



Research article

Detection algorithm for data integrity attacks in quantized FIR system identification

Jingrong Liu^{1,2}, Shuying Sun^{1,2} and Qingxiang Zhang^{1,2,*}

¹ School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

² Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, Beijing 100083, China

* **Correspondence:** Email: qxzhang1001@xs.ustb.edu.cn.

Abstract: Targeting data integrity security threats, this paper investigates the anomaly detection and performance evaluation problems of data integrity attacks for finite impulse response systems with binary quantized observations. Breaking through the limitations of existing detection methods that highly depend on precise prior models and struggle with quantization non-linearity, we start from the inherent mechanisms of parameter identification and fully exploit the persistent excitation property of the system input. By cleverly designing and partitioning the observation dataset into two independent full-rank subsets, a novel detection residual is constructed. Without relying on prior knowledge of attack types, this method can provide unified and effective anomaly detection and early warning for whether the system is suffering from denial of service attacks, tampering attacks, or replay attacks. Furthermore, to measure the detection performance of the algorithm, we strictly prove the asymptotic normality of the detection residuals under both attack-free and various attack environments and derive explicit calculation expressions for the false alarm rate and missed detection rate. Finally, the impacts of factors on the algorithm's performance are analyzed. Numerical simulations verify the effectiveness of the proposed detection algorithm and the accuracy of the theoretical derivations.

Keywords: attack detection; quantization; system identification; data integrity attacks

1. Introduction

In recent years, the widespread application of cyber-physical systems (CPSs) in fields such as industrial automation and smart grids has greatly enhanced the intelligence of physical processes [1–4]. However, the openness of communication networks also renders CPSs highly vulnerable to malicious cyberattacks [5, 6]. Within the defense-in-depth architecture of CPSs, attack detection serves as the

crucial first line of defense for sensing system anomalies and triggering active defense mechanisms.

Three typical data integrity attacks include denial of service (DoS) attacks, tampering (TA) attacks, and replay (RE) attacks [7–9]. Cyberattacks targeting data integrity can covertly undermine the authenticity and reliability of system data, leading to devastating physical consequences. In existing attack detection mechanisms, most studies operate under the assumption that the precise mathematical model of the system is already known [10, 11]. However, in practical engineering, system models are often obtained from input-output data via system identification techniques. If the measurement data utilized for parameter identification is maliciously tampered with or blocked by attackers, the identification algorithm will yield erroneous system parameters, subsequently causing all descendant detectors and controllers based on this model to fail completely [12, 13]. Therefore, achieving early detection of cyberattacks during the system parameter identification phase is the foundational prerequisite for ensuring the secure operation of the entire CPS.

Furthermore, constrained by network bandwidth and sensor costs, quantized sensors (e.g., binary sensors) are widely deployed in modern industrial systems [14, 15]. Binary quantized measurements not only discard a vast amount of amplitude information but also introduce severe data non-linearity, making it exceedingly difficult to directly apply traditional detection theories based on linear innovations or continuous residuals. Hence, the core motivation of this paper is to overcome the obstacle of binary quantization non-linearity and to explore how to design a unified detection framework for identifying the occurrence of data integrity attacks by exploiting the inherent mechanisms of the system identification algorithm itself.

Centering on the issue of attack detection, academia has conducted extensive and in-depth research in the fields of state estimation and control, primarily focusing on the three typical attack modes that compromise data integrity. Regarding packet loss and communication interruptions caused by DoS attacks, existing research mainly aims to enhance detection robustness in uncertain environments. Xu et al. [16] developed a residual change detection mechanism based on a novel probabilistic event-triggered protocol to reduce communication overhead while maintaining detection sensitivity. Liu et al. [17] thoroughly investigated the asymptotic impact mechanism of random packet dropouts on observer detection performance. Additionally, Guo et al. [18] designed a zonotope-based set-valued state observer, achieving robust detection of abnormal states through an ingenious perturbation and error decoupling approach. For TA attacks aimed at covertly manipulating system data, research focuses on the precise extraction and isolation of abnormal features. Jia et al. [19] proposed an attack identification strategy based on abnormal data change features, targeting the cooperative characteristics of multi-agent systems. Li et al. [20] introduced state isolation techniques, effectively handling the dynamic detection of tampering actions by relying on external observers. Yan et al. [21] proposed a coding transmission strategy based on bandwidth allocation from the perspective of communication channels, realizing online authentication against partial tampering attacks via a reverse data transmission mechanism. For highly stealthy RE attacks, the mainstream defense paradigm lies in actively disrupting the static system features relied upon by attackers. Zhao et al. [22] characterized residual statistics using the parity space method and proposed targeted passive defense designs to enhance detection sensitivity. Fritz and Zhang [23] stripped away the stealthiness of replay attacks by integrating time-varying permutation matrices into sensor and actuator channels to introduce system dynamics. Zhao et al. [24] proposed injecting physical watermarking with specific statistical properties into control inputs as a valid credential to distinguish authentic data from lagged replayed data.

Although the aforementioned works achieved fruitful results in attack detection at the state estimation and control levels, these methods highly depend on precise and known prior nominal system models. Once the perspective shifts back to the quantized system identification phase, existing methods become difficult to apply directly. Especially under the highly non-linear binary quantization mechanism, how to directly construct a unified detection statistic through the internal convergence mechanism of the system identification algorithm itself, without prior knowledge of the model, to effectively identify the various data integrity attacks mentioned above remains an urgent theoretical gap to be filled in related literature.

Targeting a class of finite impulse response (FIR) systems with binary quantized observations, this paper deeply investigates the detection and performance evaluation problems under data integrity attacks from the perspective of parameter identification. By fully exploiting the persistent excitation (PE) property of the system input signals, an attack detection function is reversely designed through the asymptotic convergence property of the parameter identification algorithm, thereby allowing it to serve as a unified detection statistic without requiring prior knowledge of the attack type. For DoS, TA, and RE attacks that induce a non-negligible inconsistency between the two local identification results, the proposed residual provides a unified anomaly indicator without requiring prior knowledge of the attack type. The detectability boundary associated with statistically compatible RE is also discussed explicitly. Specifically, by analyzing the characteristics of periodic inputs, a multivariate detection function is constructed based on the optimal estimation algorithm; this is then extended to quantized input scenarios to establish an implementable detection function, and subsequently, a detection algorithm is designed based on the magnitude of the detection function to determine the presence of an attack. By analyzing the impact of data integrity attacks on the detection function, and based on the established asymptotic normality, approximate calculation expressions for the missed detection rate and false alarm rate are provided using the probability density of the multivariate normal distribution as performance evaluation metrics for the detection algorithm. In addition, factors affecting the performance of the detection algorithm are discussed from multiple dimensions, including data length, threshold vectors, and attack strategies.

The main contributions of this paper can be summarized in the following three aspects:

- This paper breaks the traditional mindset of designing separate detectors for different attacks. By cleverly designing quantized input signals, two local parameter estimations are performed. The residual vector constructed from this paradigm can serve as a unified detection variable, providing unified and effective anomaly detection and early warning for whether the system is suffering from a DoS attack, a TA attack, or an RE attack.
- Under the extreme non-linear challenges brought by binary quantization, the asymptotic normality of the detection residuals under both attack-free and various attack states is rigorously proven. Based on the multivariate normal distribution, explicit theoretical calculation formulas for the false alarm rate and missed detection rate in attack detection are provided, thereby systematically measuring the performance of the attack detection algorithm.
- The internal and external factors influencing detection performance are analyzed. The impacts of data length, threshold vectors, and attack strategies on the performance of the detection algorithm are rigorously discussed, providing a basis for algorithm optimization and solid theoretical support for improving decision-making accuracy.

Recent studies have explicitly investigated parameter identification with quantized observations in the presence of cyberattacks, including tampering-resilient identification of quantized FIR systems [6],

Table 1. Comparison with closely related secure identification studies.

Work	System / observations	Attack	Primary objective	Main mechanism	Unified detection	Explicit FAR/MDR
[6]	FIR / multilevel	TA	Identification	Joint estimation	No	No
[9]	FIR / binary	RE	Identification	Replay-specific estimation	No	No
[12]	Hammerstein / quantized	RE	Identification	Compensation	No	No
This work	FIR / binary	DoS, TA, RE	Attack detection	Two-subset residual	Yes	Yes

replay-resilient identification of FIR systems with binary-valued observations [9], and intrusion-tolerant identification of quantized nonlinear Hammerstein systems under replay attacks [12]. These studies demonstrate that system identification can itself be made resilient to specific attack mechanisms. These studies and the present work address complementary questions. The former primarily ask how to maintain or recover consistent parameter estimation when a particular attack mechanism has been specified. In contrast, our objective is to determine whether the data used for identification have been attacked before the attack type is known. To this end, instead of estimating or compensating for a specific attack strategy, we exploit persistent excitation to construct two disjoint full-rank regressor subsets, obtain two local parameter estimates, and use their inconsistency as a unified detection residual. The same statistic is applied to DoS, tampering, and replay attacks. Furthermore, its asymptotic distribution is characterized to obtain explicit calculations of performance metrics. Table 1 summarizes these distinctions.

The remainder of this paper is structured as follows: Section 2 presents the quantized FIR system model and the description of the three data integrity attacks. Section 3 derives the unified attack detection mechanism based on the identification algorithm in detail. Section 4 analyzes the asymptotic normality of the detection residuals and derives the calculation methods for the missed detection rate and false alarm rate, as well as the impact of attack strategies. Section 5 verifies the correctness of the theoretical results through numerical simulations. Finally, Section 6 concludes the paper and outlines future research directions.

2. Problem formulation

2.1. System description

Consider the following single-input single-output (SISO) discrete-time FIR system:

$$\begin{aligned} y_k &= \theta_1 u_k + \theta_2 u_{k-1} + \cdots + \theta_n u_{k-n+1} + d_k \\ &= \omega_k \theta + d_k, \end{aligned} \quad (1)$$

where d_k is the system noise; u_k is the system input; $\omega_k = [u_k, \dots, u_{k-n+1}]$ is the regression vector composed of current and historical inputs; $\theta = [\theta_1, \theta_2, \dots, \theta_n]^T$ is the unknown system parameter; and y_k is the internal system output.

Assume that the output is measured by a binary sensor with a threshold C . The quantized observation s_k^0 can be represented by the following indicator function:

$$s_k^0 = I_{\{y_k \leq C\}} = \begin{cases} 1, & y_k \leq C; \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Assumption 1. The system noise sequence $\{d_k\}$ consists of independent and identically distributed Gaussian random variables with zero mean and variance σ^2 . Its cumulative distribution function and probability density function are denoted by $\Psi(\cdot)$ and $\psi(\cdot)$, respectively.

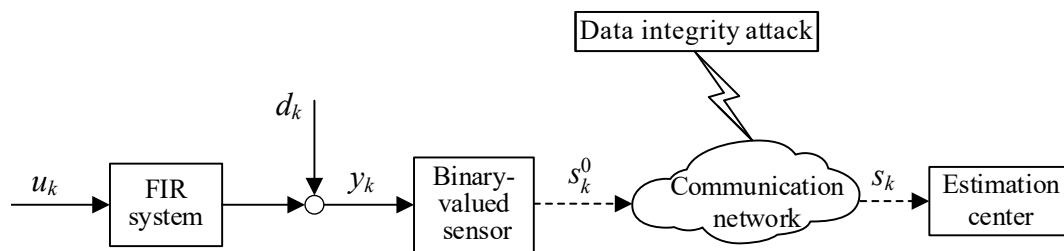


Figure 1. Schematic diagram of the quantized FIR system under network attacks.

As illustrated in Figure 1, the authentic binary measurement data s_k^0 is transmitted to the estimation center via an unsecured communication network. Let s_k denote the data ultimately received by the estimation center. During transmission, the data is highly vulnerable to malicious network attacks targeting data integrity. In an ideal attack-free environment, the received data is identical to the authentic data, i.e., $s_k \equiv s_k^0$.

2.2. Attack models

This paper considers three typical network attacks targeting data integrity. For the convenience of a unified theoretical analysis, an indicator variable $\delta_k \in \{0, 1\}$ is introduced to characterize the attack status: $\delta_k = 0$ indicates that an attack occurs at time k , while $\delta_k = 1$ indicates no attack.

1) DoS attack:

The attacker maliciously blocks data transmission within the communication network, preventing the receiver from obtaining valid information and resulting in packet loss. Since the measurement data only consists of 0 and 1, the receiver typically marks the lost data as a missing state (denoted as $\emptyset$). The attack strategy can be described by the probability pair (p_{d_0}, p_{d_1}) , meaning the probability of an attack occurrence depends on the authentic measurement data s_k^0 :

$$\begin{cases} \Pr\{\delta_k = 0 | s_k^0 = 0\} = p_{d_0}; \\ \Pr\{\delta_k = 0 | s_k^0 = 1\} = p_{d_1}. \end{cases} \quad (3)$$

2) TA attack:

This type of attack aims to alter or manipulate the transmitted data without authorization, causing the estimation center to receive false information. Its destructive mechanism is akin to that of deception attacks. Under the binary measurement framework, attacks manifest as “bit-flipping” of the original data. The attack model is established such that the received data can be expressed as $s_k = \delta_k s_k^0 + (1 - \delta_k) \hat{s}_k$, where $\hat{s}_k$ is the injected false data. The attack strategy is governed by the flipping probabilities (p_{t_0}, p_{t_1}) as follows:

$$\begin{cases} \Pr\{s_k = 0 | s_k^0 = 1\} = p_{t_0}; \\ \Pr\{s_k = 1 | s_k^0 = 0\} = p_{t_1}. \end{cases} \quad (4)$$

3) RE attack:

The attacker intercepts and records valid data within a historical time window and maliciously resends these outdated data at the current moment to replace the authentic data, thereby stealthily degrading system performance. The attack model is described as $s_k = \delta_k s_k^0 + (1 - \delta_k) s_{k-\tau}^0$, where $s_{k-\tau}^0$ represents the historical data intercepted by the attacker at time $k - \tau$. The replay delay τ is a non-negative integer

discrete random variable representing the replay intensity. Its probability distribution vector is

$$\begin{bmatrix} \Pr(\tau = 0) \\ \Pr(\tau = 1) \\ \vdots \\ \Pr(\tau = m) \end{bmatrix} = \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_m \end{bmatrix} = \mathbf{P}, \quad (5)$$

where m is the maximum replay delay. This replay attack strategy can be abbreviated as $(m, \mathbf{P})$.

Remark 1. *The three attacks considered above act only on the binary measurement channel. The system input u_k is known to the estimation center and is not modified by the attacker. Therefore, the corresponding regression vector ω_k remains available even when its binary measurement is dropped, tampered with, or replayed.*

2.3. Problem objectives

The primary task of the estimation center is to utilize the known system input sequence and the received available data (which may contain corrupted data) to design an efficient detection algorithm for identifying the occurrence of the aforementioned attacks. To this end, three core issues urgently need to be addressed:

- How can the estimation center design a unified statistical detection algorithm to identify the occurrence of attacks?
- What theoretical metrics should be employed to quantitatively evaluate the performance of the proposed detection algorithm?
- How do system factors and attack strategies affect these performance metrics, and how can decision accuracy be improved through theoretical analysis?

The subsequent sections will systematically address these questions: first, developing the unified attack detection algorithm; next, calculating the algorithm's performance evaluation metrics; and finally, discussing the influencing factors of these metrics.

3. Design of attack detection algorithm

The core of attack detection lies in constructing an appropriate statistical test function (detection variable) and establishing a decision criterion based on this function. Specifically, when the value of the detection function falls within the expected range of normal system operation, it is determined that no attack has occurred; conversely, if a significant deviation occurs, the system is deemed to be under attack. By fully exploiting the PE property of the system inputs and the asymptotic convergence nature of the parameter identification algorithm, this paper derives and designs a unified attack detection algorithm.

3.1. Baseline detection algorithm under periodic inputs

In this subsection, network attacks are temporarily excluded (i.e., assuming $s_k \equiv s_k^0$) to establish the theoretical baseline for the algorithm design.

For system (1), assume the system input sequence $\{u_k\}$ is periodic with a period of n (i.e., the input period equals the number of unknown system parameters), satisfying $u_k = u_{k+n}$ ($\forall k \geq 1$). Consequently,

the regression vectors also satisfy $\omega_k = \omega_{k+n}$. The invertible input matrix constructed from the input sequence within one period is denoted as $\Omega = [\omega_1^T, \dots, \omega_n^T]^T$. For a given positive integer $L > 0$, let the total length of observation data be $N = Ln$. The asymptotically optimal estimation algorithm for the unknown parameter vector θ in the sense of the Cramér-Rao lower bound can be expressed as follows [25]:

$$\begin{aligned}\widehat{\theta}_N &= \Omega^{-1} [\xi_{N,1}, \dots, \xi_{N,n}]^T, \\ \xi_{N,i} &= C - \mathcal{H}\left(\frac{1}{L} \sum_{l=1}^L s_{(l-1)n+i}\right), \quad i = 1, 2, \dots, n,\end{aligned}$$

where θ_N represents the estimated value of the parameter θ ; $\mathcal{H}(\cdot) \triangleq \Psi^{-1}(\cdot)$ denotes the inverse function of $\Psi(\cdot)$. As the data length approaches infinity, the estimated value converges uniformly to the true value, i.e., $\lim_{L \rightarrow \infty} \widehat{\theta}_N = \theta$ (with probability 1).

Furthermore, if the period of the system input sequence is increased to t where $t > n$, one can arbitrarily select n distinct time indices from the set $\{1, 2, \dots, t\}$, denoted as $b_1, b_2, \dots, b_n$, to construct a full-rank input submatrix $\Omega = [\omega_{b_1}^T, \omega_{b_2}^T, \dots, \omega_{b_n}^T]^T$. With an observation data length of $N = Lt$, as long as Ω is invertible, the following estimation still has the aforementioned asymptotic convergence property:

$$\begin{aligned}\widehat{\theta}_N &= \Omega^{-1} [\xi_{N,b_1}, \dots, \xi_{N,b_n}]^T, \\ \xi_{N,b_i} &= C - \mathcal{H}\left(\frac{1}{L} \sum_{l=1}^L s_{(l-1)t+b_i}\right), \quad i = 1, 2, \dots, n.\end{aligned}$$

To design a detector while avoiding data overlap and correlation, we consider setting the period of the input sequence to $2n$, satisfying $u_k = u_{k+2n}$ and $\omega_k = \omega_{k+2n}$. Within one period, the first n regressions are utilized to construct the submatrix

$$\Omega_1 = [\omega_1, \omega_2, \dots, \omega_n]^T = \begin{bmatrix} u_1 & u_{2n} & u_{2n-1} & \cdots & u_{n+2} \\ u_2 & u_1 & u_{2n} & \cdots & u_{n+3} \\ u_3 & u_2 & u_1 & \cdots & u_{n+4} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_n & u_{n-1} & u_{n-2} & \cdots & u_1 \end{bmatrix}. \quad (6)$$

Similarly, the subsequent n regressions are utilized to construct the submatrix

$$\Omega_2 = [\omega_{n+1}, \omega_{n+2}, \dots, \omega_{2n}]^T = \begin{bmatrix} u_{n+1} & u_n & u_{n-1} & \cdots & u_2 \\ u_{n+2} & u_{n+1} & u_n & \cdots & u_3 \\ u_{n+3} & u_{n+2} & u_{n+1} & \cdots & u_4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{2n} & u_{2n-1} & u_{2n-2} & \cdots & u_{n+1} \end{bmatrix}. \quad (7)$$

Based on these two independent datasets, two sets of local parameter estimates are calculated respectively:

$$\begin{aligned}\widehat{\theta}_N^{(1)} &= \Omega_1^{-1} [\xi_{N,1}, \dots, \xi_{N,n}]^T, \\ \widehat{\theta}_N^{(2)} &= \Omega_2^{-1} [\xi_{N,n+1}, \dots, \xi_{N,2n}]^T,\end{aligned}$$

$$\xi_{N,i} = C - \mathcal{H}\left(\frac{1}{L_N} \sum_{l=1}^{L_N} s_{2(l-1)n+i}\right), \quad i = 1, 2, \dots, 2n,$$

where $L_N = \lfloor \frac{N}{2n} \rfloor$, and $\lfloor \cdot \rfloor$ denotes the floor operation. We construct the following detection residual vector function:

$$\Upsilon_N = \widehat{\theta}_N^{(1)} - \widehat{\theta}_N^{(2)}.$$

In an attack-free environment, as $N \rightarrow \infty$, it is evident that $\widehat{\theta}_N^{(1)} \rightarrow \theta$ and $\widehat{\theta}_N^{(2)} \rightarrow \theta$, w.p.1. Therefore, the detection residual converges to the zero vector, i.e.,

$$\Upsilon_N = \widehat{\theta}_N^{(1)} - \widehat{\theta}_N^{(2)} \rightarrow \mathbf{0}, \quad \text{w.p.1, as } N \rightarrow \infty. \quad (8)$$

Based on the above asymptotic analysis, we establish the following attack decision criterion:

$$\Upsilon_N = \begin{cases} \mathbf{0}, & \text{without attacks;} \\ \text{others,} & \text{with attacks.} \end{cases} \quad (9)$$

3.2. Extension to general quantized inputs

To improve communication efficiency and conserve network resources, quantized input signals are frequently adopted in practical industrial control systems. Quantized inputs are typically non-periodic, thus exhibiting broader applicability compared to periodic inputs. Assume the quantized system input u_k has a discrete value of a , i.e., $u_k \in \{r_1, r_2, \dots, r_a\}$. Consequently, the input regression vector ω_k possesses $H = a^n$ possible patterns, which are collectively denoted as

$$\begin{cases} \pi_1 = [r_1, r_1, \dots, r_1]_{1 \times n}, \\ \pi_2 = [r_1, r_1, \dots, r_2]_{1 \times n}, \\ \vdots \\ \pi_H = [r_a, r_a, \dots, r_a]_{1 \times n}. \end{cases} \quad (10)$$

For a data length of N , a weighting matrix $\Lambda_{N,H}$ can be constructed based on the frequencies of each regressor, and its asymptotic limit is defined as Λ_H . Then

$$\Lambda_{N,H} = \begin{bmatrix} \sqrt{\frac{N_1}{N}} \pi_1 \\ \sqrt{\frac{N_2}{N}} \pi_2 \\ \vdots \\ \sqrt{\frac{N_H}{N}} \pi_H \end{bmatrix}_{H \times n} \rightarrow \Lambda_H = \begin{bmatrix} \sqrt{\rho_1} \pi_1 \\ \sqrt{\rho_2} \pi_2 \\ \vdots \\ \sqrt{\rho_H} \pi_H \end{bmatrix}_{H \times n}, \quad N \rightarrow \infty, \quad (11)$$

where the empirical frequency $N_\epsilon = \sum_{k=1}^N I_{\{\omega_k = \pi_\epsilon\}}$ satisfies $\sum_{\epsilon=1}^H N_\epsilon = N$; and the steady-state probability ρ_ϵ satisfies $\sum_{\epsilon=1}^H \rho_\epsilon = 1$, $\epsilon = 1, 2, \dots, H$.

If the system is not under attack, parameter estimates converging to the true value θ can be obtained by applying the following weighted least-squares-type algorithm:

$$\widehat{\theta}_N = (\Lambda_{N,H}^T \mathcal{B} \Lambda_{N,H})^{-1} \Lambda_{N,H}^T \mathcal{B} [\varpi_{N,1}, \dots, \varpi_{N,H}]^T, \quad (12)$$

$$\varpi_{N,\epsilon} = \sqrt{\frac{N_\epsilon}{N}}(C - \mathcal{H}(\frac{1}{N_\epsilon} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_\epsilon\}})) \quad \epsilon = 1, 2, \dots, H. \quad (13)$$

Here, $\mathcal{B} = \text{diag}(\mathcal{A}_\epsilon)(\mathcal{A}_\epsilon > 0)$ is a positive definite diagonal weighting matrix utilized to optimize the asymptotic variance and convergence rate of the algorithm. As $N \rightarrow \infty$, the empirical frequency $\frac{1}{N_\epsilon} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_\epsilon\}}$ converges in probability to the true distribution probability $\Psi(C - \pi_\epsilon \theta)$.

The effective execution of the Eqs (12) and (13) relies on the input sequence satisfying a specific excitation condition, defined as follows.

Definition 1 (PE Condition). *For a data sequence of length N , if there exist H_0 non-zero steady-state probabilities $\rho_l > 0$ ($l = 1, \dots, H_0$) such that the column rank of the corresponding weighting matrix Λ_{N,H_0} is full (i.e., the rank equals the number of parameters), the input sequence is said to satisfy the persistent excitation condition.*

To ensure that the detection residual can be effectively constructed, we propose the following assumption regarding the excitation intensity.

Assumption 2. *Assume the system input is sufficiently rich such that there exist at least $h \geq 2n$ regressions satisfying the persistent excitation condition, and these regressions can be partitioned into two disjoint subsets to construct two full-rank matrices with a column rank of n (denoted as $\Lambda_{N,1}$ and $\Lambda_{N,2}$, respectively).*

Remark 2. *Assumption 2 is determined entirely by the known input sequence and its regression vectors. Hence, the PE condition and the full-rank structures of $\Lambda_{N,1}$ and $\Lambda_{N,2}$ are not altered by the measurement-channel attacks defined in Section 2.2. For TA and RE attacks, only the received binary observations are modified or replaced, while the occurrence number N_i of each regressor π_i remains unchanged. Therefore, the two full-rank regressor subsets remain unchanged. For a DoS attack, some measurements associated with a regressor may be unavailable, so the number of effective observations is reduced. As long as each selected regressor retains a positive probability of successful reception, its number of valid observations still tends to infinity as $N \rightarrow \infty$. Thus, DoS affects the effective sample size but does not destroy the input-side PE or full-rank condition. Accordingly, throughout this paper, the two subsets refer to two disjoint full-rank regressor subsets constructed from the known input sequence.*

Theorem 1. *Under the condition that Assumption 2 holds, in an attack-free scenario, as $N \rightarrow \infty$, the detection residual Υ_N given by the Eqs (14)–(18) asymptotically converges to $\mathbf{0}$.*

$$\Upsilon_N = \widehat{\theta}_N^{(1)} - \widehat{\theta}_N^{(2)}, \quad (14)$$

$$\widehat{\theta}_N^{(1)} = \Phi_{N,1}[\varpi_{N,1}, \dots, \varpi_{N,\frac{h}{2}}]^T, \quad (15)$$

$$\widehat{\theta}_N^{(2)} = \Phi_{N,2}[\varpi_{N,\frac{h}{2}+1}, \dots, \varpi_{N,h}]^T, \quad (16)$$

$$\Phi_{N,\nu} = (\Lambda_{N,\nu}^T \mathcal{B} \Lambda_{N,\nu})^{-1} \Lambda_{N,\nu}^T \mathcal{B}, \quad \nu = 1, 2, \quad (17)$$

$$\varpi_{N,i} = \sqrt{\frac{N_\epsilon}{N}}(C - \mathcal{H}(\frac{1}{N_\epsilon} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_\epsilon\}})), \quad i = 1, 2, \dots, h, \quad (18)$$

where the matrices $\Lambda_{N,1}$ and $\Lambda_{N,2}$ are constructed by weighting the first $\frac{h}{2}$ and the subsequent $\frac{h}{2}$ sufficiently excited regressions, respectively.

Proof. We provide a design in Example 1 that ensures Assumption 2 holds universally. Assumption 2 guarantees that both $\Lambda_{N,1}$ and $\Lambda_{N,2}$ are full-rank PE matrices. Consequently, the inverse matrices obtained through the inversion operation are well-defined. Combining the law of large numbers and the premise that the system is attack-free, according to the Eqs (12) and (13), we obtain $\lim_{N \rightarrow \infty} \widehat{\theta}_N^{(1)} = \theta$ and $\lim_{N \rightarrow \infty} \widehat{\theta}_N^{(2)} = \theta$. Therefore, $\lim_{N \rightarrow \infty} \Upsilon_N = \mathbf{0}$. The theorem is thus proven. $\square$

To demonstrate the realizability of Assumption 2 in physical systems, we provide the following illustration.

Example 1. Consider the number of discrete quantization steps a and the regression dimension n . The total number of possible regressions generated by the system input is $H = a^n$. To satisfy Assumption 2, we require $H \geq h \geq 2n$. Due to the exponential function growing much faster than a linear function, when $a \geq 2$ (e.g., a simple binary input signal) and $n \geq 1$, the inequality $2^n \geq 2n$ always holds (equality holds for $n = 1, 2$, and is strictly greater otherwise).

This implies that as long as the input traverses all possible values, even the simplest binary input system is sufficient to provide adequate excitation. For instance, by combining the first half of the regressions, one can extract a matrix structure with rank n as follows:

$$r_1 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 1 \end{bmatrix} + r_2 \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \end{bmatrix}.$$

By symmetrically swapping the positions of r_1 and r_2 in a similar manner, a second full-rank submatrix with rank n can be extracted from the latter half of the regressions, fully meeting the theoretical requirements of Assumption 2.

In summary, under non-periodic quantized input conditions, as long as the system satisfies the basic persistent excitation requirements, the workflow and logic of its attack detection algorithm can still employ the decision criterion (9), with the residual vector Υ_N instead computed via Eqs (14)–(18).

4. Properties and performance of the detection algorithm

An ideal detection algorithm should possess perfect detection capability (i.e., an attack detection rate of 1). However, constrained by the highly nonlinear nature of the binary quantization mechanism and the system's stochastic process noise, requiring the detection residual vector Υ_N to be strictly equal to $\mathbf{0}$ to declare an attack-free state is overly rigorous and impractical. Therefore, appropriate statistical metrics must be introduced to quantitatively evaluate the performance of the detection algorithm. This section will deeply analyze the asymptotic normality of the detection function and subsequently derive approximate calculation expressions for the false alarm rate (FAR) and missed detection rate (MDR) of the algorithm.

4.1. FAR and MDR

Before evaluating the performance, we first introduce the following two lemmas.

Lemma 1 ([26]). Let $\{X_n\}$, $\{Y_n\}$, and $\{Z_n\}$ be sequences of random variables defined on a probability space (Φ, Δ, Ξ) , satisfying $X_n \xrightarrow{d} X_F$, $Y_n \xrightarrow{p} a$, and $Z_n \xrightarrow{p} b$, where a and b are finite constants, and $\xrightarrow{d}$ and $\xrightarrow{p}$ denote convergence in distribution and convergence in probability, respectively. Then it holds that

$$X_n Y_n + Z_n \xrightarrow{d} a X_F + b.$$

Lemma 2 ([27]). If an n -dimensional random vector $X = [X_1, X_2, \dots, X_n]^T$ follows a multivariate normal distribution $N(\mu_X, \Sigma_X)$, then for a hyper-rectangular integration region $A = [a, b]$ (where $-\infty \leq a_i < b_i \leq \infty, i = 1, 2, \dots, n$), its multiple integral probability can be expressed as

$$\begin{aligned} \Pr(X \in A) &= \mathcal{F}_n(a, b, \mu_X, \Sigma_X) \\ &= \Pr(a_1 \leq X_1 \leq b_1, \dots, a_n \leq X_n \leq b_n) \\ &= \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_X)}} \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \exp \left\{ -\frac{1}{2} (x - \mu_X)^T \Sigma_X^{-1} (x - \mu_X) \right\} dx. \end{aligned}$$

Due to inevitable estimation errors, we relax the strict zero-vector decision condition to an adjustable threshold vector. Accordingly, the formal definitions of the FAR and MDR are given below.

Definition 2. Consider a non-negative detection threshold vector $\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]^T$ ($\varepsilon_i > 0$). FAR P_F is the probability that the system is not under attack, but the algorithm falsely detects an attack, i.e., the probability that the absolute value of at least one element in the detection function exceeds its corresponding threshold; MDR P_M is the probability that the system is actually under attack, but the algorithm fails to effectively identify it, i.e., the probability that the absolute values of all elements in the detection function are strictly less than their corresponding thresholds.

To calculate P_F and P_M , we first need to establish the asymptotic normality of Υ_N . For brevity, the following details only focus on the non-periodic input scenario; the periodic input scenario can be derived similarly. Denote $\Psi_i = \Psi(C - \pi_i \theta)$, $\tilde{\Lambda}_1 = [\pi_1^T, \dots, \pi_{\frac{h}{2}}^T]^T$, $\tilde{\Lambda}_2 = [\pi_{\frac{h}{2}+1}^T, \dots, \pi_h^T]^T$. The corresponding pseudo-inverse weighting matrices are denoted as $\Phi_1 = (\tilde{\Lambda}_1^T \tilde{\mathcal{B}} \tilde{\Lambda}_1)^{-1} \tilde{\Lambda}_1^T \tilde{\mathcal{B}}$ and $\Phi_2 = (\tilde{\Lambda}_2^T \tilde{\mathcal{B}} \tilde{\Lambda}_2)^{-1} \tilde{\Lambda}_2^T \tilde{\mathcal{B}}$.

It should be emphasized that Assumption 2 concerns the excitation and rank properties of the known input sequence. Under the measurement-channel attack models defined in Section 2.2, these structural properties remain unchanged. The attacks affect the empirical frequencies of the received binary observations and consequently modify the limiting mean and covariance of the detection residual, which are explicitly characterized below.

Theorem 2. For system (1) and its binary quantized measurements (2), under the premise that Assumptions 1 and 2 hold, the detection residual Υ_N satisfies asymptotic normality. The specific distributions are as follows:

1) If the system is not subjected to any data integrity attacks, then

$$\sqrt{N}(\Upsilon_N - \mathbf{0}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_0), \text{ as } N \rightarrow \infty, \quad (19)$$

where the covariance matrix is $\Sigma_0 = \Phi_1 \text{diag}(\kappa_{i_1}^2) \Phi_1^T + \Phi_2 \text{diag}(\kappa_{i_2}^2) \Phi_2^T$; the individual variance is $\kappa_i^2 = \frac{\Psi_i - \Psi_i^2}{\psi^2(\mathcal{H}(\Psi_i))}$; and the indices range over $i_1 = 1, 2, \dots, \frac{h}{2}$, $i_2 = \frac{h}{2} + 1, \dots, h$, and $i = 1, \dots, h$.

2) If the system is subjected to a DoS attack (p_{d_0}, p_{d_1}) described by (3), then

$$\sqrt{N}(\mathbf{Y}_N - (\tilde{\theta}_{d,1} - \tilde{\theta}_{d,2})) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma(p_{d_0}, p_{d_1})), \text{ as } N \rightarrow \infty, \quad (20)$$

where the expected drift means are $\tilde{\theta}_{d,1} = \Phi_1[C - \mathcal{H}(\gamma_{d,1}), \dots, C - \mathcal{H}(\gamma_{d,\frac{h}{2}})]^T$ and $\tilde{\theta}_{d,2} = \Phi_2[C - \mathcal{H}(\gamma_{d,\frac{h}{2}+1}), \dots, C - \mathcal{H}(\gamma_{d,h})]^T$; the covariance is $\Sigma(p_{d_0}, p_{d_1}) = \Phi_1 \text{diag}(\kappa_{d,i_1}^2) \Phi_1^T + \Phi_2 \text{diag}(\kappa_{d,i_2}^2) \Phi_2^T$; $\kappa_{d,i}^2 = \frac{\gamma_{d,i} - \gamma_{d,i}^2}{\psi^2(\mathcal{H}(\gamma_{d,i}))}$; and the contaminated frequency is $\gamma_{d,i} = \frac{(1-p_{d_1})\Psi_i}{1-p_{d_0}-(p_{d_1}-p_{d_0})\Psi_i}$.

3) If the system is subjected to a TA attack (p_{t_0}, p_{t_1}) described by (4), then

$$\sqrt{N}(\mathbf{Y}_N - (\tilde{\theta}_{t,1} - \tilde{\theta}_{t,2})) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma(p_{t_0}, p_{t_1})), \text{ as } N \rightarrow \infty, \quad (21)$$

where the drift means and covariance structure are defined similarly to those for the DoS attack, with $\gamma_{t,i}$ as the replacement; the contaminated frequency is $\gamma_{t,i} = p_{t_1} + (1 - p_{t_0} - p_{t_1})\Psi_i$, and the variance is $\kappa_{t,i}^2 = \frac{\gamma_{t,i} - \gamma_{t,i}^2}{\psi^2(\mathcal{H}(\gamma_{t,i}))}$.

4) If the system is subjected to an RE attack $(m, \mathbf{P})$ described by (5), then

$$\sqrt{N}(\mathbf{Y}_N - (\tilde{\theta}_{\tau,1} - \tilde{\theta}_{\tau,2})) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma(m, \mathbf{P})), \text{ as } N \rightarrow \infty, \quad (22)$$

where the drift means and covariance structure are defined as above; $\gamma_{\tau,i} = \sum_{\tau=0}^m p_{\tau} \sum_{j=1}^h U_{\tau,i,j} \Psi_j$, the transition probability is $U_{\tau,i,j} = \Pr\{\omega_{k-\tau} = \pi_j \mid \omega_k = \pi_i\}$, and the variance is $\kappa_{\tau,i}^2 = \frac{\gamma_{\tau,i} - \gamma_{\tau,i}^2}{\psi^2(\mathcal{H}(\gamma_{\tau,i}))}$.

Proof. The core of the proof lies in establishing the convergence of the empirical frequencies under attack environments, and subsequently solving for the asymptotic variance of the nonlinear mapping using the delta method.

Step 1: Convergence analysis of empirical frequencies. Considering packet dropouts under a DoS attack, (18) should be modified to

$$\varpi_{N,i} = \sqrt{\frac{N_i}{N}} \left(C - \mathcal{H} \left(\frac{\sum_{k=1}^N \delta_k s_k I_{\{\omega_k = \pi_i\}}}{\sum_{k=1}^N \delta_k I_{\{\omega_k = \pi_i\}}} \right) \right).$$

By the law of total probability, the packet loss rate at time k is $\mathbb{E}[1 - \delta_k] = \Pr\{\delta_k = 0\} = p_{d_0} + (p_{d_1} - p_{d_0})\Psi_k$. The joint probability is $\Pr\{\delta_k = 1, s_k = 1\} = (1 - p_{d_1})\Psi_k$. By the law of large numbers, the conditional frequency converges in probability:

$$\frac{\sum_{k=1}^N \delta_k s_k I_{\{\omega_k = \pi_i\}}}{\sum_{k=1}^N \delta_k I_{\{\omega_k = \pi_i\}}} \xrightarrow{p} \gamma_{d,i}, \text{ as } N \rightarrow \infty. \quad (23)$$

Under a TA attack, according to the bit-flipping strategy, the expectation of receiving a 1 is

$$\begin{aligned} \mathbb{E}[s_k] &= \Pr\{s_k = 1\} \\ &= \Pr\{s_k = 1 \mid s_k^0 = 0\} \Pr\{s_k^0 = 0\} + \Pr\{s_k = 1 \mid s_k^0 = 1\} \Pr\{s_k^0 = 1\} \\ &= p_{t_1}(1 - \Psi_k) + (1 - p_{t_0})\Psi_k \\ &= p_{t_1} + (1 - p_{t_0} - p_{t_1})\Psi_k. \end{aligned}$$

By the law of large numbers, we have

$$\frac{1}{N_i} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_i\}} \xrightarrow{p} \gamma_{t,i}, \text{ as } N \rightarrow \infty. \quad (24)$$

Similarly, under an RE attack, applying the law of total probability yields

$$\begin{aligned} \mathbb{E}[s_k] &= \Pr\{s_k = 1\} \\ &= \sum_{\tau=0}^m \Pr\{s_k = s_{k-\tau}^0\} \Pr\{s_{k-\tau}^0 = 1\} \\ &= \sum_{\tau=0}^m p_\tau \sum_{j=1}^h \Pr\{\omega_{k-\tau} = \pi_j \mid \omega_k = \pi_k\} \Pr\{s_k I_{\{\omega_k=\pi_j\}} = 1\} \\ &= \sum_{\tau=0}^m p_\tau \sum_{j=1}^h U_{\tau,k,j} \Psi_j. \end{aligned}$$

Likewise, the empirical frequency converges in probability to $\gamma_{\tau,i}$.

Step 2: Derivation of asymptotic normality (taking the tampering attack as an example). Since s_k follows a 0-1 distribution, its variance satisfies $\text{Var}(s_k) = \mathbb{E}[(s_k - \mathbb{E}[s_k])^2] = \mathbb{E}[s_k] - \mathbb{E}^2[s_k]$. Based on (24) and the central limit theorem, we obtain

$$\sqrt{N_i} \left(\frac{1}{N_i} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_i\}} - \gamma_{t,i} \right) \xrightarrow{d} \mathcal{N}(0, \gamma_{t,i} - \gamma_{t,i}^2), \text{ as } N \rightarrow \infty. \quad (25)$$

Moreover, $\tilde{\theta}_1 = \Phi_1 [C - \mathcal{H}(\gamma_{t,1}), \dots, C - \mathcal{H}(\gamma_{t,\frac{h}{2}})]^T = \Phi_{N,1} \text{diag} \left(\sqrt{\frac{N_{i_1}}{N}} \right) [C - \mathcal{H}(\gamma_{t,1}), \dots, C - \mathcal{H}(\gamma_{t,\frac{h}{2}})]^T$.

From (15), we have

$$\begin{aligned} &\sqrt{N}(\hat{\theta}_N^{(1)} - \tilde{\theta}_1) \\ &= \sqrt{N} \Phi_{N,1} \text{diag} \left(\sqrt{\frac{N_{i_1}}{N}} \right) \begin{bmatrix} C - \mathcal{H} \left(\frac{1}{N_1} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_1\}} \right) - (C - \mathcal{H}(\gamma_{t,1})) \\ \vdots \\ C - \mathcal{H} \left(\frac{1}{N_{\frac{h}{2}}} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_{\frac{h}{2}}\}} \right) - (C - \mathcal{H}(\gamma_{t,\frac{h}{2}})) \end{bmatrix} \\ &= \Phi_{N,1} \begin{bmatrix} \sqrt{N_1} \left(\mathcal{H}(\gamma_{t,1}) - \mathcal{H} \left(\frac{1}{N_1} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_1\}} \right) \right) \\ \vdots \\ \sqrt{N_{\frac{h}{2}}} \left(\mathcal{H}(\gamma_{t,\frac{h}{2}}) - \mathcal{H} \left(\frac{1}{N_{\frac{h}{2}}} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_{\frac{h}{2}}\}} \right) \right) \end{bmatrix}. \end{aligned} \quad (26)$$

By the mean value theorem, we obtain

$$\begin{aligned}
 & \sqrt{N_{i_1}} \left(\mathcal{H}(\gamma_{t,i_1}) - \mathcal{H} \left(\frac{1}{N_{i_1}} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_{i_1}\}} \right) \right) \\
 &= \sqrt{N_{i_1}} \frac{d\mathcal{H}(x)}{dx} \Big|_{x=\varsigma_{i_1}} \left(\gamma_{t,i_1} - \frac{1}{N_{i_1}} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_{i_1}\}} \right) \\
 &= \sqrt{N_{i_1}} \frac{1}{\psi(\mathcal{H}(\varsigma_{i_1}))} \left(\gamma_{t,i_1} - \frac{1}{N_{i_1}} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_{i_1}\}} \right),
 \end{aligned} \tag{27}$$

where ς_{i_1} lies between γ_{t,i_1} and $\frac{1}{N_{i_1}} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_{i_1}\}}$. Combining (25) and (27), as $N \rightarrow \infty$, we obtain

$$\sqrt{N_{i_1}} \frac{1}{\psi(\mathcal{H}(\varsigma_{i_1}))} \left(\gamma_{t,i_1} - \frac{1}{N_{i_1}} \sum_{k=1}^N s_k I_{\{\omega_k=\pi_{i_1}\}} \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{\gamma_{t,i_1} - \gamma_{t,i_1}^2}{\psi^2(\mathcal{H}(\gamma_{t,i_1}))} \right). \tag{28}$$

From (26) and (28), we get

$$\sqrt{N}(\widehat{\theta}_N^{(1)} - \widetilde{\theta}_1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Phi_1 \text{diag}(\kappa_{t,i_1}^2) \Phi_1^T), \text{ as } N \rightarrow \infty. \tag{29}$$

Similarly, for the second subset, we derive

$$\sqrt{N}(\widehat{\theta}_N^{(2)} - \widetilde{\theta}_2) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Phi_2 \text{diag}(\kappa_{t,i_2}^2) \Phi_2^T), \text{ as } N \rightarrow \infty. \tag{30}$$

By Lemma 1, along with (29) and (30), since the two sample sets are independent, we obtain

$$\begin{aligned}
 & \sqrt{N}(\Upsilon_N - (\widetilde{\theta}_1 - \widetilde{\theta}_2)) \\
 &= \sqrt{N}(\widehat{\theta}_N^{(1)} - \widetilde{\theta}_1) - \sqrt{N}(\widehat{\theta}_N^{(2)} - \widetilde{\theta}_2) \\
 &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma(p_{t_0}, p_{t_1})), \text{ as } N \rightarrow \infty,
 \end{aligned} \tag{31}$$

which proves (21).

The asymptotic normality of the detection algorithm under DoS attacks and RE attacks can be proven through a process similar to that in Eqs (25)–(31).

When the system is not subjected to any data integrity attacks, we simply let the attack strategies degenerate to $(p_{d_0}, p_{d_1}) = (0, 0)$, $(p_{t_0}, p_{t_1}) = (0, 0)$, and $(m, \mathbf{P}) = (0, \mathbf{0})$, respectively, which directly proves the attack-free scenario in (19). This completes the proof. $\square$

Remark 3. Unlike DoS and TA attacks, the detectability of an RE attack depends on whether the replayed measurement remains statistically compatible with the current regressor. From Theorem 2, define $\gamma_{\tau,i} = \sum_{\ell=0}^m p_{\ell} \sum_{j=1}^h U_{\ell,i,j} \Psi_j$. If $\gamma_{\tau,i} = \Psi_i$, $i = 1, \dots, h$, then the regressor-conditioned probability distribution is unchanged by the replay attack. Consequently, both local estimates remain consistent: $\widehat{\theta}_N^{(1)} \rightarrow \theta$, $\widehat{\theta}_N^{(2)} \rightarrow \theta$. Therefore, $\Upsilon_N \rightarrow 0$. A representative case occurs for a periodic input with period T_u when the replay delay satisfies $\tau = qT_u$, $q \in \mathbb{N}^+$, because then $\omega_{k-\tau} = \omega_k$ and hence the stale measurement has the same regressor-conditioned distribution as the current one. Thus, without additional freshness or timing information, such a perfectly regressor-compatible replay attack is

intrinsically indistinguishable by the proposed identification-based detector. Nevertheless, it also preserves the consistency of the parameter estimates and hence does not degrade the parameter-identification objective considered in this paper. This statement does not imply that the replay is harmless for other freshness-sensitive estimation or control tasks.

According to Definition 2, given an n -dimensional threshold vector $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]^T$ ($\varepsilon_i > 0, i = 1, \dots, n$), the calculation methods for the FAR (P_F) and MDR (P_M) are as follows:

$$P_F = 1 - \Pr\{|\Upsilon_N| \leq \boldsymbol{\varepsilon} \mid \text{Attack-free}\};$$

$$P_M = \Pr\{|\Upsilon_N| \leq \boldsymbol{\varepsilon} \mid \text{Under Attack}\}.$$

Here, and in the following, the inequality between two vectors represents an element-wise comparison. Based on Lemma 2, P_F can be expressed by (32) due to the validity of (19); P_M can be uniformly expressed by (33), originating from the validity of (20)–(22).

$$P_F \approx 1 - \mathcal{F}_n\left(-\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}, \mathbf{0}, \frac{\Sigma_0}{N}\right), \quad (32)$$

$$P_M \approx \mathcal{F}_n\left(-\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}, \mu_A, \frac{\Sigma_A}{N}\right), \quad (33)$$

where the expected drift is $\mu_A \in \{\tilde{\theta}_{d,1} - \tilde{\theta}_{d,2}, \tilde{\theta}_{t,1} - \tilde{\theta}_{t,2}, \tilde{\theta}_{\tau,1} - \tilde{\theta}_{\tau,2}\}$, and the covariance matrix is $\Sigma_A \in \{\Sigma(p_{d_0}, p_{d_1}), \Sigma(p_{t_0}, p_{t_1}), \Sigma(m, \mathbf{P})\}$, depending on the specific type of attack experienced by the system.

4.2. Properties of the detection algorithm

The magnitudes of the FAR and MDR determine the practical effectiveness of the detection algorithm. Based on their calculation formulas, it is evident that their values are governed by multiple factors. This section focuses on discussing the impacts of factors such as data length, the threshold vector, and attack strategies on the FAR and MDR of the detection algorithm.

1) Impact of data length on detection performance

Assume Π is a hyper-rectangular region in an n -dimensional space defined by the threshold vector $\boldsymbol{\varepsilon}$, i.e.,

$$\Pi = \{x = [x_1, \dots, x_n]^T \in \mathbb{R}^n : -\varepsilon_i \leq x_i \leq \varepsilon_i, i = 1, \dots, n\} = \Pi[-\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}].$$

For data lengths N and $N + 1$, according to (19) and (32), the FAR of the detection algorithm can be expressed as

$$P_F(N, \Pi) = 1 - \Pr\left\{\left|\sqrt{N}\Upsilon_N\right| \leq \boldsymbol{\varepsilon} \mid \mathcal{N}(\mathbf{0}, \Sigma_0)\right\},$$

$$P_F(N + 1, \Pi) = 1 - \Pr\left\{\left|\sqrt{N + 1}\Upsilon_N\right| \leq \boldsymbol{\varepsilon} \mid \mathcal{N}(\mathbf{0}, \Sigma_0)\right\}.$$

Notice that

$$(-\sqrt{N + 1} - (-\sqrt{N}))\boldsymbol{\varepsilon} < \mathbf{0},$$

$$(\sqrt{N + 1} - \sqrt{N})\boldsymbol{\varepsilon} > \mathbf{0}.$$

Thus, we have the set inclusion relationship:

$$\Pi[-\sqrt{N}\boldsymbol{\varepsilon}, \sqrt{N}\boldsymbol{\varepsilon}] \subseteq \Pi[-\sqrt{N+1}\boldsymbol{\varepsilon}, \sqrt{N+1}\boldsymbol{\varepsilon}].$$

Since $\sqrt{N+1}\boldsymbol{\Upsilon}_N$ and $\sqrt{N}\boldsymbol{\Upsilon}_N$ approximately follow the same normal distribution when N is sufficiently large, it follows that

$$P_F(N+1, \Pi) < P_F(N, \Pi).$$

This indicates that as the data length N increases, the FAR of the algorithm gradually decreases.

From (20)–(22) and (33), the MDR of the detection algorithm can be expressed as

$$P_M(N) = \Pr\left\{\sqrt{N}(-\boldsymbol{\varepsilon} - \boldsymbol{\mu}_A) \leq \sqrt{N}(\boldsymbol{\Upsilon}_N - \boldsymbol{\mu}_A) \leq \sqrt{N}(\boldsymbol{\varepsilon} - \boldsymbol{\mu}_A) \mid \mathcal{N}(\mathbf{0}, \Sigma_A)\right\}. \quad (34)$$

Denote $\boldsymbol{\alpha} = -\boldsymbol{\varepsilon} - \boldsymbol{\mu}_A = [\alpha_1, \dots, \alpha_n]^T$, $\boldsymbol{b} = \boldsymbol{\varepsilon} - \boldsymbol{\mu}_A = [b_1, \dots, b_n]^T$, and $\zeta = \sqrt{N}$. When N is sufficiently large, according to (34) and Lemma 2, the MDR becomes

$$P_M(N) = \int_{\zeta\alpha_1}^{\zeta b_1} \cdots \int_{\zeta\alpha_n}^{\zeta b_n} \mathfrak{N}(X; \Sigma_A) dX = G(\zeta), \quad (35)$$

where $\mathfrak{N}(X, \Sigma_A) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_A)}} \exp\left\{-\frac{1}{2}X^T \Sigma_A^{-1} X\right\}$.

Generally, the elements chosen for the threshold vector are relatively small, and we consider the practical scenario where $P_M < 1$. In this case, $\boldsymbol{\alpha}$ and $\boldsymbol{b}$ share the same sign. Below, we only discuss the case where $\alpha > 0$ and $b > 0$; other cases can be analyzed similarly.

Define the function $g_i(\iota) = \iota \mathfrak{N}(z_i(\iota); \Sigma_A)$, where

$$z_i(\iota) = [x_1, \dots, x_{i-1}, \zeta\iota, x_{i+1}, \dots, x_n]^T \quad (\alpha_i \leq \iota \leq b_i),$$

and let $\delta(\iota) = z_i(\iota)^T \Sigma_A^{-1} z_i(\iota)$. The elements of Σ_A^{-1} are denoted as σ_{ij} ($i, j = 1, \dots, n$). The derivative of $g_i(\iota)$ with respect to ι is

$$\frac{dg_i(\iota)}{d\iota} = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_A)}} \exp\left\{-\frac{1}{2}\delta(\iota)\right\} \left(1 - \frac{\iota}{2} \frac{d\delta_i(\iota)}{d\iota}\right), \quad (36)$$

$$\begin{aligned} \frac{d\delta_i(\iota)}{d\iota} &= [0, \dots, 0, \zeta, 0, \dots, 0] \Sigma_A^{-1} z_i(\iota) + z_i^T(\iota) \Sigma_A^{-1} [0, \dots, 0, \zeta, 0, \dots, 0]^T \\ &= 2\sigma_{ii}\zeta^2 + 2\varrho_i\zeta, \end{aligned} \quad (37)$$

where $\varrho_i = \sum_{j=1, j \neq i}^n \sigma_{ij}x_j$. Substituting (37) into (36) yields

$$\frac{dg_i(\iota)}{d\iota} = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_A)}} \exp\left\{-\frac{1}{2}\delta(\iota)\right\} (1 - \sigma_{ii}\zeta^2\iota^2 - \varrho_i\zeta\iota).$$

Since Σ_A^{-1} is the inverse of a covariance matrix, it is positive definite, meaning $\sigma_{ii} > 0$. Therefore, when $\zeta > \max_i \frac{\varrho_i - \sqrt{\varrho_i^2 + 4\sigma_{ii}}}{-2\sigma_{ii}\iota}$, we have $\frac{dg_i(\iota)}{d\iota} < 0$.

The derivative of $G(\zeta)$ with respect to ζ is $\frac{dG(\zeta)}{d\zeta} = \sum_{i=1}^n \vartheta_i(\zeta)$, where

$$\vartheta_i(\zeta) = \begin{cases} \int_{\alpha_2}^{b_2} \cdots \int_{\alpha_n}^{b_n} (g_1(b_1) - g_1(\alpha_1)) dx_2 \cdots dx_n, & i = 1; \\ \int_{\alpha_1}^{b_1} \cdots \int_{\alpha_{i-1}}^{b_{i-1}} \int_{\alpha_{i+1}}^{b_{i+1}} \cdots \int_{\alpha_n}^{b_n} (g_i(b_i) - g_i(\alpha_i)) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n, & \text{others}; \\ \int_{\alpha_1}^{b_1} \cdots \int_{\alpha_{n-1}}^{b_{n-1}} (g_n(b_n) - g_n(\alpha_n)) dx_1 \cdots dx_{n-1}, & i = n. \end{cases}$$

Since $b_i > \alpha_i$, according to the monotonicity of $g_i(\iota)$, we have $\vartheta_i(\zeta) < 0$. Based on (35) and $\zeta = \sqrt{N}$, the following conclusion can be drawn:

When $N > \max_{1 \leq i \leq n} \max_{\iota} \left(\frac{\varrho_i - \sqrt{\varrho_i^2 + 4\sigma_{ii}}}{-2\sigma_{ii}} \right)^2$, it holds that $\frac{dG(\zeta)}{d\zeta} < 0$. This implies that when the data length N is sufficiently large, the MDR $P_M(N)$ of the detection algorithm decreases monotonically with respect to the data length N .

2) Impact of the threshold vector on detection performance

Consider two threshold vectors $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$, such that $\boldsymbol{\varepsilon}_1 < \boldsymbol{\varepsilon}_2$. Then we have

$$\Pi[-\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_1] \subseteq \Pi[-\boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_2]. \quad (38)$$

First, we discuss the impact of the threshold vector on the FAR. From (19) and (32), the FARs are given by

$$\begin{aligned} P_F(\boldsymbol{\varepsilon}_1) &= 1 - \Pr\{-\sqrt{N}\boldsymbol{\varepsilon}_1 \leq \sqrt{N}\mathbf{Y}_N \leq \sqrt{N}\boldsymbol{\varepsilon}_1 \mid \mathcal{N}(\mathbf{0}, \Sigma_0)\}, \\ P_F(\boldsymbol{\varepsilon}_2) &= 1 - \Pr\{-\sqrt{N}\boldsymbol{\varepsilon}_2 \leq \sqrt{N}\mathbf{Y}_N \leq \sqrt{N}\boldsymbol{\varepsilon}_2 \mid \mathcal{N}(\mathbf{0}, \Sigma_0)\}. \end{aligned}$$

From (38), it is evident that $P_F(\boldsymbol{\varepsilon}_1) > P_F(\boldsymbol{\varepsilon}_2)$. This indicates that the larger the chosen threshold, the smaller the FAR of the detection algorithm.

Next, we analyze the impact of the threshold vector on the MDR. From (20)–(22) and (33), the MDRs are

$$\begin{aligned} P_M(\boldsymbol{\varepsilon}_1) &= \Pr\{\sqrt{N}(-\boldsymbol{\varepsilon}_1 - \boldsymbol{\mu}_A) \leq \sqrt{N}(\mathbf{Y}_N - \boldsymbol{\mu}_A) \leq \sqrt{N}(\boldsymbol{\varepsilon}_1 - \boldsymbol{\mu}_A) \mid \mathcal{N}(\mathbf{0}, \Sigma_A)\}, \\ P_M(\boldsymbol{\varepsilon}_2) &= \Pr\{\sqrt{N}(-\boldsymbol{\varepsilon}_2 - \boldsymbol{\mu}_A) \leq \sqrt{N}(\mathbf{Y}_N - \boldsymbol{\mu}_A) \leq \sqrt{N}(\boldsymbol{\varepsilon}_2 - \boldsymbol{\mu}_A) \mid \mathcal{N}(\mathbf{0}, \Sigma_A)\}. \end{aligned}$$

From (38), we know that

$$\Pi[-\boldsymbol{\varepsilon}_1 - \boldsymbol{\mu}_A, \boldsymbol{\varepsilon}_1 - \boldsymbol{\mu}_A] \subseteq \Pi[-\boldsymbol{\varepsilon}_2 - \boldsymbol{\mu}_A, \boldsymbol{\varepsilon}_2 - \boldsymbol{\mu}_A].$$

Therefore, $P_M(\boldsymbol{\varepsilon}_1) < P_M(\boldsymbol{\varepsilon}_2)$. This demonstrates that the MDR of the detection algorithm increases as the threshold increases.

3) Impact of attack strategies on detection performance

The MDR (P_M) is a core metric for measuring the algorithm's capability to identify attacks. From the perspective of game theory, the attacker's objective is to seek specific attack strategies that maximize the MDR to conceal their malicious activities. Conversely, the defender expects the MDR to be as small as possible to promptly capture system anomalies and take effective response measures. Therefore, we delve into the evolution laws of the detection algorithm's performance (i.e., the asymptotic MDR) under different attack strategies.

Introduce an infinitesimally small positive number λ_1 ($0 < \lambda_1 \ll 1$) and a sufficiently large positive number λ_2 ($\lambda_2 \gg 1$). We uniformly abstract the equivalent distortions produced by the three types of attacks on the system's probability distribution into a nonlinear mapping relationship:

$$\begin{aligned} \mathcal{S}_A &\stackrel{\Delta}{\rightarrow} \Delta_A(\Psi), \quad A \in \{d, t, \tau\}, \\ \text{s.t.} \quad &p_{d_0}, p_{d_1}, p_{t_0}, p_{t_1}, p_{0}, \dots, p_m \in (0, 1), \\ &m \in \mathbb{N}^+, \\ &\sum_{\tau=0}^m p_{\tau} = 1, \end{aligned}$$

where $\mathcal{S}_A$ is the feasible region of strategy parameters for the corresponding attack type, $\Psi = [\Psi_1, \dots, \Psi_h]^T$ is the true prior probability vector under attack-free conditions, and $\Delta_A(\Psi)$ represents the measure drift induced by the attack behavior.

According to Theorem 2, the expected drift of the detection residual can be expressed as $\mu_A = [\Phi_1, -\Phi_2]\Delta_A(\Psi)$. If the attacker carefully designs a strategy that satisfies the condition

$$[\Phi_1, -\Phi_2]\Delta_A(\Psi) \in \lambda_1[-\epsilon, \epsilon], \quad A \in \{d, t, \tau\}, \quad (39)$$

meaning the expected drift is strictly confined within a microscopic hyper-rectangle far smaller than the threshold ϵ (indicating that the system anomaly caused by the attack is highly camouflaged or suppressed), as the data length $N \rightarrow \infty$, the detection residual almost completely falls within the safe threshold region with probability 1, causing the detection algorithm to fail, i.e., the asymptotic MDR $P_M \rightarrow 1$.

Conversely, if the attacker's strategy is overly aggressive, causing the expected drift to satisfy the condition

$$[\Phi_1, -\Phi_2]\Delta_A(\Psi) \notin \lambda_2(-\epsilon, \epsilon), \quad A \in \{d, t, \tau\}, \quad (40)$$

meaning the expected drift not only crosses the threshold boundary but also escapes beyond a massive hyper-rectangle scaled by λ_2 (indicating that the attack leaves highly prominent statistical footprints in the system), as the data length N increases, the detection residual will inevitably exceed the threshold bounds by a large margin, triggering an alarm and causing the asymptotic MDR $P_M \rightarrow 0$.

Remark 4. It is important to distinguish the high-concealment condition (39) from a perfectly regressor-compatible replay attack. Condition (39) only requires the projected residual drift $[\Phi_1, -\Phi_2]\Delta_A(\Psi)$ to be sufficiently small and does not necessarily imply that the underlying attacked probability measure is unchanged. In contrast, a perfectly compatible replay attack satisfies $\Delta_\tau(\Psi) = 0$, or equivalently $\gamma_{\tau,i} = \Psi_i$ for all relevant regressors. In this special case, not only does the detection residual vanish asymptotically, but the parameter estimates also retain their attack-free limits.

Remark 5. The proposed detector does not imply that every attack necessarily generates different asymptotic biases in the two local estimators. Its detectability is determined by the differential bias $\mu_A = \tilde{\theta}_{A,1} - \tilde{\theta}_{A,2} = [\Phi_1, -\Phi_2]\Delta_A(\Psi)$. For a fixed threshold vector ϵ , an attack is asymptotically distinguishable if $\mu_A \notin [-\epsilon, \epsilon]$, whereas an attack satisfying $\mu_A \in [-\epsilon, \epsilon]$ may remain undetected even as $N \rightarrow \infty$. In particular, condition (39) characterizes a highly concealed case in which the residual mean lies deeply inside the threshold region and hence $P_M \rightarrow 1$. This limitation also suggests two possible defensive directions. First, since Φ_1 and Φ_2 depend on the excitation pattern of the input regressors, changing the input excitation can reshape the mapping $[\Phi_1, -\Phi_2]\Delta_A(\Psi)$ and thereby potentially reduce the feasible set of stealthy attack strategies. Second, decreasing the threshold vector ϵ narrows the stealthy region, although an excessively small threshold may increase the FAR for finite data lengths. Therefore, input excitation and threshold selection provide potential means of improving detectability, but they do not eliminate the fundamental indistinguishability of attacks whose differential effect on the two local estimators vanishes.

The above analysis assumes that the input sequence used to construct the regressors is accurately known. If the available input is contaminated as $\tilde{u}_k = u_k + v_k$, the corresponding regressor $\tilde{\omega}_k$ contains measurement errors, leading to a typical errors-in-variables identification problem. In this case, directly

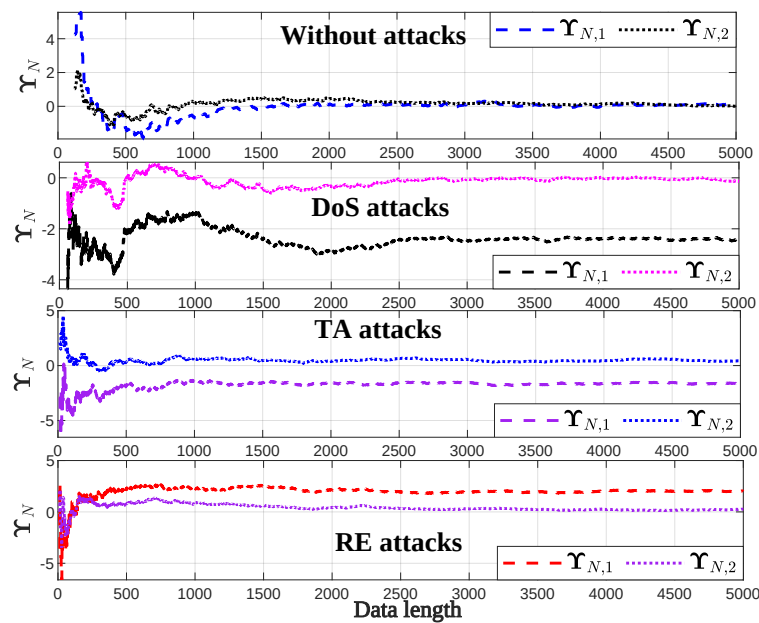


Figure 2. Convergence trajectories of the detection residual Υ_N under periodic input excitation.

applying the present estimator may result in biased local parameter estimates and altered estimation covariance, which can further cause a shift or variance inflation in the detection residual Υ_N . Therefore, the FAR and MDR expressions derived above are not guaranteed to remain valid without modification. Robust bias-compensation techniques provide a possible route to address this limitation. Recent studies have developed bias-compensated adaptive filtering methods for noisy-input environments, including robust LMS schemes under colored Gaussian input noise and impulsive observation noise [28] and bias-compensated kernel maximum-correntropy methods for nonlinear prediction with input noise and impulsive output outliers [29]. Extending the present framework to such settings would require bias-corrected identification of the two local models together with a corresponding recalibration of the residual distribution and detection thresholds.

5. Numerical simulation and performance verification

To verify the effectiveness of the proposed attack detection algorithm and the correctness of the theoretical results, this section conducts numerical simulations on a quantized FIR system subjected to three types of data integrity attacks.

Consider the following second-order discrete-time SISO FIR system:

$$\begin{cases} y_k = \omega_k \theta + d_k, \\ s_k^0 = I_{\{y_k \leq C\}}, \end{cases}$$

where the unknown parameter vector is $\theta = [3, -5]^T$; the quantization threshold of the binary sensor is $C = 0$; and the process noise is $d_k \sim \mathcal{N}(0, 10^2)$, satisfying Assumption 1. The authentic quantized observation data s_k^0 may suffer from DoS attacks, TA attacks, or RE attacks during transmission to the estimation center over the network.

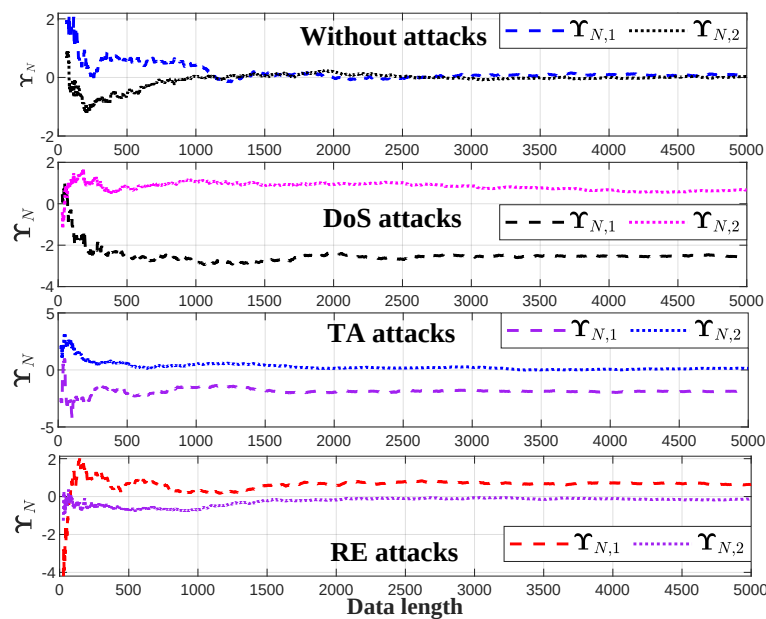


Figure 3. Convergence trajectories of the detection residual Υ_N under quantized input excitation.

Since the number of unknown system parameters is $n = 2$, a periodic input is first utilized as the benchmark testing scenario. To satisfy the persistent excitation condition, the system input is set as a sequence with a minimum positive period of 4, taking values $[2, -1, 0, -3]$ within a single period. The total observation data length is set to $N = 5000$. Simulations are conducted for the detection residual algorithm (9) under the following four scenarios: attack-free state, DoS attack strategy $(p_{d_0}, p_{d_1}) = (0.3, 0.25)$, TA attack strategy $(p_{t_0}, p_{t_1}) = (0.302, 0.335)$, and RE attack strategy $(m, \mathbf{P}) = (2, [0.6, 0.2, 0.2]^T)$. As shown in Figure 2, with the increase of data length N , the detection residual vector Υ_N gradually converges to $\mathbf{0}$ in the attack-free scenario, whereas under various data integrity attacks, the residual elements deviate significantly from zero.

Furthermore, to verify the generalizability of the algorithm under non-periodic quantized inputs, the system input is set as a binary quantized signal $u_k \in \{-1, 2\}$. Consequently, the regressor ω_k possesses $H = 4$ possible patterns: $\pi_1 = [-1, -1]$, $\pi_2 = [-1, 2]$, $\pi_3 = [2, -1]$, and $\pi_4 = [2, 2]$. Two full-rank matrices, $[\pi_1^T, \pi_2^T]^T$ and $[\pi_3^T, \pi_4^T]^T$, can be obtained, which perfectly satisfy Assumption 2. With the data length and attack strategies unchanged, the detection residuals are calculated using Eq (14). As illustrated in Figure 3, the residual convergence characteristics under the quantized input scenario are highly consistent with those in the periodic case, demonstrating the effectiveness of the generalized algorithm.

5.1. Calculation of false alarm rate and missed detection rate

To accurately evaluate the statistical performance of the algorithm, Monte Carlo simulations with $T = 1000$ independent repeated trials are conducted. Let $\Upsilon_N^{(i)}$ denote the final detection residual vector obtained in the i th trial. The detection threshold vector is set to $\boldsymbol{\varepsilon} = [0.43, 0.35]^T$. Under this threshold, the decision condition is evaluated as $|\Upsilon_N^{(i)}| \leq \boldsymbol{\varepsilon}$ (element-wise absolute value comparison).

In the attack-free state (i.e., the probabilities of all attack strategies are set to 0), the empirical FAR is

approximately calculated using the empirical frequency

$$P_F^{(T)} = 1 - \frac{1}{T} \sum_{i=1}^T I_{\{|\Upsilon_N^{(i)}| \leq \epsilon\}}.$$

Subsequently, the previously defined DoS, tampering, and replay attack strategies are injected respectively to calculate the empirical MDR

$$P_M^{(T)} = \frac{1}{T} \sum_{i=1}^T I_{\{|\Upsilon_N^{(i)}| \leq \epsilon\}}.$$

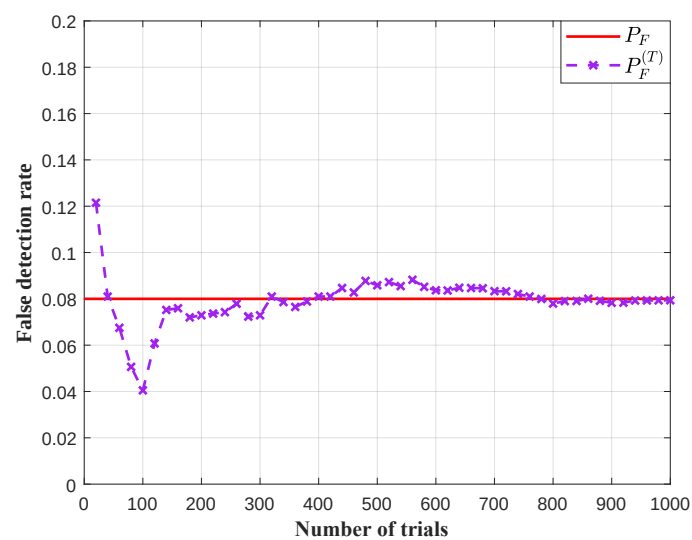


Figure 4. Convergence relationship of FAR with the number of trials T .

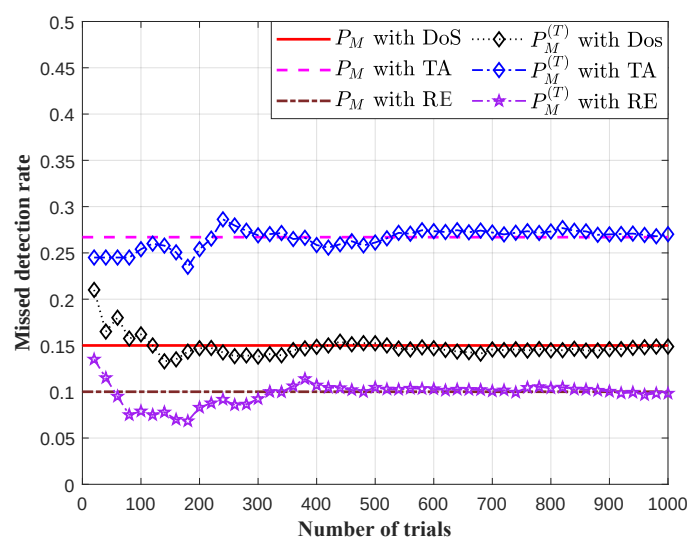


Figure 5. Convergence relationship of MDR with the number of trials T under various attacks.

The simulation results are presented in Figures 4 and 5. It can be clearly observed that as the number of independent trials T increases, the FAR and MDR of the detection algorithm converge to the theoretical values calculated by (32) and (33).

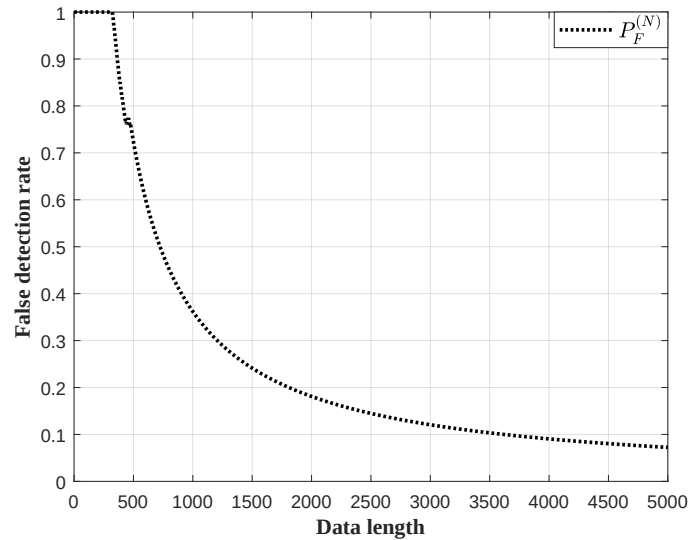


Figure 6. Evolution trend of FAR with respect to data length N .

5.2. Factors influencing detection algorithm performance

1) Impact of data length on detection performance

First, consider the attack-free scenario. Then keep the DoS, TA, and RE attack strategies, as well as the threshold vector, unchanged. A single sequential trial is conducted for the detection algorithm, yielding a sequence of residuals $\Upsilon_1, \Upsilon_2, \dots, \Upsilon_N$. The dynamic FAR $P_F^{(N)}$ and MDR $P_M^{(N)}$ are approximated by the time-averaged frequencies

$$P_F^{(N)} = 1 - \frac{1}{N} \sum_{i=1}^N I_{\{|\Upsilon_i| \leq \epsilon\}},$$

$$P_M^{(N)} = \frac{1}{N} \sum_{i=1}^N I_{\{|\Upsilon_i| \leq \epsilon\}}.$$

The experimental results are shown in Figures 6 and 7. As illustrated, with the accumulation of data length N , both the FAR and MDR of the detection algorithm decrease progressively.

2) Impact of threshold vector on detection performance

With the data volume and attack strategies constant, grid sampling is performed within the 2D space of the threshold vector $\epsilon = [\epsilon_1, \epsilon_2]^T \in [0, 1]^2$ with a step size of 0.025. Performance surfaces are plotted based on the theoretical Eqs (32) and (33). As depicted by the 3D surfaces in Figures 8–11, a smaller threshold setting (i.e., a stricter decision condition) results in a lower MDR, indicating higher system sensitivity to stealthy attacks. However, an excessively small threshold compresses the error-tolerance margin, leading to a sharp surge in FAR when the data volume is limited.

3) Impact of attack strategies on detection performance

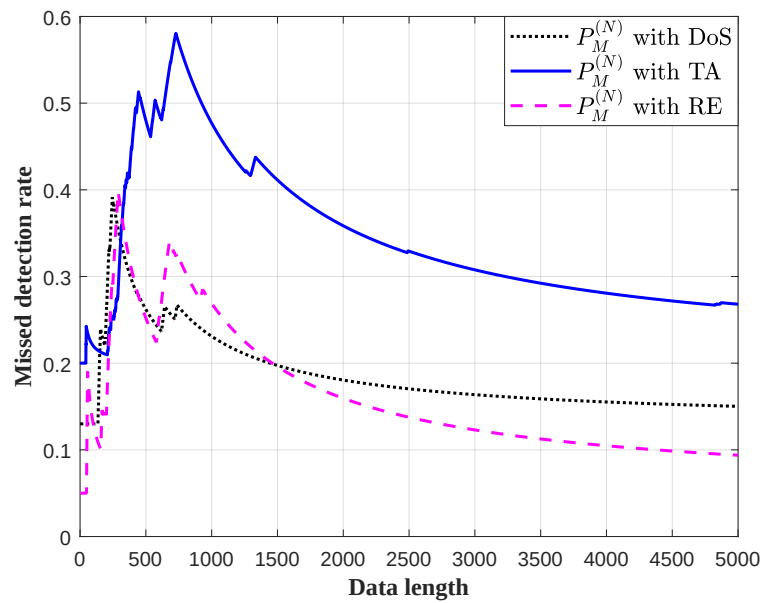


Figure 7. Evolution trend of MDR with respect to data length N .

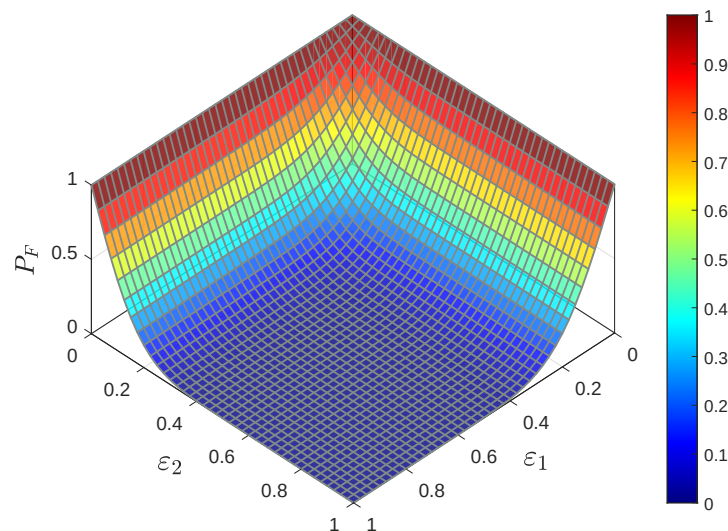


Figure 8. 3D surface relationship between FAR and threshold vector ϵ .

Set an ample data length of $N = 20,000$ and a fixed threshold vector $\epsilon = [0.43, 0.35]^T$. Two distinct attack strategies are provided for comparison. Strategy I consists of DoS (0.17, 0.18), TA (0.52, 0.53), and RE (3, $[0.82, 0.03, 0.11, 0.04]^T$), satisfying the high-concealment condition (39). Strategy II consists of DoS (0.31, 0.49), TA (0.4, 0.51), and RE (3, $[0.32, 0.29, 0.11, 0.28]^T$), satisfying the aggressive condition (40). Figures 12 and 13 illustrate the evolution trends of the MDR under these different attack strategies as the data length increases. It can be observed that when the attacker adopts Strategy I, the MDR of the detection algorithm approaches 1 as data volume N increases, rendering the detector ineffective. Conversely, when Strategy II is applied, the MDR rapidly converges to zero as N increases, successfully triggering the alarm.

To verify the detectability boundary stated in Remark 3, we consider the periodic input used

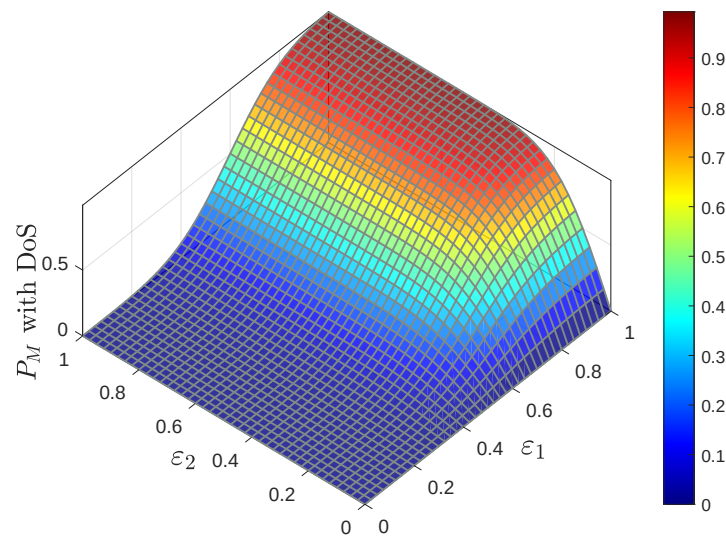


Figure 9. 3D surface relationship between MDR and threshold vector ϵ under DoS attacks.

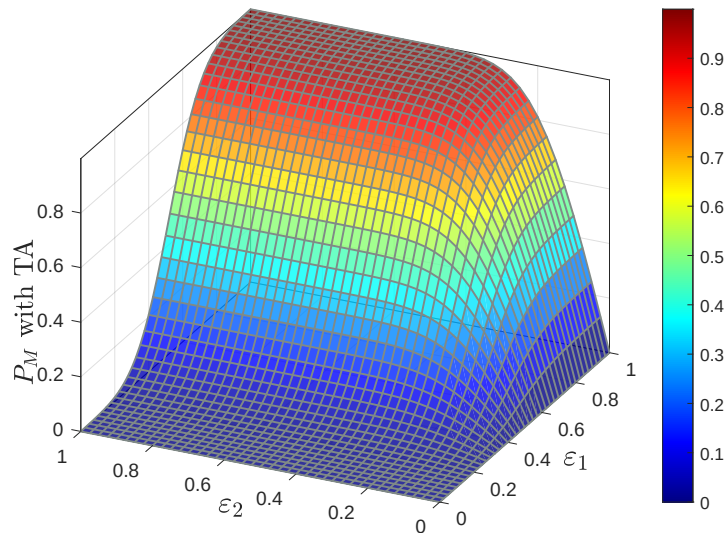


Figure 10. 3D surface relationship between MDR and threshold vector ϵ under tampering attacks.

above, whose minimum positive period is $T_u = 4$. A deterministic replay delay equal to the input period is deliberately selected: $\tau = 4$, so that $s_k = s_{k-4}^0$. Since $\omega_{k-4} = \omega_k$, the replayed observations remain statistically compatible with the current regressors and therefore satisfy $\gamma_{\tau,i} = \Psi_i$. 300 Monte Carlo simulations are conducted to avoid drawing conclusions from a particular random realization. Figure 14 compares this compatible replay with the ordinary replay strategy used previously. Under ordinary replay, the mean detection residual converges to a nonzero value. In contrast, when $\tau = T_u$, both components of the mean detection residual converge to zero. Meanwhile, the two local parameter estimates under the compatible replay converge to the true parameter vector. Therefore, although this adversarial replay cannot be reliably detected without additional freshness information, it does not destroy the consistency of the parameter-identification task considered in this paper. This experiment explicitly demonstrates the applicability boundary of the proposed replay detector.

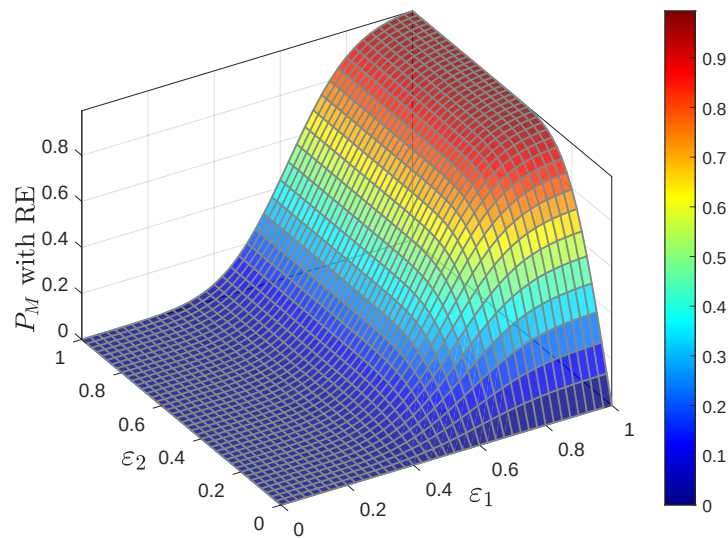


Figure 11. 3D surface relationship between MDR and threshold vector ε under replay attacks.

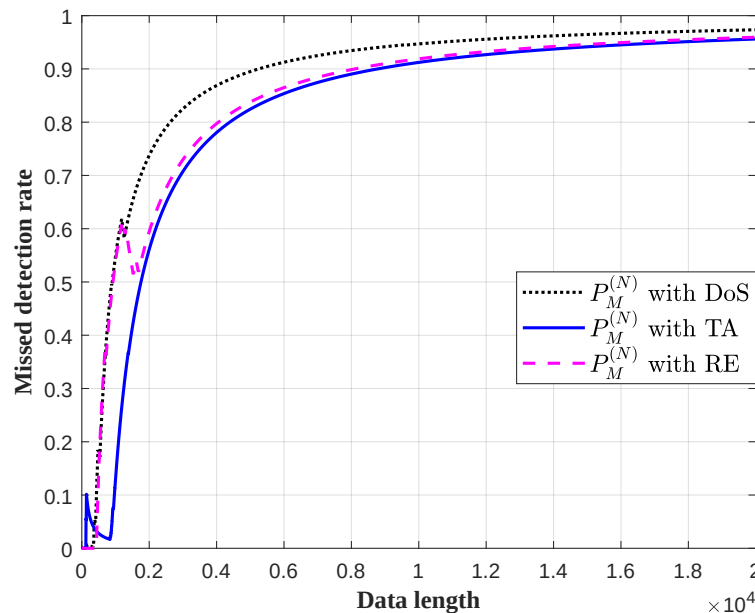


Figure 12. Evolution of MDR under attack Strategy I.

5.3. Comparative performance against alternative detection schemes

The preceding simulations have verified the convergence characteristics of the proposed detection residual and the accuracy of the theoretical FAR and MDR expressions. Nevertheless, showing that the residual changes under attacks is not, by itself, sufficient to demonstrate the practical advantage of the proposed detector. To provide a more comprehensive performance verification, this subsection further compares the proposed method with two representative alternative detection strategies under exactly the same quantized FIR system, noise environment, data lengths, and attack settings.

The purpose of the following comparison is not to claim that a unified detector must outperform every attack-specific method for every individual attack. Instead, the comparison focuses on a more

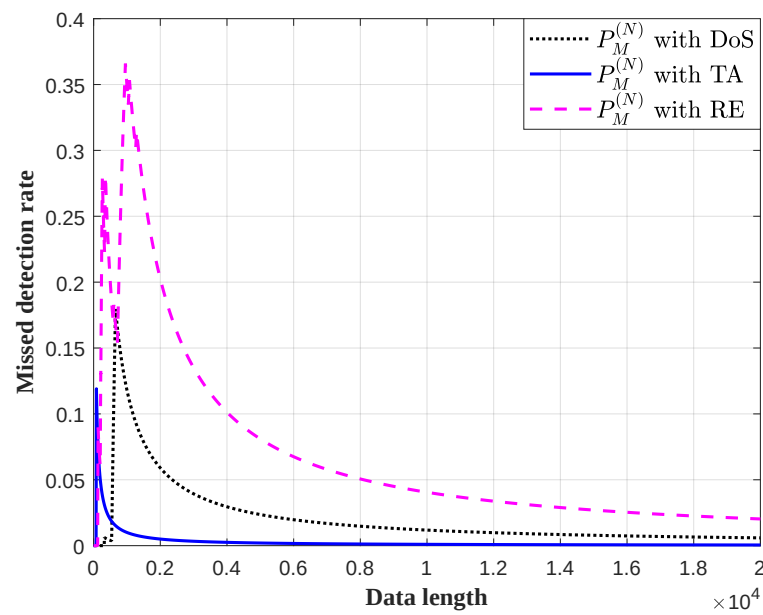
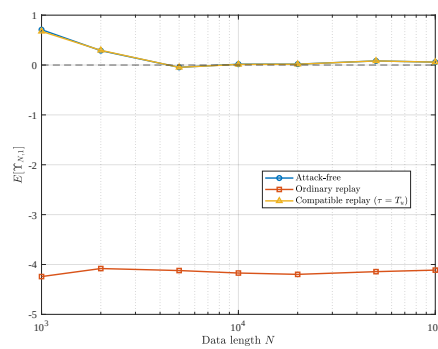
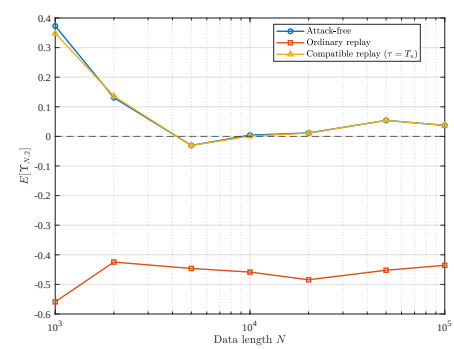


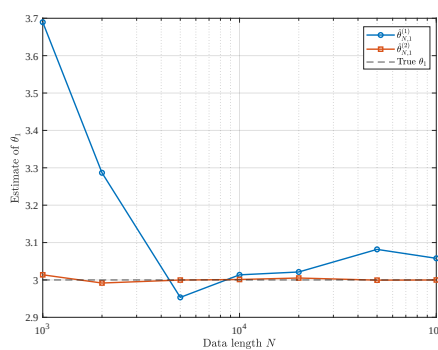
Figure 13. Evolution of MDR under attack Strategy II.



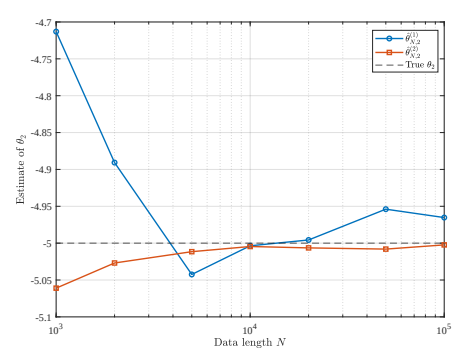
(a) First component of the mean detection residual.



(b) Second component of the mean detection residual.



(c) Convergence of the estimates of θ_1 .



(d) Convergence of the estimates of θ_2 .

Figure 14. Adversarial replay experiment with a replay delay equal to the input period.

practically relevant question: When the actual attack type is unknown in advance and an accurate nominal parameter vector may not be available, can a single detector maintain reliable detection

capability against heterogeneous data-integrity attacks?

Three detection methods are considered. 1) Proposed residual detector. The proposed method partitions the sufficiently excited regressors into two disjoint full-rank subsets and obtains two local parameter estimates. To facilitate a common scalar comparison with the two baseline detectors, the vector residual is transformed into the normalized detection score $J_P = \max_{1 \leq i \leq n} \frac{|\mathbf{Y}_{N,i}|}{\epsilon_i^{\text{ref}}}$, where ϵ^{ref} is a fixed reference vector. This scalarization is introduced only for threshold calibration and ROC evaluation and does not modify the construction mechanism of the original residual. An important characteristic of the proposed detector is that it does not compare the estimated parameter with a prescribed nominal parameter vector. Instead, the attack decision is generated from the inconsistency between two parameter estimates extracted from different sufficiently excited subsets of the received data. Consequently, neither an accurate prior value of the unknown parameter nor prior knowledge of the actual attack type is required.

2) Prior-model consistency detector (Baseline A) [30]. The first alternative method represents the conventional idea of detecting anomalies by comparing the current parameter estimate with a known nominal model. Let $\hat{\theta} * N$ denote the global parameter estimate and $\theta * \text{prior}$ denote the nominal parameter vector. A Mahalanobis-type consistency statistic is adopted: $J_A = (\hat{\theta}_N - \theta_{\text{prior}})^T \Sigma_0^{-1} (\hat{\theta}_N - \theta_{\text{prior}})$, where Σ_0 denotes the covariance matrix estimated from attack-free calibration data. For the nominal comparison reported below, the most favorable condition $\theta_{\text{prior}} = \theta$ is deliberately adopted for Baseline A. Therefore, this baseline is supplied with the exact true system parameter and represents an ideal prior-model detector. Such a setting avoids artificially weakening the baseline and allows the comparison to evaluate whether the proposed method can remain competitive even without this prior information.

3) Replay-oriented lag-consistency detector (Baseline B) [31]. The second alternative detector is designed to exploit the temporal repetition induced by replay attacks. Since replayed measurements are historical measurements reused at the current time, a replay attack may increase the similarity between the received sequence and its delayed versions. Accordingly, the detection statistic is defined as $J_B = \max_{1 \leq \ell \leq m} \frac{1}{|\mathcal{I}_\ell|} \sum_{k \in \mathcal{I}_\ell} I\{s_k = s_{k-\ell}\}$, where $\mathcal{I}_\ell$ denotes the set of valid observation pairs associated with lag ℓ . This baseline represents an attack-specific detection philosophy. It can be highly effective when the attack generates the particular temporal structure for which it is designed, but it is not explicitly constructed for DoS and tampering attacks.

The three statistics J_P , J_A , and J_B have completely different numerical scales. Hence, directly assigning the same numerical threshold to the three methods would not constitute a fair comparison. For this reason, an independent attack-free Monte Carlo calibration dataset is first generated for each detector. The threshold of each detector is independently selected according to $\gamma_j = Q_{1-\alpha}(J_j | \mathcal{H}_0)$, $j \in \{P, A, B\}$, where $Q_{1-\alpha}$ denotes the empirical $(1 - \alpha)$ -quantile and the target false-alarm probability is fixed at $\alpha = 0.05$. Therefore, all three methods are designed to operate at approximately the same FAR of 5%. Independent testing samples are subsequently used for performance evaluation, so that threshold calibration and performance testing are statistically separated. For detector j , the empirical FAR is defined as $P_F^{(j)} = \Pr\{J_j > \gamma_j | \mathcal{H}_0\}$.

For attack type $A \in \{\text{DoS}, \text{TA}, \text{RE}\}$, the detection rate is $P_D^{A,(j)} = \Pr\{J_j > \gamma_j | A\}$, and the corresponding missed detection rate is $P_M^{A,(j)} = 1 - P_D^{A,(j)}$. FAR and MDR are directly connected with the theoretical performance indices established in Section 4. However, a comparison based only on a single operating threshold may obscure the intrinsic separability of the underlying detection statistics. Therefore, the area under the receiver operating characteristic curve (ROC-AUC) is also evaluated. AUC is threshold-independent and measures the ability of a detector to distinguish attack-free and attacked

samples over all possible decision thresholds. Moreover, since the objective of the proposed method is unified attack detection, two additional cross-attack performance measures are introduced. The worst-case detection rate is defined by $P_D^{\text{worst}} = \min\{P_D^{\text{DoS}}, P_D^{\text{TA}}, P_D^{\text{RE}}\}$, whereas the macro-averaged detection rate is $P_D^{\text{macro}} = \frac{1}{3}(P_D^{\text{DoS}} + P_D^{\text{TA}} + P_D^{\text{RE}})$. The corresponding macro-AUC is defined as $\text{AUC}^{\text{macro}} = \frac{1}{3}(\text{AUC}^{\text{DoS}} + \text{AUC}^{\text{TA}} + \text{AUC}^{\text{RE}})$. In particular, P_D^{worst} is important for evaluating a unified detector. A detector may achieve nearly perfect performance for one specific attack while remaining almost ineffective against another. Such a method cannot provide reliable protection when the attack type is unknown in advance. The worst-case detection rate explicitly captures this vulnerability. To jointly account for successful attack detection and correct attack-free decisions, the macro balanced accuracy is also defined as $P_{\text{BA}}^{\text{macro}} = \frac{1}{2}[(1 - P_F) + P_D^{\text{macro}}]$.

The experiments are first conducted for data lengths $N = 5000$ and $N = 20,000$. For each detector and each data length, independent attack-free samples are used for threshold calibration, followed by 1000 independent testing trials under the attack-free, DoS, TA, and RE scenarios.

Table 2 summarizes the quantitative comparison. The notation DR/MDR reports the detection rate and missed detection rate simultaneously. All FAR, DR, MDR, worst-case DR, macro DR, and balanced-accuracy values are expressed in percent.

Table 2. Comparative detection performance under a common target FAR of 5%.

N	Detection method	FAR (%)	DoS DR/MDR (%)	TA DR/MDR (%)	RE DR/MDR (%)	Worst-case DR (%)	Macro DR (%)	Macro AUC	Macro BA (%)
5000	Proposed PE-split residual	4.7	35.0 / 65.0	19.4 / 80.6	33.5 / 66.5	19.4	29.30	0.7546	62.30
5000	Prior-model consistency	5.6	18.9 / 81.1	100 / 0	100 / 0	18.9	72.97	0.8785	83.68
5000	RE-specific lag detector	6.2	18.5 / 81.5	4.6 / 95.4	100 / 0	4.6	41.03	0.7072	67.42
20,000	Proposed PE-split residual	4.6	88.8 / 11.2	85.8 / 14.2	91.7 / 8.3	85.8	88.77	0.9762	92.08
20,000	Prior-model consistency	3.7	39.5 / 60.5	100 / 0	100 / 0	39.5	79.83	0.9334	88.07
20,000	RE-specific lag detector	6.4	30.4 / 69.6	6.4 / 93.6	100 / 0	6.4	45.60	0.7159	69.60

As shown in Table 2 and Figure 15, the empirical FARs of the three methods remain close to the prescribed 5% level. For $N = 5000$, the FARs are 4.7%, 5.6%, and 6.2% for the proposed, prior-model, and replay-specific detectors, respectively. For $N = 20,000$, the corresponding FARs are 4.6%, 3.7%, and 6.4%. The small deviations from 5% are caused by finite Monte Carlo calibration and independent finite-sample testing.

An important phenomenon can be observed when $N = 5000$. Under the ideal assumption that the exact true parameter vector is supplied to Baseline A, the prior-model consistency detector achieves perfect detection for TA and RE attacks and consequently obtains the highest macro DR of 72.97%. By comparison, the macro DR of the proposed detector is only 29.30% at this relatively limited data length. This result is retained intentionally rather than concealed, because it clarifies the different information requirements of the methods. Baseline A is given the exact parameter θ as prior information, whereas the proposed detector must infer attack-induced inconsistency directly from the binary observations without such knowledge. More importantly, the average performance of Baseline A masks a substantial weakness against DoS attacks. At $N = 5000$, its DoS DR is only 18.9%, resulting in a worst-case DR of 18.9%. The replay-specific detector exhibits an even stronger imbalance: Although its RE detection rate reaches 100%, its TA detection rate is only 4.6%. Consequently, its worst-case DR is only 4.6%. Therefore, even at $N = 5000$, the results already reveal the fundamental limitation

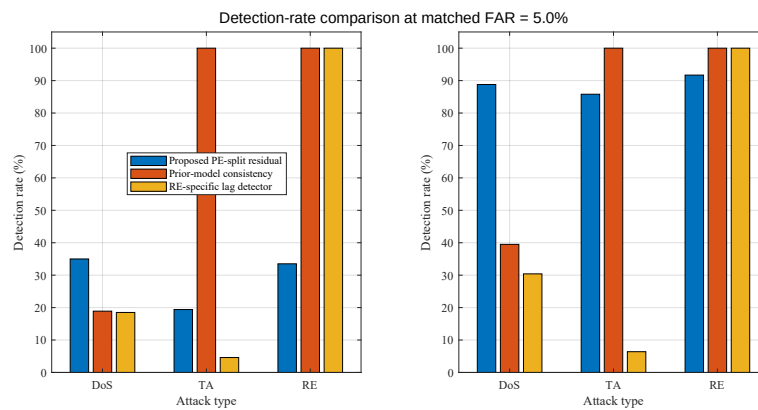


Figure 15. Detection-rate comparison of the three detectors under the matched false-alarm rate.

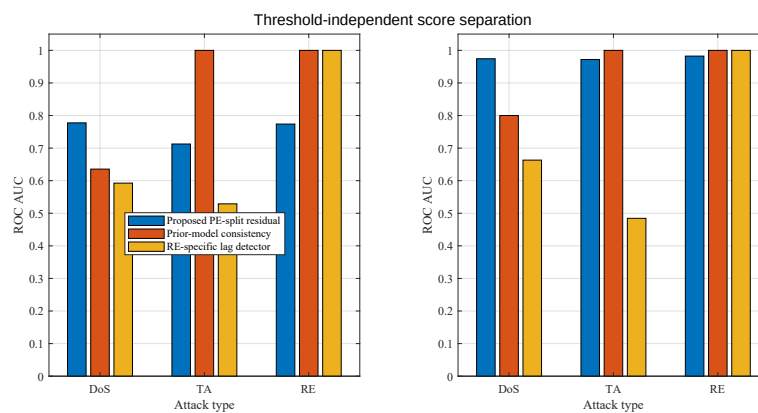


Figure 16. ROC-AUC comparison of the three detectors under different data lengths.

of an attack-specific detector: Excellent performance for the targeted attack does not imply robust detection against heterogeneous attack mechanisms. The advantage of the proposed detector becomes considerably more pronounced when the data length is increased to $N = 20,000$. The proposed method achieves detection rates of $P_D^{\text{DoS}} = 88.8\%$, $P_D^{\text{TA}} = 85.8\%$, and $P_D^{\text{RE}} = 91.7\%$. The corresponding MDRs decrease to $P_M^{\text{DoS}} = 11.2\%$, $P_M^{\text{TA}} = 14.2\%$, and $P_M^{\text{RE}} = 8.3\%$. Thus, the proposed detector achieves $P_D^{\text{worst}} = 85.8\%$ and $P_D^{\text{macro}} = 88.77\%$. In contrast, although Baseline A continues to achieve perfect detection for TA and RE attacks because it is supplied with the exact nominal parameter, its DoS DR is only 39.5%. Consequently, its worst-case DR is reduced to 39.5%, and its macro DR is 79.83%. The attack-specific nature of Baseline B is even more evident. It maintains a 100% RE detection rate but achieves only 30.4% and 6.4% detection rates for DoS and TA attacks, respectively. Its worst-case DR is therefore only 6.4%. Compared with Baseline A, the proposed method improves the worst-case DR at $N = 20,000$ by $85.8\% - 39.5\% = 46.3$ percentage points. Compared with the replay-specific baseline, the improvement reaches $85.8\% - 6.4\% = 79.4$ percentage points. These results provide direct numerical evidence that the main advantage of the proposed method lies in its balanced cross-attack detection capability, rather than in optimizing the detector for one particular attack type.

Figure 16 further compares the ROC-AUC values of the three methods. This comparison is important because AUC does not depend on the specific threshold selected in the matched-FAR experiment. At $N = 5000$, the AUCs of the proposed detector for DoS, TA, and RE attacks are 0.7775,

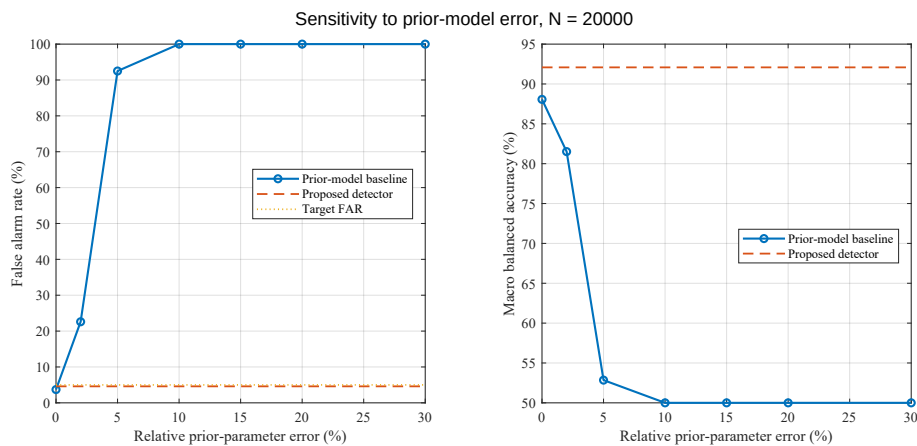


Figure 17. Sensitivity of detection performance to prior-model mismatch.

0.7126, and 0.7738, respectively, resulting in a macro-AUC of 0.7546. At this stage, the ideal prior-model detector provides a larger macro-AUC of 0.8785 because its TA and RE AUCs both reach 1. When the data length increases to $N = 20,000$, however, the three AUCs of the proposed detector increase to 0.9743, 0.9721, and 0.9824 for DoS, TA, and RE attacks, respectively. Its macro-AUC consequently reaches $\text{AUC}^{\text{macro}} = 0.9762$. The corresponding macro-AUCs of Baselines A and B are 0.9334 and 0.7159, respectively. It is particularly noteworthy that Baseline A still has perfect AUC values for TA and RE, but its DoS AUC is only 0.8001. Similarly, Baseline B provides perfect discrimination for RE, but its AUCs for DoS and TA are only 0.6632 and 0.4845, respectively. The latter value indicates almost no useful discrimination between attack-free and tampering samples. In contrast, all three AUC values of the proposed detector exceed 0.97 at $N = 20,000$. Therefore, the superior large-sample cross-attack performance is not an artifact caused by one favorable decision threshold. Instead, the attack-free and attacked detection-score distributions become intrinsically well separated for all three attack types. The macro balanced accuracy leads to the same conclusion. At $N = 20,000$, the proposed method achieves 92.08%, compared with 88.07% for Baseline A and 69.60% for Baseline B.

The preceding nominal comparison actually gives Baseline A a strong informational advantage, because the exact true parameter vector is assumed to be known: $\theta_{\text{prior}} = \theta$. However, the primary problem considered in this paper is attack detection during the parameter-identification process itself. In this situation, assuming that the unknown parameter is already exactly known may not be realistic. To explicitly evaluate this issue, the relative prior-model error is introduced as $e_{\theta} = \frac{\|\theta_{\text{prior}} - \theta\|}{\|\theta\|}$. The mismatch level is varied over $e_{\theta} \in \{0, 2\%, 5\%, 10\%, 15\%, 20\%, 30\%\}$.

Figure 17 compares the FAR and macro balanced accuracy of Baseline A under different prior-model errors with those of the proposed detector. This experiment should be interpreted by considering FAR and detection rate jointly. When the nominal parameter deviates from the true parameter, a prior-model consistency detector may generate a large statistic even under attack-free operation. In such a situation, an apparently high attack detection rate does not necessarily indicate better detection performance, because normal samples may simultaneously be classified as attacks. For this reason, macro balanced accuracy is used together with FAR. The balanced-accuracy measure penalizes both excessive false alarms and missed attacks and therefore prevents an artificially inflated DR from being interpreted as an improvement. In contrast, the proposed statistic contains no θ_{prior} . Therefore, changing e_{θ} does not alter the proposed detector itself. At $N = 20,000$, its reference FAR and macro balanced

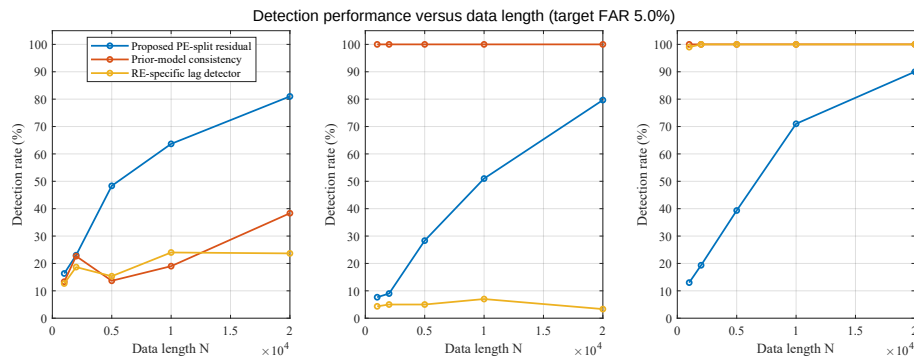


Figure 18. Detection-rate comparison versus data length under different attacks.

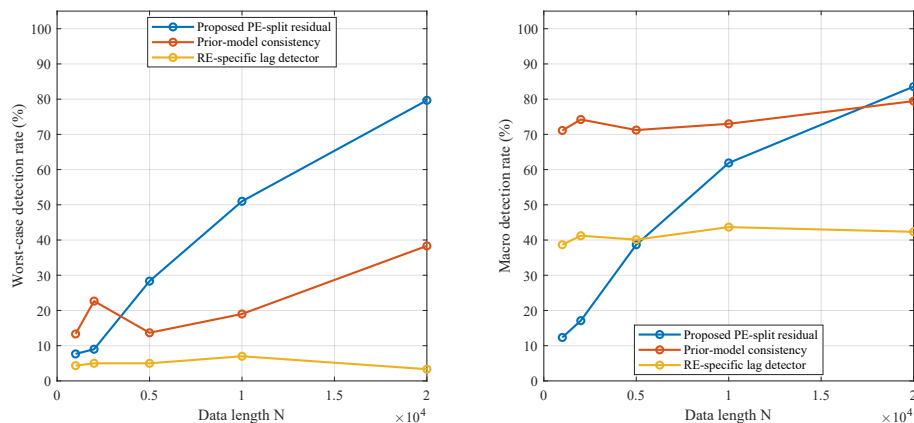


Figure 19. Unified detection performance versus data length.

accuracy remain $P_F = 4.6\%$ and $P_{BA}^{\text{macro}} = 92.08\%$, respectively. Hence, the prior-mismatch experiment illustrates an essential distinction between the two approaches: Baseline A detects deviations from an externally prescribed nominal model, whereas the proposed method detects disagreement between two independently constructed identification results obtained directly from the received data.

To further investigate the evolution of detection capability, Figure 18 compares the detection rates of the three methods over a range of data lengths. The three subfigures correspond to DoS, TA, and RE attacks, respectively. The results at $N = 5000$ and $N = 20,000$ already reveal a clear large-sample improvement of the proposed detector. Its DoS DR increases from 35.0% to 88.8%, its TA DR increases from 19.4% to 85.8%, and its RE DR increases from 33.5% to 91.7%. Correspondingly, the worst-case detection rate increases from 19.4% at $N = 5000$ to 85.8% at $N = 20,000$. This behavior is consistent with the theoretical mechanism underlying the proposed detection residual. With increasing data length, the random estimation errors of the two local parameter estimates decrease, whereas the systematic inconsistency caused by an attack becomes increasingly distinguishable from normal estimation fluctuations.

Figure 19 summarizes the comparison in terms of the worst-case and macro detection rates. These two quantities provide the most direct evidence for the unified-detection property. The replay-specific detector is able to maintain excellent RE detection but performs poorly against attacks that do not generate the temporal repetition feature for which it was designed. Therefore, its worst-case performance remains low. The prior-model detector behaves differently. When the exact parameter is supplied, it

achieves excellent TA and RE detection, but the results demonstrate that this does not guarantee equally strong performance against DoS attacks. Moreover, its applicability depends on the availability of an accurate nominal model. By contrast, at $N = 20,000$, the proposed detector simultaneously achieves high detection rates for all three considered attacks without changing its detector structure, without switching between attack-specific decision mechanisms, and without using the true parameter vector as prior information. Accordingly, the term unified detection advantage used in this paper should be understood in the following precise sense: proposed method is not designed to maximize the detection probability for one prespecified attack; instead, it provides a single identification-based detection mechanism that maintains high and balanced detection capability over heterogeneous DoS, tampering, and replay attacks without requiring prior knowledge of the attack type or an accurate nominal system parameter. This interpretation is also consistent with the numerical results. A specialized replay detector can outperform the proposed method for replay attacks alone, and an ideal prior-model detector can outperform it for certain attacks when the exact system parameter is supplied. Nevertheless, neither alternative provides the same combination of prior-model independence, attack-type independence, and high worst-case cross-attack detection performance. Therefore, the comparative simulations complement the residual-convergence and FAR/MDR analyses presented previously. The results demonstrate that the proposed method not only produces a statistically responsive residual under data-integrity attacks but also provides a practically meaningful unified detection capability when the attack mode is unknown and accurate prior model information is unavailable.

6. Conclusions

Targeting the underlying data security demands of CPSs, this paper systematically resolves the unified detection problem for a class of FIR systems with binary quantized observations facing data integrity attacks from the unique perspective of system identification. By discovering the persistent excitation property of system input signals, this paper constructs a multivariate detection residual by integrating the asymptotic convergence of the parameter identification algorithm, establishing a universal binary anomaly detection framework that does not require prior knowledge of the attack type. At the theoretical level of performance evaluation, this paper overcomes the extreme non-linear challenges introduced by binary quantization, rigorously proving the asymptotic normality of the detection residuals. Based on multivariate normal distribution theory, explicit analytical calculation methods for the FAR and MDR are derived. Furthermore, the comprehensive impacts of data length, decision threshold vectors, and attack strategies on the performance of the detection algorithm are thoroughly analyzed.

The proposed unified detection framework focuses on early anomaly warning and does not yet address attack isolation or refined classification of different attack modes. Future research will therefore proceed in several directions. First, attack isolation and classification based on higher-order statistics or pattern-recognition techniques will be investigated. Second, the present open-loop identification framework will be extended to more complex closed-loop quantized systems. Third, input-side uncertainty will be considered by studying robust and bias-compensated identification mechanisms for errors-in-variables and impulsive-noise environments, so that input-induced estimation bias can be separated from attack-induced residual changes. Finally, more comprehensive performance evaluation methods incorporating network delays, channel noise, and other practical communication uncertainties will be explored.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. Z. Yang, L. T. Yang, C. Li, L. Shan, H. Zhao, X. Nie, Collaborative Bayesian tensor factorization-based reliable traffic speed data prediction in T-CPS, *IEEE Trans. Intell. Transp. Syst.*, **26** (2025), 14393–14406. <https://doi.org/10.1109/TITS.2025.3564318>
2. P. O. F. Pires, M. L. C. Peixoto, P. H. S. Coutinho, R. M. Palhares, Joint estimation of deception attacks on sensors and actuators in cyber-physical systems, *IEEE Trans. Syst. Man Cybern.: Syst.*, **55** (2025), 7792–7803. <https://doi.org/10.1109/TSMC.2025.3600987>
3. P. Yu, Y. Hu, Y. Wang, R. Jia, J. Guo, Optimal consensus control strategy for multi-agent systems under cyber attacks via a stackelberg game approach, *IEEE Trans. Autom. Sci. Eng.*, **22** (2025), 18875–18888. <https://doi.org/10.1109/TASE.2025.3591858>
4. Z. Xu, Y. Kang, Y. Cao, Z. Li, Spatiotemporal graph convolution multifusion network for urban vehicle emission prediction, *IEEE Trans. Neural Networks Learn. Syst.*, **32** (2020), 3342–3354. <https://doi.org/10.1109/TNNLS.2020.3008702>
5. A. S. Mohamed, D. Kundur, Resilient Cyber-physical system honeypots for cyberattacker engagement, *IEEE Trans. Ind. Inf.*, **21** (2025), 8585–8595. <https://doi.org/10.1109/TII.2025.3584527>
6. W. Liu, Y. Wang, J. Guo, Consistent identification for FIR systems with multi-level quantized observations subjected to data tampering attacks: A joint estimation approach, *IEEE Trans. Ind. Electron.*, **73** (2026), 7676–7687. <https://doi.org/10.1109/TIE.2025.3642258>
7. Y. Wang, X. Zhao, W. Xing, Y. Li, B. Wang, DoS attacks on CPSs with multi-hop networks: A stochastic game with asymmetric information, *IEEE Trans. Autom. Control*, **71** (2026), 2560–2573. <https://doi.org/10.1109/TAC.2025.3631518>
8. W. Qian, D. Lu, Y. Wu, Novel Dynamic event-triggered load frequency control for networked power systems with deception attacks, *J. Syst. Sci. Complexity*, **39** (2026), 1612–1629. <https://doi.org/10.1007/s11424-025-4492-x>
9. J. Guo, Q. Zhang, Y. Zhao, Identification of FIR systems with binary-valued observations under replay attacks, *Automatica*, **172** (2025), 112001. <https://doi.org/10.1016/j.automatica.2024.112001>
10. H. Fu, P. Krishnamurthy, F. Khorrami, Combining switching mechanism with re-initialization and anomaly detection for resiliency of cyber-physical systems, *Automatica*, **172** (2025), 111994. <https://doi.org/10.1016/j.automatica.2024.111994>
11. K. Zhang, A. Kasis, T. Parisini, M. M. Polycarpou, Threat discrimination between attacks and faults for a class of nonlinear systems, *IEEE Trans. Autom. Control*, **71** (2026), 946–961. <https://doi.org/10.1109/TAC.2025.3604377>

12. Q. Zhang, J. Guo, Compensation strategy-based intrusion tolerant parameter estimation for quantized nonlinear Hammerstein systems under replay attacks, *IEEE Trans. Autom. Sci. Eng.*, **23** (2026), 1343–1359. <https://doi.org/10.1109/TASE.2025.3647889>
13. Y. Kang, Z. Li, Y. Dong, H. Xi, Markovian-based fault-tolerant control for wheeled mobile manipulators, *IEEE Trans. Control Syst. Technol.*, **20** (2011), 266–276. <https://doi.org/10.1109/TCST.2011.2109062>
14. J. Guo, J. F. Zhang, Y. Zhao, Adaptive tracking of a class of first-order systems with binary-valued observations and fixed thresholds, *J. Syst. Sci. Complexity*, **5** (2012), 1041–1051. <https://doi.org/10.1007/s11424-012-1257-0>
15. W. Zhao, E. Weyer, G. Yin, D. Dong, Y. Zhang, T. Shen, Adaptive regulation of block-oriented nonlinear systems using binary sensors with applications to automotive engine control, *IEEE Trans. Autom. Control*, **68** (2022), 1369–1382. <https://doi.org/10.1109/TAC.2022.3151217>
16. J. Xu, X. Zhang, J. Cheng, R. Liu, W. Zhang, GA-assisted event-triggered fault detection for networked systems under dos attacks, *IEEE Trans. Autom. Sci. Eng.*, **23** (2026), 5145–5154. <https://doi.org/10.1109/TASE.2026.3661045>
17. H. Liu, Y. Zhang, Y. Li, B. Niu, Proactive attack detection scheme based on watermarking and moving target defense, *Automatica*, **155** (2023), 111163. <https://doi.org/10.1016/j.automatica.2023.111163>
18. Z. Guo, X. Wang, M. Tian, J. Hu, Zonotopic state estimation and fault detection for discrete-time linear systems under DoS attacks: A switching controller design mechanism, *Automatica*, **185** (2026), 112833. <https://doi.org/10.1016/j.automatica.2026.112833>
19. R. Jia, T. Wang, W. Xue, J. Guo, Y. Zhao, Multi-time scale consensus algorithm of multi-agent systems with binary-valued data under tampering attacks, *IEEE Trans. Ind. Inf.*, **21** (2025), 9377–9388. <https://doi.org/10.1109/TII.2025.3598451>
20. Y. Li, C. N. Hadjicostis, N. Wu, Z. Li, Tamper-tolerant diagnosability analysis and tampering detectability in discrete event systems under cost constraints, *Automatica*, **171** (2025), 111971. <https://doi.org/10.1016/j.automatica.2024.111971>
21. J. Yan, Y. Xia, X. Feng, Y. Zhang, Deception attack detection based on bandwidth allocation for switched systems with quantization, *Automatica*, **154** (2023), 111094. <https://doi.org/10.1016/j.automatica.2023.111094>
22. D. Zhao, Y. Shi, S. X. Ding, Y. Li, F. Fu, Replay attack detection based on parity space method for cyber-physical systems, *IEEE Trans. Autom. Control*, **70** (2024), 2390–2405. <https://doi.org/10.1109/TAC.2024.3476309>
23. R. Fritz, P. Zhang, Detection and localization of stealthy cyberattacks in cyber-physical discrete event systems, *IEEE Trans. Autom. Control*, **68** (2023), 7895–7902. <https://doi.org/10.1109/TAC.2023.3253462>
24. L. Zhao, W. Xing, N. Xu, Stochastic event-based physical watermarks against replay attacks in cyber-physical systems, *IEEE Trans. Control Network Syst.*, **11** (2024), 1035–1045. <https://doi.org/10.1109/TCNS.2023.3331161>

25. M. Li, Y. Zhao, J. Guo, Differentially private state estimation over 1-bit communication networks: An optimal accuracy-privacy trade-off, *IEEE Trans. Autom. Control*, **2026** (2026). <https://doi.org/10.1109/TAC.2026.3723761>
26. Y. S. Chow, H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, 3rd edition, Springer-Verlag, New York, 1997.
27. A. Gut, *The Multivariate Normal Distribution*, in *An Intermediate Course in Probability*, Springer, New York, 2009. https://doi.org/10.1007/978-1-4419-0162-0_5
28. Y. R. Chien, H. E. Hsieh, G. Qian, Robust bias compensation LMS algorithms under colored Gaussian input noise and impulse observation noise environments, *Mathematics*, **13** (2025), 3348. <https://doi.org/10.3390/math13203348>
29. Y. R. Chien, E. T. Lin, G. Qian, Robust nonlinear sensor data prediction in the presence of input noise and output impulsive outliers: A bias-compensated KMCC approach, *IEEE Sens. J.*, **26** (2026), 23025–23037. <https://doi.org/10.1109/JSEN.2026.3704308>
30. J. Guo, R. Jia, R. Su, Y. Song, F. Jing, DoS attack detection in identification of FIR systems with binary-valued observations, *Asian J. Control*, **25** (2023), 2469–2481. <https://doi.org/10.1002/asjc.3005>
31. Q. Zhang, Y. Zhao, W. Xue, J. Guo, Parameter identification for FIR systems with simultaneously quantized inputs and outputs against replay attacks: a stochastic defense scheme, *IEEE Trans. Autom. Control*, **2026** (2026). <https://doi.org/10.1109/TAC.2026.3687459>



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)