A Clustering-based Approach for Secure State Estimation

Ziyang Guo*, Dawei Shi†, Daniel E. Quevedo‡, Ling Shi*

Keywords: Secure Sate Estimation, Gaussian Mixture Model, Integrity Attack, Clustering

AMS subject classifications: 93E03, 93E10, 62H30

I. Introduction

Securing cyber-physical systems (CPS) is a problem of growing importance due to the widespread applications of such systems in critical infrastructures, e.g., smart grid, intelligent transportation and health monitoring system [1]. In this context, it is crucial to ensure the performance of state estimation, an indispensable aspect of CPS, in the presence of malicious attacks. With this motivation, we focus on the problem of securely estimating the state of a linear dynamical system from a set of noisy and maliciously corrupted sensor measurements in this work.

Several recent works focus on designing detection mechanisms or building attack-resilient estimators for a specific attack strategy [2]–[4]. However, it is difficult for a system designer to determine the specific attack type beforehand in many scenarios. Hence, the development of attack detection and secure estimation schemes which are applicable to different attack scenarios is necessary. In [5], the problem of attack detection and identification in CPS was investigated in system-theoretic and graph-theoretic approaches. In [6], the maximum number of tolerable attacks that allow accurate reconstruction of the system state was characterized. The above result was further extended to noisy systems in [7]. Since localizing the compromised sensors is intrinsically a combinatorial problem, a satisfiability modulo theory based algorithm was proposed in [8] and its scalability, soundness and completeness were analyzed. In remote estimation scenario, transfer entropy based causality countermeasures for integrity attacks were proposed in [9] and convex optimization based resilient estimators were investigated in [10].

To achieve the joint objective of attack localization and secure state estimation, we propose a clustering-based detection algorithm in this work. It is able to cluster the local state estimates autonomously and provide beliefs for different sensors, based on which measurements can be fused accordingly. When a subset of the sensors are under attack, we analyze the remote estimation error covariance for the proposed clustering-based detection algorithm and evaluate the performance of the proposed algorithm through the average belief. Moreover, the effectiveness of the proposed algorithm is verified under different attack scenarios.

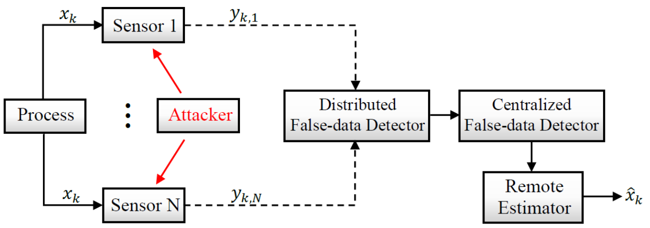

Fig. 1. System Block Diagram.

II. Problem Formulation

A. Process Model

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

$$x_{k+1} = Ax_k + w_k,$$
$$y_{k,i} = C_i x_k + v_{k,i},$$

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 $\mathbb{E}[w_k w_l'] = \delta_{kl} Q$ $(Q \geq 0)$, $\mathbb{E}[v_{k,i} v_{l,j}'] = \delta_{ij}\delta_{kl} R_i$ $(R_i > 0)$, $\mathbb{E}[w_k v_{l,i}'] = 0$, $\forall k, l \in \mathbb{N}$, $i, j = 1, 2, \dots, N$. 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.

B. Remote Estimator

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

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

the overall measurement can be represented as

$$y_k = Cx_k + v_k.$$

The work by Z. Guo and L. Shi is supported by a Hong Kong RGC GRF grant 16222716. The work by D. Shi is supported by Natural Science Foundation of China (61503027).

*: 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).

†: Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA (e-mail: dawei.shi@outlook.com).

‡: Department of Electrical Engineering (EIM-E), Paderborn University, Paderborn, Germany (email: dquevedo@ieee.org).

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, P_k^- and P_k 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

$$\begin{aligned}\hat{x}_k &= \hat{x}_k^- + P_k C' R^{-1}(y_k - C\hat{x}_k^-),\\ (P_k)^{-1} &= (P_{k-1}^-)^{-1} + C'R^{-1}C,\end{aligned}$$

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

To facilitate 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 [11]. We denote the steady-state values for local and centralized Kalman filter as

$$\begin{aligned}P_i &\triangleq \lim_{k\to+\infty} P_{k,i},\ P_i^- \triangleq \lim_{k\to+\infty} P_{k,i}^-,\\ P &\triangleq \lim_{k\to+\infty} P_k,\quad P^- \triangleq \lim_{k\to+\infty} P_k^-,\end{aligned}$$

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 of 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, false-data detectors are usually adopted to monitor system behavior and detect the existence of potential malicious attacks. Note that for 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 $k \neq l$ [11]. For 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 variance) of the innovation sequence are commonly used to diagnose the system anomalies [12]. Based on different information set, 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 variance of the 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:

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

where J_i is the detection window size of sensor i, 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. Note that $g_{k,i}$ satisfies the χ^2 distribution with mJ degrees of freedom. If $g_{k,i}$ exceeds the threshold δ_i, the detector will trigger an alarm and the measurement of sensor i will be dropped.

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

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

where g_k is χ^2 distributed with mNJ degrees of freedom. When g_k exceeds the threshold δ, the detector will trigger an alarm and all the measurements will be dropped.

D. Attack Model

Suppose that there exists a malicious attacker who is able to modify the measurement data. In practice, an attacker can launch such an 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. The index set of corrupted sensors is assumed to be time invariant. Without loss of generality, we also assume that the attack starts from time $k = 1$.

E. Problem of Interest

For the system described in the previous subsections, if no alarm is triggered at both the distributed and the centralized χ^2 detectors, the measurement data are believed to be reliable and fused at the remote estimator. If alarms are triggered at the distributed χ^2 detector, we simply remove those corrupted measurements and check 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, the remaining data will be discarded and only a time update is performed at the remote side to estimate system state. 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., reply attack [2], false-data injection attack [13] and innovation-based deception attack [14], the alarm triggered in the centralized χ^2 detector may result in a large performance degradation since the loss of uncontaminated information. To address this issue, it is necessary to develop an effective detection mechanism which is able to localize the compromised sensors and applicable to different attack scenarios.

III. Methodology

As discussed before, our interest lies in handling situations where the malicious attacker is able to deliberately design the corrupted data to bypass the distributed χ^2 detector but the centralized χ^2 detector fails to localize the compromised sensors. Inspired by the clustering algorithm used in machine learning, we propose a Gaussian-mixture-model-based detection algorithm for attack localization and secure state estimation.

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

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

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 for Gaussian random variables.

In our problem, the sensors are either uncorrupted or compromised, which leads to a 2-component mixture. It can be observed that if sensor i is uncorrupted, its local state estimate $\hat{x}_{k,i}$ is Gaussian distributed with fixed covariance P_i, i.e., $p(\hat{x}_{k,i}|\mathcal{G}_1) \sim \mathcal{N}(\mu_k^{(1)}, P_i)$. If sensor i is compromised by the attacker, then $\hat{x}_{k,i}$ may not have a Gaussian distribution, so we use the first and second moments to approximate its distribution, i.e., $p(\hat{x}_{k,i}|\mathcal{G}_2) \sim \mathcal{N}(\mu_k^{(2)}, \Sigma_k^{(2)})$. Since we consider a dynamic system, the time index k is added. Consequently, the mixture density for local state estimate $\hat{x}_{k,i}$ is obtained as

$$\begin{aligned} p(\hat{x}_{k,i}) &= \sum_{q=1}^{2} p(x|\mathcal{G}_q)\Pr(\mathcal{G}_q) \\ &= \pi_k^{(1)} f(\hat{x}_{k,i};\mu_k^{(1)},P_i) + \pi_k^{(2)} f(\hat{x}_{k,i};\mu_k^{(2)},\Sigma_k^{(2)}). \end{aligned}$$

At each time k, we adopt expectation-maximization (EM) algorithm [15] to find maximum likelihood estimates for the parameter $\Phi_k = \{\pi_k^{(q)}, \mu_k^{(q)}, \Sigma_k^{(2)}\}_{q=1}^2$ using the data $\mathcal{X}_k = \{\hat{x}_{k,i}\}_{i=1}^N$ and simultaneously fuse the measurement data with different weights. The log likelihood is given as

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

In general, Gaussian mixture model 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 [16]. 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, leading to a satisfactory estimation performance even in the presence of attacks.

References

[1] K. Kim and P. R. Kumar, "Cyber–physical systems: A perspective at the centennial," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1287–1308, 2012.

[2] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *47th Annual Allerton Conference on Communication, Control, and Computing*, 2009, pp. 911–918.

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

[4] Y. Li, L. Shi, and T. Chen, "Detection against linear deception attacks on multi-sensor remote state estimation," *IEEE Transactions on Control of Network Systems*, 2017.

[5] 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.

[6] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[7] S. Mishra, Y. Shoukry, N. Karamchandani, S. N. Diggavi, and P. Tabuada, "Secure state estimation against sensor attacks in the presence of noise," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 49–59, 2017.

[8] 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*, 2017.

[9] D. Shi, Z. Guo, K. H. Johansson, and L. Shi, "Causality countermeasures for anomaly detection in cyber-physical systems," *IEEE Transactions on Automatic Control*, 2017.

[10] Y. Mo and E. Garone, "Secure dynamic state estimation via local estimators," in *IEEE 55th Conference on Decision and Control*, 2016, pp. 5073–5078.

[11] B. D. Anderson and J. B. Moore, *Optimal Filtering*. Courier Corporation, 2012.

[12] 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.

[13] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *First Workshop on Secure Control Systems, CPS Week*, 2010.

[14] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyber-attack on remote state estimation," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 4–13, 2017.

[15] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[16] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.