Detection and isolation of routing attacks through sensor watermarking

Ferrari, Riccardo M.G.; Herdeiro Teixeira, A.M.

DOI
10.23919/ACC.2017.7963800

Publication date
2017

Document Version
Accepted author manuscript

Published in
Proceedings of the 2017 American Control Conference (ACC 2017)

Accepted Author Manuscript. Link to published article (IEEE): http://dx.doi.org/10.23919/ACC.2017.7963800

Detection and Isolation of Routing Attacks through Sensor Watermarking

Riccardo M.G. Ferrari and André M.H. Teixeira

***Abstract*— In networked control systems, leveraging the peculiarities of the cyber-physical domains and their interactions may lead to novel detection and defense mechanisms against malicious cyber-attacks. In this paper, we propose a multiplicative sensor watermarking scheme, where each sensor's output is separately watermarked by a Single Input Single Output (SISO) filter. Hence, such scheme does not require communication between multiple sensors, but can still lead to detection and isolation of malicious cyber-attacks. In particular, we analyze the benefits of the proposed watermarking scheme for two attack scenarios: the physical sensor re-routing attack and the cyber measurement re-routing one. For each attack scenario, detectability and isolability properties are analyzed with and without the proposed watermarking scheme and we show how the watermarking scheme can be leveraged to detect cyber sensor routing attacks. In order to detect compromised sensors, we design an observer-based detector with a robust adaptive threshold. Additionally, we identify the sensors involved in the re-routing attacks by means of a tailored Recursive Least Squares parameter estimation algorithm. The results are illustrated through a numerical example.**

I. Introduction

Modern control systems are increasingly relying on information and communication technology (ICT) infrastructures to exchange measurement and control signals. However, the increasing use of pervasive and open-standard ICT systems results in control systems becoming increasingly vulnerable to malicious cyberthreats, which may affect the physical processes through the control loop. Therefore, addressing cybersecurity of control systems requires both the cyber and physical domains to be taken into account. This need goes beyond capturing the effects of cyberattacks on the physical processes. On one hand, conventional cybersecurity mechanisms may be inapplicable to control systems, due to the strict functionality or performance requirements on the physical process and ICT infrastructure. On the other hand, leveraging the peculiarities of the cyber-physical domains and their interactions may lead to novel detection and defense mechanisms spanning across multiple layers, which is commonly termed as *defense-in-depth* [1].

The topic of cyber-secure control systems has been receiving increasing attention recently. An overview of existing cyberthreats and vulnerabilities in networked control systems is presented in [2], [3]. Rational adversary models are highlighted as one of the key items in security for control systems, thus making adversaries endowed with intelligence and intent, as opposed to faults. Therefore, these adversaries may exploit existing vulnerabilities and limitations in the traditional anomaly detection mechanisms and remain undetected. In fact, [4] uses such fundamental limitations to characterize a set of stealthy attack policies for networked systems modeled by differential-algebraic equations.

Recent work shows that a careful analysis of the fundamental limitations to the detectability of cyber-attacks by conventional schemes may lead to tailored detection mechanisms. Detectability conditions of stealthy false-data injection attacks to control systems are closely examined in [5], where the authors characterized modifications to the system dynamics that reveal stealthy data attacks. Recently, [6] proposed an static output coding scheme combining the outputs of multiple sensors to reveal stealthy data injection attacks on sensors. Less studied are attacks of multiplicative nature, such as replay [7] and routing attacks [8]. In particular, fundamental limitations in the detection of these attacks are not yet fully understood, and the detection and isolation of routing attacks has yet to be addressed. Within this class of attacks, replay attacks have been more extensively analyzed. In [7], the analysis of detectability conditions for replay attacks shows that, asymptotically, replay attacks are undetectable. To detect replay attacks, the authors proposed a novel detection scheme through additive watermarking, which is a well-known solution to the problem of proof of ownership verification and tampering detection in the field of multimedia data [9].

In the watermarking scheme proposed in [7], noise is purposely injected in the system by the actuators to watermark the sensor outputs through known correlations. However, such additive watermark presents some drawbacks: the performance of the system decreases and the actuators are further burdened with noisy inputs. These two drawbacks can be tackled by employing multiplicative sensor watermarks, akin to the techniques explored in [5], [6].

As main contributions of this paper, we study the fundamental limitations in detectability of routing attacks and propose tailored detection and isolation schemes to identify these attacks. In particular, to facilitate the detection and identification of routing attacks, we propose a multiplicative sensor watermarking scheme where each sensor output is separately watermarked through a SISO filter.

Two routing attack scenarios are considered, namely the cyber and physical re-routing of measurements. For each attack scenario, detectability and isolability properties are analyzed with and without the proposed watermarking scheme. Furthermore, we show how the watermarks can be leveraged

This work has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 608224 and no. 324432 (AMBI).

R. Ferrari is with the Delft Center for Systems and Controls, A. Teixeira with the Faculty of Technology, Policy and Management, both at the Delft Technical University {r.ferrari,andre.teixeira}@tudelft.nl

to detect and identify the sensors involved in the routing attacks, as well as the cyber or physical nature of the attack.

The outline of the paper is as follows. In Section II, we present the problem formulation and control system, describe the routing attack scenarios, and analyze their isolability properties without watermarking. The sensor watermarking scheme is described in Section III, where structural detectability properties are discussed for each attack scenario. To diagnose the routing attacks, an observer-based detection scheme with robust adaptive threshold is proposed in Section IV, while Section V describes an adaptive observer-based estimator that is used to diagnose the attack. Numerical results are presented in Section VI, and the paper concludes with final remarks in Section VII.

II. Problem formulation

In this section, we present the control system and describe the main problem at hand. Consider the modeling framework described in [3], where the control system is composed by a physical plant ($\mathcal{P}$), a feedback controller, and an anomaly detector ($\mathcal{R}$). The physical plant and anomaly detector are modeled in a discrete-time state-space form as, respectively,

$$\mathcal{P}: \begin{cases} x_p[k+1] = A_p x_p[k] + B_p u[k] + \eta[k] \\ y_p[k] = C_p x_p[k] + \xi[k] \end{cases} \tag{1}$$

$$\mathcal{R}: \begin{cases} x_r[k+1] = A_r x_r[k] + B_r u[k] + K_r \tilde{y}_{pw}[k] \\ y_r[k] = C_r x_r[k] + D_r u[k] + E_r \tilde{y}_{pw}[k] \end{cases} \tag{2}$$

where $x_p[k] \in \mathbb{R}^{n_p}$ and $x_r[k] \in \mathbb{R}^{n_r}$ are the state variables, $u[k] \in \mathbb{R}^{n_u}$ is the vector of control actions applied to the process, $y_p[k] \in \mathbb{R}^{n_y}$ is the vector of plant outputs, $y_{pw} \in \mathbb{R}^{n_y}$ denotes the data transmitted by the sensors, $\tilde{y}_{pw} \in \mathbb{R}^{n_y}$ the data received by the detector, and $y_r[k] \in \mathbb{R}^{n_r}$ the residual vector. The real-valued matrices A_p, B_p, C_p and A_r, B_r, C_r are of appropriate dimensions. The variables $\eta[k]$ and $\xi[k]$ denote the unknown process and measurement disturbances, respectively.

Assumption 1: The uncertainties represented by the vectors η and ξ are unknown, but their norms are upper bounded by some known and bounded sequences $\bar{\eta}[k]$ and $\bar{\xi}[k]$. □

For simplicity, we assume that each sensor measures and transmits a scalar value, where $\tilde{y}_{p,(i)}[k] \in \mathbb{R}$ denotes the the measurement of the i-th sensor. To model the fact that the sensor measurements may have been subject to physical attacks, we denote $\tilde{y}_p[k] \in \mathbb{R}^{n_y}$ as the set of measurements actually read by the sensors. Similarly, the sensor measurements are exchanged through a communication network, thus the transmitted and received data may differ due to, for instance, packet losses or data corruption. At the plant side, we denote the data transmitted by the sensors as $y_{pw}[k] \in \mathbb{R}^{n_y}$ whereas, at the detector's side, the received sensor data is denoted as $\tilde{y}_{pw}[k] \in \mathbb{R}^{n_y}$. The detector is collocated with the controller and it evaluates the behavior of the plant based only on the closed-loop models, $\tilde{y}_{pw}[k]$ and $u[k]$.

The main focus of this paper is to investigate the detection and isolation of cyber and physical sensor routing attacks,

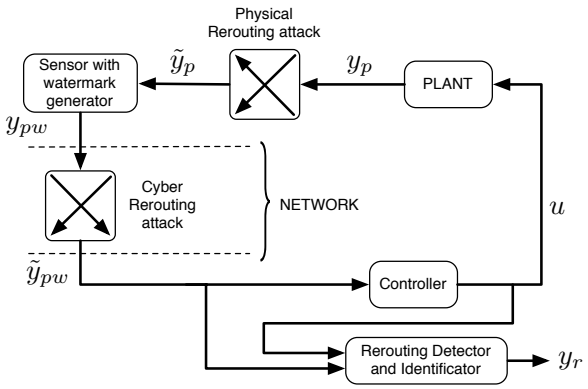

Fig. 1. A block-diagram representation of the setting considered in the present paper.

which are modeled as the multiplicative anomalies R^C and R^P such that $\tilde{y}_{pw}[k] = R^C y_{pw}[k]$ and $\tilde{y}_p[k] = R^P y_p[k]$, respectively. These attack scenarios, as well a fundamental limitation in their distinguishability, are described next.

A. Attack scenarios

Given the structure of the closed-loop system described above, we now present the attack scenarios considered in this work, which are summarized in Figure 1.

Physical measurement routing attack: In this scenario, the adversary re-routes the measurement signals read by the sensors, e.g. by physically re-wiring the sensor cables.

A physical routing attack that re-wires the measurements from sensor j to sensor i is denoted as a *physical* (j,i)*-routing attack*. More generally, multiple physical routing attacks can be characterized by a directed graph $\mathcal{G}_R = (\mathcal{V}_R, \mathcal{E}_R)$, where $\mathcal{V}_R = \{1, \ldots, n_y\}$ is the vertex set representing the set of sensors and $\mathcal{E}_R \subset \mathcal{V}_R \times \mathcal{V}_R$ is the set of directed edges representing the set of routing attacks. Furthermore, define $\mathcal{V}_O = \{v \in \mathcal{V}_R : (v,u) \in \mathcal{E}_R \text{ for some } u \in \mathcal{V}_R\}$, $\mathcal{V}_I = \{u \in \mathcal{V}_R : (v,u) \in \mathcal{E}_R \text{ for some } v \in \mathcal{V}_R\}$. Assuming the in-degree of each node is at most 1, the set of $\mathcal{E}_R$-routing attacks are described by

$$\begin{aligned} &\tilde{y}_{p,(i)}[k] = y_{p,(j)}[k],\ \forall (j,i) \in \mathcal{E}_R, \\ &\tilde{y}_{p,(l)}[k] = y_{p,(l)}[k],\ \forall l \notin \mathcal{V}_I. \end{aligned} \tag{3}$$

To obtain a more compact representation, define the Laplacian matrix of the digraph $\mathcal{G}_R$ as

$$L_{R,(i,j)} = \begin{cases} \deg(\mathrm{i}) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } (j,i) \in \mathcal{E}_R \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\deg(\mathrm{i})$ is the in-degree of $i \in \mathcal{V}_R$, which is assumed to be at most 1. Defining the *physical routing matrix* $R^P \triangleq I - L_R$, the set of physical $\mathcal{E}_R$-routing attacks are compactly described by $\tilde{y}_p[k] = R^P y_p[k]$.

Cyber measurement routing attack: In the scenario of a cyber routing attack, the adversary is able to re-route the measurements transmitted by the sensors, by modifying the respective sender identifier. Similarly to a physical routing

attack, a cyber routing attack that re-routes a measurement from sensor j to sensor i, is denoted as a *cyber* (j,i)*–routing attack*. Using the graph notation previously introduced, the set of cyber $\mathcal{E}_R$-routing attacks are described by

$$\begin{aligned} \tilde{y}_{pw,(i)}[k] &= y_{pw,(j)}[k],\ \forall (i,j) \in \mathcal{E}_R, \\ \tilde{y}_{pw,(l)}[k] &= y_{pw,(l)}[k],\ \forall l \notin \mathcal{V}_I, \end{aligned} \tag{5}$$

or by the vector form $\tilde{y}_{pw}[k] = R^C y_{pw}[k]$, where we introduced the *cyber routing matrix* R^C.

In the following, when the cyber or physical nature of a routing attack can be neglected, we refer to the attack as a $\mathcal{E}_R$-routing attack and use R instead of R^C and R^P. Next, we discuss the fundamental limitations in the isolation of the type (cyber or physical) of routing attack.

B. Indistinguishability of cyber and physical routing attacks

Identifying the cyber or physical nature of the attacks is important to devise suitable corrective measures against these attacks. Unfortunately, under the natural assumption that the sensors transmit their measurements unaltered, the following limitation is inherent to these routing scenarios.

Theorem 1: Assuming that the sensors transmit the measured outputs of the plant unaltered, i.e., $y_{pw}[k] = \tilde{y}_p[k]$, the cyber and physical $\mathcal{E}_R$–routing attacks are indistinguishable.

Proof: From (3) and (5), we have that a physical and a cyber routing attacks would respectively result in $\tilde{y}_{pw} = \tilde{y}_p = R^P y_p$ and $\tilde{y}_{pw} = R^C y_{pw}[k] = R^C y_p$, which makes the attacks indistinguishable. ■

To allow the routing detector to distinguish the nature of the attack, we propose to introduce a pre-processing step where each sensor processes the measurements through a filter before transmitting the data , which we denote as *sensor watermarking*. Furthermore, as we shall conclude in the following section, watermarking the sensors may also improve the detectability of cyber routing attacks.

III. Sensor Watermarking

Without loss of generality and in the linear case, we assume the watermark generator to be implemented through an infinite impulse response (IIR) filter of order N, which for the ith measurement is described by the difference equation

$$y_{pw,(i)}[k] = \sum_{n=1}^{N} w^i_{A,(n)} y_{pw,(i)}[k-n] + \sum_{n=0}^{N} w^i_{B,(n)} \tilde{y}_{p,(i)}[k-n], \tag{6}$$

where $w^i_A = [w^i_{A,(1)} \ \dots \ w^i_{A,(N)}]^\top \in \mathbb{R}^N$ and $w^i_B = [w^i_{B,(0)} \ \dots \ w^i_{B,(N)}]^\top \in \mathbb{R}^{N+1}$ are the filter parameters. Recall that choosing $w^i_A = 0$ retrieves a finite impulse response (FIR) filter. Furthermore, we consider the following state-space realization of the filter

$$\begin{aligned} x^i_w[k+1] &= A^i_w x^i_w[k] + B^i_w \tilde{y}_{p,(i)}[k] \\ y_{pw,(i)}[k] &= C^i_w x^i_w[k] + D^i_w \tilde{y}_{p,(i)}[k], \end{aligned} \tag{7}$$

where $x^i_w[k] \in \mathbb{R}^N$. The collection of all filters reads as

$$\begin{aligned} x_w[k+1] &= A_w x_w[k] + B_w \tilde{y}_p[k] \\ y_{pw}[k] &= C_w x_w[k] + D_w \tilde{y}_p[k], \end{aligned} \tag{8}$$

with $x_w[k] = \left[x^{1\top}_w[k] \ \dots \ x^{n_y\top}_w[k]\right]^\top$ and the matrices

$$\begin{aligned} A_w &= \text{blkdiag}\left(\{A^i_w\}_{i=1}^{n_y}\right),\ B_w = \text{blkdiag}\left(\{B^i_w\}_{i=1}^{n_y}\right), \\ C_w &= \text{blkdiag}\left(\{C^i_w\}_{i=1}^{n_y}\right),\ D_w = \text{blkdiag}\left(\{D^i_w\}_{i=1}^{n_y}\right). \end{aligned}$$

The cascade system of the plant and the filters is given by

$$\mathcal{P}_w : \begin{cases} x_{pw}[k+1] = A_{pw} x_{pw}[k] + B_{pw} u[k] + \eta_{pw}[k] \\ y_{pw}[k] = C_{pw} x_{pw}[k] + \xi_{pw}[k] \end{cases} \tag{9}$$

where $x_{pw} \in \mathbb{R}^{n_{pw}}$, with $n_{pw} \triangleq n_p + N n_y$, and we have

$$\begin{aligned} A_{pw} &\triangleq \begin{bmatrix} A_p & 0 \\ B_w C_p & A_w \end{bmatrix},\ B_{pw} \triangleq \begin{bmatrix} B_p \\ 0 \end{bmatrix},\ \eta_{pw}[k] \triangleq \begin{bmatrix} \eta[k] \\ B_w \xi[k] \end{bmatrix} \\ C_{pw} &\triangleq \begin{bmatrix} D_w C_p & C_w \end{bmatrix},\ \xi_{pw}[k] \triangleq D_w \xi[k]. \end{aligned} \tag{10}$$

For well-posedness, we need the following assumptions.

Assumption 2: No routing attacks are present for $0 \leq k < k_0$, with k_0 being the attack start time. Moreover, the variables x_p, x_{pw} and u remain bounded before and after the occurrence of an attack, i.e., there exist some stability regions $\mathscr{S} = \mathscr{S}^{x_p} \times \mathscr{S}^{x_{pw}} \times \mathscr{S}^{u} \subset \mathbb{R}^{n_p} \times \mathbb{R}^{n_{pw}} \times \mathbb{R}^m$, such that $(x_p, x_{pw}, u) \in \mathscr{S}, \forall\, k$. □

Assumption 3: (A_{pw}, C_{pw}) is a detectable pair. □

A. Models of routing attacks with watermarked sensors

Recall from Th. 1 that the cyber or physical nature of the routing attacks cannot be discerned without the watermarking scheme. Next we derive the models of cyber and physical routing attacks under the proposed sensor watermarking scheme and we analyze the influence of the watermarking filters on the detectability of each routing attack.

With the sensor watermarking scheme, the data received by the detector under a cyber routing attack is given by

$$\begin{cases} x_{pw}[k+1] = A^C_{pw} x_{pw}[k] + B_{pw} u[k] + \eta^C_{pw}[k] \\ \tilde{y}_{pw}[k] = C^C_{pw} x_{pw}[k] + \xi^C_{pw}[k] \end{cases}$$

with $A^C_{pw} = A_{pw}$, $C^C_{pw} = C_{pw} + \Delta C^C_{pw}$, $\Delta C^C_{pw} = (R^C - I) C_{pw}$, $\eta^C_{pw}[k] = \eta_{pw}[k]$, $\xi^C_{pw}[k] = \xi_{pw}[k] + \Delta\xi^C_{pw}[k]$, $\Delta\xi^C_{pw}[k] = (R^C - I)\xi_{pw}[k]$.

Instead, the physical routing effect on the dynamics can be modelled as

$$\begin{cases} x_{pw}[k+1] = A^P_{pw} x_{pw}[k] + B_{pw} u[k] + \eta^P_{pw}[k] \\ \tilde{y}_{pw}[k] = C^P_{pw} x_{pw}[k] + \xi^P_{pw}[k] \end{cases}$$

with $A^P_{pw} = A_{pw} + \Delta A^P_{pw}$, $C^P_{pw} = C_{pw} + \Delta C^P_{pw}$, and

$$\begin{aligned} \Delta A^P_{pw} &= \begin{bmatrix} 0 & 0 \\ B_w (R^P - I) C_p & 0 \end{bmatrix}, \\ \Delta C^P_{pw} &= \begin{bmatrix} D_w (R^P - I) C_p & 0 \end{bmatrix}. \end{aligned}$$

and also $\eta^P_{pw}[k] = \eta_{pw}[k] + \Delta\eta^P_{pw}[k]$, $\xi^P_{pw}[k] = \xi_{pw}[k]$, and

$$\Delta\eta^P_{pw}[k] = \begin{bmatrix} 0 \\ B_w (R^P - I)\xi[k] \end{bmatrix}.$$

B. Structural detectability of routing attacks with sensor watermarking

We start by recalling the definition of structural detectability. Consider the dynamical system $\Sigma_i \triangleq (A_i,\ B_i,\ C_i,\ D_i) = (A + \Delta A_i,\ B + \Delta B_i,\ C + \Delta C_i,\ D + \Delta D_i)$ with multiplicative anomalies and let $\Sigma = (A, B, C, D)$ be the nominal system. The detectability of attacks will be discussed according to the following definitions [10].

Definition 1: Consider two anomalies occurring at $k = k_0$, which are described by the dynamical systems Σ_1 and Σ_2, respectively. These anomalies are said to be *structurally indistinguishable* w.r.t. the input signal u if there exist non-zero initial conditions x_1 and x_2 such that $y_1[k] = y_2[k]$ for all $k \geq k_0$. Furthermore, an anomaly described by Σ_1 is said to be *structurally undetectable* w.r.t. u if it is indistinguishable w.r.t. u from the nominal system Σ. An anomaly is said to be structurally weakly-indistinguishable (undetectable) if it is structurally indistinguishable (undetectable) w.r.t. $u = 0$.

The structural indistinguishability of anomalies described by Σ_1 and Σ_2 can be analyzed by studying the zero dynamics of the system

$$\begin{bmatrix} x_1[k+1] \\ \Delta x[k+1] \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ A_1 - A_2 & A_2 \end{bmatrix} \begin{bmatrix} x_1[k] \\ \Delta x[k] \end{bmatrix} + \begin{bmatrix} B_1 \\ B_1 - B_2 \end{bmatrix} u[k]$$
$$\Delta y[k] = \begin{bmatrix} C_1 - C_2 & C_1 \end{bmatrix} \begin{bmatrix} x_1[k] \\ \Delta x[k] \end{bmatrix}, \tag{11}$$

where $\Delta x = x_1 - x_2$ and $\Delta y = y_1 - y_2$. In particular, following Definition 1, the anomalies are said to be indistinguishable if there exist initial conditions $x_1[0]$ and $\Delta x[0]$ and input u such that $\Delta y[k] = 0$ for all k, i.e., u is a zero dynamics input of (11) associated with the initial conditions $x_1[0]$ and $\Delta x[0]$.

Structural detectability of physical routing attacks: The structural detectability conditions for multiplicative anomalies naturally depend on the excitation induced by external inputs. Therefore, the analysis below focuses on structural detectability under the influence of the controlled input signal u, whereas the uncontrolled and unknown inputs (the noise terms) are not considered.

Theorem 2: Let the noise terms η and ξ be identically zero. The physical $\mathcal{E}_R$–routing attack is structurally undetectable w.r.t. u if u is a zero dynamics input signal of the system $(A_p, B_p, L_R C_p, 0)$.

Proof: Considering (11) with $\Sigma_1 = (A_{pw}, B_{pw}, C_{pw}, 0)$ and $\Sigma_2 = (A^P_{pw}, B^P_{pw}, C^P_{pw}, 0)$ the undetectability conditions can be written as the existence of initial conditions Δx and x such that

$$\begin{bmatrix} x[k+1] \\ \Delta x[k+1] \end{bmatrix} = \begin{bmatrix} A_{pw} & 0 \\ -\Delta A^P_{pw} & A^P_{pw} \end{bmatrix} \begin{bmatrix} x[k] \\ \Delta x[k] \end{bmatrix} + \begin{bmatrix} B_{pw} \\ 0 \end{bmatrix} u[k]$$
$$0 = \begin{bmatrix} -\Delta C^P_{pw} & C^P_{pw} \end{bmatrix} \begin{bmatrix} x[k] \\ \Delta x[k] \end{bmatrix}.$$

The proof concludes by choosing $\Delta x = 0$ and x and u as the state and input of the zero dynamics of $(A_p, B_p, L_R C_p, 0)$, which results in $L_R C_p x[k] = 0$, for all $k > 0$, and thus leads to $\Delta A^P_{pw} x[k] = \Delta C^P_{pw} x[k] = 0$. ∎

The previous result shows that physical routing attacks that are structurally undetectable without watermarked sensors remain so with the watermarking scheme. However, as derived in the remainder of this section, the watermarking scheme can affect the detectability of cyber routing attacks.

Structural detectability of cyber routing attacks: The necessary and sufficient conditions for structural detectability of cyber routing attacks are as follows.

Theorem 3: Let the noise terms η and ξ be identically zero. The cyber $\mathcal{E}_R$–routing attack is structurally undetectable w.r.t. u if and only if u is a zero dynamics input signal of the system $(A_{pw}, B_{pw}, L_R C_{pw}, 0)$.

Proof: The proof is similar to that of Th. 2. ∎

For a cyber (j, i)–routing attack, Th. 3 states that the attack is undetectable if the input u is constructed such that the watermarked outputs $y_{pw,(i)}$ and $y_{pw,(j)}$ are identical, which shows that detectability depends on the dynamics of the physical system and watermarking filters. Considering no external inputs, the next statement readily follows.

Corollary 1: Let the noise terms η and ξ be identically zero. The cyber $\mathcal{E}_R$–routing attack is structurally weakly-undetectable if and only if A_p has an eigenvalue λ with a corresponding eingenvector v such that $L_R \left(D_w + C_w(\lambda I - A_w)^{-1} B_w\right) C_p v = 0$.

From the above results, one can observe that watermarking the sensors' measurements can indeed facilitate the detection of cyber routing attacks. For instance, without watermarking and given the structure of L_R, a cyber (i, j)–routing attack would be undetectable to any anomaly detector if the open-loop system has a mode or input yielding equal outputs $y_{(i)}$ and $y_{(j)}$. On the other hand, suitably choosing the watermark parameters so that sensors i and j have watermark generators with different transfer functions would make such a cyber (i, j)–routing attack detectable.

In the next section, we propose an observer-based detector with a robust adaptive threshold and deriving conditions under which structurally detectable anomalies are detected.

IV. Detection of Routing Attacks

The detector $\mathcal{R}$ in (2) will be implemented as the following observer [11], modeled on the nominal dynamics of the cascade of the plant and the watermark generators (9),

$$\begin{cases} \hat{x}_{pw}[k+1] = A_{pw}\hat{x}_{pw}[k] + B_{pw}u[k] + K\left(\tilde{y}_{pw}[k] - \hat{y}_{pw}[k]\right) \\ \hat{y}_{pw}[k] = C_{pw}\hat{x}_{pw}[k] \end{cases}, \tag{12}$$

where $\hat{x}_{pw}$ and $\hat{y}_{pw}$ of suitable size are dynamic estimates of x_{pw} and y_{pw}, and the output error gain matrix K is chosen such that $A_r \triangleq A_{pw} - KC_{pw}$ is Schur. In the absence of attacks (i.e., $\tilde{y}_{pw} = y_{pw}$, and $\tilde{y}_p = y_p$), the dynamics for the estimation errors $x_r \triangleq x_{pw} - \hat{x}_{pw}$ and $y_r \triangleq \tilde{y}_{pw} - \hat{y}_{pw}$ can be derived from (9) and (12) as

$$\begin{cases} x_r[k+1] = A_r x_r[k] + \eta_{pw}[k] \\ y_r[k] = C_{pw} x_r[k] + \xi_{pw}[k] \end{cases},$$

whose solution for the output residual is

$$y_r[k] = C_{pw}\Big[\sum_{h=0}^{k-1}(A_r)^{k-1-h}\left(\eta_{pw}[h]-K\xi_{pw}[h]\right)+(A_r)^k x_r[0]\Big] + \xi_{pw}[k] \quad (13)$$

In the absence of attacks the following holds

$$|y_{r,(i)}[k]| \leq \bar{y}_{r,(i)}[k] \triangleq \alpha^i\Big[\sum_{h=0}^{k-1}\left(\delta^i\right)^{k-1-h}\left(\bar{\eta}_{pw}[h]+ \|K\|\bar{\xi}_{pw}[h]\right) + \left(\delta^i\right)^k \bar{x}_r[0]\Big] + \bar{\xi}_{pw}[k] \quad (14)$$

where $\bar{y}_{r,(i)}[k]$ is a robust detection threshold for the i–th sensor output, α^i and δ^i are two constants such that $\|C_{pw,(i)}\left(A_r\right)^k\| \leq \alpha^i\left(\delta^i\right)^k \leq \|C_{pw,(i)}\| \cdot \|\left(A_r\right)^k\|$ with $C_{pw,(i)}$ being the i–th row of matrix C_{pw} (see [11] and [12, Th. 3.5]). Furthermore, $\bar{\eta}_{pw}$, $\bar{x}_r[0]$ and $\bar{\xi}_{pw}$ are upper bounds on the norms of, respectively, η_p, $x_r[0]$ and ξ_{pw}, which can be computed thanks to Assumption 1, 2 and eq. (10).

A cyber or physical routing attack will be detected if the residual evaluation rule (14) fails for at least one time instant and one sensor.

Theorem 4 (Attack Detectability): If there exists a time index $k_d > k_0$ and a component $i \in \{1, \ldots, n_y\}$ such that during a cyber (respectively physical) routing attack the functions ϕ_1 and ϕ_2 fulfill the following inequality

$$\left|C_{pw,(i)}\left(\sum_{h=k_0}^{k_d-1}(A_r)^{k_d-1-h}\phi_1[h]\right)+\phi_2[k_d]\right| > 2\alpha^i\sum_{h=0}^{k_d-1}\left(\delta^i\right)^{k_d-1-h}\left(\bar{\eta}_{pw}[h]+\|K\|\bar{\xi}_{pw}[h]\right)+ \left(\delta^i\right)^{k_d-k_0}\left(\alpha^i\bar{x}_r[k_0]+\bar{y}_{r,(i)}[k_0]\right)+2\bar{\xi}_{pw}[k_d]$$

where $\bar{y}_{r,(i)} \triangleq \max|y_{r,(i)}|$ and ϕ_1 and ϕ_2 are defined as

$$\phi_1[h] \triangleq \begin{cases} -K\left(\Delta\xi^C_{pw}[h]+\Delta C^C_{pw}x_{pw}[h]\right) & \text{cyber} \\ \left(\Delta\eta^P_{pw}[h]-\left(\Delta A^P_{pw}+K\Delta C^P_{pw}\right)x_{pw}[h]\right) & \text{physical} \end{cases}$$

$$\phi_2[k] \triangleq \begin{cases} \Delta C^C_{pw,(i)}x_{pw}[k]+\Delta\xi^C_{pw,(i)}[k] & \text{cyber} \\ 0 & \text{physical} \end{cases}$$

then the cyber (respectively physical) routing attack will be detected at the time instant k_d. □

Proof: By noting that under an attack the residual dynamics solution can be written as

$$y_r[k] = C_{pw}\Big[\sum_{h=0}^{k-1}(A_r)^{k-1-h}\left(\eta_{pw}[h]-K\xi_{pw}[h]+\phi_1[h]\right) +(A_r)^k x_r[0]\Big]+\phi_2[k]+\xi_{pw}[k]$$

the proof then easily follows from [11, Th. 3.1]. ■

Remark 1: While Th. 2 and 3 provide conditions for structural undetectability that relate to fundamental limitations in detectability faced by any detector, Th. 4 offers a sufficient condition for detectability (of structurally detectable anomalies) that depends on the actual state trajectory of the cascaded system and on the uncertainties values.

V. Isolation and Identification of Routing Attacks

The violation of the detection inequality (14) for a component i leads to labelling the corresponding sensor as compromised, and as such belonging to $\mathcal{V}_I$. Once detection is accomplished, the next step is to isolate whether an attack is of cyber or physical nature, and identify the edges that are incident to the sensors in $\mathcal{V}_I$, that is the edge set $\mathcal{E}_R$.

The proposed isolation and identification scheme relies on two adaptive estimators, one targeted at cyber and another at physical rerouting attacks. The estimators are able to learn on-line the non-zero entries of the matrix R and their estimation error can be used to isolate between the two kinds of attacks.

A. Cyber routing attacks

The estimator dynamics are defined as

$$\begin{cases} \hat{x}^C_{pw}[k+1] = A_{pw}\hat{x}^C_{pw}[k]+B_{pw}u[k]+ K^C_{\overline{\mathcal{V}}_I}\left(\tilde{y}_{pw}[k]-\hat{y}^C_{pw}[k]\right) \\ \hat{y}^C_{pw}[k] = \hat{R}^C[k]C_{pw}\hat{x}^C_{pw}[k] \\ \hat{y}^C_p[k] = [C_p\ 0]\hat{x}^C_{pw}[k], \end{cases} \quad (15)$$

where $\hat{R}^C[k] \in [0,1]^{n_y}$ is a real valued online adaptive estimate of the routing attack matrix and $K^C_{\overline{\mathcal{V}}_I}$ is a gain matrix which stabilizes $A^C_r \triangleq A_{pw} - K^C_{\overline{\mathcal{V}}_I}C_{pw}$ while using only non-compromised sensors belonging to the set $\overline{\mathcal{V}}_I \triangleq \mathcal{V} \setminus \mathcal{V}_I$. This design constraint is to prevent the routing attacks from poisoning the estimator and the routing matrix identification. In order to obtain a stabilizing gain matrix $K^C_{\overline{\mathcal{V}}_I}$, we require that Assumption 3 holds also when the rows of C_{pw} corresponding to compromised measurements are set to zero.

Remark 2: Note that the estimation error of $\hat{x}^C_{pw}[k]$ is decoupled from the estimation error of $\hat{R}^C[k]$, since the routing matrix estimation error is non-zero only for rows corresponding to $\mathcal{V}_I$, the set of sensors previously detected as compromised, which are multiplied by zero columns of the gain matrix $K^C_{\overline{\mathcal{V}}_I}$.

In order to explain the proposed approach to learning $\hat{R}^C$, we need to note that, for the generic i–th compromised measurement, it holds $\tilde{y}_{pw,(i)} = R^C_{(i)}y_{pw}$, where $R^C_{(i)}$ is the i–th row of R^C. Furthermore, we remember that the generic j–th non-rerouted watermarked measurement $y_{pw,(j)}[k]$ fulfills eq. (6), which can be rewritten as $y_{pw,(j)}[k] = \Phi_{A,(j)}[k]w^j_A + \Phi_{B,(j)}[k]w^j_B$, where

$$\begin{aligned} \Phi_{A,(j)} &\triangleq \left[-y_{pw,(j)}[k-1],\ \ldots,\ -y_{pw,(j)}[k-N]\right], \\ \Phi_{B,(j)} &\triangleq \left[y_{p,(j)}[k],\ \ldots,\ y_{p,(j)}[k-N]\right] \end{aligned} \quad (16)$$

are the j–th rows of two matrices Φ_A and Φ_B built with values of plant outputs and their watermarked counterparts over a moving time-window .

At this point, it is straightforward to see that, in the case where the j–th measurement is rerouted to the i–th (that is $R^C_{(i)}$ has a single 1 in the j–th position), we can write $\tilde{y}_{pw,(i)}[k] = \tilde{\Phi}_{A,(i)}[k]w^j_A + \Phi_{B,(j)}[k]w^j_B$, where $\tilde{\Phi}_{A,(i)} \triangleq \left[-\tilde{y}_{pw,(i)}[k-1], \ldots, -\tilde{y}_{pw,(i)}[k-N]\right]$. It then holds

$$\begin{aligned} &\tilde{y}_{pw}[k] = R^C\Phi^C, \\ &\Phi^C_{(i,j)} \triangleq \tilde{\Phi}_{A,(i)}[k]w^j_A + \Phi_{B,(j)}[k]w^j_B \end{aligned} \tag{17}$$

where Φ^C is the *cyber routing hypothesis matrix*, whose (i,j)–th element encodes the hypothesis that the j–th measurement has been cyber rerouted to the i–th one after the watermark has been applied.

However, eq. (17) cannot be directly used to estimate R^C. While in fact the matrix $\tilde{\Phi}_A$ can be computed from received measurements even under a routing attack, the matrix Φ_B cannot be computed as the unwatermarked, unrerouted plant outputs y_p are not directly accessible. The key point of the proposed approach is to compute instead the matrix $\hat{\Phi}^C_B$, whose rows are defined as $\hat{\Phi}^C_{B,(j)} \triangleq \left[\hat{y}^C_{p,(j)}[k], \ldots, \hat{y}^C_{p,(j)}[k-N]\right]$ so that it holds

$$\begin{aligned} &\tilde{y}_{pw}[k] = R^C\hat{\Phi}^C + R^C\Delta\hat{\Phi}^C, \\ &\hat{\Phi}^C_{(i,j)} \triangleq \tilde{\Phi}_{A,(i)}[k]w^j_A + \hat{\Phi}_{B,(j)}[k]w^j_B, \end{aligned} \tag{18}$$

with $\Delta\hat{\Phi}^C_{(i,j)} \triangleq [y^C_{p,(j)}[k] - \hat{y}^C_{p,(j)}[k], \ldots, y^C_{p,(j)}[k-N] - \hat{y}^C_{p,(j)}[k-N]]w^j_B$.

With this in mind, we employ the Recursive Least Squares (RLS) algorithm [13]–[15] to update online the estimate $\hat{R}^C_{(i)}$ for each i–th compromised measurement, as follows

$$\begin{aligned} P_i[k] &= P_i[k-1] - \frac{P_i[k-1]\hat{\Phi}^C_{(:,i)}[k]\hat{\Phi}^C_{(:,i)}[k]^\top P_i[k-1]}{1+\hat{\Phi}^C_{(:,i)}[k]^\top P_i[k-1]\hat{\Phi}^C_{(:,i)}[k]} \\ \varepsilon_i[k] &= \tilde{y}^C_{pw,(i)}[k] - \hat{R}^C_{(i)}[k-1]\hat{\Phi}^C_{(:,i)}[k] \\ \hat{R}^C_{(i)}[k] &= \mathcal{P}_{[0\,1]}\left\{\hat{R}^C_{(i)}[k-1] + P_i[k]\hat{\Phi}^C_{(:,i)}[k]\varepsilon_i[k]\right\} \end{aligned}$$

where $\hat{\Phi}^C_{(:,i)}$ is the i–th column of $\hat{\Phi}^C$, $P_i \in \mathbb{R}^{n_y \times n_y}$ is semidefinite positive and initialized as αI, with $\alpha > 0$ a design scalar parameter, and $\mathcal{P}_{[0\,1]}$ is a projection operator restricting $\hat{R}^C_{(i)}[k]$ to the interval $[0\ 1]$.

In order to isolate between a cyber and a physical routing, or another anomaly not envisaged by eq. (15), we will introduce the residual $y^C_r[k] \triangleq \tilde{y}_{pw}[k] - \hat{y}^C_{pw}[k]$. Similarly as the detection case, its dynamics can be written as

$$\begin{aligned} y^C_{rw}[k] = C_{pw}\Bigg[\sum_{h=0}^{k-1}(A^C_r)^{k-1-h}\Big(\eta_{pw}[h] - K^C_{\bar{\mathcal{V}}_I}\left(\xi^C_{pw}[h] + \right. \\ \left.\Delta R^C C_{pw}x_{pw}[h]\right)\Big) + (A^C_r)^k x_r[0]\Bigg] + \Delta R^C C_{pw}x_{pw}[k] + \xi^C_{pw}[k] \end{aligned}$$

where $\Delta R^C = R^C - \hat{R}^C$, and an isolation threshold for the i–th component can be easily computed as

$$\begin{aligned} \bar{y}^C_{rw,(i)}[k] \triangleq \alpha^{C_i}\Bigg[\sum_{h=0}^{k-1}\left(\delta^{C_i}\right)^{k-1-h}\Big(\bar{\eta}_{pw}[h] + \|K^C_{\bar{\mathcal{V}}_I}\| \\ \left(\bar{\xi}^C_{pw}[h] + \Delta\bar{R}^C\|C_{pw}\|\bar{x}_{pw}\right)\Big) + \left(\delta^{C_i}\right)^k\bar{x}_r[0]\Bigg] + \\ \Delta\bar{R}^C\|C_{pw}\|\bar{x}_{pw} + \bar{\xi}^C_{pw}[k] \end{aligned} \tag{19}$$

where $\Delta\bar{R}^C \triangleq n_y \geq \|\Delta R^C\|$ is computed using the Holder's inequality and taking advantage of the fact that elements of R^C and $\hat{R}^C$ are constrained inside the set $\{0,1\}$ and the interval $[0\,1]$, respectively. Furthermore $\bar{x}_{pw} \triangleq \max\|x_{pw}\|$ over $\mathscr{S}^{x_{pw}}$ and $\bar{\xi}^C_{pw}$ is an upper bound on $\|\xi^C_{pw}\|$ that can be computed from $\bar{\xi}_{pw}$ considering the worst case reroute. Similarly to the detection threshold, this threshold by construction is robust to uncertainties and identification errors, so that it will not be crossed in the case the detected anomaly is indeed a cyber rerouting attack. The residual crossing it, conversely, will be a sufficient condition for excluding the hypothesis that a cyber rerouting attack is present.

B. Physical routing attacks

Due to space constraints this case will be only briefly sketched. It can be addressed similarly to the cyber case, provided the *physical routing hypothesis matrix* Φ^P, defined as $\Phi^P_{(i,j)} \triangleq \tilde{\Phi}_{A,(i)}[k]w^i_A + \Phi_{B,(j)}[k]w^i_B$, is used in lieu of Φ^C. It should be noted that the order by which the indexes i and j appear in the definition of $\Phi^P_{(i,j)}$ is (i,i,j,i), which differs from the ordering (i,j,j,j) of $\Phi^C_{(i,j)}$. This encodes the fact that in the physical case the rerouting happens before the watermark is applied.

VI. Numerical Example

In this section the effectiveness of the proposed sensor watermarking approach to detection, isolation and identification of rerouting attacks will be illustrated through a numerical example. The plant under attack is modeled as a discrete-time LTI system with three states, two inputs and three outputs, and can be described in state-space through the matrices

$$A = \begin{bmatrix} 0.9 & 0 & 0.1 \\ 0 & 0.9 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, C = I_3,$$

where I_3 is a 3×3 identity matrix, and the time-step has been chosen equal to 0.01 s. The plant is controlled in open-loop, the two inputs being square wave signals with amplitude equal to 1 and 0.8, and frequency 0.5 and 0.7 Hz, respectively. The model and measurement uncertainties has been implemented through two random variables uniformly distributed in the intervals $[-0.15\ 0.15]$ and $[-0.015\ 0.015]$. The uncertainty bounds occurring in the threshold definitions (14) and (19) were computed accordingly.

Watermark generators employed a bank of 4–th order IIR filters, whose coefficients have been set equal to $w^{1\top}_A = [1,\ 0.5,\ 0,\ 0]$, $w^{2\top}_A = [1,\ 0.5,\ -0.5,\ -0.5]$ and $w^{3\top}_A = [1,\ 0.5,\ 0.5,\ 0.5]$ and $w^{1\top}_B =$

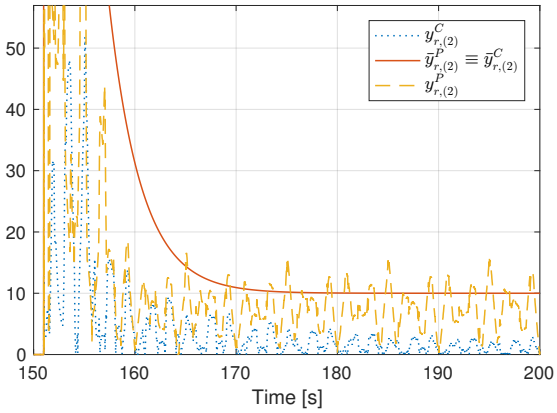

Fig. 2. Residual and thresholds computed by the cyber and the physical isolation and identification filters after a cyber routing attack, initiated at $T_0 = 150$ s.

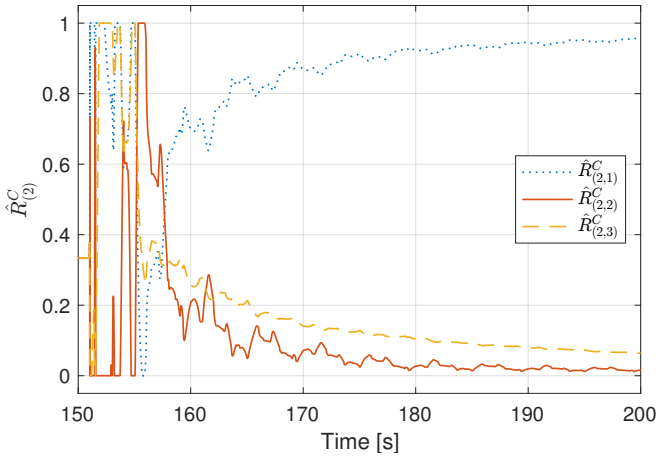

Fig. 3. Routing parameters estimate computed by the cyber isolation and identification filter after a cyber routing attack, initiated at $T_0 = 150$ s.

$[1.01, 0.51, 0.01, 0.01]$, $w_B^{2\top} = [1.01, 0.49, -0.5, -0.5]$ and $w_B^{3\top} = [1.01, 0.51, 0.49, 0.49]$.

At time $T_0 = 150$ s it is assumed that a cyber rerouting attack is carried on, leading to the sensor output 1 being rerouted to measurement 2, as described by

$$R^C = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

As can be seen in Fig. 2, the robust isolation threshold $\bar{y}^C_r$ of the cyber estimator is not crossed, while the one of the physical estimator is crossed at about 165 s, thus allowing to isolate correctly the attack as a cyber routing. Furthermore, the components of the cyber routing parameter estimate $\hat{R}^C_{(2)}$ for the second measurement converge towards the correct value $[1\,0\,0]$, thus empirically verifying the proposed approach (see Fig. 3).

VII. Conclusions

A multiplicative sensor watermarking scheme was proposed in this work, where each sensor's output is separately fed to a SISO watermark generator. As opposed to previously proposed additive watermarking schemes, no additional burden is put on physical actuators; moreover, no communication between multiple sensors is required. The benefits of the proposed scheme were analyzed for two attack scenarios: the physical sensor re-routing attack and the cyber measurement re-routing attack. For each scenario, detectability and isolability properties with and without the proposed watermarking scheme have been derived. In particular, it was shown how to design the watermarking scheme to detect both sensor attack scenarios, and identify the sensors involved in the re-routing attacks. Future work will include the extension of such scheme to other classes of attacks, as well as the ability to handle multiple concurrent attacks.

References

[1] C. S. S. P. National Cyber Security Division. (2009, Oct.) Recommended practice: Improving industrial control systems cybersecurity with defense-in-depth strategies. U.S. Department of Homeland Security. Available online: https://ics-cert.us-cert.gov/sites/default/files/recommended_practices/Defense_in_Depth_Oct09.pdf.

[2] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *First International Workshop on Cyber-Physical Systems*, June 2008.

[3] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, no. 1, pp. 135–148, 2015.

[4] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.

[5] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.

[6] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding sensor outputs for injection attacks detection," in *2014 IEEE 53rd Annual Conference on Decision and Control (CDC)*, Dec 2014, pp. 5776–5781.

[7] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *Control Systems, IEEE*, vol. 35, no. 1, pp. 93–109, Feb 2015.

[8] A. Teixeira, K. Paridari, H. Sandberg, and K. H. Johansson, "Voltage control for interconnected microgrids under adversarial actions," in *2015 IEEE 20th Conference on Emerging Technologies Factory Automation (ETFA)*, Sept. 2015, pp. 1–8.

[9] P. Bassia, I. Pitas, and N. Nikolaidis, "Robust audio watermarking in the time domain," *IEEE Transactions on Multimedia*, vol. 3, no. 2, pp. 232–241, Jun 2001.

[10] S. Ding, *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Berlin, Heidelberg: Springer-Verlag, 2008.

[11] R. M. Ferrari, T. Parisini, and M. Polycarpou, "A robust fault detection and isolation scheme for a class of uncertain input-output discrete-time nonlinear systems," in *American Control Conference, 2008*, June 2008, pp. 2804–2809.

[12] D. A. Dowler, "Bounding the norm of matrix powers," Master's thesis, Brigham Young University-Provo, 2013.

[13] T. Söderström, L. Ljung, and I. Gustavsson, "A theoretical analysis of recursive identification methods," *Automatica*, vol. 14, no. 3, pp. 231 – 244, 1978.

[14] F. Ding, Y. Shi, and T. Chen, "Performance analysis of estimation algorithms of nonstationary arma processes," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 1041–1053, March 2006.

[15] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2008.