when Red attacks the first alternative p-value.

Red procedure for attacking the first alternative p-value:

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, and sets $\tau_0 = 0$.
2. At each step t, Blue computes α_t according to Equation (20).
3. If $H_t \in \mathcal{H}^1$ and $p_t \leq \alpha_t$, Red steals p_t and the discovery is not allowed.

4. Go back to step 2 if $t < N$ or the attack has not taken place.

Figure 6 compares the statistical power of LORD without attacks (in black) and of LORD when only the first alternative p-value is attacked (in blue) for the proportion of non-nulls π_1 varying from 0.1 to 0.9, $N = 1000$, and $\alpha = 0.05$.

A single attack imposes an overall power decrease for different π_1. Notably, this decrement is more pronounced at lower π_1 values, while a single attack is largely inconsequential as π_1 gets bigger since there are many discoveries that remain to be made by Blue after Red's attack. Given the negative relationship between FDR and power, it is prudent to focus our investigation on lower π_1 values to ascertain the subsequent behavior of the FDR.

Table 2 shows the corresponding FDR and power for $\pi_1 = 0.1$.

Table 2. FDR and power of LORD without attacks and with a single attack for $\pi_1 = 0.1$. Results are based on 10^3 replications.

Method	FDR	Power
LORD without attacks	0.00910	0.28454
LORD with a single attack	0.00938	0.20463

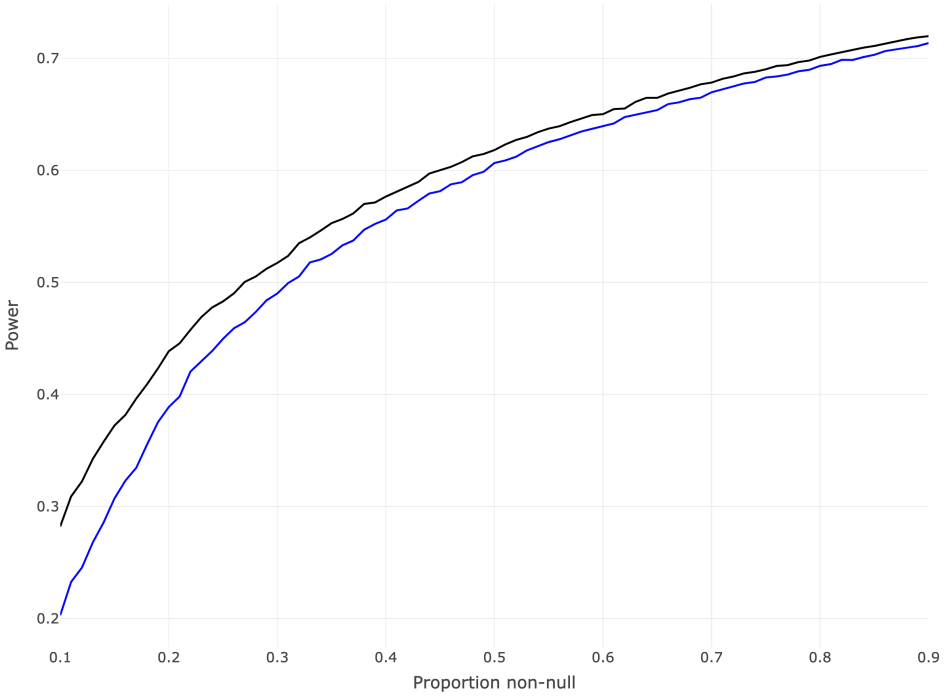


Figure 6. Power of LORD without attacks and of LORD with a single attack as the proportion of non-nulls varies. Results are based on 10^3 replications.

Stealing the first alternative p-value resulted in the average power dropping from around 0.28 to 0.20, marking a 28% decrease, while the FDR remained largely unaffected. The cascading effect resulted in missing seven extra alternative hypotheses with just a single steal, highlighting the effectiveness of this approach in undermining statistical power.

While Blue is certain of an imminent attack, the exact moment of its occurrence is undetermined. As a strategy, we propose reviewing the infinite, non-increasing sequence of positive constants $\{\gamma_t\}_{t=1}^{\infty}$ that sums to one, as the LORD algorithm does not impose any fixed formula for it.

The threshold α_t for each hypothesis H_t is a monotone decreasing function of past rejections, represented by the convolved sum of previous γ. This design implies that, as more hypotheses are tested and potentially rejected over time, the threshold for deeming

subsequent tests significant becomes progressively more restrictive. Consequently, as the testing process advances and the criterion for each test becomes more rigorous, the likelihood of achieving further discoveries diminishes. When Red prevents a discovery, the effect on the testing procedure is twofold. Firstly, the immediate outcome of such an attack is the failure to add a wealth $\alpha\gamma_1$. Secondly, the α_t value assigned to the ensuing tests becomes even more restrictive than without corruption.

Therefore, we propose as Blue's strategy to modify the original formula for the sequence $\{\gamma_t\}_{t=1}^{\infty}$ to reduce the rate of decay of each α_t until the first discovery, consequently increasing the probability of discoveries and after that go back to the default equation. Any function with a lower rate of decrease than $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$ can be applied.

Figure 7 displays the plot of the function $\gamma_t = C\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$ for $t = 1, \ldots, 1000$, when C has its default value of 0.0722 (black) and $C = 2$ (blue). As expected, for small values of t, the new function provides larger values, but as t increases, it converges toward the black curve. This behavior implies that the new α_t levels will be higher than using the default value of C, and they tend to take longer to decrease, leading to greater "wealth" until Red's attack.

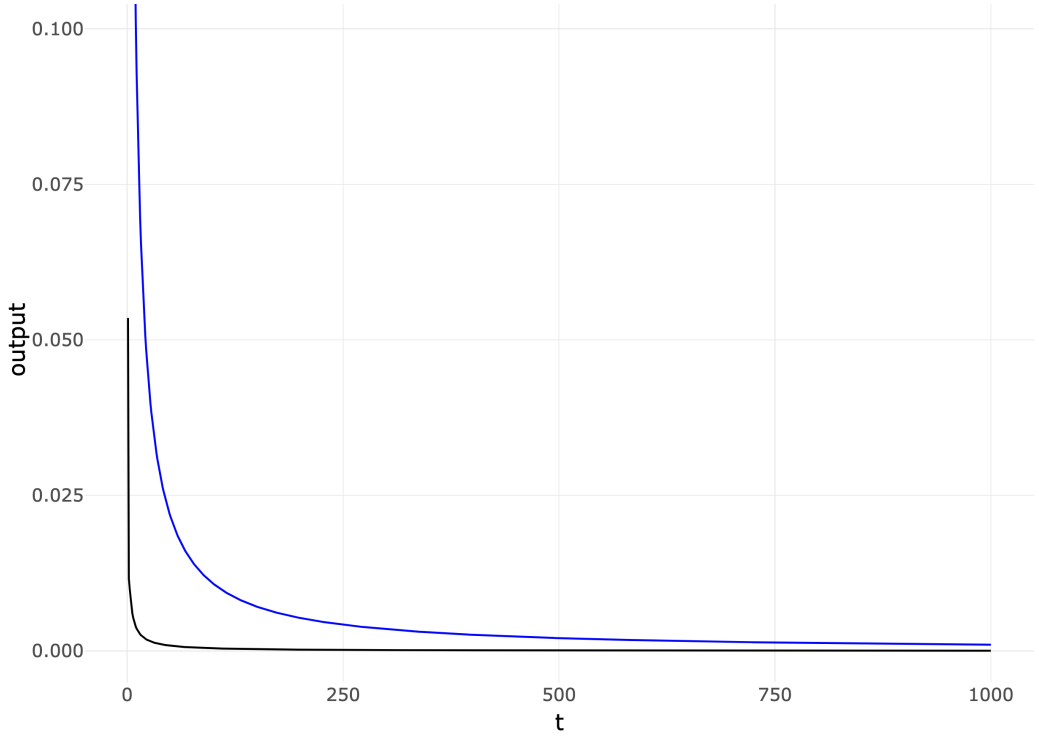


Figure 7. $\gamma_t = C\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$ for $t = 1, \ldots, 1000$, when $C = 0.0722$ and $C = 2$.

Blue procedure for defending against Red's attack at the first alternative p-value:

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = C\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, and sets $\tau_0 = 0$, and $C = 2$.
2. At each step t, Blue computes α_t according to Equation (20).
3. If $H_t \in \mathcal{H}^1$ and $p_t \leq \alpha_t$, Red steals p_t and the discovery is not allowed.
4. Blue sets $C = 0.0722$.
5. Execute step 2 till $t = N$.

Figure 8 illustrates the statistical power of LORD without attacks, of LORD attacking only the first alternative p-value, and of LORD with the defender policy implemented for the proportion of non-nulls π_1 varying from 0.1 to 0.9, N = 1000, and $\alpha = 0.05$.

Implementing the previously mentioned defensive policy by Blue results in a less pronounced reduction in power compared to scenarios lacking data corruption for every value of π_1. Specifically at $\pi_1 = 0.1$, as illustrated in Table 3, the average power diminishes from approximately 0.28 to 0.26 with the deployment of the defensive strategy, as opposed to 0.20 in the absence of any countermeasures, while the FDR remains virtually unaffected. This robustness is further exemplified by Blue's ability to recover six true discoveries out of eight lost (seven due to cascading). Without any defensive strategy, a single offensive maneuver by Red imposed seven additional discoveries, while the strategy actively limited the outcome to just one additional true discovery not being made.

Table 3. FDR and power of LORD without attacks, with the defensive strategy implemented, and with a single attack for $\pi_1 = 0.1$. Results are based on 10^3 replications.

Method	FDR	Power
LORD without attacks	0.01038	0.28254
LORD with the defensive strategy	0.00966	0.25673
LORD with a single attack	0.00824	0.19652

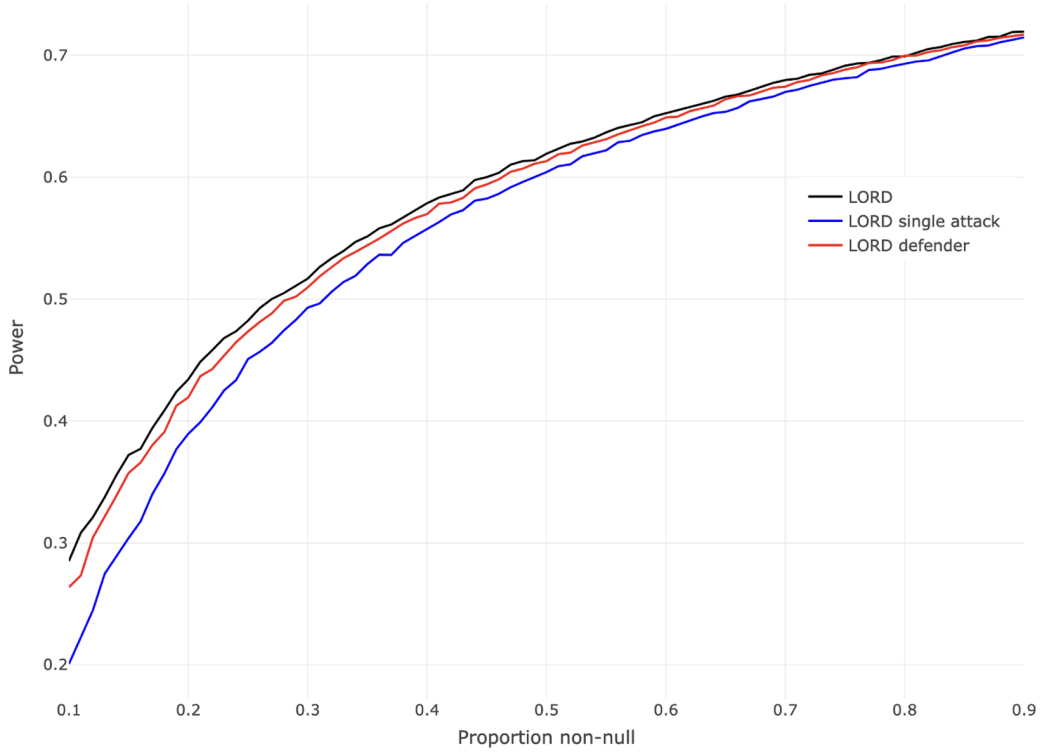


Figure 8. Power of LORD without attacks, of LORD with a single attack, and of LORD with the defender policy implemented for $\pi_1 = 0.1$. Results are based on 10^3 replications.

In conclusion, by increasing the sequence $\{\gamma_t\}_{t=1}^{\infty}$ by a constant factor up to the first discovery, Blue can protect a large fraction of the discoveries that would otherwise be lost due to cascading from a single stolen discovery, with minimal increase in FDR.

4.4. Stochastic Attacks

In this scenario, we consider a setting where Red attacks only alternative *p*-values that would otherwise be rejected, with probability ζ. This setting may arise when Red has a great intelligence capability, allowing it to steal more than just one true discovery.

Red procedure for attacking alternative *p*-values with probability ζ:

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = 0.0722 \frac{\log(t\sqrt{2})}{t \exp(\sqrt{\log t})}$ and sets $\tau_0 = 0$.
2. At each step t, Blue computes α_t according to Equation (20).
3. If $H_t \in \mathcal{H}^1$ and $p_t \leq \alpha_t$, then Red steals the p-value with probability ζ.
4. Go back to step 2 till $t = N$.

Figure 9 compares the statistical power of LORD without attacks (in black) and of LORD with a probability $\zeta = 0.1$ of attacks (in blue) for the proportion of non-nulls π_1 varying from 0.1 to 0.9, $N = 1000$, and $\alpha = 0.05$.

For a $\zeta = 0.1$ attack probability, the LORD algorithm shows a decreased power compared to its operation without any attacks for every level of non-null proportion π_1. The effect of a 10% attack is more pronounced as π_1 increases since there are more true discoveries to steal. Likewise, increasing the attack probability ζ leads to an even greater decrease in LORD's power.

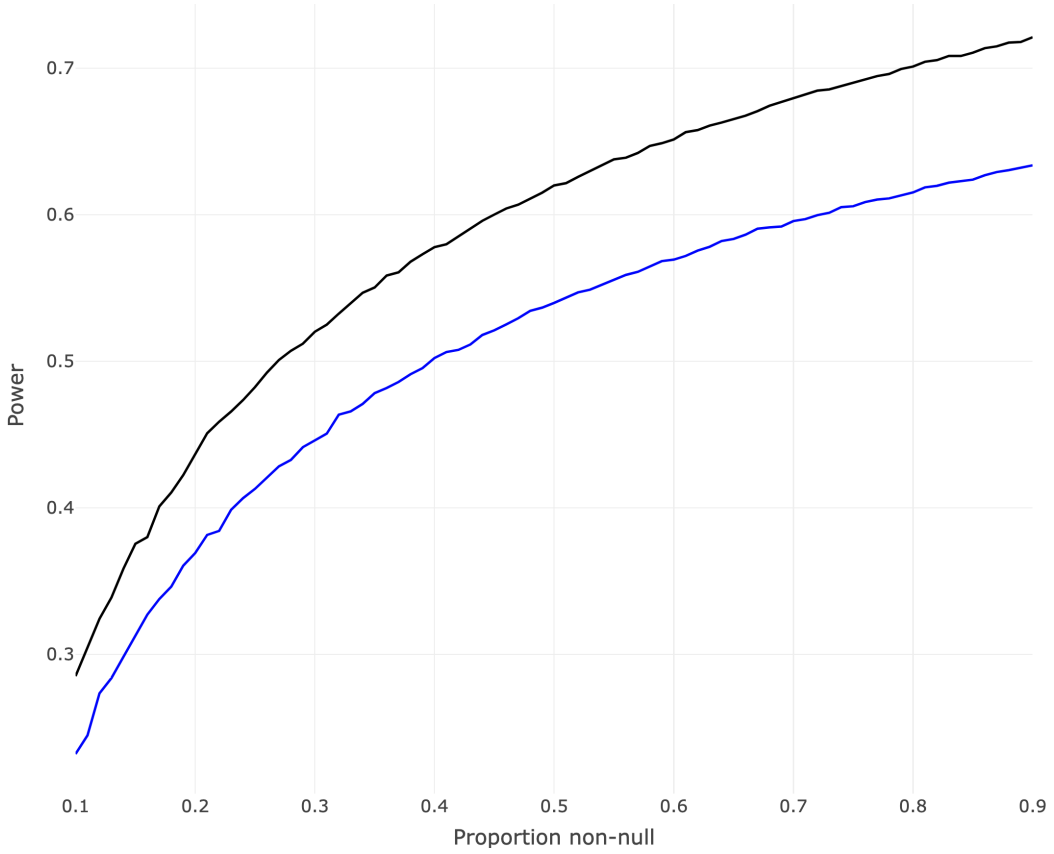


Figure 9. Power of LORD without attacks and of LORD with 10% of attack probability as the proportion of non-nulls varies. Results are based on 10^3 replications.

The joint effect of FDR and power of increasing π_1 is shown in Figure 10.

Larger values of π_1 correspond to increased power for attack and no-attack cases, alongside a reduction in the FDR. The curve associated with an attack probability of $\zeta = 0.1$ lies below and to the left of the no-attack curve. Importantly, the FDR remains below the predefined $\alpha = 0.05$. In the particular instance of $\pi_1 = 0.1$, as depicted in Table 4, there is a 21% reduction in the average power, from 0.29 to 0.23, exceeding 10%. This outcome merits emphasis: even when subjected to an attack with a ζ probability, the power reduction exceeds the scale of ζ itself, as expected due to the cascade effect described in the last section.

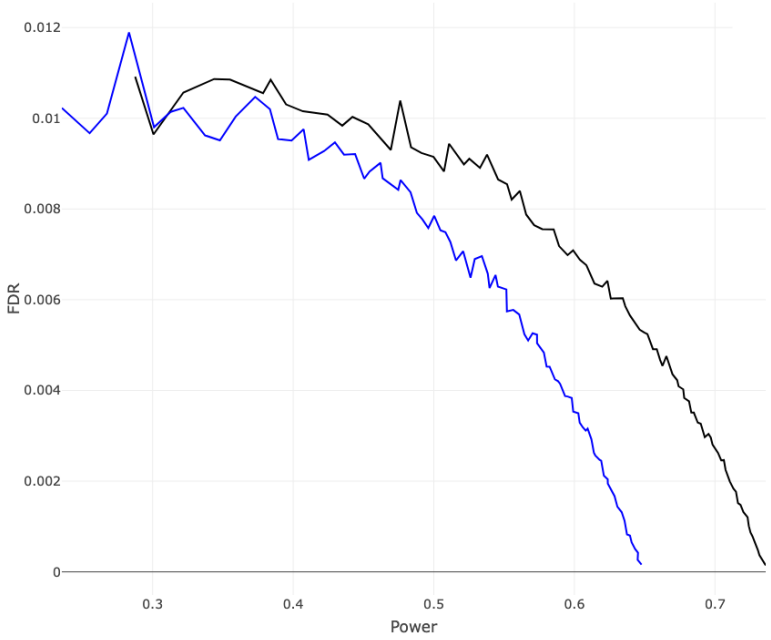


Figure 10. FDR and power of LORD without attacks (in black) and of LORD with 10% of attack probability (in blue) as the proportion of non-nulls varies. Results are based on 10^3 replications.

Table 4. FDR and power of LORD without attacks and with 10% of attack probability for $\pi_1 = 0.1$. Results are based on 10^3 replications.

Method	FDR	Power
LORD without attacks	0.00126	0.28600
LORD with 10% of attack probability	0.00868	0.22677

To counteract Red "usurping" discoveries from Blue, it becomes imperative to formulate mitigating strategies to prevent the cascade effect. Thereafter, we propose a new procedure.

4.5. Online BH Algorithm

Recall the setting where Red steals each discovery with probability ζ in Section 4.4. Note that in Step 3, Red attacks regardless of whether the p-value is in H_0 or H_1. This would be the case of a blind attacker, which steals p-values smaller than α_t without considering the ground truth. In practice, only for alternative distributions without a strong signal (e.g., μ_1 close to zero) this type of blind attack would be impactful in relation to a not-blind attack (where only null p-values are stolen). When the signals are strong, most of the rejections are from the alternative distribution, so it does not matter whether the adversary—thanks to its intelligence capability—an discriminate between true and false discoveries.

To ameliorate the "cascade effect," we tested rejecting all p-values below some small threshold. Numerical testing indicated that the power was greatly increased while the FDR was kept below the guaranteed α. Thusly motivated, we devised a so-called online BH algorithm, which applies the BH procedure—traditionally employed in offline settings—in online fashion as the p-values roll in.

Following Section 2, consider the mixed model with the null mean $\mu_0 = 0$ and the alternative mean $\mu_1 > \mu_0$ as depicted in Figure 5, and the BH algorithm presented in Equation (7):

$$i_{\max} \text{ is the greatest index for which } p_{(i)} \leq \frac{i}{N}\alpha.$$

$$\text{Reject all } H_{(i)} \text{ where: } i \leq i_{\max}.$$

The idea is to perform the BH procedure at each period t. Therefore, Blue orders all p-values received till time t and calculates the corresponding position $i_{\max}$ of the current p-value p_t. Hence, a conservative dynamic threshold is:

$$\alpha_t = \frac{i_t}{t}\alpha, \tag{26}$$

where i_t is the position of p_t in the sorted vector $p_{(1)}, p_{(2)}, \ldots, p_{(t)}$.

If we consider a stream of p-values of length N, employing the LORD algorithm till time $t = N/2$ and the Online BH for the remaining sequence has demonstrated, through simulation, an enhanced power while adhering to the FDR control.

Blue procedures in a scenario with attacking probability of ζ:

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, and sets $\tau_0 = 0$.
2. At each step $t \leq N/2$, Blue computes α_t according to the LORD algorithm.
3. For $t \geq N/2$, Blue computes α_t according to the BH algorithm.
4. If $p_t \leq \alpha_t$ (meaning that the t term would be rejected), then with probability ζ, Red eliminates the p-value, and the discovery is not allowed. The next p-value in the sequence is fed to Blue.
5. Go back to step 2 until $t = N$.

Figure 11 shows the results when we compare LORD's power with the aforementioned mixed procedure using the online BH algorithm for $N = 1000$ and $\pi_1 = 0.1$.

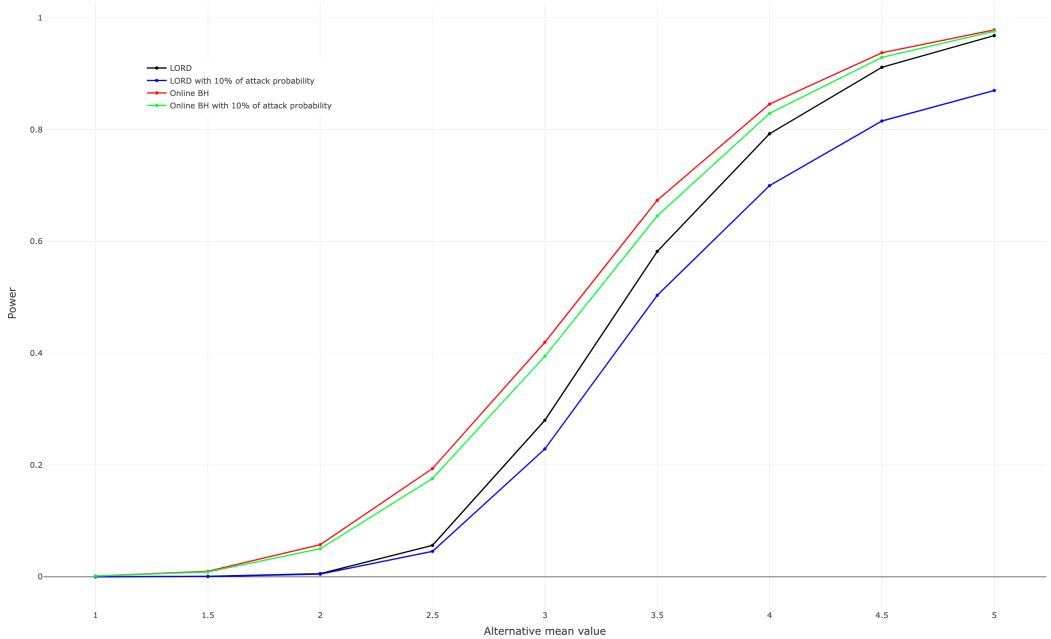


Figure 11. Power for $\mu_1 = 1, \ldots, 5$. Results are based on 2000 replications.

It is apparent that augmenting the value of μ_1, so that the alternative signal is stronger, improves the power of the hypotheses tests. Indeed, both LORD and online BH procedures exhibit a monotonic increase in power. This trend underscores the intuitive principle that

as the alternative hypothesis becomes more distinct from the null, the ability of these algorithms to identify true discoveries is enhanced, improving overall statistical power. Nevertheless, a comparative analysis between LORD (in black) and online BH (in red) reveals a substantial increase in power when Blue adopts online BH as a defender policy. Furthermore, when there is some probability of attacks $\zeta = 0.1$, the online BH (in green) shows more robustness compared to the LORD algorithm (in blue). This is evidenced by a less pronounced decrease in power, maintaining a stronger performance in the face of such adversarial conditions.

Even when we compare LORD (in black) absent of attacks with online BH (in green) with 10% of attack probability, it is clear that the latter has a better power performance for the tested values of μ_1.

Figure 12 illustrates the FDR behavior for this simulation:

As μ_1 gets larger, the FDR marginally increases with the LORD algorithm with and without 10% attack probability; this is due to more rejections inducing more wealth, resulting in more false rejections. On the other hand, the simulation indicates that the FDR trajectory experiences an ascent only up to $\mu_1 = 2$ and a decrease beyond this point when using the online BH algorithm. Importantly, the FDR remains below 0.05 even with attacks for both algorithms.

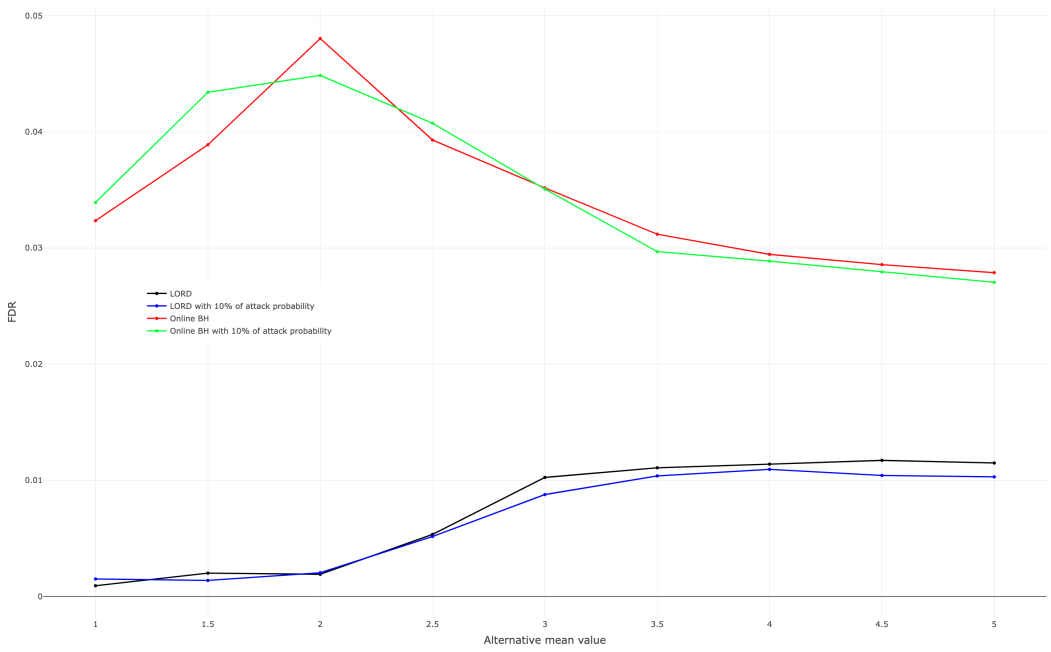


Figure 12. FDR for $\mu_1 = 1, \ldots, 5$. Results are based on 2000 replications.

Next, we investigate how the LORD and the online BH algorithms behave with greater probabilities of attacks ζ. Figure 13 depicts the power for $N = 1000$, $\pi_1 = 0.1$, and different values of ζ.

As ζ increases, the power of both algorithms diminishes, as expected. Specifically, when ζ is set to 0.5, the power of the LORD algorithm (in black) falls below 0.1, indicating a reduced capacity for making true discoveries. In contrast, the online BH algorithm (in blue) demonstrates superior performance across all simulated ζ values, maintaining a noteworthy statistical power even at $\zeta = 0.5$. Within this context, the power of the online BH approximates that of the LORD algorithm when $\zeta = 0$, showcasing its robustness in highly contested environments.

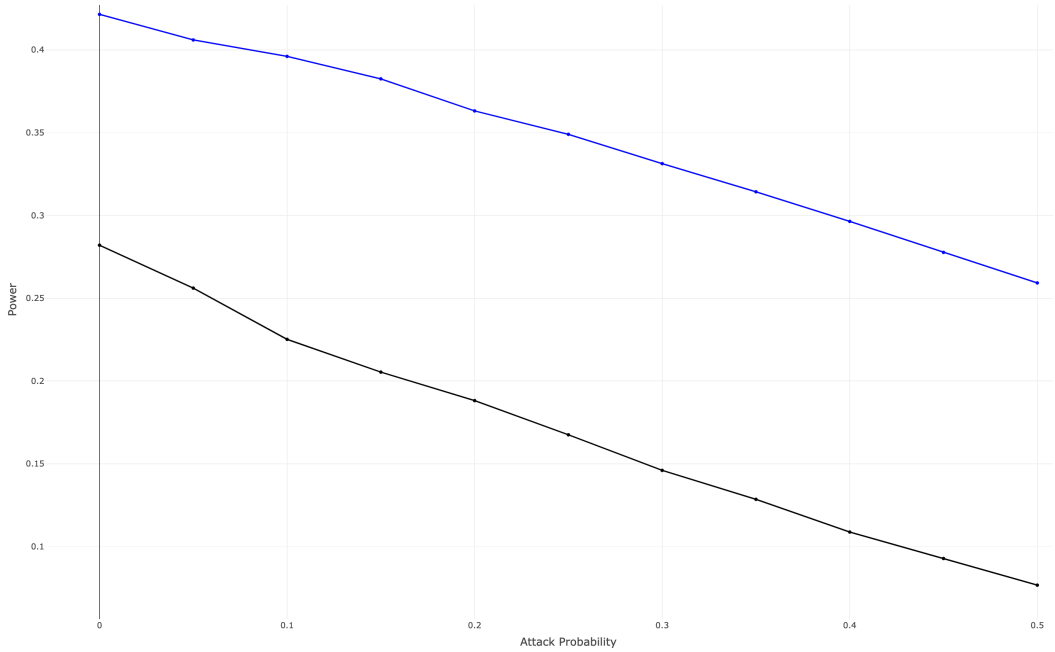


Figure 13. Power for $\zeta = 0.1, \ldots, 0.5$. Results are based on 2000 replications.

Figure 14 proves that the FDR is always below 0.05 in this simulation, which is the FDR guarantee in our case.

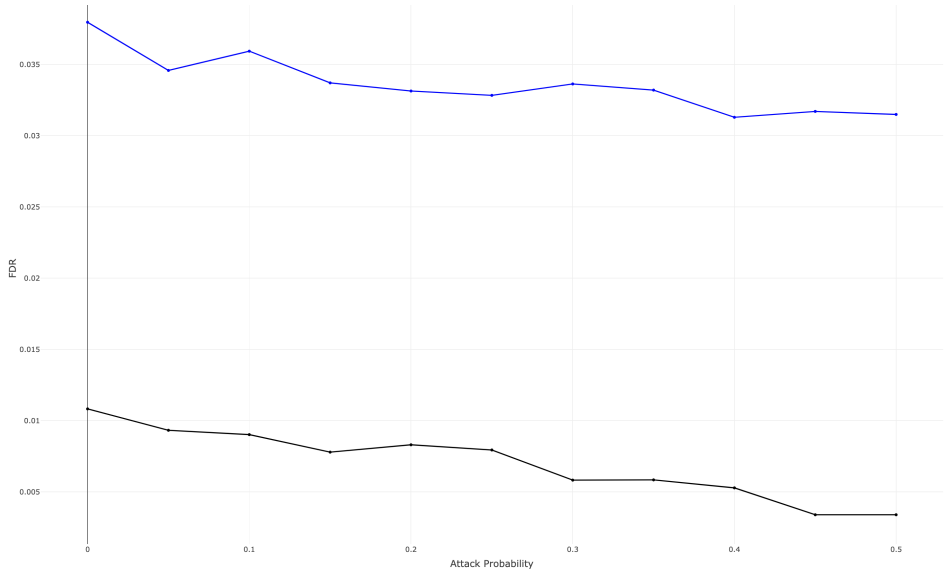


Figure 14. FDR for $\zeta = 0.1, \ldots, 0.5$. Results are based on 2000 replications.

In summary, implementing the online BH algorithm, especially when integrated with the LORD algorithm, enables Blue to protect discoveries that the cascading effect might otherwise compromise, resulting in increased statistical power with only a slight rise in the FDR. Therefore, online BH, together with LORD, is more robust against corrupted data.

5. Conclusions

This study investigated the robustness of the Levels Based On Recent Discovery (LORD) algorithm in online multiple hypothesis testing, particularly within the context of real-time data corruption. Given the increasing reliance on data-driven decision-making in maritime command and control systems such as the Brazilian Navy's SisGAAz, ensuring the integrity of these algorithms under adversarial conditions is crucial. Our research

demonstrated that while the LORD algorithm performs well in ideal conditions, it becomes vulnerable to data corruption, particularly in adversarial settings where the cascading effect can significantly degrade performance.

The results show that even a single attack on the data stream can have a pronounced cascading effect, reducing the statistical power of subsequent tests. This effect highlights a critical vulnerability in online multiple hypothesis testing algorithms like LORD, where early-stage corruption can result in a disproportionate loss of true discoveries. Our simulations of single and stochastic attacks demonstrated that data manipulation, even on a small scale, could lead to substantial loss of detection capability in maritime surveillance applications, potentially undermining mission-critical decisions.

In response, we proposed a mitigation strategy using a combination of the LORD and online BH (Benjamini–Hochberg) algorithms. This hybrid approach effectively reduced the adverse effects of data corruption, showing a considerable improvement in power recovery while maintaining false discovery rate (FDR) control.

These findings are significant for high-stakes environments like maritime operations, where real-time hypothesis testing is essential for the classification of targets and decision-making under uncertainty. The proposed enhancements offer a pathway to ensure that adversarial interference does not cripple the decision-support capabilities of systems such as SisGAAz, which rely on the rapid processing of evolving datasets. By maintaining robustness against data corruption, these systems can continue to support maritime security objectives with higher reliability. However, the study has certain limitations that should be acknowledged. The study is based on Monte Carlo simulations under specific assumptions, and no formal comparison was made with alternative online FDR procedures such as SAFFRON or GAI++, which limits the generalizability of our findings. The proposed LORD modification has not yet undergone theoretical validation, and the assumption of independence among *p*-values was adopted without further sensitivity analysis. Additionally, no real operational scenarios or military system trials were conducted, and the results remain at a proof-of-concept level.

Another limitation is the absence of formal statistical validation of our results. Although power and FDR metrics were estimated over repeated simulations (e.g., 1000 or 2000 runs), no confidence intervals or hypothesis tests were applied to assess the significance of performance differences. This aspect is particularly relevant in scenarios with marginal gains.

Future work could address these limitations by: (i) incorporating statistical tests (e.g., paired t-tests, Wilcoxon tests) and confidence intervals to validate the observed trends; (ii) formally comparing the proposed approach with other FDR algorithms; (iii) establishing theoretical guarantees for the LORD modification; (iv) testing robustness under dependent data; and (v) designing experiments with real operational data in maritime surveillance contexts. Moreover, a formal mathematical proof of the proposed mitigation techniques would provide a more rigorous foundation for their use in real-world systems.

Author Contributions: Conceptualization, V.B.A.d.A.A., C.F.S.G., M.Â.L.M., M.d.S.; methodology, V.B.A.d.A.A.; software, V.B.A.d.A.A.; validation, G.C.R. and I.P.d.A.C.; formal analysis, V.B.A.d.A.A.; investigation, V.B.A.d.A.A.; writing—original draft preparation, V.B.A.d.A.A.; writing—review and editing, G.C.R. and I.P.d.A.C.; visualization, G.C.R.; supervision, I.P.d.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported, without funding, by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brazil (CAPES).

Data Availability Statement: All data supporting the reported results are available in the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AJB	Brazilian Jurisdictional Waters
ANP	National Agency of Petroleum, Natural Gas and Biofuels
BAF	Brazilian Air Force
BH	Benjamini and Hochberg
BN	Brazilian Navy
CDF	Cumulative Density Function
FDP	False Discovery Proportion
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
GAI	Generalized Alpha-Investing
LORD	Levels Based On Recent Discovery
mFDR	Marginal False Discovery Rate
PDF	Probability Density Function
PEM	Strategic Plan of the Brazilian Navy
SAR	Search and Rescue
SisGAAz	Blue Amazon Management System

References

1. Huang, I.-L.; Lee, M.-C.; Chang, L.; Huang, J.-C. Development and Application of an Advanced Automatic Identification System (AIS)-Based Ship Trajectory Extraction Framework for Maritime Traffic Analysis. *J. Mar. Sci. Eng.* **2024**, *12*, 1672. [CrossRef]
2. Vaidya, A.; Sharma, S. Anomaly detection in the course evaluation process: A learning analytics–based approach. *Interact. Technol. Smart Educ.* **2024**, *21*, 168–187. [CrossRef]
3. Zhang, C.; Hu, D.; Yang, T. Research of artificial intelligence operations for wind turbines considering anomaly detection, root cause analysis, and incremental training. *Reliab. Eng. Syst. Saf.* **2024**, *241*, 109634. [CrossRef]
4. Xie, G.; Wang, J.; Liu, J.; Lyu, J.; Liu, Y.; Wang, C.; Zheng, F.; Jin, Y. Im-iad: Industrial image anomaly detection benchmark in manufacturing. *IEEE Trans. Cybern.* **2024**, *54*, 2720–2733. [CrossRef]
5. Belis, V.; Odagiu, P.; Aarrestad, T.K. Machine learning for anomaly detection in particle physics. *Rev. Phys.* **2024**, *12*, 100091. [CrossRef]
6. Xu, H.; Wang, Y.; Jian, S.; Liao, Q.; Wang, Y.; Pang, G. Calibrated one-class classification for unsupervised time series anomaly detection. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 5723–5736. [CrossRef]
7. Pereira, D.A.M.; dos Santos, M.; Costa, I.P.A.; Moreira, M.A.L.; Terra, A.V.; Rocha, C.S.; Gomes, C.F.S. Multicriteria and statistical approach to support the outranking analysis of the OECD countries. *IEEE Access* **2022**, *10*, 69714–69726. . [CrossRef]
8. Costa, I.P.A.; Basílio, M.P.; Maêda, S.M.N.; Rodrigues, M.V.G.; Moreira, M.A.L.; Gomes, C.F.S.; dos Santos, M. Algorithm Selection for Machine Learning Classification: An Application of the MELCHIOR Multicriteria Method. In *Modern Management Based on Big Data II and Machine Learning and Intelligent Systems III*; IOS Press: Amsterdam, The Netherlands, 2021; pp. 154–161. [CrossRef]
9. Drumond, P.; Basílio, M.P.; Costa, I.P.A.; Pereira, D.A.M.; Gomes, C.F.S.; dos Santos, M. Multicriteria Analysis in Additive Manufacturing: An ELECTRE-MOr Based Approach. In *Modern Management Based on Big Data II and Machine Learning and Intelligent Systems III*; IOS Press: Amsterdam, The Netherlands, 2021; pp. 126–132.
10. Maêda, S.M.N.; Basílio, M.P.; Costa, I.P.A.; Moreira, M.A.L.; dos Santos, M.; Gomes, C.F.S.; de Almeida, I.D.P.; Costa, A.P.A. Investments in Times of Pandemics: An Approach by the SAPEVO-M-NC Method. In *Modern Management Based on Big Data II and Machine Learning and Intelligent Systems III*; IOS Press: Amsterdam, The Netherlands, 2021; pp. 162–168. [CrossRef]
11. de Almeida, I.D.P.; Hermogenes, L.R.S.; Costa, I.P.A.; Moreira, M.A.L.; Gomes, C.F.S.; dos Santos, M.; Costa, D.O.; Gomes, I.J.A. Assisting in the choice to fill a vacancy to compose the PROANTAR team: Applying VFT and the CRITIC-GRA-3N methodology. *Procedia Comput. Sci.* **2022**, *214*, 478–486. [CrossRef]
12. Erhan, L.; Ndubuaku, M.; Di Mauro, M.; Song, W.; Chen, M.; Fortino, G.; Bagdasar, O.; Liotta, A. Smart anomaly detection in sensor systems: A multi-perspective review. *Inf. Fusion* **2021**, *67*, 64–79. [CrossRef]
13. Van Slyke, C.; Kamis, A. The Limits of Empiricism: A Critique of Data-Driven Theory Development. *ACM SIGMIS Database* **2024**, *55*, 120–146. [CrossRef]
14. Bakr, M.E. On basic life testing issues in medical research using non-parametric hypothesis testing. *Qual. Reliab. Eng.* **2024**, *40*, 1002–1013. [CrossRef]

15. Wu, Y.; Anwar, A.; Quynh, N.N.; Abbas, A.; Cong, P.T. Impact of economic policy uncertainty and renewable energy on environmental quality: Testing the LCC hypothesis for fast growing economies. *Environ. Sci. Pollut.* **2024**, *31*, 36405–36416. [CrossRef] [PubMed]
16. Wijaya, O.; Said, H. The mediating role of sustainable supply chain management on entrepreneurship strategy, social capital and SMEs' financial and non-financial performance. *Uncertain Supply Chain Manag.* **2024**, *12*, 557–566. [CrossRef]
17. Quin, F.; Weyns, D.; Galster, M.; allen, C.C. A/B testing: A systematic literature review. *J. Syst. Softw.* **2024**, *211*, 112011. [CrossRef]
18. Rahman, S.U.; Zhao, S.; Junaid, D. The FDI inflows in low-income and lower-middle-income countries: The moderating role of military expenditure. *Int. J. Sustain.* **2024**, *16*, 131–153.
19. Costa, A.P.A.; Choren, R.; Pereira, D.A.M.; Terra, A.V.; Costa, I.P.A.; Junior, C.S.R.; dos Santos, M.; Gomes, C.F.S.; Moreira, M.A.L. Integrating multicriteria decision making and principal component analysis: A systematic literature review. *Cogent Eng.* **2024**, *11*, 2374944. [CrossRef]
20. Allen, L.; Lu, H.; Cordiner, J. Knowledge-Enhanced Spatiotemporal Analysis for Anomaly Detection in Process Manufacturing. *Comput. Ind.* **2024**, *161*, 104111. [CrossRef]
21. Javanmard, A.; Montanari, A. Online rules for control of false discovery rate and false discovery exceedance. *Ann. Stat.* **2018**, *46*, 526–554. [CrossRef]
22. Rogerson, P.A. Testing Hypotheses When You Have More Than a Few. *Ann. Inst. Stat. Math.* **2024**, *57*, 175–190. [CrossRef]
23. Chen, S.; Arias-Castro, E. On the power of some sequential multiple testing procedures. *Ann. Inst. Stat. Math.* **2024**, *73*, 311–336. [CrossRef]
24. Li, A.; Wang, J.; Baruah, S.; Sinopoli, B.; Zhang, N. An empirical study of performance interference: Timing violation patterns and impacts. In Proceedings of the 30th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS'24), Hong Kong, China, 13–16 May 2024.
25. Zhang, H.; Zuo, Z.; Li, Z.; Ma, L.; Liang, S.; Lü, Q.; Zhou, H. Leak detection for natural gas gathering pipelines under corrupted data via assembling twin robust autoencoders. *Process Saf. Environ.* **2024**, *188*, 492–513. [CrossRef]
26. Narayanan, S.; Maheswari, S.; Zephan, P. Real-Time Monitoring of Data Pipelines: Exploring and Experimentally Proving that the Continuous Monitoring in Data Pipelines Reduces Cost and Elevates Quality. *EAI Endorsed Trans. Scalable Inf. Syst.* **2024**, *11*, 1. [CrossRef]
27. Andrade, I.; Franco, L.G. Blue Amazon as Brazil's maritime frontier: Strategic importance and imperatives for national defense. In *Brazilian Borders: A Public Policy Assessment*; IPEA: Brasilia, Brazil, 2018; pp. 151–178.
28. Andrade, I.; Rocha, A.; Franco, L.G. Blue Amazon Management System (SisGAAz): Sovereignty, Surveillance and Defense of the Brazilian Jurisdictional Waters; Discussion Paper; pp. 1–35. Available online: http://repositorio.ipea.gov.br/handle/11058/10978 (accessed on 4 August 2025).
29. Tracing the History of Exploration in the Brazilian Pre-Salt Oil Region. Available online: https://www.offshore-technology.com/features/pre-salt-oil-region-brazil/ (accessed on 31 January 2024).
30. *Brazil. Law No. 8,617, of January 4, 1993. Provides for the Territorial Sea, the Contiguous Zone, the Exclusive Economic Zone, and the Brazilian Continental Shelf, and Other Provisions*; Official Gazette of the Union: Brasília, Brazil, 1993; p. 57.
31. Brazil's Pre-Salt Oil Gains Unprecedented Global Popularity. Available online: https://finance.yahoo.com/news/brazils-pre-salt-oil-gains-210000522.html/ (accessed on 31 August 2023).
32. Rodrigues, S. Strategic plan of the Brazilian Navy. *J. Braz. Naval War Coll.* **2021**, 13–30.
33. Gerhardinger, L.; Gorris, P.; Gonçalves, L.; Herbst, D.; Vila Nova, D.; de Carvalho, F.G.; Glaser, M.; Zondervan, R.; Glavovic, B. Healing Brazil's Blue Amazon: The Role of Knowledge Networks in Nurturing Cross-Scale Transformations at the Frontlines of Ocean Sustainability. *Front. Mar. Sci.* **2018**, *4*, 395–412. [CrossRef]
34. Devore, J.L. Hypotheses and Test Procedures. In *Probability and Statistics for Engineering and the Sciences*; Cengage Learning: Boston, MA, USA, 2006; pp. 311–324.
35. Fisher, R.A. Statistical Methods for Research Workers. In *Breakthroughs in Statistics*; Springer: New York, NY, USA, 1992; pp. 66–70.
36. Efron, B. False Discovery Rate Control. In *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*; Cambridge University Press: New York, NY, USA, 2010; pp. 46–52.
37. Austin, S.R.; Dialsingh, I.; Altman, N. Multiple hypothesis testing: A review. *J. Indian Soc. Agric. Stat.* **2014**, *68*, 303–314.
38. Ramdas, A.; Yang, F.; Wainwright, M.J.; Jordan, M.I. Online control of the false discovery rate with decaying memory. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 9 December 2017.
39. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
40. Robertson, D.; Wason, J.M.S.; Ramdas, A. Online Multiple Hypothesis Testing. *Statist. Sci.* **2023**, *38*, 557–575. [CrossRef]
41. Benjamini, Y.; Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing Under Dependency. *Ann. Stat.* **2001**, *29*, 1165–1188. [CrossRef]

42. Optional Material: Online False Discovery Rate Control. Available online: https://data102.org/fa20/assets/notes/notes_online_FDR.pdf (accessed on 23 September 2024).
43. Foster, D.; Stine, R. Alpha-Investing: A Procedure for Sequential Control of Expected False Discoveries. *J. R. Stat. Soc. Ser. B* **2008**, *70*, 429–444. [CrossRef]
44. Aharoni, E.; Rosset, S. Generalized α-investing: Definitions, optimality results and application to public databases. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2014**, *76*, 771–794. [CrossRef]
45. Ramdas, A.; Zrnic, T.; Wainwright, M.; Jordan, M. SAFFRON: An adaptive algorithm for online control of the false discovery rate. *Proc. Mach. Learn. Res.* **2018**, *80*, 4286–4294.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.