Resilient State Estimation against Switching Attacks on Stochastic Cyber-Physical Systems

Sze Zheng Yong [a] Minghui Zhu [b] Emilio Frazzoli [a]

Abstract— **In this paper, we address the resilient state estimation problem for some relatively unexplored security issues for cyber-physical systems, namely switching attacks and the presence of stochastic process and measurement noise signals, in addition to attacks on actuator and sensor signals. We model the systems under attack as hidden mode stochastic switched linear systems with unknown inputs and propose the use of the multiple model inference algorithm developed in [1] to tackle these issues. We also furnish the algorithm with the lacking asymptotic analysis. Moreover, we characterize fundamental limitations to resilient estimation (e.g., upper bound on the number of tolerable attacks) and discuss the issue of attack detection under this framework. Simulation examples of switching attacks on benchmark and power systems show the efficacy of our approach to recover unbiased state estimates.**

I. Introduction

Cyber-physical systems (CPS) are systems in which computational and communication elements collaborate to control physical entities. Such systems include the power grid, autonomous vehicles, medical devices, etc. Most of these systems are *safety-critical* and if compromised or malfunctioning, can cause serious harm to the controlled physical entities and the people operating or utilizing them. Recent incidents of attacks on CPS, e.g., the Maroochy water breach, the StuxNet computer worm and various industrial security incidents [2], [3], highlight a need for CPS security and for new designs of resilient estimation and control.

Much of the early research focus has been on the characterization of undetectable attacks and on attack detection and identification techniques, which range from a simple application of data time-stamps in a previous work [4] to hypothesis tests using residuals (e.g., [5]–[8]). However, the ability to reliably estimate the true system states despite attacks is just as desirable, if not more than purely attack detection; thus, this problem has garnered considerable interest in recent years because the availability of resilient state estimates would, among others, allow for continued operation with the same controllers as in the case without attacks or for locational marginal pricing of electricity based on the real unbiased state information despite attacks.

Literature review. For deterministic linear systems under actuator and sensor signal attacks (e.g., via data injection [5]–[7]), the resilient state estimation problem has been mapped onto an ℓ_0 optimization problem, which is NP-hard [7], [9]; thus, a relaxation of the problem to a convex problem is considered in [9]. A further extension [10] computes a worst-case bound on the state estimate error in the presence of additive modeling errors with known bounds, but the optimization problem remains NP-hard. More importantly, these approaches do not apply in the presence of additive stochastic (unbounded) noise signals, which is one of the security issues we consider in this paper.

In addition, attacks that exploit the switching vulnerability of CPS or that alter its network topology have been recently identified as a serious CPS security concern. Some instances of such vulnerability are attacks on the circuit breakers of a smart grid [11] or on the logic mode (e.g., failsafe mode) of a traffic infrastructure [12], on the meter/sensor data network topology [13] and on the power system network topology [8]. However, to the best of our knowledge, no resilient state estimators for dynamic systems have been developed to deal with this new class of attacks.

Another set of relevant literature pertains to that of simultaneous input and state estimation (e.g., [14]–[16]). Of particular importance are the stability and optimality properties that are investigated in detail in [16], as well as the relationship between strong detectability and filter existence that is recently discovered in a related work [17]. Inspired by the multiple model approach (see, e.g., [18], [19] and references therein), our previous work [1] introduced an inference algorithm that estimates hidden modes, unknown inputs and states simultaneously, which we now propose as the key tool to achieve resilient estimation.

Contributions. In this paper, we propose a resilient state estimation algorithm that solves some previously unaddressed issues in ensuring secure estimation of cyber-physical systems: (i) *switching attacks* (attacks of switching mechanisms altering system- and data-level network topologies *and* time-varying attack strategies), and (ii) presence of stochastic process and measurement noise signals, in addition to the commonly studied (iii) actuator and sensor signal attacks. We model cyber-physical systems under attack as hidden mode stochastic switched linear systems with unknown inputs and hence, the inference algorithm developed in a previous work [1] for such systems can be applied for asymptotically recovering unbiased state estimates (i.e., resilient state estimates). We then study the asymptotic behavior of the approach in [1] and provide sufficient conditions for *asymptotically* achieving convergence to the true model (*consistency*), or to the closest model according to some information-theoretic measure (*convergence*). In addition, we characterize funda-

[a] S.Z Yong and E. Frazzoli are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA (e-mail: szyong@mit.edu, frazzoli@mit.edu).

[b] M. Zhu is with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA, USA (e-mail: muz16@psu.edu).

mental security limitations to resilient estimation, such as the upper bound on the number of tolerable attacks, and discuss the subject of attack detection associated with our approach.

II. Motivating Example

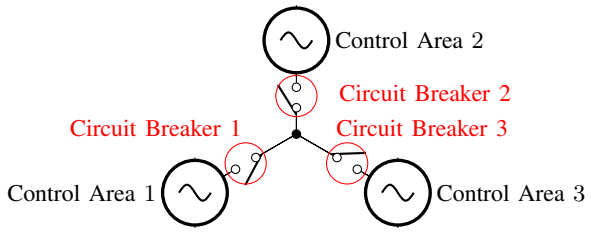

Fig. 1: Example of a three-area power station in a radial topology (corresponding to *node/bus attack)*.

To motivate the problem of resilient state estimation of stochastic cyber-physical systems under switching attacks, we consider a *power system* [20] with multiple control areas, each consisting of generators and loads, with tie-lines providing interconnections between areas (see example of a 3-area system in Figure 1). A simplified model of the control areas and the tie-lines is given by (see also parameter definitions in [20, Chap. 10]):

Control area i: ($i \in \{1, 2, \dots, N_{ca}\}$)

$$\begin{aligned} \frac{d\Delta\omega_i}{dt} + \frac{D_i\Delta\omega_i}{M_i} - \frac{\Delta P_{mech_i}}{M_i} + \frac{\sum_{j\neq i}\Delta P_{tie}^{ij}}{M_i} &= -\frac{\Delta P_{L_i}}{M_i}, \\ \frac{d\Delta P_{mech_i}}{dt} + \frac{\Delta P_{mech_i}}{T_{CH_i}} - \frac{\Delta P_{v_i}}{T_{CH_i}} &= 0, \\ \frac{d\Delta P_{v_i}}{dt} + \frac{\Delta P_{v_i}}{T_{G_i}} + \frac{\Delta\omega_i}{R_i^f T_{G_i}} &= \frac{\Delta P_{ref_i}}{T_{G_i}}; \end{aligned} \tag{1}$$

Tie-line power flow, P_{tie}^{ij}, between areas i and j (no attack):

$$\begin{aligned} \frac{d\Delta P_{tie}^{ij}}{dt} &= T_{ij}(\Delta\omega_i - \Delta\omega_j), \\ \Delta P_{tie}^{ji} &= -\Delta P_{tie}^{ij}, \end{aligned} \tag{2}$$

where $\Delta\omega_i$, ΔP_{mech_i} and ΔPv_i represent deviations of the angular frequency, mechanical power and steam-valve position from their nominal operating values.

A malicious agent is assumed to have access to circuit breakers that control the tie-lines, and is thus able to sever the connection between control areas. Depending on the topology of the tie-line interconnection graph, such attacks may correspond to a *node/vertex/bus attack* (disconnection of a control area from all others) or a *link/edge/line attack* (disabling of a specific tie-line between two control areas), i.e., the power flow across the tie lines is altered:

Attack on circuit breaker i (node/bus attack):

$$\Delta P_{tie}^{ij} = -\Delta P_{tie}^{ji} = 0, \ \forall j \neq i; \tag{3}$$

Attack on circuit breaker (i, j) (link/line attack):

$$\Delta P_{tie}^{ij} = -\Delta P_{tie}^{ji} = 0. \tag{4}$$

In addition, we assume that the system dynamics and measurements are subject to random noise and attacks via additive data injection in the actuator and sensor signals. The goal of resilient state estimation is thus to obtain unbiased state estimates despite switching attacks, i.e., attacks on switches/circuit breakers and the switching/time-varying nature of the attack strategy on switches, actuators and sensors.

III. Problem Formulation

A. System Description and Attack Modeling

We consider two different classes of possibly time-varying (switching) attacks on cyber-physical systems (CPS):

Mode Attack: Attacks on the switching mechanism that changes the system's *mode* of operation, or on the sensor data or interconnection network *topology*. *Examples:* Attack on circuit breakers [11], the power network topology [8] and the sensor data network [13]; attack on the logic switch of a traffic infrastructure [12].

Signal Attack: Attacks on actuator and sensor signals of unknown *magnitude* and *location* (i.e., subset of attacked actuators or sensors). *Examples:* Denial of service, deceptive attacks via data injection [5], [7].

Moreover, we allow the system to be perturbed by random process and measurement noise signals. Thus, we can represent the above attacks on a noisy dynamic system using a hidden mode, switched linear discrete-time stochastic system with unknown inputs governed by:

$$\begin{aligned} (x_{k+1}, q_k) &= (A_k^{q_k}x_k + B_k^{q_k}u_k^{q_k} + G_k^{q_k}d_k^{q_k} + w_k^{q_k}, q_k), & x_k \in \mathcal{C}_{q_k}, \\ (x_k, q_k)^+ &= (x_k, \delta^{q_k}(x_k)), & x_k \in \mathcal{D}_{q_k}, \\ y_k &= C_k^{q_k}x_k + D_k^{q_k}u_k^{q_k} + H_k^{q_k}d_k^{q_k} + v_k^{q_k}, \end{aligned} \tag{5}$$

where $x_k \in \mathbb{R}^n$ is the continuous system state and $q_k \in \mathcal{Q} = \{1, 2, \dots, \mathfrak{N}\}$ is the hidden discrete state or *mode*, which a malicious attacker has access to. The hidden modes include the modes of operation that attacked switching mechanisms (e.g., circuit breakers, relays) operate via access of the jump set $\mathcal{D}_{q_k}$ and the mode transition function $\delta^{q_k}(\cdot)$, or the possible interconnection network topologies that dictate the system matrices, $A_k^{q_k}$ and $B_k^{q_k}$, and the sensor data network topologies, $C_k^{q_k}$ and $D_k^{q_k}$, that an attacker can choose (*mode attack*), as well as the different hypotheses about which actuators and sensors are attacked that determine the true G_k^* and H_k^* (*signal location attack*).

For each mode q_k, $u_k^{q_k} \in U_{q_k} \subset \mathbb{R}^m$ is the known input, $d_k^{q_k} \in \mathbb{R}^p$ the unknown input or *attack signal* and $y \in \mathbb{R}^l$ the output, where the corresponding process noise $w_k^{q_k} \in \mathbb{R}^n$ and measurement noise $v_k^{q_k} \in \mathbb{R}^l$ are mutually uncorrelated, zero-mean Gaussian white random signals with known covariance matrices, $Q_k^{q_k} = \mathbb{E}[w_k^{q_k} w_k^{q_k\top}] \succeq 0$ and $R_k^{q_k} = \mathbb{E}[v_k^{q_k} v_k^{q_k\top}] \succ 0$, respectively. x_0 is independent of $v_k^{q_k}$ and $w_k^{q_k}$ for all k.

B. Assumptions on System and Attacker

1) System Assumptions: The matrices $A_k^{q_k}$, $B_k^{q_k}$, $G_k^{q_k}$, $C_k^{q_k}$, $D_k^{q_k}$ and $H_k^{q_k}$ are known. Moreover, $G_k^{q_k} := G^{q_k} I_G^{q_k}$ and $H_k^{q_k} := H^{q_k} I_H^{q_k}$ for some matrices G^{q_k} and H^{q_k} of appropriate dimensions, where $I_G^{q_k}$ and $I_H^{q_k}$ are such that $d_k^{a,q_k} := I_G^{q_k} d_k$ and $d_k^{s,q_k} := I_H^{q_k} d_k$ represent the subvectors of d_k representing *signal magnitude attacks* on the actuators and sensors, respectively, according to each hypothesis about the signal attack location, while G^{q_k} and H^{q_k} provide a means of incorporating information about the way the attacks affect the system (G^{q_k} and H^{q_k} are identity matrices if no such prior knowledge is available). For simplicity and without loss of generality, we assume

that the actuator and sensor signal attacks are distinct and hence, $\begin{bmatrix} G_k^{q_k} \\ H_k^{q_k} \end{bmatrix} := \begin{bmatrix} G^{q_k}\tilde{I}_G^{q_k} & 0 \\ 0 & H^{q_k}\tilde{I}_H^{q_k} \end{bmatrix}$ with $G^{q_k} \in \mathbb{R}^{n\times t_a}$, $\tilde{I}_G^{q_k} \in \mathbb{R}^{t_a\times p_a^{q_k}}$, $H^{q_k} \in \mathbb{R}^{\ell\times t_s}$, $\tilde{I}_H^{q_k} \in \mathbb{R}^{t_s\times p_s^{q_k}}$ and $p_s^{q_k} + p_a^{q_k} = p$ for each model $q \in \mathcal{Q}$, when there are t_a actuators and t_s sensors under signal attacks, and the maximum total number of attacks is $p \leq p^*$, where p^* is the maximum number of asymptotically correctable signal attacks (cf. Theorem 1 for its characterization). We also require that the system is *strongly detectable*[1] in each mode. In fact, strong detectability is *necessary* for each mode in order to asymptotically correct the unknown attack signals (also necessary for deterministic systems [21, Theorem 6]). Note also that strongly detectable systems need not be stable (cf. example in the proof of Theorem 1), but rather that the strongly undetectable modes of such systems are stable.

2) Attacker Assumptions: We do not constrain the malicious attack signals d_k to be a signal of any type (random or strategic) nor to follow any model, thus no prior 'useful' knowledge of the dynamics of d_k is available (uncorrelated with $\{d_\ell\}$ for all $k \neq \ell$, $\{w_\ell\}$ and $\{v_\ell\}$ for all ℓ). The only assumptions concerning the malicious attacker will be about the knowledge of (i) the upper bound on the *number* of actuators/sensors that can be attacked and (ii) the switching mechanisms/topologies that may be compromised (hence, the number of possible modes of operation when under mode attack), as well as (iii) that the strategy switching frequency for both mode and signal attacks is limited. Note that the final assumption is reasonable and realistic if the time a malicious agent takes to regain access/control is large compared to the time scale for the convergence of the inference algorithm and/or when the intention of the malicious agent is to confuse or avoid detection through intermittent attacks.

C. Problem Statement

The *objective* of this paper is to develop a *resilient state estimator* for system (5), i.e., a state filter that asymptotically recovers *unbiased* state estimates of the system irrespective of the location or magnitude of attacks on its actuators and sensors as well as switching mechanism/topology attacks, based on the multiple model approach given in [1]. We would also like to characterize fundamental limitations to attack resilience: (i) the maximum number of asymptotically correctable signal attacks and (ii) the maximum number of required models with this estimator. Furthermore, we want to analyze the asymptotic behavior (*model identifiability*) of the multiple model inference algorithm proposed in [1] and study its implication on the optimality of its state and input estimates, as well as on attack detection.

IV. Resilient State Estimation

To achieve resilient state estimation against switching attacks in the presence of stochastic process and measurement noise signals, we note that the system under switching attack is representable as a hidden mode, switched linear system with unknown inputs given in (5). Since we do not know the true model (i.e., the attack strategy corresponding to the true *mode attack* and *signal location attack*), combinations of possible attack strategies need to be considered, and as such, the multiple model estimation approach is a natural choice for solving this problem. Thus, we propose the use of the multiple model algorithm that we previously designed for vehicle collision avoidance [1] to solve this problem.

We will begin with a brief summary of the multiple model inference algorithm in [1]. Then, we characterize some fundamental limitations to resilient estimation. Finally, we explore the asymptotic properties of the resilient state inference algorithm for *model identification* in Section V.

A. Multiple Model State and Input Filter

We now provide an abbreviated review of the multiple model approach for simultaneous mode, state and unknown input estimation given in [1]. Two variants of the multiple model inference algorithm—static and dynamic—were proposed in that work. The latter provides a possibility of incorporating prior knowledge about the switching strategy of the attack. However, we assume no such knowledge about the malicious agent and will consider only the static variant (cf. Algorithm 2) in this paper, which consists of two components: (i) a bank of mode-matched filters, and (ii) a likelihood-based approach for computing model probability.

1) Mode-Matched Filters: The bank of filters is comprised of $\mathfrak{N}$ simultaneous state and input filters, one for each mode, based on the optimal recursive filter developed in [16] (with omitted superscript q_k for brevity; cf. Algorithm 1):

Unknown Input Estimation:

$$\begin{aligned} \hat{d}_{1,k} &= M_{1,k}(z_{1,k} - C_{1,k}\hat{x}_{k|k} - D_{1,k}u_k), \\ \hat{d}_{2,k-1} &= M_{2,k}(z_{2,k} - C_{2,k}\hat{x}_{k|k-1} - D_{2,k}u_k), \\ \hat{d}_{k-1} &= V_{1,k-1}\hat{d}_{1,k-1} + V_{2,k-1}\hat{d}_{2,k-1}, \end{aligned} \tag{6}$$

Time Update:

$$\begin{aligned} \hat{x}_{k|k-1} &= A_{k-1}\hat{x}_{k-1|k-1} + B_{k-1}u_{k-1} + G_{1,k-1}\hat{d}_{1,k-1}, \\ \hat{x}^\star_{k|k} &= \hat{x}_{k|k-1} + G_{2,k-1}\hat{d}_{2,k-1}, \end{aligned} \tag{7}$$

Measurement Update:

$$\hat{x}_{k|k} = \hat{x}^\star_{k|k} + \tilde{L}_k(z_{2,k} - C_{2,k}\hat{x}^\star_{k|k} - D_{2,k}u_k), \tag{8}$$

where $\hat{x}_{k-1|k-1}$, $\hat{d}_{1,k-1}$, $\hat{d}_{2,k-1}$ and $\hat{d}_{k-1}$ denote the optimal estimates of x_{k-1}, $d_{1,k-1}$, $d_{2,k-1}$ and d_{k-1}. Due to space constraints, the filter derivation along with its notations and definitions as well as necessary and sufficient conditions for filter stability and optimality are omitted; the reader is referred to [16] for a detailed discussion.

2) Mode Probability Computation: To compute the probability of each mode, the multiple model approach exploits the whiteness property [1, Theorem 1] of the generalized innovation sequence, ν_k, defined as

$$\nu_k := \Gamma_k(z_{2,k} - C_{2,k}\hat{x}^\star_{k|k} - D_{2,k}u_k) := \Gamma_k\overline{\nu}_k, \tag{9}$$

i.e., $\nu_k \sim \mathcal{N}(0, S_k)$ with covariance $S_k := \mathbb{E}[\nu_k\nu_k^\top] = \Gamma_k\tilde{R}^\star_{2,k}\Gamma_k^\top$ and where Γ_k is chosen such that S_k is invertible

[1] A linear system is *strongly detectable* if $y_k = 0\ \forall k \geq 0$ implies $x_k \to 0$ as $k \to \infty$ for all initial states x_0 and input sequences $\{d_i\}_{i\in\mathbb{N}}$ (see [16, Section 3.2] for necessary and sufficient conditions for this property).

Algorithm 1 Opt-Filter (q_k, $\hat{x}^{q_k}_{k-1|k-1}$, $\hat{d}^{q_k}_{1,k-1}$, $P^{x,q_k}_{k-1|k-1}$, $P^{xd,q_k}_{1,k-1}$, $P^{d,q_k}_{1,k-1}$) [superscript q_k omitted in the following]

▷ Estimation of $d_{2,k-1}$ and d_{k-1}

$\hat{A}_{k-1} = A_{k-1} - G_{1,k-1}M_{1,k-1}C_{1,k-1}$;

$\hat{Q}_{k-1} = G_{1,k-1}M_{1,k-1}R_{1,k-1}M^\top_{1,k-1}G^\top_{1,k-1} + Q_{k-1}$;

$\tilde{P}_k = \hat{A}_{k-1}P^x_{k-1|k-1}\hat{A}^\top_{k-1} + \hat{Q}_{k-1}$;

$\tilde{R}_{2,k} = C_{2,k}\tilde{P}_k C^\top_{2,k} + R_{2,k}$;

$P^d_{2,k-1} = (G^\top_{2,k-1}C^\top_{2,k}\tilde{R}^{-1}_{2,k}C_{2,k}G_{2,k-1})^{-1}$;

$M_{2,k} = P^d_{2,k-1}G^\top_{2,k-1}C^\top_{2,k}\tilde{R}^{-1}_{2,k}$;

$\hat{x}_{k|k-1} = A_{k-1}\hat{x}_{k-1|k-1} + B_{k-1}u_{k-1} + G_{1,k-1}\hat{d}_{1,k-1}$;

$\hat{d}_{2,k-1} = M_{2,k}(z_{2,k} - C_{2,k}\hat{x}_{k|k-1} - D_{2,k}u_k)$;

$\hat{d}_{k-1} = V_{1,k-1}\hat{d}_{1,k-1} + V_{2,k-1}\hat{d}_{2,k-1}$;

$P^d_{12,k-1} = M_{1,k-1}C_{1,k-1}P^x_{k-1|k-1}A^\top_{k-1}C^\top_{2,k}M^\top_{2,k} - P^d_{1,k-1}G^\top_{1,k-1}C^\top_{2,k}M^\top_{2,k}$;

$P^d_{k-1} = V_{k-1}\begin{bmatrix} P^d_{1,k-1} & P^d_{12,k-1} \\ P^{d\top}_{12,k-1} & P^d_{2,k-1}\end{bmatrix}V^\top_{k-1}$;

▷ Time update

$\hat{x}^\star_{k|k} = \hat{x}_{k|k-1} + G_{2,k-1}\hat{d}_{2,k-1}$;

$P^{\star x}_{k|k} = G_{2,k-1}M_{2,k}R_{2,k}M^\top_{2,k}G^\top_{2,k} + (I - G_{2,k-1}M_{2,k}C_{2,k})\tilde{P}_k(I - G_{2,k-1}M_{2,k}C_{2,k})^\top$;

$\tilde{R}^\star_{2,k} = C_{2,k}P^{\star x}_{k|k}C^\top_{2,k} + R_{2,k} - C_{2,k}G_{2,k-1}M_{2,k}R_{2,k} - R_{2,k}M^\top_{2,k}G^\top_{2,k-1}C_{2,k}$;

▷ Measurement update

$\tilde{L}_k = (P^{\star x}_{k|k}C^\top_{2,k} - G_{2,k-1}M_{2,k}R_{2,k})\tilde{R}^{\star\dagger}_{2,k}$;

$\hat{x}_{k|k} = \hat{x}^\star_{k|k} + \tilde{L}_k(z_{2,k} - C_{2,k}\hat{x}^\star_{k|k} - D_{2,k}u_k)$;

$P^x_{k|k} = (I - \tilde{L}_kC_{2,k})G_{2,k-1}M_{2,k}R_{2,k}\tilde{L}^\top_k + \tilde{L}_kR_{2,k}M^\top_{2,k}G^\top_{2,k-1}(I - \tilde{L}_kC_{2,k})^\top + (I - \tilde{L}_kC_{2,k})P^{\star x}_{k|k}(I - \tilde{L}_kC_{2,k})^\top + \tilde{L}_kR_{2,k}\tilde{L}^\top_k$;

▷ Estimation of $d_{1,k}$

$\tilde{R}_{1,k} = C_{1,k}P^x_{k|k}C^\top_{1,k} + R_{1,k}$;

$M_{1,k} = \Sigma^{-1}_k$;

$P^d_{1,k} = M_{1,k}\tilde{R}_{1,k}M_{1,k}$;

$\hat{d}_{1,k} = M_{1,k}(z_{1,k} - C_{1,k}\hat{x}_{k|k} - D_{1,k}u_k)$;

and $\tilde{R}^\star_{2,k}$ is given in Algorithm 1. Specifically, the *likelihood function* for each mode q at time k conditioned on all prior measurements Z^{k-1} is obtained as

$$\begin{aligned}&\mathcal{L}(q_k|z_{2,k}) := f_{z_{2,k}|Z^{k-1},q_k}(z_{2,k}|Z^{k-1},q_k)\\ &= f_{\nu_k|Z^{k-1},q_k}(\nu_k|Z^{k-1},q_k) = \mathcal{N}(\nu^{q_k}_k;0,S^{q_k}_k).\end{aligned} \tag{10}$$

Then, using Bayes' rule, the posterior probability for each mode j can be computed using

$$\mu^j_k = P(q_k = j|z_{1,k},z_{2,k},Z^{k-1}) = \frac{\mathcal{N}(\nu^j_k;0,S^j_k)\mu^j_{k-1}}{\sum_{i=1}^{\mathfrak{N}}\mathcal{N}(\nu^i_k;0,S^i_k)\mu^j_{k-1}}. \tag{11}$$

Note that a heuristic lower bound on all mode probabilities needs to be imposed such that the modes are kept alive in case of a switch in the strategy of the attacker. Finally, based on these posterior mode probabilities, the most probable mode at each time k is determined and thus the associated state and input estimates and covariances, as follows:

$$\begin{aligned}&\hat{q}_k = j^* = \arg\max \mu^j_k,\\ &\hat{x}_{k|k} = \hat{x}^{j^*}_{k|k},\quad \hat{d}_k = \hat{d}^{j^*}_k,\ P^x_{k|k} = P^{x,j^*}_{k|k},\ P^d_k = P^{d,j^*}_k.\end{aligned} \tag{12}$$

B. Fundamental Limitations of Attack-Resilient Estimation

1) Number of Asymptotically Correctable Signal Attacks: More formally, we introduce the following definition:

Definition 1 (Asymptotically/exponentially correctable signal attacks)**.** *We say that p actuators and sensors signal attacks are asymptotically/exponentially correctable, if for any initial state $x_0 \in \mathbb{R}^n$ and signal attack sequence $\{d_j\}_{j\in\mathbb{N}}$ in $\mathbb{R}^p$, we have an estimator such that the estimate bias asymptotically/exponentially tends to zero, i.e., $\mathbb{E}[\hat{x}_k - x_k] \to 0$ (and $\mathbb{E}[\hat{d}_{k-1} - d_{k-1}] \to 0$) as $k \to \infty$.*

Algorithm 2 Static-MM-Estimator ( )

Initialize $\forall j \in \{1,2,\dots,\mathfrak{N}\}$: $\hat{x}^j_{0|0}$; μ^j_0; $\hat{d}^j_{1,0} = (\Sigma^j_0)^{-1}(z^j_{1,0} - C^j_{1,0}\hat{x}^j_{0|0} - D^j_{1,0}u_0)$; $P^{d,j}_{1,0} = (\Sigma^j_0)^{-1}(C^j_{1,0}P^{x,j}_{0|0}C^{j\top}_{1,0} + R^j_{1,0})(\Sigma^j_0)^{-1}$;

for $k = 1$ to N **do**

  for $j = 1$ to $\mathfrak{N}$ **do**

    ▷ Mode-Matched Filtering

    Run Opt-Filter(j,$\hat{x}^j_{k-1|k-1}$, $\hat{d}^j_{1,k-1}$, $P^{x,j}_{k-1|k-1}$, $P^{d,j}_{1,k-1}$);

    $\overline{\nu}^j_k := z^j_{2,k} - C^j_{2,k}\hat{x}^{\star,j}_{k|k} - D^j_{2,k}u_k$;

    $\mathcal{L}(j|z^j_{2,k}) = \frac{1}{(2\pi)^{p^j_{\tilde{R}}/2}|\tilde{R}^{j,\star}_{2,k}|^{1/2}_+}\exp\left(-\frac{\overline{\nu}^{j\top}_k\tilde{R}^{j,\star\dagger}_{2,k}\overline{\nu}^j_k}{2}\right)$;

  end for

  for $j = 1$ to $\mathfrak{N}$ **do**

    ▷ Mode Probability Update (small $\epsilon > 0$)

    $\overline{\mu}^j_k = \max\{\mathcal{L}(j|z^j_{2,k})\mu^j_{k-1},\epsilon\}$;

  end for

  for $j = 1$ to $\mathfrak{N}$ **do**

    ▷ Mode Probability Update (normalization)

    $\mu^j_k = \frac{\overline{\mu}^j_k}{\sum_{\ell=1}^{\mathfrak{N}}\overline{\mu}^\ell_k}$;

    ▷ Output

    Compute (12);

  end for

end for

Remark 1. *Note the distinction in the definitions of asymptotically/exponentially correctable signal attacks in Definition 1 and of correctable signal attacks in [9, Definition 1]. Their definition implies finite-time estimation and is related to strong observability [9]. Due to the new challenges of further considering stochastic noise signals and mode switching, we adopt the weaker notion of asymptotic estimation, which only requires a 'weaker' condition of strong detectability (implied by strong observability [16]). This is mainly for the sake of theoretical analysis. Simulation results demonstrate that our algorithm is fast enough.*

To derive an estimation-theoretic upper bound on the maximum number of signal attacks that can be asymptotically tolerated, we assume that the true model or mode ($q_k = *$) is known. Thus, the resilient state estimation problem is identical to the state and input estimation problem in [16], where the unknown inputs represent the attacks on the actuator and sensor signals. It has been shown in [16] that unbiased states (and also unknown inputs) can be obtained asymptotically (exponentially) if the system is strongly detectable (cf. [16], [17] for more details, e.g. regarding filter stability and existence). With this in mind, the upper bound on the maximum number of signal attacks that can be asymptotically (exponentially) corrected is:

Theorem 1. *The maximum number of asymptotically (exponentially fast) correctable actuators and sensors signal attacks, p^*, for system* (5) *is equal to the number of sensors, l, i.e., $p^* \leq l$ and the upper bound is achievable.*

Proof. A necessary and sufficient condition for strong detectability (with the true model $q_k = *$) is given in [16] as

$$\operatorname{rk}\begin{bmatrix} zI - A^* & -G^* \\ C^* & H^* \end{bmatrix} = n + p^*, \ \forall z \in \mathbb{C}, |z| \geq 1. \qquad (13)$$

Since the above system matrix has only $n+l$ rows, it follows that its rank is at most $n + l$. Thus, from the necessary condition for (13), we obtain $n + p^* \leq n + l \Rightarrow p^* \leq l$. We show that the upper bound is achievable using the discrete-time equivalent model (with time step $\Delta t = 0.1s$) of the smart grid case study in [11], where in both circuit breaker modes, $A = \begin{bmatrix} 0.9520 & 0.0936 \\ -0.9358 & 0.8584 \end{bmatrix}$ and $G = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. If the first state is measured but compromised (e.g., $C = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $H = 1 \Rightarrow p^* = l$), it can be verified that the system is strongly detectable, i.e., with two invariant zeros at $\{0.9945 \pm 0.0311j\}$ that are strictly in the unit circle in the complex plane. Similarly, it can be verified that the unstable system with matrices $A = \begin{bmatrix} 1.5 & 1 \\ 0 & 0.1 \end{bmatrix}$, $G = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $H = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ (i.e., with $p^* = l$) has an invariant zero at $\{0.1\}$ and is hence strongly detectable. Thus, in both cases, the optimal filter in [16], [17] can be applied and unbiased state estimates can be asymptotically achieved when $p^* = l$. ■

Moreover, the necessity of strong detectability can serve as a guide to determine which actuators or sensors need to be safeguarded to guarantee resilient estimation. Since strong detectability is a system property that is independent of the filter design, the necessity of this property can be viewed as a fundamental limitation for resilient estimation, i.e., the ability to asymptotically/exponentially obtain unbiased estimates.

2) Number of Required Models for Estimation Resilience: Then, in a similar spirit as the attack set identification approach of [7], [8] in which a bank of deterministic residuals are computed to determine the true attack set (but not the magnitude of the attacks), we consider a bank of filters to find the most probable model/mode. We now characterize the maximum number of models $\mathfrak{N}^*$ that need to be considered with the multiple model approach in Section IV-A:

Theorem 2. *Suppose there are t_a actuators and t_s sensors, and at most $p \leq l$ of these signals are attacked. Suppose also that there are t_m possible attack modes (*mode attack*). Then, the combinatorial number of all possible models, and hence the maximum number of models that need to be considered with the multiple model approach, is*

$$\mathfrak{N}^* = t_m \binom{t_a + t_s}{p} = t_m \binom{t_a + t_s}{t_a + t_s - p}.$$

Proof. The maximum number of required models is the number of combinations of p attacks among $t_a + t_s$ sensors and actuators for each of the t_m attack modes of operation/topologies. Note that this number is the maximum because resilience may be achievable with less models: For instance, when $t_m = 1$, $t_a = 0$ and $t_s = 2 = l$, $p = 1$, $A = \begin{bmatrix} 0.1 & 1 \\ 0 & 0.2 \end{bmatrix}$ and $C = I_2$, we have $\mathfrak{N}^* = 2$, but it can be verified that with $G = 0_{2\times 2}$ and $H = I_2$ (only one model, i.e., $1 = \mathfrak{N} < \mathfrak{N}^*$), the system is strongly detectable. ■

Remark 2. *If $\mathfrak{N} > 1$, the multiple model approach requires that the number of attacks is strictly less than the number of sensor measurements, i.e., $p < l$. Otherwise, the generalized innovation* (9) *is empty and we have no means of selecting the 'best' model, i.e., of computing mode probabilities.*

If more information about the attacks is known, then one may expect that less models need to be considered. For instance, if there are at most $n_a \leq t_a$ and $n_s \leq t_s$ attacks on the actuators and sensors, respectively, with a total of p attacks (where $p \leq l$ and $p \leq n_a + n_s$), then the maximum number of models that are required is

$$\mathfrak{N}^* = t_m \textstyle\sum_{i=0}^{\min\{n_a,p\}} \binom{t_a}{i} \binom{t_s}{\min\{p-i,n_s\}}.$$

However, it turns out that more information may also increase the number of models. Nonetheless, with more information, the problem with more attacks that was previously not solvable because the (fewer) models are not strongly detectable, may now become solvable because although the number of models is increased, each of these models is strongly detectable. An example of this is with $t_m = 1$, $A = \begin{bmatrix} 0.1 & 1 \\ 0 & 1.2 \end{bmatrix}$ and $C = I$. If we assume that $n_a = 0$ and $n_s = p = 2$, then with $G = 0$ and $H = I$ (only one model is required), the system is not strongly detectable with zeros at $\{0.1, 1.2\}$. However, if $n_a = 0$ and $n_s = p = 1 < l = 2$, we have 2 models with $G = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $H_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $H_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, where both models can be verified to be strongly detectable.

V. Estimator Properties

A. Model Identification

We furnish the multiple model algorithms summarized in Section IV-A with an asymptotic analysis (not given in [1]) and thus its *model identification* property, which consists of either *model consistency* or *model convergence*, i.e., the convergence of the mode probability of the true model (if the true model is in the model set $\mathcal{Q}$) or of the 'closest' model (if the true model is not in the model set, with closeness defined in some information-theoretic sense), respectively, to 1. Note that model identifiability is a property of the inference algorithm in [1] and in resilient state estimation, this property refers to the ability to asymptotically infer the true attack strategy (i.e., the true mode attack and signal location attack). Throughout this section, the true model is assumed fixed in the time scale of interest given the assumption of limited strategy switching frequency in the attacker model in Section III (needed for the sake of analysis).

We first find the ratio of model probabilities from (11):

$$\frac{\mu_k^j}{\mu_k^i} := \frac{P(q_k=j|Z^k)}{P(q_k=i|Z^k)} = \frac{\mathcal{N}(\nu_k^j;0,S_k^j)\mu_{k-1}^j}{\mathcal{N}(\nu_k^i;0,S_k^i)\mu_{k-1}^i} = \frac{\mu_0^j}{\mu_0^i}\prod_{\ell=1}^k \frac{\mathcal{N}(\nu_\ell^j;0,S_\ell^j)}{\mathcal{N}(\nu_\ell^i;0,S_\ell^i)}$$
$$= \frac{\mu_0^j}{\mu_0^i}\exp\textstyle\sum_{\ell=1}^k \ln\frac{\mathcal{N}(\nu_\ell^j;0,S_\ell^j)}{\mathcal{N}(\nu_\ell^i;0,S_\ell^i)} = \frac{\mu_0^j}{\mu_0^i}\exp\textstyle\sum_{\ell=1}^k \ln\frac{f_\ell^j}{f_\ell^i}, \qquad (14)$$

where $\frac{\mu_0^j}{\mu_0^i}$ is the ratio of priors and we have used f_ℓ^j as a shorthand for $f_{z_{2,\ell}|q_\ell,Z^{\ell-1}}(z_{2,\ell}|q_\ell = j, Z^{\ell-1})$. From the

above ratio, we observe that the exponential rate at which the models are distinguishable depends on the sequence $\left\{\ln \frac{f_\ell^j}{f_\ell^i}\right\}_{\ell=1}^k$. Thus, in the following, we examine the behavior of this sequence, specifically the average behavior of this sequence (averaged over all possible states) and show that the posterior model mean probabilities converge to their true values if the true model is in the set of models. Otherwise, we show that the multiple model approach converges to the model that is 'closest' to the true model in an information-theoretic sense (i.e., with the minimum Kullback-Leibler (KL) divergence [22]) from within the model set. The case when the sequence is ergodic can also be considered but has been omitted due to space limitations.

To analyze the average model probability behavior (studied in part by [23]), we note that the mean of $\ln \frac{f_k^j}{f_k^i}$ is given by $\mathbb{E}_{f_k^*}\left[\ln \frac{f_k^j}{f_k^i}\right] = \mathbb{E}_{f_k^*}\left[\ln \frac{f_k^*}{f_k^i}\right] - \mathbb{E}_{f_k^*}\left[\ln \frac{f_k^*}{f_k^j}\right]$, where f_k^* is the distribution associated with the *true* model while $\mathbb{E}_{f_k^*}\left[\ln \frac{f_k^*}{f_k^q}\right]$ for $q \in \{i, j\}$ coincides with the definition of the Kullback-Leibler (KL) divergence (denoted $D(f_k^*\|f_k^q)$) that is widely recognized as an important measure of 'distance' between two probability distributions f_k^* and f_k^q in information theory [22]. With this, the ratio of the geometric means of model probabilities (14) can be computed and expressed as:

$$\begin{aligned}\frac{\overline{\mu}_k^j}{\overline{\mu}_k^i} &:= \frac{\mu_0^j}{\mu_0^i} \exp \textstyle\sum_{\ell=1}^k \mathbb{E}_{f_\ell^*}\left[\ln \frac{f_\ell^j}{f_\ell^i}\right] \\ &= \frac{\mu_0^j}{\mu_0^i} \exp \textstyle\sum_{\ell=1}^k \big(D(f_\ell^*\|f_\ell^i) - D(f_\ell^*\|f_\ell^j)\big).\end{aligned} \tag{15}$$

The KL divergence for each model $q \in \mathcal{Q}$ of the multiple model approach in this paper can be computed as:

$$\begin{aligned}&D(f_\ell^*\|f_\ell^q) := \mathbb{E}_{f_\ell^*}\left[\ln \frac{f_\ell^*}{f_\ell^q}\right] \\ &= \tfrac{1}{2}(p_{\tilde{R}_\ell^q} - p_{\tilde{R}_\ell^*}) \ln 2\pi + \tfrac{1}{2}\ln|\tilde{R}_{2,\ell}^{q,\star}|_+ - \tfrac{1}{2}\ln|\tilde{R}_{2,\ell}^{*,\star}|_+ \\ &\quad + \tfrac{1}{2}\mathbb{E}_{f_\ell^*}[\mathrm{tr}(\overline{\nu}_\ell^q \overline{\nu}_\ell^{q\top}(\tilde{R}_{2,\ell}^{q,\star})^\dagger)] - \tfrac{1}{2}\mathbb{E}_{f_\ell^*}[\mathrm{tr}(\overline{\nu}_\ell^* \overline{\nu}_\ell^{*\top}(\tilde{R}_{2,\ell}^{*,\star})^\dagger)] \\ &= \tfrac{1}{2}(p_{\tilde{R}_\ell^q} - p_{\tilde{R}_\ell^*}) \ln 2\pi + \tfrac{1}{2}\ln|\tilde{R}_{2,\ell}^{q,\star}|_+ - \tfrac{1}{2}\ln|\tilde{R}_{2,\ell}^{*,\star}|_+ \\ &\quad + \tfrac{1}{2}\mathrm{tr}(\tilde{R}_{2,\ell}^{q|*,\star}(\tilde{R}_{2,\ell}^{q,\star})^\dagger) - \tfrac{1}{2}\mathrm{tr}(\tilde{R}_{2,\ell}^{*,\star}(\tilde{R}_{2,\ell}^{*,\star})^\dagger),\end{aligned} \tag{16}$$

where $\tilde{R}_{2,\ell}^{q|*,\star} := \mathbb{E}_{f_\ell^*}[\overline{\nu}_\ell^q \overline{\nu}_\ell^{q\top}]$ and we have used the fact that $(I - C_{2,\ell}^q G_{2,\ell-1}^q M_{2,\ell}^q)$ is idempotent such that $(I - C_{2,\ell}^q G_{2,\ell-1}^q M_{2,\ell}^q)^\top (\tilde{R}_{2,\ell}^{q,\star})^\dagger (I - C_{2,\ell}^q G_{2,\ell-1}^q M_{2,\ell}^q) = (\tilde{R}_{2,\ell}^{q,\star})^\dagger$ to simplify the above expression. Note that the unknown inputs of each model need not have the same dimension; thus $p_{\tilde{R}_\ell}^q := \mathrm{rank}(\tilde{R}_{2,\ell}^{q,\star})$ can be different for all $q \in \{\mathcal{Q} \cup *\}$.

Inspired by the sufficient conditions for systems without unknown inputs [24], [25], we consider two conditions:

Condition (i) The true model $*$ is in the set of models, i.e., $* \in \mathcal{Q}$ and there exists a time step $T \in \mathbb{N}$ such that $f_\ell^* \neq f_\ell^q$, or equivalently,

$$\begin{aligned}&\tfrac{1}{2}p_{\tilde{R}_\ell^q}\ln 2\pi + \tfrac{1}{2}\ln|\tilde{R}_{2,\ell}^{q,\star}|_+ + \tfrac{1}{2}\mathrm{tr}(\tilde{R}_{2,\ell}^{q|*,\star}(\tilde{R}_{2,\ell}^{q,\star})^\dagger) \\ &\quad \neq \tfrac{1}{2}p_{\tilde{R}_\ell^*}\ln 2\pi + \tfrac{1}{2}\ln|\tilde{R}_{2,\ell}^{*,\star}|_+ + \tfrac{1}{2}\mathrm{tr}(\tilde{R}_{2,\ell}^{*,\star}(\tilde{R}_{2,\ell}^{*,\star})^\dagger),\end{aligned}$$

for all $q \in \mathcal{Q}, q \neq *$ for all $\ell \geq T$,

Condition (ii) The true model $*$ is not in the model set of models, i.e., $* \notin \mathcal{Q}$, but there exist a time step $T \in \mathbb{N}$ and a model $q \in \mathcal{Q}$ such that $D(f_\ell^*\|f_\ell^q) < D(f_\ell^*\|f_\ell^{q'})$, or equivalently,

$$\begin{aligned}&\tfrac{1}{2}p_{\tilde{R}_\ell^q}\ln 2\pi + \tfrac{1}{2}\ln|\tilde{R}_{2,\ell}^{q,\star}|_+ + \tfrac{1}{2}\mathrm{tr}(\tilde{R}_{2,\ell}^{q|*,\star}(\tilde{R}_{2,\ell}^{q,\star})^\dagger) \\ &\quad < \tfrac{1}{2}p_{\tilde{R}_\ell^{q'}}\ln 2\pi + \tfrac{1}{2}\ln|\tilde{R}_{2,\ell}^{q',\star}|_+ + \tfrac{1}{2}\mathrm{tr}(\tilde{R}_{2,\ell}^{q'|*,\star}(\tilde{R}_{2,\ell}^{q',\star})^\dagger),\end{aligned}$$

for all $q' \in \mathcal{Q}, q' \neq q$ for all $\ell \geq T$.

Condition (i) implies that the likelihood functions for all other models $q \neq *$ are not identical to the likelihood function for the true model $q = *$ for all $\ell \geq T$. In contrast, when the true model is not in the set of models, Condition (ii) implies that there exists a *unique* model $q \in \mathcal{Q}$ for all $\ell \geq T$ with a likelihood function that is closest to the true model and the other models are strictly less similar to the true model, measured in terms of their KL divergences.

Theorem 3 (Mean Consistency)**.** *Suppose Condition (i) holds; then, the multiple model approach is, on average, consistent, i.e., the model (geometric) mean probability of the true model converges to 1 (cf.* (15)*).*

Proof. Since $D(f_\ell^*\|f_\ell^q) \geq 0$ with equality if and only if $f_\ell^* = f_\ell^q$ ([22, Lemma 3.1]), then with $i = * \in \mathcal{Q}$ as the true model and $j \in \mathcal{Q}, j \neq *$, the summand in the exponent of (15) is always strictly negative, i.e., $D(f_\ell^*\|f_\ell^*) - D(f_\ell^*\|f_\ell^j) = -D(f_\ell^*\|f_\ell^j) < 0$ for all $\ell \geq T$ since $f_\ell^* \neq f_\ell^j$ by assumption. This means that, the ratios of model mean probabilities of all other models ($j \in \mathcal{Q}, j \neq *$) to the true model mean probability converge exponentially to zero, i.e., the mean probability of the true model converges to 1. ■

Note that even if for some $q \in \mathcal{Q}$, $f_\ell^q = f_\ell^*$ for all $\ell \in \mathbb{N}$ (i.e., Condition (i) fails to hold), the posterior model mean probabilities will be no worse than their prior probabilities.

Theorem 4 (Mean Convergence)**.** *Suppose Condition (ii) holds, i.e., the true model in not the set of models $\mathcal{Q}$, but there exists a model $q \in \mathcal{Q}$ with minimum KL divergence; then, with the multiple model approach, the identified model converges on average to the closest model in the set of models, i.e., to the model $q \in \mathcal{Q}$.*

Proof. Since Condition (ii) holds by assumption, then with $j = q'$ and $i = q$, the summand in the exponent of (15) is always strictly negative, which result in the exponential convergence to zero of the ratios of model mean probabilities of all other models ($q' \in \mathcal{Q}, q' \neq q$) to model q. ■

B. Optimality of State and Input Estimates

The following corollary characterizes the optimality of the state and input estimates when using the multiple model approach with the assumption that the true model is in the model set, i.e., $* \in \mathcal{Q}$. (Otherwise, the state and input estimates may be biased.)

Corollary 1. *If Condition (i) holds, then the state and input estimates in* (12) *converge on average to optimal state and input estimates in the minimum variance unbiased sense.*

Proof. For the true model, the filter gains are chosen such that the error variance is minimized and that the estimates are unbiased (cf. [16, Section 5] for a detailed derivation and

discussion). Hence, the state and input estimates are optimal in the minimum variance unbiased sense. If Condition (i) holds, by Theorem 3, the state and input estimates given by (12) also converge on average to the state and input estimates of the true model, which are optimal. ■

C. Attack Detection

While model consistency is quintessential for establishing the soundness of the multiple model approach, it may not be necessary for resilience. For instance, in the trivial case that there are no attacks $d_k = 0$ for all k, the state estimates of all models would perform equally well. In other words, the attacks need not be detected for obtaining resilient estimates. Moreover, if the estimator is not mean consistent but the true mode is in the set of models, then by Theorem 3, there exist some models with generalized innovations that have identical probability distributions as the generalized innovation of the true model (since their KL-divergences are identically zero), and that are hence Gaussian white sequences [1]. Since this is an indication that the input and state filters for these modes are optimal, these attack modes would be *undetectable* and better estimates cannot be achieved; thus, we regard our resilient state estimator as *optimal*.

VI. Simulation Examples

A. 3-Area Power System (Mode & Signal Magnitude Attacks)

We return to the motivating example in Section II and consider specifically the discrete-time equivalent (with a time step $\Delta t = 0.1s$) of a 3-area system in a radial topology corresponding to a node attack (as depicted in Figure 1) with $D_1 = 3$, $R_1^f = 0.03$, $M_1^a = 4$, $T_{CH_1} = 5$, $T_{G_1} = 4$, $D_2 = 0.275$, $R_2^f = 0.07$, $M_2^a = 40$, $T_{CH_2} = 10$, $T_{G_2} = 25$, $D_3 = 2$, $R_3^f = 0.04$, $M_3^a = 35$, $T_{CH_3} = 20$, $T_{G_3} = 15$, $T_{12} = 2.54$, $T_{23} = 1.5$ and $T_{31} = 2.5$. We assume that all inputs ΔP_{L_i} and ΔP_{ref_i} are identically zero and that all states are measured (i.e., $C_k^{q_k} = I$) where only measurements of $\Delta\omega_i$ are corrupted by additive errors d_i for $i = 1, 2, 3$ ($t_a = 0$, $t_s = 3$, $p = 3$) and the system is affected by additive zero mean Gaussian white process and measurement noise signals with known covariances $Q = 10^{-2} \times \mathrm{diag}(1, 1.6, 2, 1.2, 2.5, 1.4, 0.3, 2.11, 3, 0.2, 0.9, 1.8)$ and $R = 10^{-2} \times \mathrm{diag}(2.1, 0.6, 2.2, 0.2, 1.9, 1.4, 1.3, 1.1, 2.3, 1.2, 0.3, 1.8)$. For this tie-line interconnection topology, the circuit breaker attacks result in $\mathfrak{N} = t_m = 5$ possible modes of operation: all switches are safe/"on" ($q = 1$), only circuit breaker i is attacked/"off" ($q = i + 1$, $i = 1, 2, 3$) and two or more circuit breakers are attacked/"off" ($q = 5$).

For conciseness, we only show the results for the case when the attacker is assumed to switch from $q = 2$ to $q = 5$ at $t = 500s$, although our approach can also be successfully employed for all possible switching sequences. We observe from Figure 2 that the resilient state estimation algorithm is able to estimate the hidden mode, i.e., the true switching mode. Furthermore, we observe from Figure 3 that the system states (including those from unattacked measurements; not depicted) and unknown attack magnitudes are successful estimated, i.e., the signal attacks d_i that can be observed to affect the *raw* measurements of $\Delta\omega_i$ for $i = 1, 2, 3$ have been corrected/removed.

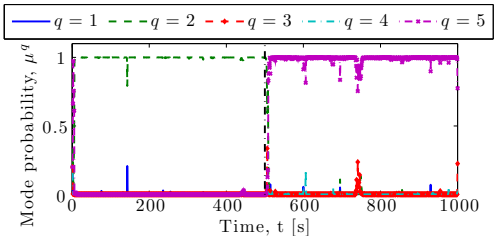

Fig. 2: Mode probabilities for Example VI-A.

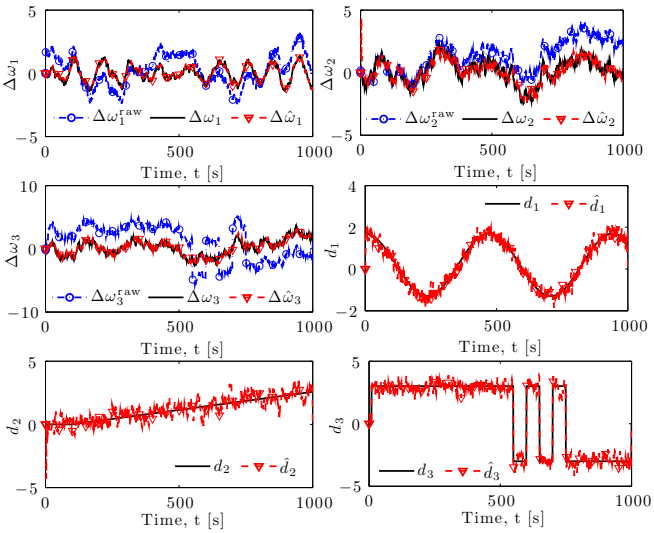

Fig. 3:State and attack magnitude estimates in ExampleVI-A.

B. Benchmark System (Signal Magnitude & Location Attacks)

In this example, we consider the resilient state estimation problem for a system (modified from [16]) that has been used as a benchmark for many state and input filters, with only one mode of operation ($t_m = 1$) and with possible attacks on the actuator and 4 of the 5 sensors ($t_a = 1$, $t_s = 4$):

$$A = \begin{bmatrix} 0.5 & 2 & 0 & 0 & 0 \\ 0 & 0.2 & 1 & 0 & 1 \\ 0 & 0 & 0.3 & 0 & 1 \\ 0 & 0 & 0 & 0.7 & 1 \\ 0 & 0 & 0 & 0 & 0.1 \end{bmatrix}; B = G = \begin{bmatrix} 1 \\ 0.1 \\ 0.1 \\ 1 \\ 0 \end{bmatrix}; C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -0.1 & 0 & 0 \\ 0 & 0 & 1 & -0.5 & 0.2 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0.25 & 0 & 0 & 1 \end{bmatrix};$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}; Q = 10^{-4} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.5 & 0 & 0 \\ 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}; R = 10^{-4} \begin{bmatrix} 1 & 0 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0 & 0.3 \\ 0 & 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0 & 1 & 0 \\ 0 & 0.3 & 0 & 0 & 1 \end{bmatrix}.$$

The known input u_k is 2 for $100 \leq k \leq 300$, -2 for $500 \leq k \leq 700$ and 0 otherwise, whereas the unknown inputs are as given in Figure 5. We also assume that there are at most $p = 4$ attacks with no constraints on n_a and n_s; as a result, we have to consider $\mathfrak{N} = 1 \cdot \binom{5}{4} = 5$ models.

Due to space limitation, we only provide simulation results for the case when the signal attack locations are switched from $q = 3$ (attack on actuator and sensors 1,3,4) to $q = 2$ (attack on actuator and sensors 1,2,4) at time $t = 500s$. From Figure 4, we observe that the mode probabilities converge to their true values. Figure 5 shows the estimates of states as well as the unknown attack signal magnitudes. The

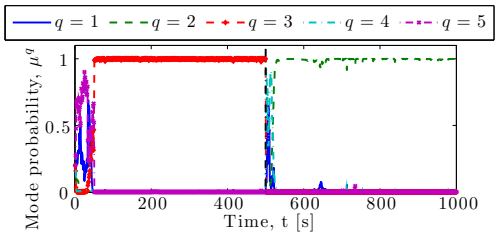

Fig. 4: Mode probabilities for Example VI-B.

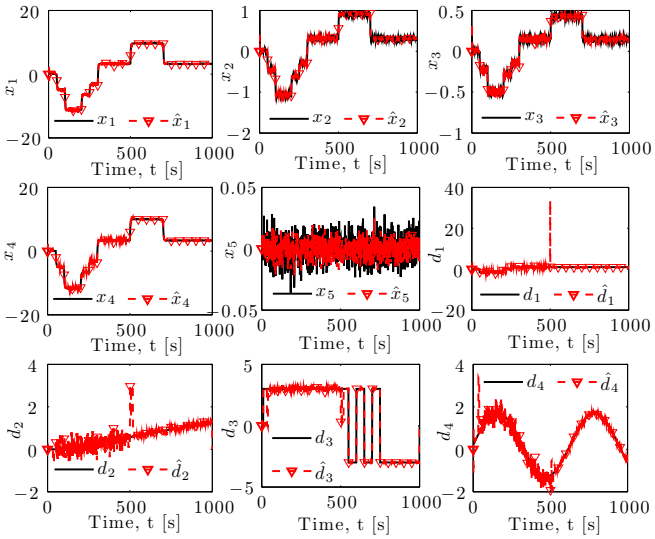

Fig. 5:State and attack magnitude estimates in ExampleVI-B.

state estimates, which are our main concern, are seen to be good even before the mode probabilities converge, while the unknown attack signals are also reasonably well estimated, with the exception of little jumps in its estimates during the switch in attack locations at $t = 500s$. Similar results (not shown) are obtained for all other attack modes, $q = 1$ (attack on actuator and sensors 1,2,3), $q = 4$ (attack on actuator and sensors 2,3,4) and $q = 5$ (attack on sensors 1,2,3,4).

VII. Conclusion

We addressed the problem of resilient state estimation for switching (mode/topology) attacks and attacks on actuator and sensor signals of stochastic cyber-physical systems. We first modeled the problem as a hidden mode switched linear stochastic system with unknown inputs and showed that the multiple model inference algorithm in [1] is a suitable solution to these issues and furnished the algorithm with an asymptotic analysis. Moreover, we provided an achievable upper bound on the maximum number of asymptotically correctable signal attacks and also the maximum number of required models for the multiple model approach. Simulation examples demonstrated the effectiveness of our approach.

Acknowledgments

This work was supported by the National Science Foundation, grant #1239182. M. Zhu is partially supported by ARO W911NF-13-1-0421 (MURI) and NSF CNS-1505664.

References

[1] S.Z. Yong, M. Zhu, and E. Frazzoli. Generalized innovation and inference algorithms for hidden mode switched linear stochastic systems with unknown inputs. In *IEEE Conference on Decision and Control (CDC)*, pages 3388–3394, December 2014.

[2] A.A. Cárdenas, S. Amin, and S. Sastry. Research challenges for the security of control systems. In *Proceedings of the 3rd Conference on Hot Topics in Security*, HOTSEC'08, pages 6:1–6:6, 2008.

[3] J.P. Farwell and R. Rohozinski. Stuxnet and the future of cyber war. *Survival*, 53(1):23–40, 2011.

[4] M. Zhu and S. Martínez. On distributed constrained formation control in operator-vehicle adversarial networks. *Automatica*, 49(12):3571–3582, 2013.

[5] A.A. Cardenas, S. Amin, and S. Sastry. Secure control: Towards survivable cyber-physical systems. In *International Conference on Distributed Computing Systems Workshops*, pages 495–500, 2008.

[6] Y. Mo and B. Sinopoli. False data injection attacks in control systems. In *Workshop on Secure Control Systems*, 2010.

[7] F. Pasqualetti, F. Dörfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11):2715–2729, November 2013.

[8] J. Weimer, S. Kar, and K.H. Johansson. Distributed detection and isolation of topology attacks in power networks. In *Proceedings of the 1st International Conference on High Confidence Networked Systems*, HiCoNS '12, pages 65–72, New York, NY, USA, 2012. ACM.

[9] H. Fawzi, P. Tabuada, and S. Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic Control*, 59(6):1454–1467, June 2014.

[10] 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)*, pages 163–174, April 2014.

[11] S. Liu, S. Mashayekh, D. Kundur, T. Zourntos, and K. Butler-Purry. A framework for modeling cyber-physical switching attacks in smart grid. *IEEE Transactions on Emerging Topics in Computing*, 1(2):273–285, December 2013.

[12] B. Ghena, W. Beyer, A. Hillaker, J. Pevarnek, and J.A. Halderman. Green lights forever: Analyzing the security of traffic infrastructure. In *8th USENIX Workshop on Offensive Technologies*, August 2014.

[13] J. Kim and L. Tong. On topology attack of a smart grid: Undetectable attacks and countermeasures. *IEEE Journal on Selected Areas in Communications*, 31(7):1294–1305, July 2013.

[14] S. Gillijns and B. De Moor. Unbiased minimum-variance input and state estimation for linear discrete-time systems. *Automatica*, 43(1):111–116, January 2007.

[15] S.Z. Yong, M. Zhu, and E. Frazzoli. Simultaneous input and state estimation for linear discrete-time stochastic systems with direct feedthrough. In *Conference on Decision and Control (CDC)*, pages 7034–7039, 2013.

[16] S.Z. Yong, M. Zhu, and E. Frazzoli. A unified filter for simultaneous input and state estimation of linear discrete-time stochastic systems. *Automatica*, 2015. Provisionally accepted. Available from: `http://arxiv.org/abs/1309.6627`.

[17] S.Z. Yong, M. Zhu, and E. Frazzoli. On strong detectability and simultaneous input and state estimation with a delay. In *IEEE Conference on Decision and Control (CDC)*, 2015. To appear.

[18] Y. Bar-Shalom, T. Kirubarajan, and X.-R. Li. *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[19] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan. Interacting multiple model methods in target tracking: a survey. *IEEE Transactions on Aerospace and Electronic Systems*, 34(1):103–123, January 1998.

[20] A.J. Wood, B.F. Wollenberg, and G.B. Sheble. *Power generation, operation, and control*. John Wiley & Sons, 2013.

[21] S. Sundaram and C.N. Hadjicostis. Delayed observers for linear systems with unknown inputs. *IEEE Transactions on Automatic Control,*, 52(2):334–339, February 2007.

[22] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[23] Z. Zhao and X.R. Li. The behavior of model probability in multiple model algorithms. In *8th International Conference on Information Fusion*, volume 1, pages 331–336, July 2005.

[24] Y. Baram and N.R. Sandell. An information theoretic approach to dynamical systems modeling and identification. *IEEE Transactions on Automatic Control*, 23(1):61–66, February 1978.

[25] Y. Baram and N.R. Sandell. Consistent estimation on finite parameter sets with application to linear systems identification. *IEEE Transactions on Automatic Control*, 23(3):451–454, June 1978.