2019 American Control Conference (ACC)
Philadelphia, PA, USA, July 10-12, 2019

Kalman Filter Based Secure State Estimation and Individual Attacked Sensor Detection in Cyber-Physical Systems

Mohammad Hossein Basiri, John G. Thistle, John W. Simpson-Porco, and Sebastian Fischmeister

Abstract— **In this paper we propose two real-time attack detection and secure state estimation algorithms, namely Rolling Window Detector (RWD) and Novel Residual Detector (NRD). These algorithms are basically developed based on Kalman state estimation. In the former, we present a statistical testing approach which is handled over a finite time horizon T to detect individual attacked sensors. The latter extends the χ^2–detector to be able to detect individual compromised sensors. Both methods then will be employed together with a modified version of Kalman filter to perform a secure state estimation with a relatively low estimation error. Efficiency of the algorithms will be assessed in both unstealthy and stealthy scenarios. Productivity of the methods will be underlined in the stealthy case, which is of much more significance among cyber-security challenges. Simulation results on an IEEE 14-bus power grid test system along with a comprehensive comparison between the performance of RWD and NRD with a recently introduced tool, which is the only other method that tries to detect individual attacked sensors, proves the effectiveness of the algorithms.**

I. Introduction

Development of control systems in terms of their reliability and trustworthiness has become a major concern within the control community. In primal systems, exchanging sensor measurements and control inputs throughout the system was largely handled over wired and well-maintained infrastructures. Cyber-Physical Systems (CPS), provide the opportunity not only to have large scale wide spread control systems but also to take advantage of wireless data communication approaches. When CPS comes into view, control, computation and networking bind together to form a suitable infrastructure for control and systems purposes. Although, benefiting from communication channels has enhanced the solicitation for CPSs, the vulnerability of such systems to malicious external attackers has attracted a great amount of attention [1] and made the problem of secure state estimation to a highly attractive problem in both control and communication systems. Hence, Confrontation of CPSs with an intelligent intruder who has access to an arbitrary subset of the sensors requires fruitful techniques to overcome any possible performance degradations. Several well-known attacks on CPSs include Stuxnet on a Supervisory Control and Data Acquisition (SCADA) system [2], attacks on the wireless network channels in smart power grid systems [3], and compromising Anti-lock Braking System (ABS) sensors of a vehicle [4].

Research works such as [5] proved that in deterministic systems, detectability of attacks can be equivalently interpreted by a control-theoretic notion of *invariant dynamics*. More rigorously, in a deterministic system, an attacker is undetectable if and only if it stimulates only the *zero dynamics* of the system [6]. In this paper we focus on stochastic systems contaminated by both process and measurement noise.

One of the most common methodologies for detecting attacks is Kalman filter-based attack detection [7]–[11]. In [7], the authors use a hypothetical testing based on the residual vector to detect the attacks. Their method is basically established on the Cumulative Summation (CUSUM) algorithm. In [8], the attacker's action is formulated in terms of an optimization problem maximizing the degradation on the system while not exceeding certain threshold to remain undetected. The authors use [12] to convert their first problem to a quadratically constrained quadratic problem which is proved in [12] to have an analytical solution; however, this method requires computation of a certain quantity for each combination of the sensors to decide which subset of the sensors need be secured. Hence, this approach is largely an off-line method, whereas real-time attack detection is of more interest in real-world applications. Authors of [9] study the well-known χ^2–detector based on Kalman filtering but for a particular case of sinusoidal signals in power grids. An intelligent malicious intruder often prefers to conceal himself from the controller/detector while performing his attack action over the system. The level of stealthiness of an attacker has been formulated in several recent studies. Particularly, the notion of ϵ–stealthiness is generalized in [10] based on an information-theoretic concept. We will use a milder version of the stealthiness notion in this paper since our objective is developing effective algorithms for unstealthy and stealthy cases rather than largely to focus on the information-theoretic details of this notion.

Explicitly, the contributions in this paper are as follows. We introduce two attack detection and secure state estimation algorithms, namely Rolling Window Detector (RWD) and Novel Residual Detector (NRD).[1] These algorithms are based on Kalman state estimation. In the former, we present a statistical testing approach which is handled over a finite time horizon T to detect individual attacked sensors. The latter extends the χ^2–detector to be able to detect individual compromised sensors. Both methods then will be employed together with a modified version of Kalman filter to perform a secure state estimation with a relatively low estimation error. The efficiency of the algorithms will be assessed in

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada, {mh.basiri, jthistle, jwsimpson, sfischme}@uwaterloo.ca

[1] In the rest of the paper, RWD and NRD denote the proposed algorithms.

both unstealthy and stealthy scenarios. Specifically, in the stealthy case, that is of great significance among cyber-security challenges, we see that our methods can effectively detect almost all of the attacked sensors while a powerful recently introduced tool in the literature, namely Imhotep-SMT [13] can detect *none* of the compromised sensors. The authors in [13] proposed a Satisfiability Modulo Theory (SMT) approach which contains solving a combinatorial problem with a relatively high computational complexity. To the best of our knowledge, that method is the only other that attempts to identify specific, individual attacked sensors. As such, this paper focuses on a comparison with that specific method. Although this method has been proved to be able to detect individual attacked sensors, it turns out that Imhotep-SMT fails to perform well in the stealthy case. In addition, due to the high computational complexity of the combinatorial aspect of the ℓ_0 optimization and that of SMT solver, these estimators have delays which is not favorable in real-time applications.

The rest of this paper is organized as follows. Sec. II describes the model of the system under study in normal and attacked cases along with the attack modeling. Sec. III is devoted to some preliminary concepts for attack detection via Kalman filtering, and to the χ^2–detector which will be exploited afterwards. In Sec. IV we explain our developed real-time algorithms. The algorithms will be described in three cases: the stealthy, the unstealthy, and the very unstealthy cases. Simulation results on an IEEE 14-bus power grid test system will be presented in Sec. V. The attack detection and the secure estimation results will be compared to the recently introduced estimator and several detection rate analyses will be conducted. Finally, Sec. VI concludes the paper.

II. System and Attack Modeling

We model the system under study as a state-space dynamical model driven by process and measurement noise as follows,

$$\mathcal{P}: \begin{cases} x_{k+1} = Ax_k + Bu_k + \nu_k, \\ y_k = Cx_k + w_k, \end{cases} \tag{1}$$

where, $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$ are the state(s), input(s) and output(s) of the system, respectively. ν and w are the process and measurement noise, assumed to have zero-mean Gaussian distributions with noise covariance matrices Q and R, respectively, i.e., $\nu \sim \mathcal{N}(0, Q)$ and $w \sim \mathcal{N}(0, R)$. The initial state x_0 is assumed to be a random Gaussian variable $x_0 \sim \mathcal{N}(0, \Sigma)$ which is independent of process and measurement noise.

In this paper we focus on the random attack as a false data injection into the sensor measurements. The attacked system can be represented as follows,

$$\mathcal{P}_a: \begin{cases} x^a_{k+1} = Ax^a_k + Bu_k + \nu_k, \\ y^a_k = Cx^a_k + w_k + Da_k, \end{cases} \tag{2}$$

where x^a_k and y^a_k denote the state and the output of the system under attack, respectively.[2] a_k is the attack vector consisting of arbitrary signals injected by the intruder to corrupt the sensors. Sensors accessed by the intruder are determined by D which is a diagonal matrix. Specifically, sensor j is under attack if $D_{jj} = 1$, otherwise $D_{jj} = 0$. We assume that the pair (A, C) is observable which is a reasonable assumption since otherwise the state of the system is not reconstructable even in the absence of attacks. In this paper we focus on the case in which the attack vector is represented by zero-mean Independent and Identically Distributed (i.i.d.) Gaussian components. With the model (2) in mind, two possible important cases might occur, namely the unstealthy case and the stealthy case. In the former, having striven to fool the sensors, the attacker just attempts to corrupt the measurements by injecting false random sequences to the output sensors and achieves a performance disruption in the system without trying to remain undetected. In the latter, which is of more concern, the intruder aims to inject false data to the measurements while subsisting undetected. Getting to that goal, the attacker injects false noise components to the measurements with a relatively comparable standard deviation to that of the measurement noise components. Particularly, as the standard deviation[3] of the attack gets closer to that of the measurement noise, the attack is regarded as a stealthier one. Mathematically, the random attack injected by the intruder in the measurements, the process and the measurement noise can be expressed as follows,

$$\nu_k = \sigma_\nu X_\nu, \quad w_k = \sigma_w Y_w, \quad a_k = \sigma_a Z_a, \tag{3}$$

where σ_ν, σ_w, and σ_a represent the process noise power, the measurement noise power and the attack power, respectively. For the i.i.d. noise and random attack model assumed in this paper, $X_\nu, Y_w, Z_a \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Hence, $\nu \sim \mathcal{N}(0, Q = diag(\sigma^2_\nu))$ and $w \sim \mathcal{N}(0, R = diag(\sigma^2_w))$ where $diag(\cdot)$ is a diagonal matrix (with appropriate dimension) with the argument on its diagonal elements and zero elsewhere. The stealthy case studied in the simulation results of the paper is modeled as $\sigma_a \to \sigma_w$.[4] In this respect, the intelligent attacker tries to deteriorate the system performance while being undetected by hiding himself in the measurement noise. For a deeper definition and analysis of the stealthiness notion, the reader is referred to [14] and [10]. We will investigate these two cases with the simple aforementioned definition in subsequent sections through various simulations.

III. Attack Detection Preliminaries

In this section we review some basic and preliminary concepts which we will modify/generalize and upon which we establish our algorithms in the subsequent section.

A. Kalman Filtering

Kalman filtering is a well-known procedure to provide the optimal estimate of the states of a dynamical system composed of a contaminated Gaussian process and measurements

[2] In the sequel, we drop the superscript a for simplicity in the notations.

[3] In the rest of the paper, we abuse the terminology of "noise power" and "attack power" to denote the standard deviation of w_k and a_k, respectively.

[4] We call the situation $\sigma_a = \sigma_w$ as the *strictly* stealthy case.

driven by Gaussian noise. In particular, Kalman recursive estimation equations can be formulated as follows,

- **Measurement Update**

$$K_k = P_{k|k-1}C^T\left(CP_{k|k-1}C^T + R\right)^{-1}, \quad (4)$$
$$P_{k|k} = P_{k|k-1} - K_kCP_{k|k-1}, \quad (5)$$
$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k\left(y_k - C\hat{x}_{k|k-1}\right), \quad (6)$$

- **Time Update**

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k, \quad (7)$$
$$P_{k+1|k} = AP_{k|k}A^T + Q, \quad (8)$$

where $\hat{x}_{k+1|k}$, $P_{k+1|k}$ and K_k denote the estimated state of the system at time $k+1$ using the information up to time k, the error covariance matrix predicted at time $k+1$ using the information up to time k and Kalman gain matrix at time k, respectively. Since the system is assumed to be observable and typically it has been running for a long period of time, the Kalman filter converges in a few steps and the system can be assumed to reach the steady state before the attacks begin. Consequently, before an attack starts, $P \triangleq \lim_{k\to\infty} P_{k|k-1}$ and $K \triangleq \lim_{k\to\infty} K_k = PC^T(CPC^T + R)^{-1}$.

B. Conventional χ^2–Detector

This detector has been widely used in both fault detection problems and recently in security in cyber-physical context. In fact, this method is originally highly applicable in fault detection problems but due to its simplicity and effectiveness, researchers have recently started using it in cyber-security problems as well.

Particularly, the residual vector is calculated based on the estimation resulted from the Kalman filter,

$$r_k = y_k - \hat{y}_k = y_k - C\hat{x}_k{}^{[5]}, \quad (9)$$

which has a white Gaussian distribution in the absence of an attack. The "power" of the residual vector is then calculated at each time instant k as follows,

$$g_k = r_k^T\Sigma_{r,k}^{-1}r_k, \quad (10)$$

where $\Sigma_{r,k}$ is the covariance matrix of the residual vector,

$$\Sigma_{r,k} = CP_{k|k}C^T + R. \quad (11)$$

The χ^2–detector then contains a scalar statistical testing which compares g_k with a pre-specified threshold. The threshold can be determined from the χ^2 table corresponding to the desired confidence interval and the degree of freedom. If,

$$g_k > \textit{threshold}, \quad (12)$$

then an alarm is triggered indicating that an attack is underway.

In this paper we propose two different algorithms comprising modified Kalman filtering. In the first one, namely the RWD, we employ a modified version of Kalman filter along with a simple developed statistical testing to reveal individual random attacks if any is underway. In the latter, namely the NRD, we utilize the modified Kalman filter along with a modified χ^2–detector in order to detect individual attacked sensors. We show that our developed algorithms are also able to detect attacks in real-time in the stealthy attack case which is of more interest to the cyber-security community. These two algorithms are then invoked to perform a secure state estimation with a relatively low estimation error.

[5]For notational convenience, we sometimes denote $\hat{x}_{k|k}$ and $\hat{y}_{k|k}$ by $\hat{x}_k$ and $\hat{y}_k$, respectively in the rest of the paper.

IV. Proposed Attack Detection Algorithms

As was described before, cyber-physical systems are prone to be attacked by different adversaries who compromise the sensor readings to deceive the estimation procedure. More sophisticated attackers have access to a subset of the sensors and are able to inject false data to those measurements causing deterioration in the system performance.

In this section, we will explain our two novel developed algorithms which firstly provide the knowledge of whether or not any attack is underway. Subsequently, the proposed algorithms try to detect individual sensors under attack. The algorithms are based on Kalman filtering and can be regarded as modified Kalman filter-based attack detection. Then this allows us to securely estimate the states of the system with a relatively low estimation error. A key advantage of the proposed methods is their functionality in the particularly interesting stealthy case, together with their simplicity and low computational complexity mainly compared to the recently proposed method, Imhotep-SMT [13]. This advantage is highly notable to be exploited in the real-time applications. Besides, Imhotep-SMT uses only a finite number of measurements to perform the secure estimation, namely n last measurements where n is the order of the dynamical system, while our proposed methods are based on Kalman filtering which incorporates *all* previous information from the beginning up to the current point. This essentially results in much smoother estimates of the states of the system which empowers the algorithms to detect the attacks more efficiently.

A. RWD Attack Detector

1) RWD Overview: Here we explain our first developed algorithm which employs a modified version of the Kalman filter to effectively detect the individual sensors under attack. One of the key advantages of this algorithm is its simplicity and flexibility which provides the ground for the user to specify his own window length depending on a trade-off between the speed and accuracy of the detection.

The main idea behind RWD is to compare the cumulative sum of the matrices in the form $\bar{P}_{k|k} = \mathbb{E}\left[\left(y_k - \hat{y}_k\right)\left(y_k - \hat{y}_k\right)^T\right]$ with the corresponding $\Sigma_{r,k}$ matrix predicted by the Kalman filter over a rolling window with finite length T. Intuitively, as the attacker injects an arbitrary random sequence to the sensor measurements, this will corrupt the estimated states (followed by the estimated outputs) at the attack time. As the Kalman filter is a recursive algorithm, it takes time for the filter to reconstruct the

estimates using the error covariance matrix $P_{k|k}$. The error covariance matrix $P_{k|k}$ represents the following expected value representing how close the estimated states are to the actual states in the presence of the attack and is predicted over the whole time horizon on which the Kalman recursions (4)-(8) perform,

$$P_{k+1|k+1} = \mathbb{E}\left[(x_{k+1} - \hat{x}_{k+1})(x_{k+1} - \hat{x}_{k+1})^T\right]. \quad (13)$$

Having projected the error covariance matrix onto the output subspace ($\mathbb{R}^p$), this results in the quantity called S in the following. Mathematically, RWD adopts a limited size window over which the following $\hat{\Sigma}_k$ is calculated as the time passes. Then, the following quantity, S, is evaluated in each time instant k and is compared with a threshold matrix H,

$$S(T, k_0) = \underbrace{\frac{1}{T}\sum_{k=k_0}^{k_0+T-1}(y_k - \hat{y}_k)(y_k - \hat{y}_k)^T}_{\hat{\Sigma}_k} - \underbrace{\left(CP_{k|k}C^T + R\right)}_{\Sigma_{r,k}} > H. \quad (14)$$

Although this is a heuristic detection procedure, it is a drastically easy evaluation and is of much lower computational complexity compared to several recently introduced detectors, e.g., [13], [15].

2) RWD Algorithm Explanation: At first (and at each time instant k), RWD makes a backup of the estimated state $\hat{x}_k$ (for the first time instant, this will be the initial estimate which is chosen to be $\hat{x}_{0|-1} = \mathbf{0}_{n\times 1}$). Then it applies the standard Kalman filter to obtain the state estimate $\hat{x}_k$. For each sensor output, the RWD algorithm starts to calculate the cumulative sum of the form $\hat{\Sigma}_k$ defined in (14) up to time T at each time instant $k = k_0$. At the same time, $\Sigma_{r,k}$ is obtained by the Kalman estimation. Subsequently, RWD compares the diagonal elements of $S = \hat{\Sigma}_k - \Sigma_{r,k}$ corresponding to each sensor output j with the corresponding element of the threshold matrix H. If $\left(\hat{\Sigma}_k - \Sigma_{r,k}\right)_{jj} > H_{jj}$, then RWD concludes that sensor j is under attack. Thereupon, RWD modifies the measurement vector y_k by replacing the j^{th} element of the measurement vector y_k with the j^{th} element of the estimated measurement vector $\hat{y}_{k-1}$. Subsequently, the Kalman filter re-estimates $\hat{x}_k$ using the backup variable made at the beginning of the algorithm and the modified measurement vector y_k. With this new modified estimation, RWD modifies the previous state estimate and the output estimate. This procedure continues until all the sensor outputs are met. Finally, as in the standard Kalman filter, the state estimate and error covariance matrix are projected to the next time instant via time update equations (7) and (8). The RWD algorithm with the above procedure runs over a period of time to provide the attack detection and a secure state estimation with a relatively low estimation error. RWD algorithm is summarized in Algorithm 1.

Remark IV.1. It is worth noting that the selection of the window length T entails a trade-off between the speed of detection and its accuracy. The longer the window length T is, the more accurate attack detection would be since it incorporates more information. In fact, a larger T will cover a longer period of time in which RWD has more time to do the statistical comparison and reveals the compromised sensors with more certainty. On the other hand, it takes a longer time, which slows the detection procedure. Hence, the selection of window length depends mainly on the application. If a more accurate detection is desired and the resulting delay in the detection does not matter too much, a larger T is suitable, otherwise a shorter T need be chosen. Furthermore, the speed of the dynamics of the plant has to be taken into account while choosing T. In essence, the slower the dynamical system is, the larger the value of T needed.

Algorithm 1 RWD Attack Detection

Initialize:
$x_0 \sim \mathcal{N}(0, \Sigma)$, $\hat{x}_{0|-1} = \mathbf{0}$, $P_{0|-1} = \Sigma$, Q, R, T, H.
while $k \leq k_{\text{final}}$ **do**
  Make a backup of $\hat{x}_{k|k-1}$ in $\hat{x}_{k,\text{backup}}$.
  for Each sensor j **do**
    Apply standard Kalman filter and estimate $\hat{x}_{k|k}$
    calculate: $\hat{\Sigma}_k = \frac{1}{T}\sum_{k_0}^{k_0+T-1}(y_k - \hat{y}_k)(y_k - \hat{y}_k)^T$
    calculate: $\Sigma_{r,k} = CP_{k|k}C^T + R$
    if $\left(\hat{\Sigma}_k - \Sigma_{r,k}\right)_{jj} > H_{jj}$ **then**
      Modify j^{th} component of vector y_k by replacing it with the j^{th} component of vector $\hat{y}_{k-1}$.
      Re-estimate $\hat{x}_{k|k}$ via (6) using $\hat{x}_{k,\text{backup}}$ and the new y_k.
      Re-estimate $\hat{y}_k$ with the new $\hat{x}_{k|k}$.
    end if
  end for
  Predict $\hat{x}_{k+1|k}$ and $P_{k+1|k}$ via (7) and (8).
end while

Remark IV.2. With the above trade-off in mind, selection of the window length has to be short enough that $\bar{P}_{k_0+T-1} \simeq \bar{P}_{k_0}$ (where $\bar{P}_\bullet$ denotes $\bar{P}_{\bullet|\bullet}$). On the other hand, it has to be long enough for $\hat{\Sigma}_k$ to be statistically meaningful depending on the entire time horizon for the detector. This remark together with the previous one suggest a lower and an upper bound on T. Specifically, the lower bound is basically entailed by the statistical significance of S and $\bar{P}$ while the upper bound is determined based on the user-specified response time of the detection scheme to the attacks. Also, the speed of the dynamics of the system has effect on both the lower and the upper bounds.

B. NRD Attack Detector

1) NRD Overview: In the NRD algorithm we exploit the modified Kalman filtering as in RWD. In addition, we make use of an extended version of the χ^2–detector for individual sensor attack detection.

2) NRD Algorithm Explanation: NRD exploits both the conventional and an extended version of the χ^2–detector to mitigate the attack effects and perform a secure detection and

Algorithm 2 NRD ATTACK DETECTION

1: **Initialize:**
$x_0 \sim \mathcal{N}(0, \Sigma)$, $\hat{x}_{0|-1} = \mathbf{0}$, $P_{0|-1} = \Sigma$, Q, R.
2: **while** $k \leq k_{\text{final}}$ **do**
3: Make a backup of $\hat{x}_{k|k-1}$ in $\hat{x}_{k,\text{backup}}$.
4: Apply standard Kalman filter and estimate $\hat{x}_{k|k}$
5: **calculate:** $g_k = r_k^T \Sigma_{r,k}^{-1} r_k$
6: **if** $g_k = r_k^T \Sigma_{r,k}^{-1} r_k > threshold_1$ **then**
7: **for** Each sensor j **do**
8: **calculate:** $g_{j,k} = r_k^2 / (\Sigma_{r,k})_{jj}$
9: **if** $g_{j,k} = r_k^2 / (\Sigma_{r,k})_{jj} > threshold_2$ **then**
10: Modify j^{th} component of vector y_k by replacing it with the j^{th} component of vector $\hat{y}_{k-1}$.
11: Re-estimate $\hat{x}_{k|k}$ via (6) using $\hat{x}_{k,\text{backup}}$ and the new y_k.
12: Re-estimate $\hat{y}_k$ with the new $\hat{x}_{k|k}$.
13: **end if**
14: **end for**
15: **end if**
16: Predict $\hat{x}_{k+1|k}$ and $P_{k+1|k}$ via (7) and (8).
17: **end while**

state estimation. first of all, the residual vector defined by (9) is calculated at each time instant k and its power expressed by (10) is compared to the threshold obtained from the χ^2 distribution table (denoted by threshold$_1$ in the Algorithm 2). This discloses the occurrence of an attack underway if there exists any. The next step is to detect the individual attacked sensors. To aim at this objective, we make use of a scalar statistical testing which exploits the power of the residual in a scalar manner. Specifically, we focus on the following quantity corresponding to the j^{th} sensor output,

$$g_{j,k} = r_k^2 / (\Sigma_{r,k})_{jj} . \tag{15}$$

Evaluation and comparison of (15) with the corresponding threshold can effectively help us detect individual attacked sensors. This is followed by a similar approach described in Algorithm 1. In particular, after having individual sensors detected, a crucial step is to modify the previous corrupted estimation. NRD is outlined in Algorithm 2.

Remark IV.3. It is worth mentioning that choosing both the threshold matrix H (in RWD) and the threshold to which (15) has to be compared, essentially implies a trade-off between the false alarm rate and the sensitivity of the system against attacks. Particularly, a higher threshold results in a lower false alarm rate while in a less sensitive system to the attacks, i.e., a lower true positive rate.

Remark IV.4. In RWD and NRD, we have not used the last measurement in lines 9 and 10 of the algorithms, respectively. This is because of the possible existence of False Negatives (FNs). In fact, there might have been an undetected attack on the j^{th} sensor output of the previous time instant. Thus, the algorithms safely assure a correct modification by using the last estimated value for the j^{th} sensor output.

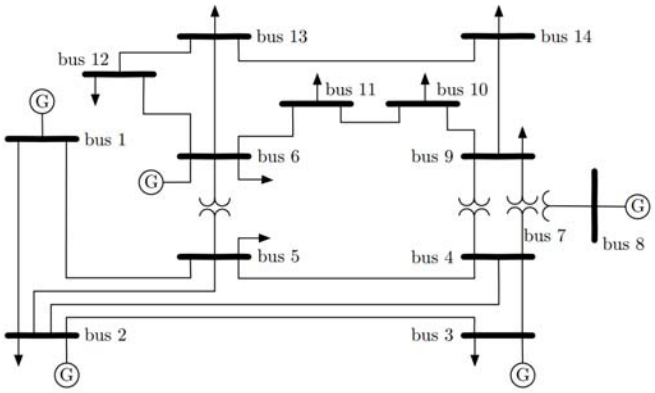
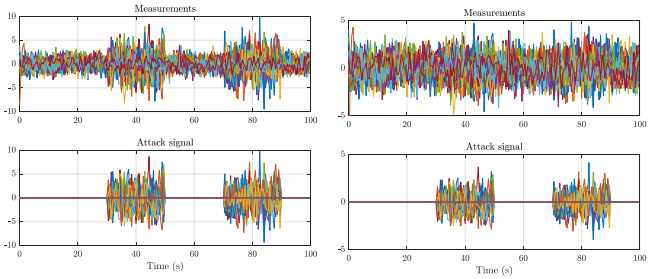


Fig. 1: IEEE 14-bus power grid test system [16]

V. SIMULATION RESULTS

IEEE 14-bus power grid system: In this section we verify the effectiveness of RWD and NRD on an IEEE 14-bus power grid system shown in Fig. 1. This system has $n = 10$ states equipped with $p = 35$ output sensors and the input is assumed to be zero. The system is composed of 5 synchronous generators and 14 buses. States of the system denote rotor angles δ_i and the frequencies $\omega_i = d\delta_i/dt$ of the generators. Similar to [16] we assume 14 sensors are measuring the real power injected at the 14 buses and 20 sensors are deployed to measure real power flows along every branch. One sensor is also measuring the rotor angle at generator number 1. With some simplifying assumptions, the evolution of the states of this system can be captured with a linear discrete-time state-space model of the form (1).[6] It is also assumed that a random subset of sensors are attacked by an intruder by injecting false data in the form of Gaussian distributed random variables to those sensors during time intervals $30 \leq t \leq 50$ and $70 \leq t \leq 90$. We investigate three different scenarios, namely the stealthy, the unstealthy, and the very unstealthy scenarios (see Fig. 2 for unstealthy and stealthy cases). This subset has been chosen as sensors 2, 8, 15, 16, 19, 21, 22, 24, 25, 26, 27, 29, 30, 31 and is fixed during the simulations. It is shown in [6] that the rotor angle sensor must be secured for the system to remain stable. As a test bench, we compare our results with the recently proposed powerful estimator, namely Imhotep-SMT [13]. The authors in [13] proposed a Satisfiability Modulo Theory (SMT) approach which contains solving a combinatorial problem with a relatively high computational complexity. Opposed to [13], both RWD and NRD provide a much faster detection and estimation solution as they are using modified/extended version of conventional methods such as Kalman filtering and χ^2–detector. It is remarkable that the main reason for the comparison of the proposed methods with Imhotep-SMT in the subsequent simulations is that Imhotep-SMT turns out to be one of the most prominent recent methods which is proved to be able to detect individual attacked sensors followed by secure state estimation. It is also important to keep in mind that although the proposed methods are much faster, they

[6]The interested reader is referred to [6], [17] for the details of the model derivation.



(a) Measurement and attack signal for the unstealthy case (b) Measurement and attack signal for the stealthy case

Fig. 2: Measurement and attack signal for the unstealthy and stealthy cases

are not always as precise as Imhotep-SMT. This will be discussed in more details through False Positive (FP) and False Negative (FN) notions in the simulations.

A. The Unstealthy Case

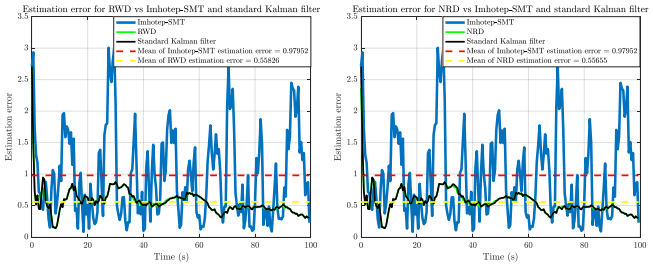
For an unstealthy attacker we assume that the measurement noise power and the attack power equal $\sigma_w = 1.1$ and $\sigma_a = 2.6$, respectively (Fig. 2).

1) RWD Algorithm: To simulate the RWD algorithm the threshold matrix H has been chosen with diagonal elements equal to 10. Running RWD with $T = 10$ along with the Imhotep-SMT and also a standard Kalman filter not equipped with a detection scheme on the system under attack results in the estimation error shown in Fig. 3a. The estimation error is calculated as the Euclidean norm of the difference between the vector of actual states and its estimation,

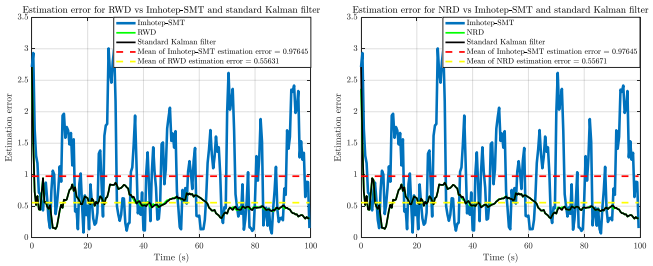
$$\text{Estimation error} = \|x_k - \hat{x}_k\|_2. \tag{16}$$

As one can easily see from Fig. 3a, RWD gives a more accurate estimation. In this case, the detected sensors with RWD and Imhotep-SMT are sensors 2, 8, 15, 16, 19, 21, 22, 25, 26, 27, 29, 30, 31 and sensors 2, 8, 15, 19, 21, 22, 24, 25, 26, 29, 30, 31, respectively. It can be seen that RWD has one FN (sensor 24) with no FPs and Imhotep-SMT has two FNs (sensors 16 and 27) with no FPs. It can be seen that RWD could effectively improve the precision of the estimation. The quantity calculated via (14) for the individual attacked sensor detection is shown in Fig. 4 for several compromised and uncompromised sensors. As seen form this figure, the algorithm could detect the attacks immediately after the attacks occurred. We also note that one may ask why he can see no evident sign of the attacks performed on the system in Fig. 3. This will be answered with complete details in Sec. V-C.

2) NRD Algorithm: The test system for NRD algorithm is the same as for RWD with the same attacked sensors and the same attack time intervals. In this case the confidence interval has been chosen as 95%, i.e., the error rate is chosen to be less than 5%. The detected sensors with NRD are sensors 2, 8, 15, 16, 19, 21, 22, 24, 25, 26, 27, 29, 30, 31 with no FNs and no FPs. Detected sensors by Imhotep-SMT are the same as before. Fig. 3b depicts the estimation error of NRD and Imhotep-SMT. It is seen from the figure that in this case NRD can improve the state estimation even better than RWD. The scalar statistic calculated via (15) for the individual attacked sensor detection is shown in Fig. 4 for several compromised and uncompromised sensors. As seen form this figure, the algorithm could detect the attacks immediately after the attacks occurred.

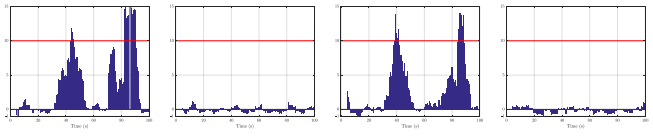


(a) Estimation error for RWD vs Imhotep-SMT on the attacked IEEE 14-bus power grid system for the unstealthy case (b) Estimation error for NRD vs Imhotep-SMT on the attacked IEEE 14-bus power grid system for the unstealthy case

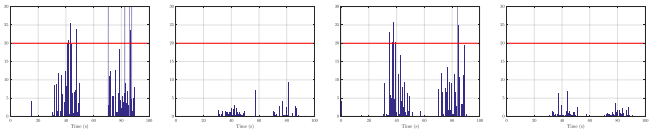


(c) Estimation error for RWD vs Imhotep-SMT on the attacked IEEE 14-bus power grid system for the stealthy case (d) Estimation error for NRD vs Imhotep-SMT on the attacked IEEE 14-bus power grid system for the stealthy case

Fig. 3: Estimation error for RWD and NRD vs Imhotep-SMT on the attacked IEEE 14-bus power grid system for the unstealthy and stealthy cases



(a) Compromised sensor 8 with RWD (b) Uncompromised sensor 9 with RWD (c) Compromised sensor 15 with RWD (d) Uncompromised sensor 34 with RWD

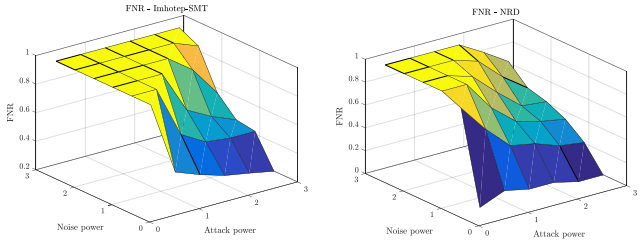


(e) Compromised sensor 8 with NRD (f) Uncompromised sensor 9 with NRD (g) Compromised sensor 15 with NRD (h) Uncompromised sensor 34 with NRD

Fig. 4: The quantity and the individual residuals calculated via (14) and (15) for the individual attacked sensor detection for several compromised and uncompromised sensors for RWD and NRD algorithms for the unstealthy case

B. The Stealthy Case

To model a stealthy attacker we assume the measurement noise power and the attack power are equal. Let us choose



(a) FNR for Imhotep-SMT (b) FNR for NRD

Fig. 5: FNR for Imhotep-SMT and NRD

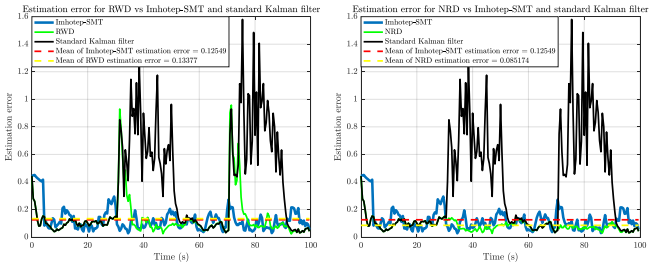
$\sigma_w = \sigma_a = 1.1$, (a strictly stealthy attacker) (Fig. 2). In this way, the stealthy attacker tries to remain undetected by having himself hidden in the measurement noise.

1) RWD Algorithm: In this case, the threshold matrix H has been chosen with diagonal elements equal to 2. T is fixed to 10 as in the unstealthy case. The estimation error of RWD versus Imhotep-SMT is illustrated in Fig. 3c. The detected sensors with RWD in this case are sensors 1, 2, 3, 6, 7, 8, 15, 16, 19, 21, 22, 24, 25, 26, 27, 29, 31, 32 which contain five FPs (sensors 1, 3, 6, 7, and 32) and one FNs (sensor 30). The substantial feature of RWD (and NRD) comes in here. In the stealthy case which is of high significance, Imhotep-SMT could not detect *any* of the attacked sensors. Besides, although RWD contains some FPs (together with a single false alarm at the beginning of time course for compromised sensor 15 and one FN, it still could detect a majority number of attacked sensors along with reducing the estimation error.

2) NRD Algorithm: The estimation error for this case is depicted in Fig. 3d. In this case, the detected corrupted sensors are 2, 4, 8, 10, 15, 16, 19, 21, 22, 24, 25, 26, 27, 29, 31 (despite there is no alarm for compromised sensor 8 during the time interval $30 \leq t \leq 50$. This includes two FPs (sensors 4 and 10) and one FN (sensor 30). Again, Imhotep-SMT was not able to detect *any* of the attacked sensors. The quantity calculated via (14) and the scalar statistic calculated via (15) for the individual attacked sensor detection are omitted due to space limitations.

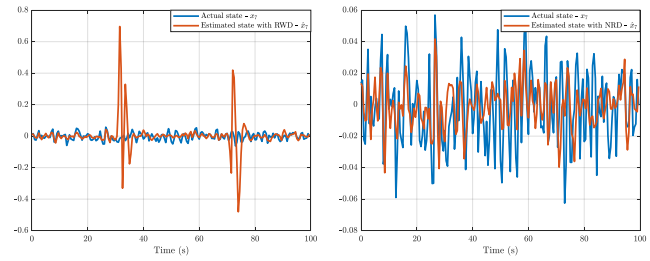
C. The Very Unstealthy Case

Although this case might not be a very favorable situation from the detection point of view, it will answer the question of why one can not see an obvious indication of the attacks in Fig. 3. It might be questionable that why there is no apparent sign of the attacks in the aforementioned time intervals in Fig. 3. This makes really sense for the stealthy case, as in this case the attack power is close to the measurement noise power (in the strictly stealthy case they are equal) which prevents to reveal any substantial estimation error or state estimation deviation during the attack intervals. In fact, as was explained in deep details before, in this case the stealthy attacker attempts to remain undetected by concealing himself inside the measurement noise. This in turn results in no apparent jump neither in the estimation error nor in the state estimation trajectories. This holds for even a huge amount of attack power, i.e., a powerful stealthy attacker where again Imhotep-SMT can detect *none* of the compromised sensors



(a) Estimation error for RWD vs Imhotep-SMT and standard Kalman filter on the attacked IEEE 14-bus power grid system for a very unstealthy attacker

(b) Estimation error for NRD vs Imhotep-SMT and standard Kalman filter on the attacked IEEE 14-bus power grid system for a very unstealthy attacker



(c) Actual state x_7 and its estimation $\hat{x}_7$ with RWD for a very unstealthy attacker

(d) Actual state x_7 and its estimation $\hat{x}_7$ with NRD for a very unstealthy attacker

Fig. 6: Estimation error and state estimation for RWD and NRD vs Imhotep-SMT and standard Kalman filter on the attacked IEEE 14-bus power grid system for a very unstealthy attacker

whereas the proposed methods can. Hence we need to focus on the unstealthy case. For this case, it has to be noted that we have not chosen very striking unstealthy attacks during the former simulations, since such an attacker might be much easy to be detected and is not of much significance from the detection outlook. In spite of all that, to make our discussion comprehensive and to highlight a worth noting point, we look at the estimation error and the state estimation trajectories of a *very* unstealthy attacker, i.e., $\sigma_a \gg \sigma_w$. For this, let us choose $\sigma_w = 0.1$ and $\sigma_a = 5$. Like before, in this case we also compare the estimation error of the algorithms with that of a standard Kalman filter not equipped with a detection scheme. To suppress the effects of the initial conditions and to have a clearer insight on the comparisons, we have let the system reach its steady state, and then performed the attacks in the same intervals as before. The results are shown in Fig. 6. As one can easily see from Fig. 6a and 6c, the RWD begins *mitigating* the effect of the attacks immediately after they occur. Besides, the mitigation procedure lasts for a relatively much shorter period of time compared to the attack intervals. This again highlights the real-time feature of the proposed algorithm. Even better than that, Fig. 6b and 6d show that the NRD has the ability to perfectly mitigate the attacks and therefore there are no spikes in the estimation error and the state estimation trajectory.

Remark V.1. Although a very unstealthy attacker might not be of much interest from the detection perspective, this is of high importance from the viewpoint of the amount of performance degradation and system destruction that he

may cause. In other words, in this case it does not matter for the attacker to remain undetected while performing his action whereas his main objective is to carry out as much devastation as possible on the system. From Fig. 6 we saw that (although RWD has a little bit larger error than Imhotep-SMT) both of our algorithms are able to mitigate this huge attack efficiently with a much less computational complexity compared to Imhotep-SMT which is regarded as another advantage of RWD and NRD in this special case.

Remark V.2. It is notable that based on the mathematical formulation of Imhotep-SMT [13], while it exploits a fixed number of previous measurements (n previous measurements where n is the order of the system) to perform the secure estimation, RWD and NRD exploit *all* former measurements from the beginning of the time horizon. This is due to the nature of the Kalman filter recursive structure. This fact along with the developed algorithms for RWD and NRD results in a lower estimation error in both unstealthy and stealthy scenarios.

Detection Rate Analysis: Having gotten a clearer insight on the detection rate analysis, we can also carry out our analysis in terms of other well-known metrics, namely False Positive Rate (FPR) and False Negative Rate (FNR). Here we have a brief investigation on the ability of Imhotep-SMT and NRD for the attack detection based on the aforementioned metrics. As was used in the previous results, some of the most common metrics that can be applied here are FPR and FNR. These metrics are defined as the following,

$$FPR = \frac{FP}{FP + TN}, \quad FNR = \frac{FN}{FN + TP}. \tag{17}$$

As a sample comparison of detection performance between NRD and Imhotep-SMT we depict the FNR[7] for these two methods for different combinations of attack power and noise power. The FNRs are shown in Fig. 5. From Fig. 5 one can perceive that the NRD exhibits fewer false negatives for almost all the combinations of the attack power and the noise power. This is of much more particular importance that in the stealthy scenarios, in which σ_w and σ_a are close to each other, the NRD algorithm has a better detection performance (lower FN) in comparison to Imhotep-SMT.

VI. Conclusion

In this paper we proposed two Kalman filter-based algorithms for attacked sensor detection and secure state estimation. The algorithms adopt a modified version of Kalman filter along with simple statistical testings aiming to detect individual sensors under attack. The performance of the algorithms were assessed in both unstealthy and stealthy cases. Using an IEEE 14-bus power grid system as the test system, the ability of the algorithms in attack detection and secure state estimation was demonstrated compared to a newly developed tool in the literature which is the only other method that tries to detect individual attacked sensors. In addition, as a deeper attack detection analysis, the FNR was depicted for one of the algorithms compared to that of for the previously presented tool for different combinations of attack power and measurement noise power. In the simulation results, the capability of the algorithms in both the unstealthy and stealthy cases was demonstrated and their functionality in the stealthy case, which is of much more significance in the cyber-security context, was highlighted that could be regarded as a key advantage of the developed algorithms.

[7]Counterparts of these notions, particularly True Positive Rate (TPR) and True Negative Rate (TNR) are also applicable as follows,

$$TPR = \frac{TP}{TP + FN}, \quad TNR = \frac{TN}{TN + FP}.$$

References

[1] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *HotSec*, 2008.

[2] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[4] Y. Shoukry, P. Martin, P. Tabuada, and M. Srivastava, "Non-invasive spoofing attacks for anti-lock braking systems," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2013, pp. 55–72.

[5] G. Basile and G. Marro, *Controlled and conditioned invariants in linear system theory*. Prentice Hall Englewood Cliffs, 1992.

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

[7] S. Lee, Y. Cho, and B.-C. Min, "Attack-aware multi-sensor integration algorithm for autonomous vehicle navigation systems," *arXiv preprint arXiv:1709.02456*, 2017.

[8] J. Miloševič, T. Tanaka, H. Sandberg, and K. H. Johansson, "Analysis and mitigation of bias injection attacks against a kalman filter," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 8393–8398, 2017.

[9] K. Manandhar, X. Cao, F. Hu, and Y. Liu, "Detection of faults and attacks including false data injection attack in smart grid using kalman filter," *IEEE transactions on control of network systems*, vol. 1, no. 4, pp. 370–379, 2014.

[10] C.-Z. Bai, V. Gupta, and F. Pasqualetti, "On kalman filtering with compromised sensors: Attack stealthiness and performance bounds," *IEEE Transactions on Automatic Control*, vol. 62, no. 12, pp. 6641–6648, 2017.

[11] Y. H. Chang, Q. Hu, and C. J. Tomlin, "Secure estimation based kalman filter for cyber–physical systems against sensor attacks," *Automatica*, vol. 95, pp. 399–412, 2018.

[12] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

[13] Y. Shoukry, P. Nuzzo, A. Puggelli, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Secure state estimation for cyber-physical systems under sensor attacks: A satisfiability modulo theory approach," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 4917–4932, 2017.

[14] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs," *Automatica*, vol. 82, pp. 251–260, 2017.

[15] Y. Shoukry, A. Puggelli, P. Nuzzo, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Sound and complete state estimation for linear dynamical systems under sensor attacks using satisfiability modulo theory solving," in *American Control Conference (ACC)*. IEEE, 2015, pp. 3818–3823.

[16] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*. IEEE, 2011, pp. 2195–2201.

[17] F. Pasqualetti, A. Bicchi, and F. Bullo, "A graph-theoretical characterization of power network vulnerabilities," in *American Control Conference (ACC), 2011*. IEEE, 2011, pp. 3918–3923.