Secure State Estimation Against Integrity Attacks: A Gaussian Mixture Model Approach

Ziyang Guo, Dawei Shi, Daniel E. Quevedo, *Senior Member, IEEE*, and Ling Shi

***Abstract*—We consider the problem of estimating the state of a linear time-invariant Gaussian system using N sensors, where a subset of the sensors can potentially be compromised by an adversary. In this case, locating the compromised sensors is of crucial importance for obtaining an accurate state estimate. Inspired by the clustering algorithm in machine learning, we propose a Gaussian-mixture-model-based (GMM-based) detection mechanism. It clusters the local state estimate autonomously and provides a belief for each sensor, based on which measurements from different sensors can be fused accordingly. When a subset of the sensors are under the optimal innovation-based deception attacks, we derive the remote estimation error covariance recursions under different detection mechanisms, e.g., distributed χ^2 false-data detector, centralized χ^2 false-data detector, and GMM-based detection algorithm. The performance of the proposed GMM-based detection algorithm is further evaluated through average belief in the same attack scenario. Moreover, we discuss applications of GMM-based detection algorithm on other attack scenarios, e.g., false-data injection attack, replay attack, and ϵ-stealthy attack. Simulation examples are provided to demonstrate the developed results.**

***Index Terms*—Secure sate estimation, Gaussian mixture model, integrity attack, clustering.**

I. Introduction

CYBER-PHYSICAL systems (CPS) are systems that integrate sensing, computing and control techniques with physical plants. Due to its widespread applications in critical infrastructures ranging from national power grids to personal health monitoring systems, CPS security becomes a problem of growing importance [1]. Typically, sensors in CPS are vulnerable to malicious attacks since in most cases they are spatially distributed and cannot be fully protected [2]. Any successful attack may cause severe consequences on national economy, social security or even loss of human lives, e.g., StuxNet malware [3], Maroochy water bleach [4].

According to the available system knowledge, disclosure resources and disruption resources of an adversary, different attacks can be launched on the system [5]. Integrity attack, which tampers the sensor measurement and compromises the data integrity, is one particular category and has received considerable research attention. There are a number of works focusing on specific types of integrity attacks and aiming to design detection schemes or build attack-resilient estimators [6]–[14]. False-data injection attack with full system knowledge was considered in two different scenarios for electric power grids and its stealthiness was demonstrated numerically [6]. When the system parameters were unknown for the attacker, a data-driven method was proposed to learn the operating subspace and launch attacks accordingly [7]. Moreover, the effect of false-data injection attack on state estimation of dynamic Gaussian systems was investigated in [8]. Replay attack attempts to inject exogenous control inputs and simultaneously repeat historical measurements to maintain stealthiness to the false-data detector. The feasibility conditions of such an attack were investigated in [9] and a watermark authentication signal was introduced as a countermeasure in [10]. The trade-off between estimation quality and detection rate was studied in a stochastic game framework for replay attacks [11]. The innovation-based deception attack was initially studied in [12] for remote state estimation as an affine operation on the innovation process, based on which the attack policy was generalized to an arbitrary function of the innovation sequence and the worst-case attack strategy was obtained [13]. Under different information sets, three sequential data verification and fusion procedures were proposed in [14] for multi-sensor systems under innovation-based deception attacks.

Note that it is difficult to determine the specific attack type in many scenarios. Hence, the development of attack detection and secure estimation policies which are applicable to different attack scenarios is necessary. In [15], the problem of attack detection and identification in CPS was characterized using system-theoretic and graph-theoretic approaches and centralized and distributed attack detection and identification monitors were proposed. In [16], the maximum number of tolerable attacks that allow successful detection of a binary random state was characterized and the optimal detector was shown to be a threshold rule. Similar results were obtained in the area of distributed detection [17], [18] and accurate state reconstruction of noiseless dynamic systems [19], which was further extended to systems with noise and modeling errors in [20], [21]. Since

Manuscript received December 11, 2017; revised July 13, 2018, September 26, 2018, and October 20, 2018; accepted October 26, 2018. Date of publication November 1, 2018; date of current version November 30, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pierre Borgnat. The work by Z. Guo and L. Shi was supported by a Hong Kong RGC General Research Fund 16222716. The work by D. Shi was supported by National Natural Science Foundation of China under Grant 61503027. *(Corresponding author: Dawei Shi.)*

Z. Guo and L. Shi are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: zguoae@connect.ust.hk; eesling@ust.hk).

D. Shi is with the Key Laboratory of Intelligent Control and Decision of Complex Systems, the MIIT Key Laboratory of Drive and Control of Motion Systems, and the School of Automation, Beijing Institute of Technology, Beijing 100081, China (e-mail: daweishi@bit.edu.cn).

D. E. Quevedo is with the Department of Electrical Engineering (EIM-E), Paderborn University, Paderborn, Germany (e-mail: dquevedo@ieee.org).

Digital Object Identifier 10.1109/TSP.2018.2879037

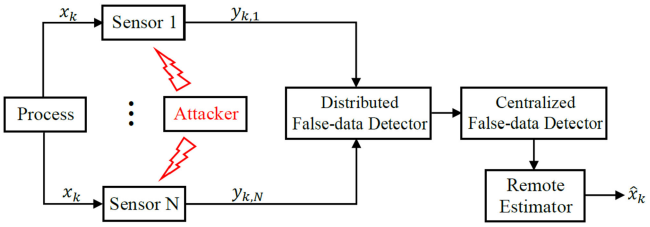


Fig. 1. System Block Diagram.

locating the compromised sensors is intrinsically a combinatorial problem, a satisfiability modulo theory based approach for secure state estimation was proposed [22], where the scalability, soundness and completeness of the proposed algorithm were analyzed. In [23], transfer entropy based causality countermeasures were investigated for both sensor measurements and innovation sequences. The countermeasures were shown to be related to the system parameters, based on which the convergence condition was analyzed and the effectiveness was evaluated. In [24], when a subset of the sensors were compromised by an adversary, a convex optimization based data fusion approach was adopted to combine the local estimates into a more secure state estimate. A more general convex optimization based estimator with l_1 penalty was considered in [25] and sufficient and necessary conditions for resilience were analyzed with a trivial gap. A finite-state stochastic modeling framework for secure state estimation was introduced in [26]. Further results and developments related to attack detection and secure state estimation can be found in [27]–[29].

In the current work, we consider the problem of estimating the state of a linear dynamical system when an unknown subset of sensors may be corrupted by an adversary. The system architecture is shown in Fig. 1. Each sensor measures outputs of an underlying physical process and transmits noisy measurements to a remote estimator. There may exist a malicious attacker who is capable to compromise a subset of the sensors and modify the transmitted data packets to degrade the estimation performance. In this case, to diagnose the existence of malicious attacks and obtain a secure state estimate, a distributed and a centralized false-data detector are equipped at the remote side. Many practical applications can be formulated as this system setup [30], such as autonomous vehicles [31], health monitoring systems [32], smart grid [33], etc. One specific example is autonomous driving of vehicles. To perceive the environment and navigate without much human input, different sensors are deployed on a vehicle to monitor the position, velocity, acceleration and other information. In such systems, data transmissions are performed through wireless communication networks and malicious attacks may lead to malfunction of the speed control or braking systems. For safety considerations, it is necessary to adopt efficient false-data detection schemes to ensure the data security and accuracy. The classical χ^2 false-data detector, which is widely used for fault detection in the process industry and studied in the research community [6], [8], [9], [11]–[14], [34], is considered here. Specifically, if no alarm is triggered at both the distributed and the centralized χ^2 detectors, then the measurement data are believed to be reliable and fused at the remote estimator. If alarms are triggered at the distributed χ^2 detector, then one simply removes those corrupted measurements and checks the remaining data through the centralized χ^2 detector: If no alarm is triggered at the centralized χ^2 detector, the remaining data will be fused at the remote estimator. Otherwise, it will be discarded. In this case, for those carefully designed attacks which are able to bypass the distributed χ^2 detector, but fail to remain stealthy to the centralized χ^2 detector (e.g., innovation-based deception attack [12], false-data injection attack [8] and reply attack [9]), the alarm triggered in the centralized χ^2 detector may result in a large performance degradation since uncontaminated information is discarded. Consequently, it is necessary to develop a detection mechanism which is able to locate the compromised sensor and is applicable to different attack scenarios.

Inspired by the clustering algorithm in machine learning, we propose a GMM-based detection mechanism for attack localization and secure state estimation in this work. Gaussian mixture model is a probabilistic model for representing normally distributed subpopulations within an overall population. In general, it does not require knowing which subpopulation a data point belongs to, allowing the model to learn the subpopulations automatically, which constitutes a form of unsupervised learning [35]. Hence, the proposed detection algorithm is able to divide the sensors into two categories autonomously and weight the measurements from different sensors with different beliefs, often leading to a satisfactory estimation performance. The main contributions of this paper are summarized as follows:

1) We propose a GMM-based detection algorithm (Algorithm 1) when a subset of the sensors are under stealthy integrity attack. This method is effective to the problem that the centralized χ^2 false-data detector cannot locate the attacks in cases where the distributed χ^2 false-data detector cannot detect their existence. To improve the performance and obtain a more accurate state estimate, we propose a receding horizon GMM-based detection algorithm (Algorithm 2).
2) We derive the remote estimation error covariance recursions under distributed χ^2 false-data detector, centralized χ^2 false-data detector and GMM-based detection algorithm when a subset of the sensors are under the optimal innovation-based deception attack, respectively (Theorem 1 and Theorem 2). Moreover, we provide expressions for the average belief used to evaluate the performance of the proposed GMM-based detection algorithm (Theorem 3).
3) We discuss applications of the GMM-based detection algorithm on other attack scenarios, namely, false-data injection attack, replay attack and ϵ-stealthy attack. Simulation examples are also provided to compare the performance of different detection mechanisms under different attack scenarios.

The remainder of the paper is organized as follows. Section II introduces the problem formulation. Section III proposes the modified and receding horizon GMM-based detection algorithms for secure state estimation. Section IV derives the remote estimation error covariances and average belief for the

Algorithm 1: Modified GMM for Secure State Estimation.

/*Reach steady state*/
Initialize $\hat{x}_{-\infty,i}=0$, $P_{-\infty,i}=\Pi_i$, $\hat{x}_{-\infty}=0$, $P_{-\infty}=\Pi$;
for $k=-\infty:0$ **do**
  for $i=1:N$ **do**
    $P_{k,i}=[(AP_{k-1,i}A'+Q)^{-1}+C_i'R_i^{-1}C_i]^{-1}$;
    $\hat{x}_{k,i}=A\hat{x}_{k-1,i}+P_{k,i}C_i'R^{-1}$ $(y_{k,i}-C_iA\hat{x}_{k-1,i})$;
  end for
  $P_k=[(AP_{k-1}A'+Q)^{-1}+C'R^{-1}C]^{-1}$;
  $\hat{x}_k=A\hat{x}_{k-1}+P_kC'R^{-1}(y_k-CA\hat{x}_{k-1})$;
end for
/*GMM clustering and data fusion*/
Set $P_i=P_{0,i}$, $\Sigma_i^{(1)}=P_{0,i}$;
for $k=1:\infty$ **do**
  for $i=1:N$ **do**
    $\hat{x}_{k,i}=A\hat{x}_{k-1,i}+P_iC_i'R_i^{-1}(y_{k,i}-C_iA\hat{x}_{k-1,i})$;
  end for
  Initialize $\pi_k^{(q)}$, $\mu_k^{(q)}$, $\Sigma_k^{(2)}$, where $q=1,2$;
  while Termination condition not reached **do**
    Expectation: for each i, compute
    $\gamma_{k,i}^{(1)}=\frac{\pi_k^{(1)}f(\hat{x}_{k,i};\mu_k^{(1)},\Sigma_i^{(1)})}{\pi_k^{(1)}f(\hat{x}_{k,i};\mu_k^{(1)},\Sigma_i^{(1)})+\pi_k^{(2)}f(\hat{x}_{k,i};\mu_k^{(2)},\Sigma_k^{(2)})}$;
    $\gamma_{k,i}^{(2)}=1-\gamma_{k,i}^{(1)}$;
    Maximization: re-estimate parameters as follows
    $\pi_k^{(q)}=\frac{\sum_{i=1}^N\gamma_{k,i}^{(q)}}{N}$;
    $\mu_k^{(q)}=\frac{\sum_{i=1}^N\gamma_{k,i}^{(q)}\hat{x}_{k,i}}{\sum_{i=1}^N\gamma_{k,i}^{(q)}}$;
    $\Sigma_k^{(2)}=\frac{\sum_{i=1}^N\gamma_{k,i}^{(2)}(\hat{x}_{k,i}-\mu_k^{(2)})(\hat{x}_{k,i}-\mu_k^{(2)})'}{\sum_{i=1}^N\gamma_{k,i}^{(2)}}$;
  end while
  $P_k=[(AP_{k-1}A'+Q)^{-1}+\sum_{i=1}^N\gamma_{k,i}^{(1)}C_i'R_i^{-1}C_i]^{-1}$;
  $\hat{x}_k=A\hat{x}_{k-1}+\sum_{i=1}^N\gamma_{k,i}^{(1)}P_kC_i'R_i^{-1}$ $(y_{k,i}-C_iA\hat{x}_{k-1})$;
end for

Algorithm 2: Receding Horizon GMM for Secure State Estimation.

/*Reach steady state*/
Initialize $\hat{x}_{-\infty,i}=0$, $P_{-\infty,i}=\Pi_i$, $\hat{x}_{-\infty}=0$, $P_{-\infty}=\Pi$, W;
for $k=-\infty:0$ **do**
  for $i=1:N$ **do**
    $P_{k,i}=[(AP_{k-1,i}A'+Q)^{-1}+C_i'R_i^{-1}C_i]^{-1}$;
    $\hat{x}_{k,i}=A\hat{x}_{k-1,i}+P_{k,i}C_i'R^{-1}$ $(y_{k,i}-C_iA\hat{x}_{k-1,i})$;
  end for
  $P_k=[(AP_{k-1}A'+Q)^{-1}+C'R^{-1}C]^{-1}$;
  $\hat{x}_k=A\hat{x}_{k-1}+P_kC'R^{-1}(y_k-CA\hat{x}_{k-1})$;
end for
/*Receding horizon GMM clustering and data fusion*/
Set $P_i=P_{0,i}$, $\Sigma_i^{(1)}=P_{0,i}$;
for $k=1:\infty$ **do**
  for $i=1:N$ **do**
    $\hat{x}_{k,i}=A\hat{x}_{k-1,i}+P_iC_i'R_i^{-1}$ $(y_{k,i}-C_iA\hat{x}_{k-1,i})$;
  end for
  Initialize $\pi_k^{(q)}$, $\mu_j^{(q)}$, $\Sigma_k^{(2)}$, where $q=1,2$, $j\in J$ with $J=[k-W+1:k]$;
  while Termination condition not reached **do**
    Expectation: for each i, compute
    $\gamma_{k,i}^{(1)}=\frac{\pi_k^{(1)}\prod_{j\in J}f(\hat{x}_{j,i};\mu_j^{(1)},\Sigma_i^{(1)})}{\pi_k^{(1)}\prod_{j\in J}f(\hat{x}_{j,i};\mu_j^{(1)},\Sigma_i^{(1)})+\pi_k^{(2)}\prod_{j\in J}f(\hat{x}_{j,i};\mu_j^{(2)},\Sigma_k^{(2)})}$;
    $\gamma_{k,i}^{(2)}=1-\gamma_{k,i}^{(1)}$;
    Maximization: re-estimate parameters as follows
    $\pi_k^{(q)}=\frac{\sum_{i=1}^N\gamma_{k,i}^{(q)}}{N}$;
    $\mu_j^{(q)}=\frac{\sum_{i=1}^N\gamma_{k,i}^{(q)}\hat{x}_{j,i}}{\sum_{i=1}^N\gamma_{k,i}^{(q)}}$;
    $\Sigma_k^{(2)}=\frac{\sum_{i=1}^N\sum_{j=k-W+1}^W\gamma_{k,i}^{(2)}(\hat{x}_{j,i}-\mu_j^{(2)})(\hat{x}_{j,i}-\mu_j^{(2)})'}{W\sum_{i=1}^N\gamma_{k,i}^{(2)}}$;
  end while
  $P_k=[(AP_{k-1}A'+Q)^{-1}+\sum_{i=1}^N\gamma_{k,i}^{(1)}C_i'R_i^{-1}C_i]^{-1}$;
  $\hat{x}_k=A\hat{x}_{k-1}+\sum_{i=1}^N\gamma_{k,i}^{(1)}P_kC_i'R_i^{-1}$ $(y_{k,i}-C_iA\hat{x}_{k-1})$;
end for

proposed detection algorithm when a subset of the sensors are under the optimal innovation-based deception attack. Numerical examples are also provided to demonstrate the developed results. Applications of GMM-based detection algorithm on other attack scenarios are discussed in Section V. Section V draws conclusions.

Notations: $\mathbb{N}$ and $\mathbb{R}$ denote the sets of positive integers and real numbers, respectively. $\mathbb{R}^n$ is the n-dimensional Euclidean space. $\mathbb{S}_+^n$ ($\mathbb{S}_{++}^n$) is the set of $n\times n$ positive semi-definite (definite) matrices. When $X\in\mathbb{S}_+^n$ ($\mathbb{S}_{++}^n$) , we simply write $X\geq 0$ ($X>0$). $\mathcal{N}(\mu,\Sigma)$ denotes Gaussian distribution with mean μ and covariance matrix Σ. For a matrix X, $\mathrm{Tr}(X)$, X' and $X+(\cdot)'$ stand for the trace, transpose and $X+X'$, respectively. $\mathrm{Diag}\{\cdot\}$ represents a block diagonal matrix. $\mathbb{E}[\cdot]$ refers to the expectation of a random variable.

II. Problem Formulation

A. Process Model

Consider a networked system consisting of N sensors and one remote estimator as depicted in Fig 1. Each sensor $i\in\mathcal{N}\triangleq\{1,2,\ldots,N\}$ measures an output of a linear time-invariant process:

$$x_{k+1}=Ax_k+w_k, \tag{1}$$

$$y_{k,i}=C_ix_k+v_{k,i}, \tag{2}$$

where $k \in \mathbb{N}$ is the time index, $x_k \in \mathbb{R}^n$ is the system state, and $y_{k,i} \in \mathbb{R}^{m_i}$ is the measurement obtained sensor i. Both $w_k \in \mathbb{R}^n$ and $v_{k,i} \in \mathbb{R}^{m_i}$ are zero-mean i.i.d. Gaussian noises with $\mathbb{E}[w_k w_l'] = \delta_{kl} Q\ (Q \geq 0)$, $\mathbb{E}[v_{k,i} v_{l,j}'] = \delta_{ij}\delta_{kl} R_i\ (R_i > 0)$, $\mathbb{E}[w_k v_{l,i}'] = 0,\ \forall k, l \in \mathbb{N},\ i, j = 1, 2, \ldots, N$. The initial state x_0 is zero-mean Gaussian with covariance matrix $\Pi_0 > 0$ and independent of w_k and $v_{k,i}$ for all $k \geq 0$. The pairs (A, C_i) are detectable, and $(A, \sqrt{Q})$ is controllable.

Remark 1: The system parameters (A, C, Q, R, Π_0) are assumed to be known a priori. Typically, there are two ways to acquire these parameters: one is to obtain state-space models from specific physical problems [36]; the other is to adopt system identification techniques to build mathematical models and learn system parameters from observed input and output signals [37].

Remark 2: According to [38], it is a combination of experiment, analysis and pragmatism that suggests the Gaussian assumption of the initial state. To be specific, experiments establish that many naturally occurring processes are Gaussian; by modeling certain natural process as resulting from the sum of some individual, possibly non-Gaussian processes, the central limit theorem suggests an approximately Gaussian character for the sum; and finally, the filtering problem is generally easier to solve with the Gaussian assumption. Hence, we assume that the initial state is zero-mean Gaussian distributed, i.e., $x_0 \sim \mathcal{N}(0, \Pi_0)$.

B. Remote Estimator

At time instant k, each sensor transmits its local measurement to a remote estimator. By defining

$$\begin{aligned} y_k &\triangleq \begin{bmatrix} y_{k,1}' & y_{k,2}' & \ldots & y_{k,N}' \end{bmatrix}', \\ v_k &\triangleq \begin{bmatrix} v_{k,1}' & v_{k,2}' & \ldots & v_{k,N}' \end{bmatrix}', \\ C &\triangleq \begin{bmatrix} C_1' & C_2' & \ldots & C_N' \end{bmatrix}', \\ R &\triangleq \mathrm{Diag}\{R_1, R_2, \ldots, R_N\}, \end{aligned}$$

the overall measurement can be represented as

$$y_k = C x_k + v_k. \tag{3}$$

To estimate the system state based on the received measurements, a Kalman filter is adopted at the remote estimator:

$$\begin{aligned} \hat{x}_k^- &= A\hat{x}_{k-1}, \\ P_k^- &= AP_{k-1}A' + Q, \\ K_k &= P_k^- C'(CP_k^- C' + R)^{-1}, \\ \hat{x}_k &= \hat{x}_k^- + K_k(y_k - C\hat{x}_k^-), \\ P_k &= (I - K_k C)P_k^-, \end{aligned}$$

where $\hat{x}_k^-$ and $\hat{x}_k$ are the *a priori* and the *a posteriori* minimum mean squared error (MMSE) estimates of the state x_k, and P_k^- and P_k are the corresponding estimation error covariances. The recursion starts from $\hat{x}_0 = 0$ and $P_0 = \Pi_0 > 0$. An alternative form for the measurement update is

$$\hat{x}_k = \hat{x}_k^- + P_k C' R^{-1}(y_k - C\hat{x}_k^-), \tag{4}$$

$$(P_k)^{-1} = (P_{k-1}^-)^{-1} + C' R^{-1} C, \tag{5}$$

which is known as the information-form Kalman filter. Similar to 4 and 5, the local Kalman filter for sensor i, $i = 1, 2 \ldots, N$ can also be obtained.

To simplify the subsequent discussion, we define the Lyapunov and Riccati operators $h, g_i, g : \mathbb{S}_{++}^n \to \mathbb{S}_{++}^n$ as

$$\begin{aligned} h(X) &\triangleq AXA' + Q, \\ g_i(X) &\triangleq X - XC_i'(C_i X C_i' + R_i)^{-1} C_i X, \\ g(X) &\triangleq X - XC(CXC' + R)^{-1} CX. \end{aligned}$$

It is well known that the Kalman filter converges from any initial condition exponentially fast when (A, C_i) is detectable and $(A, \sqrt{Q})$ is controllable [38]. We denote the steady-state values for the local and centralized Kalman filter as

$$P_i \triangleq \lim_{k \to +\infty} P_{k,i},\ P_i^- \triangleq \lim_{k \to +\infty} P_{k,i}^-, \tag{6}$$

$$P \triangleq \lim_{k \to +\infty} P_k,\ P^- \triangleq \lim_{k \to +\infty} P_k^-, \tag{7}$$

where P_i, P_i^-, P and P^- are the unique positive definite solution of $g_i \circ h(X) = X$, $h \circ g_i(X) = X$, $g \circ h(X) = X$ and $h \circ g(X) = X$, respectively. Without loss of generality, we assume that the system starts from the steady state with $P_{i,0} = P_i$ and $P_0 = P$, which results in fixed-gain local and centralized Kalman filters, i.e.,

$$\begin{aligned} K_i &= P_i C_i' R_i^{-1} = P_i^- C_i'(C_i P_i^- C_i' + R_i)^{-1}, \\ K &= PC' R^{-1} = P^- C'(CP^- C' + R)^{-1}. \end{aligned}$$

C. False-Data Detectors

To ensure the data integrity in CPS, a false-data detector is usually adopted to monitor system behavior and detect the existence of potential malicious attacks. Note that for a local Kalman filter, the innovation $z_{k,i} = y_{k,i} - C_i\hat{x}_{k,i}^-$ has a steady-state Gaussian distribution $\mathcal{N}(0, C_i P_i^- C_i' + R_i)$ and $\mathbb{E}[z_{k,i} z_{l,i}'] = 0$ for all $k \neq l$ [38]. For a centralized Kalman filter, the innovation $z_k = y_k - C\hat{x}_k^-$ has a steady-state Gaussian distribution $\mathcal{N}(0, CP^- C' + R)$ and $\mathbb{E}[z_k z_l'] = 0$ for all $k \neq l$. Hence, the statistical characteristics (mean and covariance) of the innovation sequence can be used to diagnose the system anomalies [39]. Based on different information sets, the following distributed and centralized χ^2 false-data detectors are considered.

In the distributed case, the false-data detector diagnoses the existence of cyber attacks by parallelly checking the sum of the normalized innovation sequence for every single sensor, i.e., at time k, the detection criterion of sensor i, $i = 1, 2, \ldots, N$ follows the hypothesis test

$$g_{k,i} = \sum_{j=k-J_i+1}^{k} z_{j,i}'(C_i P_i^- C_i' + R_i)^{-1} z_{j,i} \underset{H_1}{\overset{H_0}{\lessgtr}} \delta_i, \tag{8}$$

where J_i is the detection window size of sensor i, and δ_i is the threshold of sensor i. The null hypothesis H_0 means that the system is operating normally, while the alternative hypothesis H_1 means that the system is under attack. The normalized sum in (8) satisfies the χ^2 distribution with mJ degrees of freedom. If $g_{k,i}$ exceeds the threshold, the detector will trigger an alarm and the measurement of sensor i will be dropped. Note that

for a given threshold δ_i, a larger window size J_i leads to a more accurate detector but a slower response speed. For a given window size J_i, the smaller the threshold δ_i, the lower the tolerance of the detector to the attack.

In the centralized case, the false-data detector checks the system anomalies based on the innovation calculated by the centralized Kalman filter, i.e., at time k, the detection criterion follows the hypothesis test

$$g_k = \sum_{j=k-J+1}^{k} z_j'(CP^{-}C' + R)^{-1} z_j \underset{H_1}{\overset{H_0}{\lessgtr}} \delta. \tag{9}$$

The normalized sum in (9) is χ^2 distributed with mNJ degrees of freedom. When g_k exceeds the threshold, an alarm will be triggered and all the measurements will be dropped.

D. Problem of Interest

Suppose that there exists an attacker who is able to modify the measurement data of the system described in the previous subsections. In practice, an attacker can launch such an integrity attack in different fashions. For example, it can change the physical environment to mislead the sensors or hack the on-board sensor chip or manipulate the data packet during the sensor-to-estimator transmission. The ability of an attacker in the real world is usually limited, so we assume that it can only compromise a subset of the sensors. For convenience, we denote the set of compromised sensors as $\mathcal{A}$ with cardinality $|\mathcal{A}| = M < N$ and the set of sensors without attacks as $\mathcal{S}$ with cardinality $|\mathcal{S}| = N - M$. The index set of malicious sensors is assumed to be time invariant. Without loss of generality, we also assume that the attack starts from time $k = 1$. In this case, if the attack can be detected by the distributed χ^2 detector and no alarm is triggered at the centralized χ^2 detector, the remote estimator is able to remove those corrupted measurements and obtain a relatively good estimate of the system state. However, if a carefully designed attack can successfully bypass the distributed χ^2 detector but fails to remain stealthy to the centralized χ^2 detector, e.g., innovation-based deception attack [12], false-data injection attack [8] and reply attack [9], the alarm triggered at the centralized χ^2 detector will result in abandoning of all measurements, i.e., only performing a time update at the remote estimator, since the centralized χ^2 detector cannot locate the corrupted measurements. This may lead to a large estimation error due to discarding uncontaminated information. To address this issue, we wish to find an effective detection mechanism which is able to conquer the problem that the centralized χ^2 detector cannot locate the attacks in cases where the distributed χ^2 detector cannot detect their existence. The detailed mathematical formulation and analysis of the proposed detection algorithm will be introduced in the following two sections.

III. Gaussian Mixture Model Based Detection Algorithm

As discussed in the previous section, our interest lies in handling situations where the malicious attacker is able to deliberately design the corrupted data and successfully bypass the distributed χ^2 detector. When a subset of sensors are under attack, the centralized χ^2 detector is able to know the existence of the attack but fails to diagnose which sensor is under attack. Motivated by the above observations, in this section, we propose a GMM-based detection algorithm which is able to locate the attack and yield a more secure and robust state estimate. Moreover, a receding horizon GMM-based detection algorithm is proposed to improve the estimation performance.

Gaussian mixture model is a probabilistic model for representing normally distributed subpopulations within an overall population. It is parameterized by two types of values, the mixture component weights and the component means and covariances. For a Gaussian mixture model with $\mathcal{Q}$ components, the q-th component $\mathcal{G}_q$ has mean $\mu^{(q)}$ and covariance $\Sigma^{(q)}$. The mixture component weights are defined as $\pi^{(q)}$ for component $\mathcal{G}_q$, with the constraint $\sum_{q=1}^{Q} \pi^{(q)} = 1$. In this case, the mixture density can be represented as

$$\begin{aligned} p(x) &= \sum_{q=1}^{\mathcal{Q}} p(x|\mathcal{G}_q) \Pr(\mathcal{G}_q) \\ &= \sum_{q=1}^{\mathcal{Q}} \pi^{(q)} f(x; \mu^{(q)}, \Sigma^{(q)}), \end{aligned} \tag{10}$$

where $f(x; \mu, \Sigma) \triangleq \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu))$ denotes the probability density function (pdf) for Gaussian random variables.

Note that in our problem, the local state estimate $\hat{x}_{k,i}$ follows different distributions when sensor i is uncorrupted or compromised. The situation at hand can be described by the following 2-component mixture model:

$$\begin{aligned} p(\hat{x}_{k,i}) &= \sum_{q=1}^{2} p(\hat{x}_{k,i}|\mathcal{G}_q) \Pr(\mathcal{G}_q) \\ &= \pi^{(1)} p(\hat{x}_{k,i}|\mathcal{G}_1) + \pi^{(2)} p(\hat{x}_{k,i}|\mathcal{G}_2). \end{aligned} \tag{11}$$

According to Kalman filtering analysis [38], it can be shown that when sensor i is uncorrupted (namely, belongs to the first cluster), the local state estimate $\hat{x}_{k,i}$ is Gaussian distributed with fixed covariance P_i, i.e., $p(\hat{x}_{k,i}|\mathcal{G}_1) \sim \mathcal{N}(\mu_k^{(1)}, P_i)$, $\forall i \in \mathcal{N}$. When sensor i is compromised by an adversary (namely, belongs to the second cluster), the exact distribution of local estimate $\hat{x}_{k,i}$ is unknown since we do not know the specific attack type and the attack starting time. In this case, we use the first and second moments to approximate the distribution of all local estimates in the second cluster, i.e., $p(\hat{x}_{k,i}|\mathcal{G}_2) \sim \mathcal{N}(\mu_k^{(2)}, \Sigma_k^{(2)})$, $\forall i \in \mathcal{N}$. Consequently, the mixture density (11) for the local state estimate $\hat{x}_{k,i}$ is further obtained as

$$p(\hat{x}_{k,i}) = \pi_k^{(1)} f\left(\hat{x}_{k,i}; \mu_k^{(1)}, P_i\right) + \pi_k^{(2)} f\left(\hat{x}_{k,i}; \mu_k^{(2)}, \Sigma_k^{(2)}\right). \tag{12}$$

Based on the modified Gaussian mixture model (12), at each time k, we adopt expectation-maximization (EM) algorithm [35], [40] to find the maximum likelihood estimates for the parameter $\Phi_k = \{\pi_k^{(q)}, \mu_k^{(q)}, \Sigma_k^{(2)}\}_{q=1}^{2}$ using the data

$\mathcal{X}_k = \{\hat{x}_{k,i}\}_{i=1}^N$. The log likelihood is given as

$$\mathcal{L}(\Phi_k; \mathcal{X}_k) = \sum_{i=1}^{N} \log\left(\pi_k^{(1)} f\left(\hat{x}_{k,i}; \mu_k^{(1)}, P_i\right) + \pi_k^{(2)} f\left(\hat{x}_{k,i}; \mu_k^{(2)}, \Sigma_k^{(2)}\right)\right). \tag{13}$$

The whole detection algorithm is summarized in Algorithm 1. To be specific, before time instant $k = 0$ (lines 2 to 10), the remote estimator runs a Kalman filter and enters steady state. Starting from time instant $k = 1$ (lines 12 to 29), the remote estimator clusters the local state estimates (lines 14 to 25) and fuses measurement data using different weights (lines 27 and 28). It can be observed that at each time instant k, the expectation step generates a belief $\gamma_{k,i}^{(q)}$, which indicates the probability of sensor i belonging to component $\mathcal{G}_q$. Then, the maximization step re-estimates the parameters based on the belief obtained in the expectation step. This iterative procedure maximizes the log likelihood in (13).

Remark 3: For the case where the specific attack type and the attack starting time are known a priori, the exact distribution of the local state estimate $\hat{x}_{k,i}$ when sensor i is compromised by the attacker can be obtained. In this case, we can directly use the probability density function $p(\hat{x}_{k,i}|\mathcal{G}_2)$, $\forall i \in \mathcal{N}$ to model the mixture density instead of approximating the distribution of all local estimates in the second cluster to be Gaussian, which leads to a better performance. To verify this point, we provide simulation examples in Section IV-D.

Remark 4: Note that we use local state estimates $\{\hat{x}_{k,i}\}_{i=1}^N$ instead of sensor measurements $\{y_{k,i}\}_{i=1}^N$ to cluster the sensors although they are both Gaussian distributed in the absence of attacks. The main reason is that when applying the EM algorithm to sensor measurements, it is difficult to obtain the closed-form update rules for $\mu_k^{(2)}$ and $\Sigma_k^{(2)}$, i.e., lines 24 and 25 of Algorithm 1.

Remark 5: The EM algorithm can be viewed as a special case of the majorize-minimization (MM) algorithm. It maximizes a concave lower bound of the log likelihood (13). Therefore, it guarantees the convergence to a local optimum. However, the EM algorithm is sensitive to the initial condition. Heuristic methods, e.g., k-means or l_1 minimization, are usually used to initialize the parameters [35], [40].

Remark 6: It is worth noting that although the notation P_k has been retained for simplicity in Algorithm 1, this quantity does not represent the covariance of the state estimation error $x_k - \hat{x}_k$ any more, since the use of compromised data is not taken into account in the recursion of this matrix. For a given attack type, the remote estimation error covariance can be approximately calculated. This will be illustrated in the next section.

In Algorithm 1, the GMM-based detection algorithm classifies the sensor based on the local state estimate at current time k. Intuitively, the more data that is available, the better the performance that can be obtained. Hence, to improve the performance of Algorithm 1, we enlarge the data set from $\mathcal{X}_k = \{\hat{x}_{k,i}\}_{i=1}^N$ to $\mathcal{X}_k^W = \{\{\hat{x}_{j,i}\}_{i=1}^N\}_{j=k-W+1}^k$, with W being the window size, and propose a receding horizon GMM-based detection algorithm, which is summarized in Algorithm 2. Different from Algorithm 1, at time k, a history of the local state estimates within time window $[k - W + 1, k]$ are adopted by the false-data detector to do clustering; see lines 14 to 26 of Algorithm 2.

IV. Performance Analysis for Innovation-Based Deception Attack

In this section, we consider the scenario where a subset of sensors are under the optimal innovation-based deception attack, and we analyze the remote estimation error covariance under a distributed χ^2 detector, centralized χ^2 detector and GMM-based detection algorithm, respectively. Moreover, we characterize the average belief for the GMM-based detection algorithm under the same attack scenario. In the subsequent discussions, we use the superscript to represent the quantities in the presence of an attack.

A. Distributed χ^2 False-Data Detector

The innovation-based deception attack was studied in [12], [13] under χ^2 false-data detector for single-sensor systems. Based on [12], [13], we consider such an attack in a multi-sensor system (1)–(2) equipped with a distributed χ^2 false-data detector at the remote estimator and assume that only a subset of sensors are under attack. To be specific, when sensor $i \in \mathcal{A}$ is under the innovation-based deception attack, the attack policy is defined as

$$\tilde{z}_{k,i} = f_{k,i}(z_{k,i}), \tag{14}$$

where $z_{k,i}$ is the original innovation of sensor i, $\tilde{z}_{k,i}$ is the corrupted innovation provided by the malicious attacker, and $f_{k,i}(\cdot)$ is an arbitrary function. To avoid being detected by the distributed χ^2 detector, the corrupted innovation $\tilde{z}_{k,i}$ should preserve the same distribution as $z_{k,i}$, i.e., $\tilde{z}_{k,i} \sim \mathcal{N}(0, C_i P_i^- C_i' + R_i)$. Here, for each sensor $i \in \mathcal{A}$, we simply adopt the optimal innovation-based deception attack obtained in the single-sensor scenario [13] in the form of

$$\tilde{z}_{k,i} = -z_{k,i}, \tag{15}$$

i.e., at each time instant k, the measurement of sensor i is modified from $y_{k,i}$ to $\tilde{y}_{k,i}$ such that the corrupted innovation $\tilde{z}_{k,i}$ is the negative of the original innovation $z_{k,i}$. In this case, the distributed χ^2 detector fails to diagnose the existence of the attack and the remote estimator uses the compromised measurement to update its estimate according to

$$\tilde{x}_k^- = A\tilde{x}_{k-1}, \tag{16}$$

$$\tilde{x}_k = \tilde{x}_k^- + P\sum_{i\in\mathcal{A}} C_i R_i^{-1}(\tilde{y}_{k,i} - C_i\tilde{x}_k^-) + P\sum_{j\in\mathcal{S}} C_j R_j^{-1}(y_{k,j} - C_j\tilde{x}_k^-), \tag{17}$$

where $\tilde{x}_k^-$ and $\tilde{x}_k$ are the *a priori* and the *a posteriori* state estimate at the remote estimator in the presence of an attack. Hence, the remote state estimate will deviate from its true value $\hat{x}_k$, and the remote estimation error covariance under innovation-based deception attack is obtained as follows.

Theorem 1: Consider the multi-sensor system (1)–(2) with a distributed χ^2 detector under the optimal innovation-based deception attack (15). When a subset $\mathcal{A}$ of the sensors are under attack, the estimation error covariance at the remote estimator follows the recursion

$$\begin{aligned}\tilde{P}_k &= (I-KC)(A\tilde{P}_{k-1}A'+Q)(I-KC)'+KRK' \\ &+\sum_{i\in\mathcal{A}}(I-KC)(P_{k,ci}^{an-}+P_{k,ci}^{aa-})C_i'R_i^{-1}C_iP \\ &+\sum_{i\in\mathcal{A}}PC_i'R_i^{-1}C_i(P_{k,ci}^{an-}+P_{k,ci}^{aa-})'(I-KC)' \\ &+\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{A}}PC_i'R_i^{-1}C_i(P_{k,ij}^{nn-}+P_{k,ij}^{na-} \\ &+P_{k,ij}^{an-}+P_{k,ij}^{aa-})C_j'R_j^{-1}C_jP, \end{aligned} \tag{18}$$

where the definitions and iterations of $P_{k,ij}^{nn-}$, $P_{k,ij}^{na-}$, $P_{k,ij}^{an-}$, $P_{k,ij}^{aa-}$, $P_{k,ci}^{an-}$, $P_{k,ci}^{aa-}$ are given in (30)–(35). The initial value of $P_{1,ij}^{nn-}$, $P_{1,ij}^{na-}$, $P_{1,ij}^{an-}$ and $P_{1,ij}^{aa-}$ is the solution of

$$X=A(I-K_iC_i)X(I-K_jC_j)A'+Q+\delta_{ij}AK_iR_iK_i'A',$$

and the initial value of $P_{1,ci}^{an-}$ and $P_{1,ci}^{aa-}$ is the solution of

$$X=A(I-KC)X(I-K_iC_i)'A'+Q+APC_i'K_i'A'.$$

Proof: See Appendix. ■

B. Centralized χ^2 False-Data Detector

In this subsection, we consider the scenario where a centralized χ^2 detector is adopted together with the distributed χ^2 detector. When a subset $\mathcal{A}$ of the sensors are under the optimal innovation-based deception attack (15), the corrupted innovation $\tilde{z}_k$ cannot preserve the same distribution of the original innovation z_k due to the existence of the uncontaminated data. As a result, an alarm will be triggered by the centralized χ^2 detector although the distributed χ^2 detector fails to detect such an attack. However, the centralized χ^2 detector is only able to detect the existence of the attack rather than locating the attack. Hence, the remote estimator will drop all the received measurements and perform a time update as follows:

$$\tilde{x}_k = A\tilde{x}_{k-1}, \tag{19}$$

$$\tilde{P}_k = A\tilde{P}_{k-1}A'+Q. \tag{20}$$

When the attack is sparse enough, the estimation error covariance using both the centralized and distributed χ^2 detector (20) may be even larger than that using the distributed χ^2 detector only (18). This is shown in simulation examples in the next section. Intuitively, performing time update loses a lot of useful information when the attack is only launched on a small number of sensors.

C. GMM-Based Detection Algorithm

To overcome the disadvantage of the centralized χ^2 detector and locate the compromised sensors, in this subsection we consider the case where the GMM-based detection algorithm is used at the remote estimator in addition to the distributed χ^2 detector. When M out of N sensors are under the optimal innovation-based deception attack, we analyze the remote estimation error covariance and evaluate the performance of the GMM-based detection algorithm.

Note that $\gamma_{k,i}^{(1)}$ obtained in Algorithm 1 is closely related to the measurement y_k, which makes the calculation of the actual estimation error covariance at the remote estimator intractable. Hence, we assume that $\gamma_{k,i}^{(1)}$ and y_k are uncorrelated and obtain an approximate remote estimation error covariance in the following theorem. In fact, the difference between the approximate and the actual estimation error covariances, or equivalently the correlation between $\gamma_{k,i}^{(1)}$ and y_k, is negligible when the window size W is large enough. We will show this through simulation examples in the next subsection.

Theorem 2: Consider the multi-sensor system (1)–(2) with the GMM-based detection algorithm (Algorithm 1) under the optimal innovation-based deception attack (15). When a subset $\mathcal{A}$ of the sensors are under attack, the approximate estimation error covariance at the remote estimator follows the recursion

$$\begin{aligned}\tilde{P}_k &= (I-K_kC)(A\tilde{P}_{k-1}A'+Q)(I-K_kC)'+K_kRK_k' \\ &+\sum_{i\in\mathcal{A}}(I-K_kC)(P_{k,ci}^{an-}+P_{k,ci}^{aa-})C_i'R_{k,i}^{-1}C_iP_k \\ &+\sum_{i\in\mathcal{A}}P_kC_i'R_{k,i}^{-1}C_i(P_{k,ci}^{an-}+P_{k,ci}^{aa-})'(I-K_kC)' \\ &+\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{A}}P_kC_i'R_{k,i}^{-1}C_i(P_{k,ij}^{nn-}+P_{k,ij}^{na-} \\ &+P_{k,ij}^{an-}+P_{k,ij}^{aa-})C_j'R_{k,j}^{-1}C_jP_k, \end{aligned} \tag{21}$$

where $K_k \triangleq P_kC'R_k^{-1}$ with $R_k^{-1} \triangleq \mathrm{Diag}\{R_{k,1}^{-1},\ldots,R_{k,N}^{-1}\} = \mathrm{Diag}\{R_1^{-1}\gamma_{k,1}^{(1)},\ldots,R_N^{-1}\gamma_{k,N}^{(1)}\}$ and $\gamma_{k,i}^{(1)}$ being the output of the GMM-based detection algorithm. P_k follows the recursion

$$P_k=\left[(AP_{k-1}A'+Q)^{-1}+\sum_{i=1}^{N}\gamma_{k,i}^{(1)}C_i'R_i^{-1}C_i\right]^{-1} \tag{22}$$

with initial condition $P_0=P$. $P_{k,ci}^{an-}$ and $P_{k,ci}^{aa-}$ follow the recursions (37)–(38) with the same initial values in Theorem 1. $P_{k,ij}^{nn-}$, $P_{k,ij}^{na-}$, $P_{k,ij}^{an-}$ and $P_{k,ij}^{aa-}$ follow the recursions (30)–(33) with the same initial values in Theorem 1.

Proof: See Appendix. ■

To provide more analysis and insight into the GMM-based detection algorithm, we characterize the average belief of Algorithm 1 when a subset $\mathcal{A}$ of the sensors are under the optimal innovation-based deception attack (15). The obtained results are summarized in the following theorem, before which the definition of the average belief and some other notations are introduced first.

Let $\mathbb{E}[\gamma_i^{(1)}]$ be the belief that sensor i is uncontaminated and $\mathbb{E}[\gamma_i^{(2)}]$ be the belief that sensor i is compromised. Then, we define $\mathbb{E}[\gamma_{\mathcal{S}}^{(1)}]=\frac{1}{N-M}\sum_{i\in\mathcal{S}}\mathbb{E}[\gamma_i^{(1)}]$ as the average belief in the sensor being uncontaminated when it is indeed uncorrupted and $\mathbb{E}[\gamma_{\mathcal{S}}^{(2)}]=\frac{1}{N-M}\sum_{i\in\mathcal{S}}\mathbb{E}[\gamma_i^{(2)}]$ as the average belief in the sensor being compromised when it is actually uncontaminated. Similarly, $\mathbb{E}[\gamma_{\mathcal{A}}^{(1)}]=\frac{1}{M}\sum_{i\in\mathcal{A}}\mathbb{E}[\gamma_i^{(1)}]$ and $\mathbb{E}[\gamma_{\mathcal{A}}^{(2)}]=\frac{1}{M}\sum_{i\in\mathcal{A}}\mathbb{E}[\gamma_i^{(2)}]$ represent the average belief in

the sensor being uncontaminated and compromised, respectively, when it is corrupted by the attacker. Recall that for sensor $i \in \mathcal{S}$, its steady-state estimation error covariance is denoted as P_i. According to [13], when senor $j \in \mathcal{A}$ is under the optimal innovation-based attack (15), its estimation error covariance follows the recursion $\tilde{P}_{k,j} = A\tilde{P}_{k-1}A' + Q + 3P_j^- C_j'(C_j P_j^- C_j' + R_j)^{-1} C_j P_j^-$ with initial value $\tilde{P}_{0,j} = P_j$. For stable systems, $\tilde{P}_{k,j}$ converges to a steady-state value, which is denoted as $\tilde{P}_j$. For unstable systems, $\tilde{P}_{k,j}$ diverges to infinity as k increases. Thus, we choose a sufficiently large T and denote $\tilde{P}_j = \tilde{P}_{T,j}$ as its steady-state value.

Theorem 3: Consider the multi-sensor system (1)–(2) with the GMM-based detection algorithm (Algorithm 1) under the optimal innovation-based deception attack (15). When a subset $\mathcal{A}$ of the sensors are under attack, the average belief that sensor i is not compromised is

$$\mathbb{E}[\gamma_{\mathcal{S}}^{(1)}] = \frac{1}{\alpha}\sum_{i\in\mathcal{S}}\int_{j\in\mathcal{A}}\int_{i\in\mathcal{S}} \frac{\alpha f(\xi_i;\mu,P_i)}{\alpha f(\xi_i;\mu,P_i)+\beta f(\xi_i;\nu,\Sigma)} f(\xi_i;0,P_i)\,\mathrm{d}\xi_i f(\zeta_j;0,\tilde{P}_j)\,\mathrm{d}\zeta_j \tag{23}$$

if $i \in \mathcal{S}$, or

$$\mathbb{E}[\gamma_{\mathcal{A}}^{(1)}] = \frac{1}{\beta}\sum_{j\in\mathcal{A}}\int_{j\in\mathcal{A}}\int_{i\in\mathcal{S}} \frac{\beta f(\zeta_j;\mu,P_i)}{\alpha f(\zeta_j;\mu,P_i)+\beta f(\zeta_j;\nu,\Sigma)} f(\xi_i;0,P_i)\,\mathrm{d}\xi_i f(\zeta_j;0,\tilde{P}_j)\,\mathrm{d}\zeta_j \tag{24}$$

if $i \in \mathcal{A}$, where $\alpha = N - M$, $\beta = M$, $\mu = \frac{1}{\alpha}\sum_{i\in\mathcal{S}}\xi_i$, $\nu = \frac{1}{\beta}\sum_{j\in\mathcal{A}}\zeta_j$ and $\Sigma = \frac{1}{\beta}\sum_{j\in\mathcal{A}}(\zeta_j-\nu)(\zeta_j-\nu)'$. Moreover, the average belief that sensor i is compromised is $\mathbb{E}[\gamma_{\mathcal{S}}^{(2)}] = 1 - \mathbb{E}[\gamma_{\mathcal{S}}^{(1)}]$ if $i \in \mathcal{S}$, or $\mathbb{E}[\gamma_{\mathcal{A}}^{(2)}] = 1 - \mathbb{E}[\gamma_{\mathcal{A}}^{(1)}]$ if $i \in \mathcal{A}$.

Proof: See Appendix. ■

Remark 7: Note that $\mathbb{E}[\gamma_{\mathcal{S}}^{(1)}] \in [0,1]$ and $\mathbb{E}[\gamma_{\mathcal{A}}^{(2)}] \in [0,1]$ can be adopted as a performance measure of the proposed GMM-based detection algorithm. The closer the values are to one, the better the performance.

D. Simulation Example

In this subsection, we further evaluate the effectiveness of the GMM-based detection algorithm through numerical examples and comparisons. To do this, we consider a linear time-invariant dynamic process which is measured by 20 sensors. The system parameters A, Q, C_i and R_i are randomly generated from intervals $[0.4, 0.99]$, $[0.5, 2]$, $[1, 2]$ and $[1, 2]$, respectively, for all $i \in \mathcal{N}$. Without loss of generality, we assume that the first 5 sensors are under the optimal innovation-based deception attack. The receding horizon GMM-based algorithm (Algorithm 2) with window size $W = 5$ is used to cluster the sensors. The detection window size and the false-alarm rate of both the distributed and centralized χ^2 false-data detector are set as $J_i = 5, \forall i \in \mathcal{N}$, $J = 5$ and 5%, respectively. The system runs Kalman filter and enters steady state during time period $[1, 26]$. The attack starts from time instant $k = 27$. In order to show the robustness of the proposed algorithm, we report the averaged results over 100 randomly generated systems in the following, see Fig. 2 to Fig. 4.

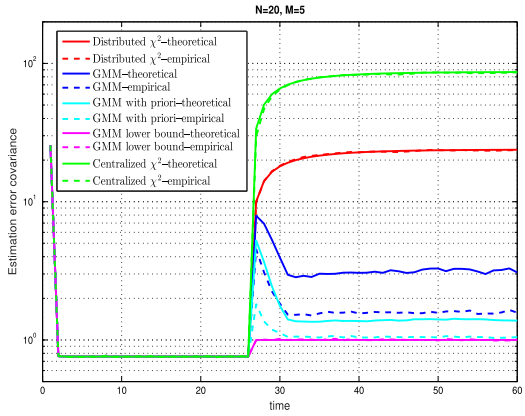


Fig. 2. Remote estimation error covariance using different detection mechanisms when a stable system is under innovation-based deception attack.

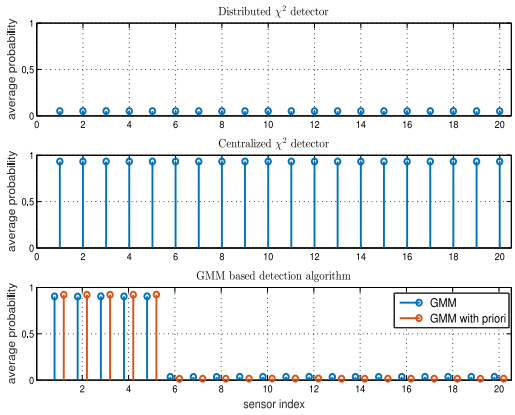


Fig. 3. Average belief on each sensor using different detection mechanisms.

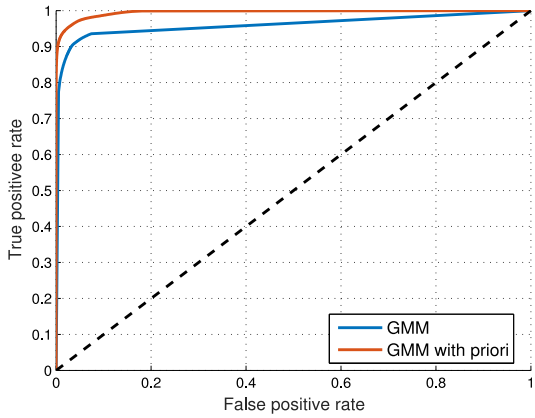


Fig. 4. Receiver operating characteristic curve of GMM-based detection algorithm.

The normalized estimation error covariances at the remote estimator under different detection mechanisms are shown in Fig. 2. When the system is equipped with the distributed χ^2 false-data detector, centralized χ^2 false-data detector and receding horizon GMM-based detection algorithm, the remote estimation error covariances correspond to the red, green and blue lines in the figure, respectively. To facilitate the comparison, we present the estimation error covariance of the GMM-based detection algorithm with prior information, namely, knowing

the specific attack type and attack starting time, using a cyan-colored line. We also provide a performance lower bound for the GMM-based detection algorithm, i.e., running a Kalman filer based on all the uncontaminated sensors, using a magenta-colored line. All the theoretical values are plotted with solid lines while the empirical values are shown in dashed lines. It can be observed that the proposed GMM-based detection algorithm almost achieve the performance bound, which is much better than the traditional distributed and centralized χ^2 false-data detectors. Moreover, the estimation error covariance calculated using Theorem 2 provides a good approximation of its true value. Actually, according to the simulation results, the approximate error covariance provides a performance upper bound of the GMM-based detection algorithm. It is also worth noting that the prior information, i.e., the specific attack type and attack starting time, leads to a better mixture model of local state estimate $\hat{x}_{k,i}$, which further leads to a better estimation performance. However, the performance degradation without this (rather idealized) prior information is not significant, which indicates that the Gaussian approximation of the second cluster in (12) is reasonable.

The average belief on each sensor using different detection mechanisms is presented in Fig. 3. Since the optimal innovation-based deception attack strategy (15) preserves the Gaussian distribution of the original innovation sequence, the distributed χ^2 false-data detector cannot detect the existence of the attack. As a result, the average belief in sensor i being compromised equals the false-alarm rate of the χ^2 false-data detector for all $i \in \mathcal{N}$, which is consistent with Fig. 3. The centralized χ^2 false-data detector is able to notice the existence of the attack due to the existence of uncorrupted data. However, it cannot locate the compromised sensors, which leads to a large alarm rate for each sensor, as shown in Fig. 3. The average beliefs of the proposed GMM-based detection algorithm and the GMM-based detection algorithm with prior information are shown by the blue and red stems in the third subplot of Fig. 3, respectively. It can be observed that, even though the GMM-based detection algorithm with prior information performs slightly better than that without prior information, both of them can successfully cluster the sensors into two categories, i.e., the belief in the sensor being compromised is high when it is indeed under attack and is low when it is actually uncontaminated. This indicates that $\gamma_{\mathcal{S}}^{(2)}$ and $\gamma_{\mathcal{A}}^{(2)}$ obtained in Theorem 3 can be adopted as a performance measure for the GMM-based detection algorithm. Our finding also illustrates why a smaller estimation error covariance in Fig. 2 can be obtained using the GMM-based detection algorithm.

To provide more insights into the proposed GMM-based detection algorithm, its receiver operating characteristic curve is numerically evaluated by taking different values of a threshold η. Specifically, if $\gamma_{k,i}^{(1)} > \eta$, we regard sensor i as a good one. Otherwise, we believe that sensor i is compromised by an adversary. When the threshold η takes different values in the interval $[0, 1]$, the true positive rate versus the false positive rate of the GMM-based detection algorithm with and without the prior information is shown in Fig. 4. "True positive" means that the algorithm believes the sensor is compromised when it is indeed compromised. However, if the sensor is actually uncontaminated, it is said to be "false positive". There is no doubt that a smaller false positive rate and a larger true positive rate are expected. Hence, the closer to the upper left corner of the curve, the better the classification result. The ROC curve in Fig. 4 shows that the proposed GMM-based detection algorithm has a good performance in sensor clustering. When the specific attack type and the attack starting time are known as a priori, the performance of GMM-based detection algorithm becomes better.

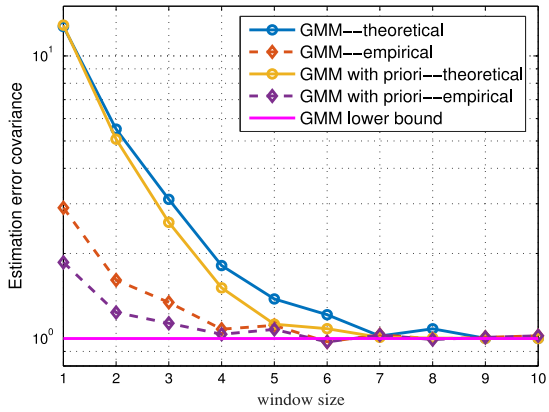


Fig. 5. Remote estimation error covariance of GMM-based detection algorithm with different window sizes when the first 5 sensors are under attack.

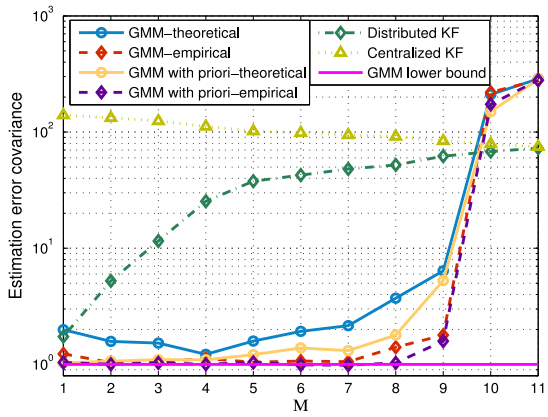


Fig. 6. Remote estimation error covariance of GMM-based detection algorithm with $W = 5$ when different number of sensors are compromised by the attacker.

Moreover, to evaluate the performance of the GMM-based detection algorithm, the remote estimation performance with respect to different window size and different number of compromised sensors are provided. In this case, we consider a system with parameters $A = 0.7$, $Q = 1.2$ and randomly generated $C_i \in [1, 2]$, $R_i \in [0.5, 1.5]$ for all $i \in \mathcal{N}$. Fig. 5 shows the normalized estimation error covariance of GMM-based detection algorithm with different window sizes when the first 5 sensors are under optimal innovation-based deception attacks, while Fig. 6 shows the normalized estimation error covariance of GMM-based detection algorithm with $W = 5$ when different number of sensors are compromised by the attacker. It can be observed from Fig. 5 that both the approximate and empirical estimation error covariances become smaller as the window size

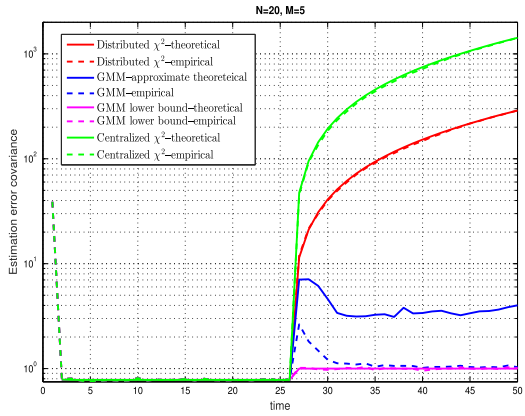


Fig. 7. Remote estimation error covariance using different detection mechanisms when an unstable system is under innovation-based deception attack.

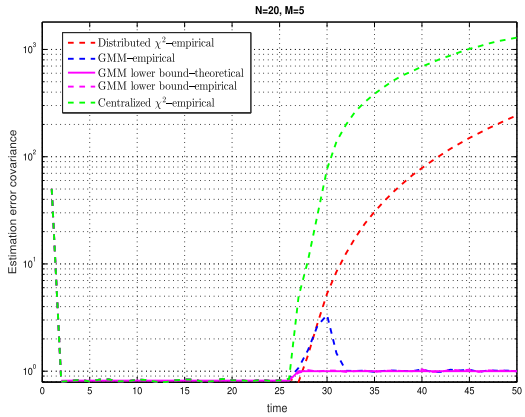


Fig. 8. Remote estimation error covariance using different detection mechanisms when an unstable system is under false-data injection attack.

increases. When the window size is large, they are close to the performance lower bound. It is also worth noting that the difference between the approximate and the empirical estimation error covariances, or equivalently the correlation between $\gamma_{k,i}^{(1)}$ and y_k, becomes smaller as the window size becomes larger. Hence, in practical applications, it is reasonable to assume that $\gamma_{k,i}^{(1)}$ and y_k are uncorrelated by choosing an appropriate window size. According to Fig. 6, the remote estimation quality of GMM-based detection algorithm remains good when less than half of the sensors are under attack. This is consistent to the results obtained in [19], [21], which claim that it is impossible to accurately reconstruct the state of a system if more than half the sensors are attacked. It is also worth noting that knowing the attack type and the attack starting time beforehand improves the performance of GMM-based detection algorithm.

Finally, we consider the same attack scenario for an unstable system with parameters $A = 1.01$, $Q = 1.2$, and randomly generated $C_i \in [1, 2]$, $R_i \in [0.5, 1.5]$ for all $i \in \mathcal{N}$. The normalized estimation error covariances at the remote estimator under different detection mechanisms are shown in Fig. 7. It can be observed that the estimation error goes to infinity when using the distributed and the centralized χ^2 false-data detector, however, it remains bounded under the GMM-based detection algorithm. Actually, it is pretty close to the performance lower bound, i.e., the Kalman filter based on all the uncorrupted sensors.

V. Application on Other Attack Scenarios

Although we focus on the innovation-based deception attack in the previous section, the proposed GMM-based detection algorithm can be applied to many other scenarios. Actually, the GMM-based detection algorithm is expected to be effective for almost all existing sparse integrity attacks [8], [9], [12], [41] which are undetectable using local detectors. Here, the local detector may not be the distributed χ^2 false-data detector (8); it could be in other forms, e.g., the Kullback-Leibler divergence adopted in [41]–[43]. In this section, we discuss the application of the proposed GMM-based detection algorithm on false-data injection attack [8], replay attack [9], and ϵ-stealthy attack [41], and evaluate its effectiveness through simulation examples.

A. False-Data Injection Attack

We first investigate the performance of GMM-based detection algorithm when a subset of the sensors are under false-data injection attack in this subsection. To do this, we consider a system with parameters $A = 1$, $Q = 1.2$ and randomly generated $C_i \in [1, 2]$, $R_i \in [0.5, 1.5]$ for all $i \in \mathcal{N}$. According to Theorem 2 in [8], for sensor $i \in \mathcal{A}$, if the injected attack sequence $y_{k,i}^a$ satisfies

$$\|y_{0,i}^a\| \leq 1,$$

$$y_{k+1,i}^a = y_{k,i}^a + C_i K_i \frac{y_{0,i}^a}{\|y_{0,i}^a\|},$$

the system is perfectly attackable and the attack can be launched stealthily under distributed χ^2 false-data detector. However, due to the existence of the redundant information from sensor $j \in \mathcal{S}$, such an attack cannot bypass the centralized χ^2 false-data detector. When the alarm is triggered in the centralized χ^2 detector, all the sensory data are dropped and the one-step predicted state estimate is obtained at the remote estimator. To compare the above two detection mechanisms with the proposed GMM-based detection algorithm, we provide the normalized remote estimation error covariance in Fig. 8. Observed from the figure, the estimation error under GMM-based detection algorithm is almost the same as that of Kalman filter using all uncontaminated sensors, while the estimation error covariances under distributed and centralized χ^2 detectors diverge to infinity exponentially fast.

B. Replay Attack

Replay attacks prevent the remote estimator from knowing the true data and normally have two phase. In the first phase, the malicious attacker records the sensory data for a certain period of time and replays the recorded data repeatedly during the second phase. To evaluate the effectiveness of the proposed detection scheme, we consider a system with parameters $A = 0.7$, $Q = 1.2$ and randomly generated $C_i \in [1, 2]$, $R_i \in [0.5, 1.5]$ for all $i \in \mathcal{N}$. According to [9], if $A(I - K_i C_i)$ is stable for sensor $i \in \mathcal{A}$, the detection rate of distributed χ^2 detector in the presence of attack will converge to the false alarm rate in the

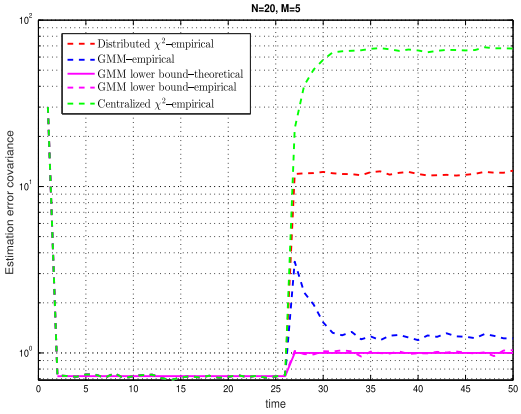


Fig. 9. Remote estimation error covariance using different detection mechanisms when a stable system is under replay attack.

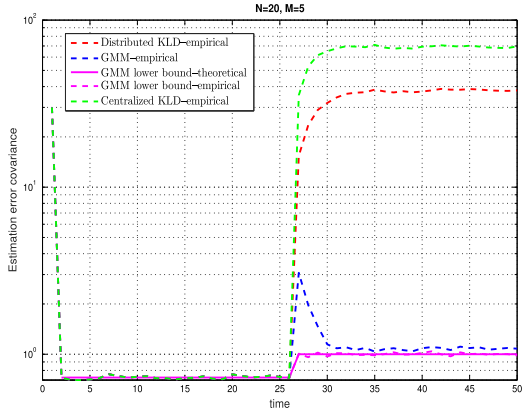


Fig. 10. Remote estimation error covariance using different detection mechanisms when a stable system is under ϵ-stealthy attack.

absence of attack. Hence, replay attack is able to successfully bypass the distributed χ^2 false-data detector. The remote estimation error covariances under different detection mechanisms when a subset of the sensors are under replay attack are shown in Fig. 9. It can be observed that the GMM-based detection algorithm outperforms the distributed and centralized χ^2 detectors significantly.

C. ϵ-Stealthy Attack

Besides the distributed χ^2 false-data detector, other criteria can be adopted as local detection schemes. Kullback-Leibler divergence (KLD), a non-negative measure of the distance between two probability distributions, is widely used in detection theory [41]–[44]. In this subsection, we consider the scenario where the distributed and centralized χ^2 detectors are replaced by the corresponding KLD. Similar to [41], KLD between the corrupted innovation and the nominal innovation is adopted as a stealthiness metric. In the distributed case, KLD detection criterion for sensor $i \in \mathcal{N}$ is defined as

$$\lim_{k\to+\infty} \frac{1}{k} D(\tilde{z}_{1:k,i} \| z_{1:k,i}) \underset{H_1}{\overset{H_0}{\lessgtr}} \epsilon_i,$$

where ϵ_i is the threshold for sensor i and $z_{1:k,i}$ stands for the innovation sequence $\{z_{1,i}, z_{2,i}, \ldots, z_{k,i}\}$. In the centralized case, KLD detection criterion is defined as

$$\lim_{k\to+\infty} \frac{1}{k} D(\tilde{z}_{1:k} \| z_{1:k}) \underset{H_1}{\overset{H_0}{\lessgtr}} \epsilon.$$

If the attack sequence $\tilde{z}_{1:k}$ satisfies hypothesis H_0, the attack is called ϵ-stealthy. The optimal ϵ-stealthy attack policy that maximizes the remote estimation error covariance is investigated in [41], based on which we set $\epsilon_i = \epsilon = 0.1$, $\forall i \in \mathcal{N}$ and consider a system with parameters $A = 0.7$, $Q = 1.2$ and randomly generated $C_i \in [1, 2]$, $R_i \in [0.5, 1.5]$ for all $i \in \mathcal{N}$. The remote estimation error covariances under distributed KLD detection criterion, centralized KLD detection criterion and GMM-based detection algorithm are shown in Fig. 10. It can be observed that the estimation quality under the proposed GMM-based approach is much better than that under KLD criteria.

VI. Conclusion

This paper studies a secure state estimation problem when a subset of sensors are compromised by a malicious attacker. To locate the compromised sensors and obtain an accurate state estimate, a modified GMM-based detection algorithm was proposed. It was able to autonomously cluster the sensors and fuse the measurements based on different beliefs. To further improve the performance, a receding horizon GMM-based detection algorithm was developed. For the innovation-based deception attack, expressions for the remote estimation error covariance recursions under the distributed χ^2 false-data detector, centralized χ^2 false-data detector and GMM-based detection algorithm were obtained. Moreover, to evaluate the performance of the proposed GMM-based detection algorithm, its average belief was characterized when a subset of sensors were under such an attack. Numerical simulations were provided to demonstrate the developed results. Finally, applications of the proposed GMM-based detection algorithm on other types of integrity attacks were discussed and its effectiveness was demonstrated through simulation examples.

The proposed GMM-based detection algorithm depends on the perfect knowledge of the system parameters. For systems with parametric uncertainties, the estimation error covariance of the Kalman filter becomes larger [45], which further degrades the performance of the GMM-based detection algorithm. Hence, the problem of developing a detection scheme which is robust to uncertainties is worth investigating. Another possible direction of the future work is to apply the GMM-based detection algorithm to detect intermittent attacks, where the adversary has energy constraints and cannot launch attacks all the time, or more intelligent attacks, where the adversary is able to interact with the detector.

Appendix

A. Proof of Theorem 1

Proof: When sensor $i \in \mathcal{A}$ is under the optimal innovation-based deception attack, the corrupted measurement satisfies

$$\tilde{y}_{k,i} - C_i \tilde{x}^-_{k,i} = \tilde{z}_{k,i} = -z_{k,i} = -(y_{k,i} - C_i \hat{x}^-_{k,i}). \quad (25)$$

According to (16)–(17), the *a priori* and the *a posterior* estimation error at the remote estimator can be obtained as

$$x_k - \tilde{x}_k^- = A(x_{k-1} - \tilde{x}_{k-1}) + w_{k-1} \tag{26}$$

$$\begin{aligned} x_k - \tilde{x}_k &= x_k - \tilde{x}_k^- - P\sum_{i\in\mathcal{A}} C_i' R_i^{-1}[C_i(x_k - \tilde{x}_k^-) \\ &\quad - C_i(x_k - \hat{x}_{k,i}^-) - C_i(x_k - \tilde{x}_{k,i}^-) - v_{k,i}] \\ &\quad - P\sum_{j\in\mathcal{S}} C_j' R_j^{-1}[C_j(x_k - \tilde{x}_k^-) + v_{k,j}] \\ &= (I - KC)(x_k - \tilde{x}_k^-) \\ &\quad + P\sum_{i\in\mathcal{A}} C_i' R_i^{-1} C_i[(x_k - \hat{x}_{k,i}^-) + (x_k - \tilde{x}_{k,i}^-)] \\ &\quad + P\sum_{i\in\mathcal{A}} C_i' R_i^{-1} v_{k,i} - P\sum_{j\in\mathcal{S}} C_j' R_j^{-1} v_{k,j}, \end{aligned} \tag{27}$$

where the second equality follows from (25). Based on (26)–(27), the estimation error covariance at the remote estimator can be represented as

$$\tilde{P}_k^- = AP_{k-1}A' + Q, \tag{28}$$

$$\begin{aligned} \tilde{P}_k &= (I - KC)\tilde{P}_k^-(I - KC)' + KRK' \\ &\quad + \sum_{i\in\mathcal{A}} \left[(I - KC)(P_{k,ci}^{an-} + P_{k,ci}^{aa-})C_i' R_i^{-1} C_i P + (\cdot)'\right] \\ &\quad + \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{A}} PC_i' R_i^{-1} C_i (P_{k,ij}^{nn-} + P_{k,ij}^{na-} \\ &\quad + P_{k,ij}^{an-} + P_{k,ij}^{aa-}) C_j' R_j^{-1} C_j P. \end{aligned} \tag{29}$$

Due to (27) and the fact that

$$\begin{aligned} x_k - \hat{x}_{k,i}^- &= A(I - K_i C_i)(x_{k-1} - \hat{x}_{k-1,i}^-) + w_{k-1} \\ &\quad - AK_i v_{k-1,i}, \\ x_k - \tilde{x}_{k,i}^- &= A(x_{k-1} - \tilde{x}_{k-1,i}^-) + w_{k-1} \\ &\quad + AK_i C_i(x_{k-1} - \hat{x}_{k-1,i}^-) + AK_i v_{k-1,i}, \end{aligned}$$

it can be easily obtained that

$$\begin{aligned} P_{k,ij}^{nn-} &\triangleq \mathbb{E}[(x_k - \hat{x}_{k,i})(x_k - \hat{x}_{k,j}^-)'] \\ &= A(I - K_i C_i) P_{k-1,ij}^{nn-} (I - K_j C_j)' A' + Q \\ &\quad + \delta_{ij} AK_i R_i K_i' A', \end{aligned} \tag{30}$$

$$\begin{aligned} P_{k,ij}^{na-} &\triangleq \mathbb{E}[(x_k - \hat{x}_{k,i})(x_k - \tilde{x}_{k,j}^-)'] \\ &= A(I - K_i C_i) P_{k-1,ij}^{na-} A' + Q \\ &\quad + A(I - K_i C_i) P_{k-1,ij}^{nn-} C_j' K_j' A' - \delta_{ij} AK_i R_i K_i' A', \end{aligned} \tag{31}$$

$$\begin{aligned} P_{k,ij}^{an-} &\triangleq \mathbb{E}[(x_k - \tilde{x}_{k,i})(x_k - \hat{x}_{k,j}^-)'] \\ &= AP_{k-1,ij}^{an-} (I - K_j C_j)' A' + Q \\ &\quad + AK_i C_i P_{k-1,ij}^{nn-} (I - K_j C_j)' A' - \delta_{ij} AK_i R_i K_i' A', \end{aligned} \tag{32}$$

$$\begin{aligned} P_{k,ij}^{aa-} &\triangleq \mathbb{E}[(x_k - \tilde{x}_{k,i})(x_k - \tilde{x}_{k,j}^-)'] \\ &= AP_{k-1,ij}^{aa-} A' + Q + AK_i C_i P_{k-1,ij}^{na-} A' \\ &\quad + AP_{k-1,ij}^{an-} C_j' K_j' A' + AK_i C_i P_{k-1,ij}^{nn-} C_j' K_j' A' \\ &\quad + \delta_{ij} AK_i R_i K_i' A', \end{aligned} \tag{33}$$

$$\begin{aligned} P_{k,ci}^{an-} &\triangleq \mathbb{E}[(x_k - \tilde{x}_k)(x_k - \hat{x}_{k,i}^-)'] \\ &= A(I - KC) P_{k-1,ci}^{an-} (I - K_i C_i)' A' + Q \\ &\quad - APC_i' K_i' A' + AP\sum_{j\in\mathcal{A}} C_j' R_j^{-1} C_j (P_{k-1,ji}^{nn-} \\ &\quad + P_{k-1,ji}^{an-})(I - K_i C_i)' A', \end{aligned} \tag{34}$$

$$\begin{aligned} P_{k,ci}^{aa-} &\triangleq \mathbb{E}[(x_k - \tilde{x}_k)(x_k - \tilde{x}_{k,i}^-)'] \\ &= A(I - KC) P_{k-1,ci}^{aa-} A' + A(I - KC) P_{k-1,ci}^{an-} C_i' K_i' A' \\ &\quad + APC_i' K_i' A' + Q + AP\sum_{j\in\mathcal{A}} C_j' R_j^{-1} C_j (P_{k-1,ji}^{na-} \\ &\quad + P_{k-1,ji}^{aa-} + P_{k-1,ji}^{nn-} C_i' K_i' + P_{k-1.ji}^{an-} C_i' K_i') A'. \end{aligned} \tag{35}$$

Since the attack starts from steady state, the initial value of recursions (30)–(33) is the steady-state value of $\mathbb{E}[(x_k - \hat{x}_{k,i})(x_k - \hat{x}_{k,j}^-)']$, which corresponds the solution of $X = A(I - K_i C_i)X(I - K_j C_j)'A' + Q + \delta_{ij} AK_i R_i K_i' A'$. The initial value of recursions (34)–(35) is the steady-state value of $\mathbb{E}[(x_k - \hat{x}_k)(x_k - \hat{x}_{k,i}^-)']$, which corresponds the solution of $X = A(I - KC)X(I - K_i C_i)'A' + Q + APC_i' K_i' A'$. ■

B. Proof of Theorem 2

Proof: When a subset $\mathcal{A}$ of the sensors are under the optimal innovation-based deception attack, according to the update rules used at the remote estimator in Algorithm 1, the state estimation error under GMM-based detection algorithm can be represented as

$$\begin{aligned} x_k - \tilde{x}_k^- &= A(x_{k-1} - \tilde{x}_{k-1}) + w_{k-1}, \\ x_k - \tilde{x}_k &= x_k - \tilde{x}_k^- - P_k\sum_{i\in\mathcal{A}} \gamma_{k,i}^{(1)} C_i' R_i^{-1}[C_i(x_k - \tilde{x}_k^-) \\ &\quad - C_i(x_k - \hat{x}_{k,i}^-) - C_i(x_k - \tilde{x}_{k,i}^-) - v_{k,i}] \\ &\quad - P_k\sum_{j\in\mathcal{S}} \gamma_{k,j}^{(1)} C_j' R_{k,j}^{-1}[C_j(x_k - \tilde{x}_k^-) + v_{k,j}] \\ &= (I - K_k C)(x_k - \tilde{x}_k^-) \\ &\quad + P_k\sum_{i\in\mathcal{A}} C_i' R_{k,i}^{-1} C_i[(x_k - \hat{x}_{k,i}^-) + (x_k - \tilde{x}_{k,i}^-)] \\ &\quad + P_k\sum_{i\in\mathcal{A}} C_i' R_{k,i}^{-1} v_{k,i} - P_k\sum_{j\in\mathcal{S}} C_j' R_{k,j}^{-1} v_{k,j}, \end{aligned} \tag{36}$$

where $K_k = P_k C' R_k^{-1}$, $R_k^{-1} = \mathrm{Diag}\{R_{k,1}^{-1}, \ldots, R_{k,N}^{-1}\} = \mathrm{Diag}\{R_1^{-1}\gamma_{k,1}^{(1)}, \ldots, R_N^{-1}\gamma_{k,N}^{(1)}\}$, $\gamma_{k,i}^{(1)}$ is calculated according to Algorithm 1, and P_k evolves according to (22). Similar to the proof of Theorem 1, it can be obtained that the remote estimation error covariance follows 21. One difference is the

recursions of $P^{an-}_{k,ci}$ and $P^{aa-}_{k,ci}$ now become

$$\begin{aligned} P^{an-}_{k,ci} &= A(I-K_{k-1}C)P^{an-}_{k-1,ci}(I-K_iC_i)'A'+Q \\ &\quad -\gamma^{(1)}_{k-1,i}AP_{k-1}C_i'K_i'A'+AP_{k-1}\sum_{j\in\mathcal{A}}C_j'R^{-1}_{k-1,j}C_j \\ &\quad \times(P^{nn-}_{k-1,ji}+P^{an-}_{k-1,ji})(I-K_iC_i)'A', \end{aligned} \tag{37}$$

$$\begin{aligned} P^{aa-}_{k,ci} &= A(I-K_{k-1}C)P^{aa-}_{k-1,ci}A'+\gamma^{(1)}_{k-1,i}AP_{k-1}C_i'K_i'A' \\ &\quad +Q+A(I-K_{k-1}C)P^{an-}_{k-1,ci}C_i'K_i'A' \\ &\quad +AP_{k-1}\sum_{j\in\mathcal{A}}C_j'R^{-1}_{k-1,j}C_j(P^{na-}_{k-1,ji}+P^{aa-}_{k-1,ji} \\ &\quad +P^{nn-}_{k-1,ji}C_i'K_i'+P^{an-}_{k-1.ji}C_i'K_i')A', \end{aligned} \tag{38}$$

since the remote estimator no longer adopts the steady-state gain to update its estimates. ■

C. Proof of Theorem 3

Proof: To evaluate the performance of Algorithm 1 under the optimal innovation-based attack, we denote $\hat{x}_i$ as a sample data from Gaussian distribution $\mathcal{N}(x_k,P_i)$ for all $i\in\mathcal{S}$ and $\tilde{x}_j$ as a sample data from Gaussian distribution $\mathcal{N}(x_k,\tilde{P}_j)$ for all $j\in\mathcal{A}$. Note that the EM algorithm maximizes likelihood and converges to a local optimum. Hence, we assume that the GMM-based detection algorithm converges to its optimal solution, i.e., the sample mean and sample variance of each cluster, which is denoted as $\mu^{(1)}=\frac{1}{N-M}\sum_{i\in\mathcal{S}}\hat{x}_i$, $\Sigma^{(1)}_i=P_i$, $\mu^{(2)}=\frac{1}{M}\sum_{j\in\mathcal{A}}\tilde{x}_j$, and $\Sigma^{(2)}=\frac{1}{M}\sum_{j\in\mathcal{A}}(\tilde{x}_j-\mu^{(2)})(\tilde{x}_j-\mu^{(2)})'$. In this case, the belief of being uncorrupted for each sensor $i\in\mathcal{S}$ can be obtained as

$$\begin{aligned} &\mathbb{E}[\gamma^{(1)}_i]= \\ &\quad\times\int_{j\in\mathcal{A}}\int_{i\in\mathcal{S}}\frac{\pi^{(1)}f(\hat{x}_i;\mu^{(1)},\Sigma^{(1)}_i)}{\pi^{(1)}f(\hat{x}_i;\mu^{(1)},\Sigma^{(1)}_i)+\pi^{(2)}f(\hat{x}_i;\mu^{(2)},\Sigma^{(2)})} \\ &\quad f(\hat{x}_i;x_k,P_i)\mathrm{d}\hat{x}_i f(\tilde{x}_j;x_k,\tilde{P}_j)\mathrm{d}\tilde{x}_j, \end{aligned} \tag{39}$$

where $\pi^{(1)}=\frac{N-M}{N}$, $\pi^{(2)}=\frac{M}{N}$, and the integral is taken with respect to every $i\in\mathcal{S}$ and every $j\in\mathcal{A}$. By changing variables, one has

$$\begin{aligned} &\mathbb{E}[\gamma^{(1)}_i] \\ &\quad=\int_{j\in\mathcal{A}}\int_{i\in\mathcal{S}}\frac{\pi^{(1)}f(y_i;\mu^{(1)}_y,\Sigma^{(1)}_i)}{\pi^{(1)}f(y_i;\mu^{(1)}_y,\Sigma^{(1)}_i)+\pi^{(2)}f(y_i;\mu^{(2)}_y,\Sigma^{(2)})} \\ &\quad f(y_i;0,P_i)\mathrm{d}y_i f(y_j;0,\tilde{P}_j)\mathrm{d}y_j, \end{aligned} \tag{40}$$

where $y_i=\hat{x}_i-x_k$, $\mu^{(1)}_y=\frac{1}{N-M}\sum_{i\in\mathcal{S}}y_i=\mu^{(1)}-x_k$, $y_j=\tilde{x}_j-x_k$ and $\mu^{(2)}_y=\frac{1}{M}\sum_{j\in\mathcal{A}}y_j=\mu^{(2)}-x_k$. Then, taking average over all the sensors in subset $\mathcal{S}$ yields the average belief of being uncontaminated,

$$\mathbb{E}[\gamma^{(1)}_{\mathcal{S}}]=\frac{1}{N-M}\sum_{i\in\mathcal{S}}\mathbb{E}[\gamma^{(1)}_i], \tag{41}$$

when sensor $i\in\mathcal{S}$, which is the same as (23). Similarly, the average belief that sensor i is under attack, $\mathbb{E}[\gamma^{(1)}_{\mathcal{A}}]$, when sensor $i\in\mathcal{A}$ can be obtained in the form of (24). ■

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and the anonymous reviewers for their suggestions which have improved the quality of the work.

REFERENCES

[1] K. Kim and P. R. Kumar, "Cyber–physical systems: A perspective at the centennial," *Proc. IEEE* (Special Centennial Issue), vol. 100, pp. 1287–1308, May 2012.
[2] A. Vempaty, O. Ozdemir, K. Agrawal, H. Chen, and P. K. Varshney, "Localization in wireless sensor networks: Byzantines and mitigation techniques," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1495–1508, Mar. 2013.
[3] J. P. Farwell and R. Rohozinski, "Stuxnet and the future of cyber war," *Survival*, vol. 53, no. 1, pp. 23–40, 2011.
[4] J. Slay and M. Miller, "Lessons learned from the Maroochy water breach," in *Proc. Int. Conf. Crit. Infrastructure Protection*, 2007, pp. 73–82.
[5] 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.
[6] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, 2011, Art. no. 13.
[7] J. Kim, L. Tong, and R. J. Thomas, "Subspace methods for data attack on state estimation: A data driven approach," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1102–1114, Mar. 2015.
[8] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *Proc. 1st Workshop Secure Control Syst., CPS Week*, 2010.
[9] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput.*, 2009, pp. 911–918.
[10] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Syst.*, vol. 35, no. 1, pp. 93–109, Feb. 2015.
[11] F. Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," in *Proc. 52nd IEEE Conf. Decis. Control*, 2013, pp. 1854–1859.
[12] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyber-attack on remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 4–13, Mar. 2017.
[13] S. Wu, Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal innovation-based deception attack on remote state estimation," in *Proc. Amer. Control Conf.*, 2017, pp. 3017–3022.
[14] Y. Li, L. Shi, and T. Chen, "Detection against linear deception attacks on multi-sensor remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 846–856, Sep. 2018.
[15] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.
[16] Y. Mo, J. P. Hespanha, and B. Sinopoli, "Resilient detection in the presence of integrity attacks," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 31–43, Jan. 2014.
[17] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of byzantine attacks," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 16–29, Jan. 2009.
[18] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Distributed Bayesian detection in the presence of byzantine data," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5250–5263, Oct. 2015.
[19] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1454–1467, Jun. 2014.
[20] M. Pajic *et al.*, "Robustness of attack-resilient state estimators," in *Proc. 5th Int. Conf. Cyber-Physical Syst.*, 2014, pp. 163–174.

[21] S. Mishra, Y. Shoukry, N. Karamchandani, S. N. Diggavi, and P. Tabuada, "Secure state estimation against sensor attacks in the presence of noise," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 49–59, Mar. 2017.

[22] 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 Trans. Autom. Control*, vol. 62, no. 10, pp. 4917–4932, Oct. 2017.

[23] D. Shi, Z. Guo, K. H. Johansson, and L. Shi, "Causality countermeasures for anomaly detection in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 63, no. 2, pp. 386–401, Feb. 2018.

[24] Y. Mo and E. Garone, "Secure dynamic state estimation via local estimators," in *Proc. IEEE 55th Conf. Decis. Control*, 2016, pp. 5073–5078.

[25] D. Han, Y. Mo, and L. Xie, "Resilience and performance analysis for state estimation against integrity attacks," *IFAC-PapersOnLine*, vol. 49, no. 22, pp. 55–60, 2016.

[26] D. Shi, R. J. Elliott, and T. Chen, "On finite-state stochastic modeling and secure estimation of cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 62, no. 1, pp. 65–80, Jan. 2017.

[27] J. He, P. Cheng, L. Shi, and J. Chen, "SATS: Secure average-consensus-based time synchronization in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6387–6400, Dec. 2013.

[28] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *Proc. IEEE Amer. Control Conf.*, 2015, pp. 2439–2444.

[29] D. Shi, T. Chen, and M. Darouach, "Event-based state estimation of linear dynamic systems with unknown exogenous inputs," *Automatica*, vol. 69, pp. 275–288, 2016.

[30] A. Banerjee, K. K. Venkatasubramanian, T. Mukherjee, and S. K. Gupta, "Ensuring safety, security, and sustainability of mission-critical cyber-physical systems," *Proc. IEEE*, vol. 100, no. 1, pp. 283–299, Jan. 2012.

[31] C. Berger and B. Rumpe, "Autonomous driving-5 years after the urban challenge: The anticipatory vehicle as a cyber-physical system," 2014, arXiv:1409.0413.

[32] I. Korhonen, J. Parkka, and M. Van Gils, "Health monitoring in the home of the future," *IEEE Eng. Med. Biol. Mag.*, vol. 22, no. 3, pp. 66–73, May/Jun. 2003.

[33] Y. Mo *et al.*, "Cyber–physical security of a smart grid infrastructure," *Proc. IEEE*, vol. 100, no. 1, pp. 195–209, 2012.

[34] A. Pouliezos and G. S. Stavrakakis, *Real Time Fault Monitoring of Industrial Processes*. Berlin, Germany: Springer Science & Business Media, 2013, vol. 12.

[35] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[36] K. J. Aström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton, NJ, USA: Princeton Univ. Press, 2010.

[37] G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis*. New York, NY, USA: Academic, 1977, vol. 26.

[38] B. D. Anderson and J. B. Moore, *Optimal Filtering*. Chelmsford, MA, USA: Courier Corp., 2012.

[39] R. K. Mehra and J. Peschon, "An innovations approach to fault detection and diagnosis in dynamic systems," *Automatica*, vol. 7, no. 5, pp. 637–640, 1971.

[40] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.

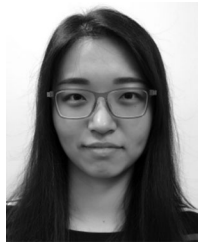
[41] C.-Z. Bai, V. Gupta, and F. Pasqualetti, "On Kalman filtering with compromised sensors: Attack stealthiness and performance bounds," *IEEE Trans. Autom. Control*, vol. 62, no. 12, pp. 6641–6648, Dec. 2017.

[42] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," in *Proc. Amer. Control Conf.*, 2015, pp. 195–200.

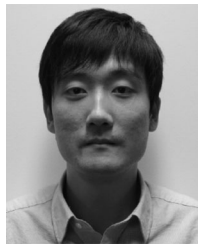
[43] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Worst-case stealthy innovation-based linear attack on remote state estimation," *Automatica*, vol. 89, pp. 117–124, 2018.

[44] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Berlin, Germany: Springer Science & Business Media, 2013.

[45] Q. Ge, T. Shao, Z. Duan, and C. Wen, "Performance analysis of the Kalman filter with mismatched noise covariances," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4014–4019, Dec. 2016.



Ziyang Guo received the B.Eng. degree (Hons.) from the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2014 and the Ph.D. degree from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2018. From September 2016 to December 2016, she was a visiting student in the School of Engineering and Applied Sciences, Harvard University. Her research interests include cyber-physical system security, networked state estimation, and wireless sensor networks.



Dawei Shi received the B.Eng. degree in electrical engineering and its automation from the Beijing Institute of Technology, Beijing, China, in 2008, and the Ph.D. degree in control systems from the University of Alberta, Edmonton, AB, Canada, in 2014. In December 2014, he was appointed as an Associate Professor at the School of Automation, Beijing Institute of Technology. From February 2017 to Jul 2018, he was with the Harvard John A. Paulson School of Engineering and Applied Sciences, as a Postdoctoral Fellow in bioengineering. Since July 2018, he has been with the School of Automation, Beijing Institute of Technology, where he is currently a Professor. His research focuses on analysis and synthesis of complex sampled-data control systems, with applications to biomedical engineering, robotics and motion systems.

Daniel E. Quevedo (S'97–M'05–SM'14) received Ingeniero Civil Electrónico and the M.Sc. degrees from the Universidad Técnica Federico Santa María, Chile, in 2000 and the Ph.D. degree from the University of Newcastle, Callaghan, NSW, Australia. He is the Head of the Chair of Automatic Control (Regelungs- und Automatisierungstechnik), Paderborn University, Germany. He was supported by a full scholarship from the alumni association during his time at the Universidad Técnica Federico Santa María and received several university-wide prizes upon graduating. He received the IEEE Conference on Decision and Control Best Student Paper Award in 2003 and was also a finalist in 2002. In 2009, he was awarded a five-year Research Fellowship from the Australian Research Council. He is an Associate Editor of the IEEE CONTROL SYSTEMS MAGAZINE, the Editor of the *International Journal of Robust and Nonlinear Control*, and is the Chair of the IEEE Control Systems Society Technical Committee on Networks & Communication Systems. His research interests include control of networked systems and power converters.

Ling Shi received the B.S. degree in electrical and electronic engineering from Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2002 and the Ph.D. degree in control and dynamical systems from California Institute of Technology, Pasadena, CA, USA, in 2008. He is currently an Associate Professor with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology. His research interests include cyber-physical systems security, networked control systems, sensor scheduling, and event-based state estimation. He was an Editorial Board Member of The European Control Conference 2013–2016. He has been a Subject Editor for the *International Journal of Robust and Nonlinear Control* since March 2015, an Associate Editor for the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS since July 2016, and an Associate Editor for the IEEE CONTROL SYSTEMS LETTERS since February 2017. He was also an Associate Editor for a Special Issue on Secure Control of Cyber Physical Systems in the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS in 2015–2017. He is the General Chair of the 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS 2018).