Secure State Estimation: Optimal Guarantees against Sensor Attacks in the Presence of Noise

Shaunak Mishra*, Yasser Shoukry*, Nikhil Karamchandani*, Suhas Diggavi* and Paulo Tabuada*
*Electrical Engineering Department, University of California, Los Angeles

***Abstract*—Motivated by the need to secure cyber-physical systems against attacks, we consider the problem of estimating the state of a noisy linear dynamical system when a subset of sensors is arbitrarily corrupted by an adversary. We propose a secure state estimation algorithm and derive (optimal) bounds on the achievable state estimation error. In addition, as a result of independent interest, we give a coding theoretic interpretation for prior work on secure state estimation against sensor attacks in a noiseless dynamical system.**

I. Introduction

Cyber-physical systems (CPS) manage the vast majority of today's critical infrastructure and securing such CPS against malicious attacks is a problem of growing importance [1]. As a stepping stone towards securing complex CPS deployed in practice, several recent works have studied security problems in the context of linear dynamical systems [1], [2], [3], [4], [5], [6] leading to a fundamental understanding of how the system dynamics can be leveraged for security guarantees. With this motivation, in this paper we focus on securely estimating the state of a linear dynamical system from a set of noisy and maliciously corrupted sensor measurements. We restrict the sensor attacks to be sparse in nature, *i.e.*, an adversary can arbitrarily corrupt a subset of sensors in the system.

Prior work related to secure state estimation against sensor attacks in linear dynamical systems can be broadly categorized into three classes depending on the noise model for sensor measurements: 1) noiseless 2) bounded non-stochastic noise, and 3) Gaussian noise. For the noiseless setting, the work reported in [1], [2], [3] shows that, under a strong notion of observability, sensor attacks (modeled as a sparse attack vector) can always be detected and isolated, and hence the state of the system can be exactly estimated. In contrast, when the sensor measurements are affected by noise as well as maliciously corrupted, the problem of distinguishing between noise and attack vector arises. Results reported in [5], [6], [7] are representative of the second class: bounded non-stochastic noise. They provide sufficient conditions for distinguishing the sparse attack vector from bounded noise but do not guarantee the optimality of their estimation algorithm. The work reported in this paper falls in the third class: Gaussian noise. Prior work in this class includes [8], [9], [10], [11]. In [8], the analysis is restricted to detecting a class of sensor attacks called *replay* attacks (*i.e.*, attacks in which legitimate sensor outputs are replaced with outputs from previous time instants). In [9], the authors focus on the performance degradation of a scalar Kalman filter (*i.e.*, scalar state and a single sensor) when the sensor is under attack. Since they consider a single sensor setup, attack sparsity across multiple sensors is not studied, and in addition, they focus on an adversary whose objective is to degrade the estimation performance and stay undetected at the same time (thereby restricting the class of sensor attacks). In [10] and [11], robustification approaches for state estimation against sparse sensor attacks are proposed, but they lack optimality guarantees against arbitrary sensor attacks.

In contrast to prior work in the Gaussian noise setup, we consider a general linear dynamical system and give (optimal) guarantees on the achievable state estimation error against arbitrary sensor attacks. The following toy example is illustrative of the nature of the problem addressed in this paper and some of the ideas behind our solution.

Example 1: Consider a linear dynamical system with a scalar state $x(t)$ such that $x(t+1) = x(t) + w(t)$, where $w(t)$ is the process noise following a Gaussian distribution with zero mean and is instantiated i.i.d. over time. The system has three sensors (indexed by d) with outputs $y_d(t) = x(t) + v_d(t)$, where $v_d(t)$ is the sensor noise at sensor d. Similarly to the process noise, $v_d(t)$ is Gaussian distributed with zero mean and is instantiated i.i.d. over time. The sensor noise is also independent across sensors. Now, consider an adversary which can attack any one of the sensors and arbitrarily change its output. In the absence of sensor noise, it is trivial to detect such an attack since the two good sensors (not attacked by the adversary) will have the same output. Hence, a *majority* based rule on the outputs leads to the exact state. However, in the presence of sensor noise, even the good sensors may not have the same output and a simple majority based rule cannot be used for estimation. In this paper, we build on the intuition that we may still be able to identify sensors whose outputs can lead to a *good* state estimate by leveraging the noise statistics over a large enough time window. In particular, our approach for this example would be to hypothesize a subset of two sensors as good, and then check whether the outputs from the two sensors are *consistent* with the Kalman state estimate based on outputs from the same subset of sensors. Furthermore, we show in this paper that such an approach leads to the optimal state estimation error for the given adversarial setup.

In this paper, we generalize the Kalman filter based approach in the above example to a general linear dynamical system with sensor and process noise. Our main contributions can be listed as follows:

- We give optimal guarantees on the achievable state estimation error against arbitrary sensor attacks and propose an algorithm to achieve the same guarantees;

The work was supported by NSF grant 1136174 and DARPA under agreement number FA8750-12-2-0247.

- As a result of independent interest, we give a coding theoretic interpretation (alternate proof) for the necessary and sufficient conditions for secure state estimation in the absence of noise [2], [3], [6] (known as the sparse observability condition).

The remainder of this paper is organized as follows. Section II deals with the setup. The main results are stated in Section III. Section IV considers the simpler setting of a scalar state and illustrates the main ideas behind our estimation algorithm and Section V considers its generalization to a vector state. Finally, we discuss the coding theoretic view of the sparse observability condition [3] in Section VI.

II. Setup

A. System model

We consider a linear dynamical system with sensor attacks as shown below:

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{w}(t), \quad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{v}(t) + \boldsymbol{\phi}(t), \quad (1)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ denotes the state of the plant at time $t \in \mathbb{N}$, $\mathbf{w}(t) \in \mathbb{R}^n$ denotes the process noise at time t, $\mathbf{y}(t) \in \mathbb{R}^p$ denotes the output of the plant at time t and $\mathbf{v}(t) \in \mathbb{R}^p$ denotes the sensor noise at time t. The process noise $\mathbf{w}(t) \sim \mathcal{N}\left(\mathbf{0}, \sigma_w^2 \mathbf{I}_n\right)$, *i.e.*, $\mathbf{w}(t)$ is Gaussian distributed with zero mean and covariance matrix $\sigma_w^2 \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of dimension n and $\sigma_w \in \mathbb{R}$. Similarly, sensor noise $\mathbf{v}(t) \sim \mathcal{N}\left(\mathbf{0}, \sigma_v^2 \mathbf{I}_p\right)$. Both $\mathbf{v}(t)$ and $\mathbf{w}(t)$ are instantiated i.i.d. over time, and $\mathbf{v}(t)$ is independent of $\mathbf{w}(t)$.

The sensor attack vector $\boldsymbol{\phi}(t) \in \mathbb{R}^p$ in (1) is introduced by a k-adversary defined as follows. A k-adversary has access to any k out of the p sensors in the system. Specifically, let $\boldsymbol{\kappa} \subseteq \{1, 2, \dots p\}$ denote the set of attacked sensors (with $|\boldsymbol{\kappa}| = k$). The k-adversary can observe the actual outputs in the k attacked sensors and change them arbitrarily. Specifically, the output of an attacked sensor $j \in \boldsymbol{\kappa}$ can be expressed as

$$y_j(t) = \mathbf{c}_j^T \mathbf{x}(t) + v_j(t) + \phi_j(t), \quad (2)$$

where T denotes the matrix transpose operation, $\mathbf{c}_j^T$ is the jth row of $\mathbf{C}$, $v_j(t)$ is the noise at sensor j and $\phi_j(t)$ is the adversarial corruption introduced at sensor j. For $j \notin \boldsymbol{\kappa}$, $\phi_j(t) = 0$. The adversary's choice of $\boldsymbol{\kappa}$ is unknown but is assumed to be constant over time (static adversary). The adversary is assumed to have unbounded computational power, and knows the system parameters (*e.g.,* $\mathbf{A}$ and $\mathbf{C}$) and noise statistics (*e.g.,* σ_w^2 and σ_v^2). However, the adversary is limited to have only causal knowledge of the process noise and the sensor noise in good sensors (not attacked by the adversary). We discuss this assumption in more detail in Section II-C.

B. State estimation: prediction and filtering

In this paper, we address two state estimation problems: (1) state prediction and (2) state filtering.

In the state prediction problem, the goal is to estimate the state at time t based on outputs till time $t-1$. In the absence of sensor attacks, using a Kalman filter for predicting the state in (1) leads to the optimal (MMSE) error covariance asymptotically [12]. In particular, the Kalman filter update rule can be written as:

$$\hat{\mathbf{x}}(t+1) = \mathbf{A}\hat{\mathbf{x}}(t) + \mathbf{L}(t)\left(\mathbf{y}(t) - \mathbf{C}\hat{\mathbf{x}}(t)\right), \quad (3)$$

where $\hat{\mathbf{x}}(t+1)$ is the state estimate at time $t+1$ and $\mathbf{L}(t)$ is the Kalman filter gain. For a Kalman filter in steady state [12], the steady state gain satisfies $\mathbf{L}(t) = \mathbf{L}$. Also, we use $P_{opt,\mathbf{s}}$ to denote the trace of steady state (prediction) error covariance matrix [12] obtained by using a Kalman filter on a sensor subset $\mathbf{s} \subseteq \{1, 2, \dots p\}$.

In contrast to the prediction problem, the goal in the state filtering problem is to estimate the state at time t based on outputs till time t. In the absence of sensor attacks, a Kalman filter update rule similar to (3) can be used for the filtering problem [12], and we use $F_{opt,\mathbf{s}}$ to denote the trace of steady state (filtering) error covariance matrix obtained by using a Kalman filter on a sensor subset $\mathbf{s}$ (details in [13]).

C. Causal knowledge assumptions

At time t, the attack vector $\boldsymbol{\phi}(t)$ in (1) depends on the knowledge of the adversary at time t, and in this context, we limit the adversary's knowledge of the process and sensor noise along the lines of causality. In particular, for the prediction problem we assume the following for a k-adversary:

(A1) The adversary's knowledge at time t is statistically independent of $\mathbf{w}(t')$ for $t' > t$, *i.e.*, $\boldsymbol{\phi}(t)$ is statistically independent of $\{\mathbf{w}(t')\}_{t'>t}$;

(A2) For a *good* sensor $d \in \{1, 2, \dots p\} - \boldsymbol{\kappa}$, the adversary's knowledge at time t (and hence $\boldsymbol{\phi}(t)$) is statistically independent of $\{v_d(t')\}_{t'>t}$.

Intuitively, assumptions (A1) and (A2) limit the adversary to have only causal knowledge of the process noise and the sensor noise in good sensors (not attacked by the adversary). Note that, apart from (A1) and (A2), we do not impose any restrictions on the statistical properties, boundedness and the time evolution of the corruptions introduced by the k-adversary. In the filtering problem, we replace assumptions (A1) and (A2) with (A3) and (A4) as described below:

(A3) The adversary's knowledge at time t is statistically independent of $\mathbf{w}(t')$ for $t' \geq t$, *i.e.*, $\boldsymbol{\phi}(t)$ is statistically independent of $\{\mathbf{w}(t')\}_{t' \geq t}$;

(A4) For a good sensor $d \in \{1, 2, \dots p\} - \boldsymbol{\kappa}$, the adversary's knowledge at time t (and hence $\boldsymbol{\phi}(t)$) is statistically independent of $\{v_d(t')\}_{t' \geq t}$.

Clearly, (A3) is a stronger version of (A1), requiring $\boldsymbol{\phi}(t)$ to be independent of $\mathbf{w}(t)$. Similarly, (A4) is a stronger version of (A2).

D. Sparse observability condition

For the matrix pair $(\mathbf{A}, \mathbf{C})$, the observability matrix $\mathbf{O}$ with observability index μ is defined as shown below:

$$\mathbf{O} = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{\mu-1} \end{bmatrix}. \quad (4)$$

In this context, a linear dynamical system, characterized by the pair $(\mathbf{A},\mathbf{C})$, is said to be observable if there exists a positive integer μ such that $\mathbf{O}$ has full column rank. In the absence of sensor and process noise, the conditions under which state estimation can be done despite sensor attacks have been studied in [2], [3], [6]. In particular, a linear dynamical system as shown in (1) is called θ-sparse observable if for every subset $\mathbf{s} \subseteq \{1,\dots p\}$ of size θ, the pair $(\mathbf{A},\mathbf{C_s})$ is observable (where $\mathbf{C_s}$ is formed by the rows of $\mathbf{C}$ corresponding to sensors indexed by the elements of $\mathbf{s}$). Also, θ is the smallest positive integer to satisfy the above observability property. The condition:

$$\theta \leq p - 2k, \tag{5}$$

is necessary and sufficient for *exact* state estimation against a k-adversary in the absence of process and sensor noise [3]; we will refer to this condition as the sparse observability condition. We provide a coding theoretic interpretation for the same in Section VI.

III. Main results

We first state our achievability result followed by an impossibility result.

*Theorem 1 (**Achievability**):* Consider the linear dynamical system defined in (1) satisfying the sparse observability condition (5) against a k-adversary. Assuming (A1) and (A2), and a time window $G = \{t_1, t_1+1, \dots t_1+N-1\}$ for the state prediction problem, the following bound on the prediction error is achievable against a k-adversary. For any $\varepsilon > 0$ and $\delta > 0$, there exists a large enough N such that:

$$\mathbb{P}\left(\frac{1}{N} \sum_{t \in G} \mathbf{e}^T(t)\mathbf{e}(t) \leq \max_{\mathbf{s} \subset \{1,2,\dots p\},\, |\mathbf{s}| = p-k} (P_{opt,\mathbf{s}}) + \varepsilon \right) \geq 1 - \delta, \tag{6}$$

where $\mathbf{e}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t)$ is the estimation error for the state estimate $\hat{\mathbf{x}}(t)$. In other words, with high probability (w.h.p.), the bound $\limsup_{N\to\infty} \frac{1}{N} \sum_{t \in G} \mathbf{e}^T(t)\mathbf{e}(t) \leq \max_{\mathbf{s} \subset \{1,2,\dots p\},\, |\mathbf{s}| = p-k} (P_{opt,\mathbf{s}})$ is achievable. Similarly, for the state filtering problem, assuming (A3) and (A4) against a k-adversary, the following bound on the corresponding filtering error $\mathbf{e}(t)$ is achievable w.h.p.:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t \in G} \mathbf{e}^T(t)\mathbf{e}(t) \leq \max_{\mathbf{s} \subset \{1,2,\dots p\},\, |\mathbf{s}| = p-k} (F_{opt,\mathbf{s}}). \tag{7}$$

The achievability in Theorem 1 is through our proposed algorithms, which we discuss in the following sections. The impossibility result can be stated as follows.

*Theorem 2 (**Impossibility**):* Consider the linear dynamical system defined in (1) and an oracle MMSE estimator that has knowledge of $\boldsymbol{\kappa}$, *i.e.*, the set of sensors attacked by a k-adversary. Then, there exists an attack sequence $\boldsymbol{\phi}(t)$ such that the trace of the prediction error covariance of the oracle estimator is bounded from below as follows:

$$tr\left(\mathbb{E}\left(\mathbf{e}(t)\mathbf{e}^T(t)\right)\right) \geq P_{opt,\mathbf{s}}, \tag{8}$$

where $\mathbf{e}(t)$ above is the oracle estimator's prediction error and $\mathbf{s} = \{1,2,\dots p\} - \boldsymbol{\kappa}$. Similarly, for the filtering problem,

$$tr\left(\mathbb{E}\left(\mathbf{e}(t)\mathbf{e}^T(t)\right)\right) \geq F_{opt,\mathbf{s}}. \tag{9}$$

Proof: Consider the attack scenario where the outputs from all attacked sensors are equal to zero, *i.e.*, the corruption $\phi_j(t) = -\mathbf{c}_j^T\mathbf{x}(t) - v_j(t),\ \forall j \in \boldsymbol{\kappa}$. Hence, the information collected from the attacked sensors cannot enhance the estimation performance. Accordingly, the estimation performance from the remaining sensors is the best one can expect to achieve.

Clearly, for the adversary's *best* choice of $\boldsymbol{\kappa}$, the guarantees given in our achievability match the impossibility bound (in an empirical average sense), and hence, we consider our guarantees *optimal*. We measure the performance of our proposed algorithms in terms of empirical average (and not expectation) since the resultant error in the presence of attacks may not be ergodic.

IV. Secure state estimation: scalar state

In this section, we illustrate the main ideas behind our general scheme in the simpler setting of estimating a scalar state variable against a k-adversary. In particular, we focus on the state prediction problem for the system in (1) when the state is a scalar and there are $p \geq 2k+1$ sensors (*i.e.*, 1-sparse observability condition against k-adversary). For clarifying the presence of scalar terms in our analysis, we use the scalar version (regular instead of bold face) of the notation developed in Section II, *i.e.*, $x(t)$ for the plant's state, $\hat{x}(t)$ for the estimate, and $y_d(t) = c_d x(t) + v_d(t)$ for the output of a good sensor $d \in \{1,2,\dots p\} - \boldsymbol{\kappa}$. We first describe our proposed algorithm for a time window $G = \{t_1, t_1+1, \dots t_1+N-1\}$ of size N, and then analyze its performance.

Secure scalar state prediction algorithm: Considering a time window G, Algorithm 1 shows the secure state prediction algorithm for the case when the state is a scalar. The algorithm runs a bank of $\binom{p}{p-k}$ Kalman filters in parallel; one Kalman filter associated with each distinct set of $p-k$ sensors. For each distinct set $\mathbf{s}$ of $p-k$ sensors, the corresponding Kalman filter fuses all the measurements from these sensors in order to calculate (prediction) estimate $\hat{x}_{\mathbf{s}}(t)$. Using the calculated estimate $\hat{x}_{\mathbf{s}}(t)$, we calculate the individual residues for each sensor as shown in (10). The algorithm, then, exhaustively searches for the set $\mathbf{s}$ of $p-k$ sensors which satisfy the residue test shown in (11). If a set $\mathbf{s}^\star$ satisfies the residue test, it is declared *good* and the corresponding Kalman estimate $\hat{x}_{\mathbf{s}^\star}(t)$ is used as the state estimate for the given time window. Intuitively, the residue test checks if the outputs from a given sensor set $\mathbf{s}$ are *consistent* with the corresponding Kalman estimate over the time window G.

Performance analysis: Consider the set $\mathbf{s}$ of $p-k$ sensors which are not attacked by the k-adversary. Assuming that the Kalman filter corresponding to set $\mathbf{s}$ is in steady state, it can be shown that $\mathbb{E}\left(r_d^2(t)\right) = c_d^2 P_{opt,\mathbf{s}} + \sigma_v^2,\ \forall d \in \mathbf{s}$ [12] (where residue $r_d(t)$ is as defined in (10)). For large enough N, due to the (strong) law of large numbers (LLN), the residue test will be satisfied w.h.p. for at least this set of good sensors. This ensures that w.h.p., the algorithm will not return an empty set. Also, the estimate $\hat{x}_{\mathbf{s}}(t)$ from this set of good sensors trivially achieves the error bound (6). But, since the algorithm can return any set of size $p-k$ which satisfies the residue test, it may be possible that some of the sensors in the returned set are corrupt. In the remainder of our analysis, we show that for *any* set returned by the algorithm, the corresponding Kalman estimate achieves (6).

Algorithm 1 SECURE STATE PREDICTION - SCALAR CASE

1: Enumerate all sets $\mathbf{s} \in \mathbf{S}$ such that:
$$\mathbf{S} = \{\mathbf{s} | \mathbf{s} \subset \{1,2,\dots p\},\ |\mathbf{s}| = p-k\}.$$
2: For each $\mathbf{s} \in \mathbf{S}$, run a Kalman filter that uses all sensors indexed by $\mathbf{s}$ and returns estimate $\hat{x}_{\mathbf{s}}(t) \in \mathbb{R}$.
3: For each $\mathbf{s} \in \mathbf{S}$, calculate the residues for all sensors $d \in \mathbf{s}$ over a time window $G = \{t_1, t_1+1, \dots t_1+N-1\}$ as:
$$r_d(t) = y_d(t) - c_d \hat{x}_{\mathbf{s}}(t) \qquad \forall d \in \mathbf{s}, \quad \forall t \in G. \tag{10}$$
4: Pick the set $\mathbf{s}^\star \in \mathbf{S}$ which satisfies the following residue test:
$$\frac{1}{N}\sum_{t\in G} r_d^2(t) \le c_d^2 P_{opt,\mathbf{s}^\star} + \sigma_v^2 + \varepsilon \qquad \forall d \in \mathbf{s}^\star, \tag{11}$$
where $\varepsilon \ge 0$ is a design parameter and can be made arbitrarily small for large enough N.
5: Return $\mathbf{s}^\star$ and $\hat{x}(t) := \hat{x}_{\mathbf{s}^\star}(t) \quad \forall t \in G$.

Suppose the algorithm returns a set $\mathbf{s}$ of $p-k$ sensors. There is definitely one good sensor (say sensor d) in this set because there can be at most k attacked sensors and $p-k > k$. Since the residue test is satisfied for this sensor, we have the following constraint:

$$\begin{aligned}\frac{1}{N}\sum_{t\in G} r_d^2(t) &\overset{(a)}{=} \frac{1}{N}\sum_{t\in G}\left(c_d x(t) + v_d(t) - c_d\hat{x}_{\mathbf{s}}(t)\right)^2 \\ &= \frac{1}{N}\sum_{t\in G}\left(c_d e(t) + v_d(t)\right)^2 \\ &= \frac{c_d^2}{N}\sum_{t\in G} e^2(t) + \frac{1}{N}\sum_{t\in G} v_d^2(t) + \frac{2c_d}{N}\sum_{t\in G} e(t)v_d(t) \\ &\overset{(b)}{\le} c_d^2 P_{opt,\mathbf{s}} + \sigma_v^2 + \varepsilon, \end{aligned} \tag{12}$$

where (a) follows from $y_d(t) = c_d x(t) + v_d(t)$ for a good sensor d and (b) follows from the residue test. The error $e(t)$ above is the state estimation (prediction) error at time t (in the presence of a k-adversary) when $\hat{x}_{\mathbf{s}}(t)$ is used as the state estimate. Using LLN, we can make an additional simplification as follows. For any $\varepsilon > 0$, there exists a large enough N such that:

$$\frac{c_d^2}{N}\sum_{t\in G} e^2(t) + \frac{2c_d}{N}\sum_{t\in G} e(t)v_d(t) \tag{13}$$

$$\overset{(a)}{\le} c_d^2 P_{opt,\mathbf{s}} + \left|\sigma_v^2 - \frac{1}{N}\sum_{t\in G} v_d^2(t)\right| + \varepsilon \overset{(b)}{\le} c_d^2 P_{opt,\mathbf{s}} + 2\varepsilon, \tag{14}$$

where (a) follows from (12), and (b) follows w.h.p. due to LLN. Our next step will be to show that the cross term $\frac{2c_d}{N}\sum_{t\in G} e(t)v_d(t)$ in (13) is vanishingly small w.h.p. as $N \to \infty$; this leads to the required bound on $\frac{1}{N}\sum_{t\in G} e^2(t)$ using (14). We do so in two steps: first we show that the mean of the cross term $\frac{2c_d}{N}\sum_{t\in G} e(t)v_d(t)$ is zero and then show that its variance is vanishingly small as $N \to \infty$.

The mean of the cross term $\frac{2c_d}{N}\sum_{t\in G} e(t)v_d(t)$ can be computed as shown below:

$$\mathbb{E}\left(\frac{2c_d}{N}\sum_{t\in G} e(t)v_d(t)\right) \overset{(a)}{=} \frac{2c_d}{N}\sum_{t\in G}\mathbb{E}\left(e(t)\right)\mathbb{E}\left(v_d(t)\right) = 0, \tag{15}$$

where (a) follows from the independence of $e(t)$ from $v_d(t)$ (due to assumption (A2), $\hat{x}_{\mathbf{s}}(t)$ is independent of good sensor noise $v_d(t)$ despite sensor attacks). Also, using (15) and taking the expectation in (13):

$$\mathbb{E}\left(\frac{1}{N}\sum_{t\in G} e^2(t)\right) \le P_{opt,\mathbf{s}} + \frac{2\varepsilon}{c_d^2}. \tag{16}$$

As the final step in our analysis, we will now show that the variance of cross term $\frac{2c_d}{N}\sum_{t\in G} e(t)v_d(t)$ is vanishingly small as $N \to \infty$. For any $\varepsilon_1 > 0$, there exists a large enough N such that:

$$\begin{aligned}&\mathbb{E}\left(\left(\frac{1}{N}\sum_{t\in G} e(t)v_d(t)\right)^2\right) \\ &= \frac{\sum_{t\in G}\mathbb{E}\left(e^2(t)v_d^2(t)\right)}{N^2} + \frac{2}{N^2}\sum_{t,\,t'\in G,\,t<t'}\mathbb{E}\left(e(t)v_d(t)e(t')v_d(t')\right) \\ &\overset{(a)}{=} \frac{1}{N^2}\sum_{t\in G}\mathbb{E}\left(e^2(t)\right)\mathbb{E}\left(v_d^2(t)\right) \\ &\quad + \frac{2}{N^2}\sum_{t,\,t'\in G,\,t<t'}\mathbb{E}\left(e(t)v_d(t)e(t')\right)\mathbb{E}\left(v_d(t')\right) \\ &= \frac{\sigma_v^2}{N}\mathbb{E}\left(\frac{\sum_{t\in G} e^2(t)}{N}\right) \overset{(b)}{\le} \varepsilon_1, \end{aligned} \tag{17}$$

where (a) follows from the independence of $e(t)$ from $v_d(t)$ and the independence of $v_d(t')$ from $e(t)v_d(t)e(t')$ (for $t' > t$), (b) follows from (16). The above result implies that the cross term $\frac{2c_d}{N}\sum_{t\in G} e(t)v_d(t)$ (with zero mean) has vanishingly small variance as $N \to \infty$. As a result, using Chebyshev's inequality and (14), we have the error bound (6).

V. SECURE STATE ESTIMATION: VECTOR STATE

In this section, we consider the state estimation problem (against a k-adversary) for the general linear dynamical system described in (1), when the state is a vector. We focus on the prediction problem in this section; see [13] for the filtering problem. We assume that the system is θ-sparse observable such that it satisfies the sparse observability condition (5) against a k-adversary. We first introduce some additional notation required for our proposed algorithm.

Additional notation: Consider a set $\mathbf{s}$ of $p-k$ sensors. Such a set has $\binom{p-k}{\theta}$ sensor subsets of size θ, and we index these subsets of $\mathbf{s}$ by i. Due to the θ-sparse observability condition, each subset i forms an observable pair $(\mathbf{A}, \mathbf{C}_i)$ with observability matrix $\mathbf{O}_i$ and observability index μ_i; $\mathbf{C}_i$ is formed by rows of $\mathbf{C}$ corresponding to subset i of $\mathbf{s}$. We define matrices $\mathbf{J}_i$ and $\mathbf{M}_{\mu_i}$ as shown below:

$$\mathbf{J}_i = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{C}_i & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{C}_i\mathbf{A} & \mathbf{C}_i & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_i\mathbf{A}^{\mu_i-2} & \mathbf{C}_i\mathbf{A}^{\mu_i-3} & \dots & \mathbf{C}_i \end{bmatrix},\ \mathbf{M}_{\mu_i} = \sigma_w^2\mathbf{J}_i\mathbf{J}_i^T + \sigma_v^2\mathbf{I}_{\mu_i}. \tag{18}$$

The pseudo-inverse of $\mathbf{O}_i$ is denoted by $\mathbf{O}_i^\dagger$. The output from sensor subset i (of size θ) at time t is denoted by $\mathbf{y}_i(t) \in \mathbb{R}^\theta$. We consider the state estimation problem for a time window G

Algorithm 2 SECURE STATE PREDICTION - VECTOR CASE

1: Enumerate all sets $\mathbf{s} \in \mathbf{S}$ such that:
$$\mathbf{S} = \{\mathbf{s} | \mathbf{s} \subset \{1,2,\dots p\},\ |\mathbf{s}| = p-k\}.$$
2: For each $\mathbf{s} \in \mathbf{S}$, run a Kalman filter that uses all sensors indexed by $\mathbf{s}$ and returns estimate $\hat{\mathbf{x}}_{\mathbf{s}}(t) \in \mathbb{R}^n$.
3: For each set $\mathbf{s} \in \mathbf{S}$, enumerate all subsets of size θ and index them by i. Let μ_i be the observability index associated with sensor subset i. For each subset i of $\mathbf{s}$ (subset of size θ), calculate the *block* residue:
$$\mathbf{r}_i(t) = \begin{bmatrix} \mathbf{y}_i(t) \\ \mathbf{y}_i(t+1) \\ \vdots \\ \mathbf{y}_i(t+\mu_i-1) \end{bmatrix} - \mathbf{O}_i \hat{\mathbf{x}}_{\mathbf{s}}(t) \quad \forall t \in G.$$
4: Pick the set $\mathbf{s}^\star \in \mathbf{S}$ which satisfies the following block residue test for each subset i of $\mathbf{s}^\star$ (subset of size θ). Partition G into μ_i groups $G_0, G_1, \dots G_{\mu_i - 1}$ of size N_B such that $G_l = \{t | ((t-t_1)\ mod\ \mu_i) = l\}$ and check that for each G_l:
$$\frac{1}{N_B} \sum_{t \in G_l} tr\left(\mathbf{O}_i^\dagger \mathbf{r}_i(t) \mathbf{r}_i^T(t) \mathbf{O}_i^{\dagger T}\right) \leq P_{opt,\mathbf{s}^\star} + tr\left(\mathbf{O}_i^\dagger \mathbf{M}_{\mu_i} \mathbf{O}_i^{\dagger T}\right) + \varepsilon, \quad (19)$$
where $\varepsilon \geq 0$ is a design parameter which can be made arbitrarily small for large enough N_B.
5: Return $\mathbf{s}^\star$ and $\hat{\mathbf{x}}(t) := \hat{\mathbf{x}}_{\mathbf{s}^\star}(t) \quad \forall t \in G$.

of size N and assume without loss of generality that μ_i divides N such that $\mu_i N_B = N$.

Secure state prediction algorithm: Similar to the scalar setting, Algorithm 2 runs a bank of $\binom{p}{p-k}$ Kalman filters in parallel. For each distinct set $\mathbf{s}$ of $p-k$ sensors, the corresponding Kalman filter fuses all the measurements from these sensors in order to calculate an estimate $\hat{\mathbf{x}}_{\mathbf{s}}(t)$. For a sensor set $\mathbf{s}$ of size $p-k$ to satisfy the block residue test, each of its $\binom{p-k}{\theta}$ subsets should satisfy (19) for each group G_l. If a set $\mathbf{s}^\star$ satisfies the residue test, it is declared good and the corresponding Kalman estimate $\hat{\mathbf{x}}_{\mathbf{s}^\star}(t)$ is used as the state estimate for the given time window. Intuitively, the residue test checks if the outputs from every *observable* sensor subset of size θ within set $\mathbf{s}$ are *consistent* with the corresponding Kalman estimate over the time window G (see [13] for performance analysis).

VI. SPARSE OBSERVABILITY: CODING THEORETIC VIEW

In this section, we revisit the sparse observability condition (5) against a k-adversary and give a coding theoretic interpretation for the same. We first describe our interpretation for a linear system, and then discuss how it can be generalized for non-linear systems. Consider the linear dynamical system in (1) without the process and sensor noise. If the system's initial state is $\mathbf{x}(0) \in \mathbb{R}^n$ and the system is θ-sparse observable, then clearly in the absence of sensor attacks, by observing the outputs from any θ out of p sensors for n time instants ($t = 0,1,\dots n-1$) we can exactly recover $\mathbf{x}(0)$ and hence, *exactly* estimate the state of the plant. A coding theoretic view of this can be given as follows. Consider the outputs from sensor $d \in \{1,2,\dots p\}$ for n time instants as a symbol $\mathcal{Y}_d \in \mathbb{R}^n$. Thus, in the (symbol) observation vector $\mathcal{Y} = [\mathcal{Y}_1 \quad \mathcal{Y}_2 \dots \mathcal{Y}_p]$, any θ symbols are sufficient to recover the initial state $\mathbf{x}(0)$. Now, let us consider the case of a k-adversary which can arbitrarily corrupt any k sensors. In the coding theoretic view, this corresponds to arbitrarily corrupting any k (out of p) symbols in the observation vector. Recovery of the initial state despite such corruptions depends on the (symbol) Hamming distance between the observation vectors corresponding to two distinct initial states (say $\mathbf{x}^{(1)}(0)$ and $\mathbf{x}^{(2)}(0)$ with $\mathbf{x}^{(1)}(0) \neq \mathbf{x}^{(2)}(0)$). In the case of θ-sparse observability, at most $\theta - 1$ symbols in observation vectors corresponding to $\mathbf{x}^{(1)}(0)$ and $\mathbf{x}^{(2)}(0)$ can be identical; if θ of the symbols are identical, this would imply $\mathbf{x}^{(1)}(0) = \mathbf{x}^{(2)}(0)$. Hence, the Hamming distance between the observation vectors corresponding to $\mathbf{x}^{(1)}(0)$ and $\mathbf{x}^{(2)}(0)$ is at least $p - \theta + 1$ symbols. This implies that we can correct up to $k < \frac{p-\theta+1}{2}$ corruptions (equivalent to $\theta \leq p - 2k$), which is precisely the sparse observability condition required against a k-adversary (details in [13]). By analogously defining θ-sparse observability for a (noiseless) non-linear system, the same interpretation holds for the non-linear case; this leads to an alternative proof for the necessary and sufficient conditions for secure state estimation in any noiseless dynamical system.

REFERENCES

[1] F. Pasqualetti, F. Dorfler, and F. Bullo, "Control-theoretic methods for cyber-physical security," *IEEE Control Systems Magazine*, Aug. 2014, to appear. [Online]. Available: http://motion.me.ucsb.edu/pdf/2013u-pdb.pdf

[2] 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, June 2014.

[3] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *arXiv pre-print*, Sep. 2013. [Online]. Available: http://arxiv.org/abs/1309.3511

[4] S. Mishra, N. Karamchandani, P. Tabuada, and S. Diggavi, "Secure state estimation and control using multiple (insecure) observers," in *IEEE Conference on Decision and Control (CDC)*, 2014.

[5] 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," *arXiv pre-print*, Dec. 2014.

[6] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *American Control Conference (ACC)*, 2015.

[7] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. Pappas, "Robustness of attack-resilient state estimators," in *ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, 2014.

[8] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Allerton Conference on Communication, Control, and Computing*, 2009.

[9] C.-Z. Bai and V. Gupta, "On kalman filtering in the presence of a compromised sensor: fundamental performance bounds," in *American Control Conference (ACC)*, 2014.

[10] J. Mattingley and S. Boyd, "Real-time convex optimization in signal processing," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 50–61, May 2010.

[11] S. Farahmand, G. B. Giannakis, and D. Angelosante, "Doubly robust smoothing of dynamical processes via outlier sparsity constraints," *IEEE Trans. on Signal Processing*, vol. 59, no. 10, pp. 4529–4543, Oct. 2011.

[12] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[13] S. Mishra, Y. Shoukry, N. Karamchandani, S. Diggavi, and P. Tabuada, "Secure state estimation: optimal guarantees against sensor attacks in the presence of noise," *arXiv pre-print*, 2015. [Online]. Available: http://arxiv.org/abs/1504.05566