Received: 28 February 2018 | Revised: 15 June 2019 | Accepted: 22 July 2019

DOI: 10.1002/rnc.4724

PRIVACY AND SECURITY OF CYBER-PHYSICAL SYSTEMS

WILEY

Joint attack detection and secure state estimation of cyber-physical systems

Nicola Forti[1,2] | **Giorgio Battistelli**[1] | **Luigi Chisci**[1] | **Bruno Sinopoli**[3]

[1]Dipartimento di Ingegneria dell'Informazione, Università di Firenze, Firenze, Italy

[2]NATO-STO Centre for Maritime Research and Experimentation (CMRE), viale San Bartolomeo 400, 19126, La Spezia, Italy

[3]Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, Missouri

Correspondence
Nicola Forti, NATO-STO Centre for Maritime Research and Experimentation (CMRE), viale San Bartolomeo 400, 19126 La Spezia, Italy.
Email: nicola.forti@cmre.nato.int

Funding information
National Science Foundation, Grant/Award Number: CCF 1646526; U.S. Department of Energy, Grant/Award Number: DE-OE0000779

Summary

This paper deals with secure state estimation of cyber-physical systems subject to switching (on/off) attack signals and injection of fake packets (via either packet substitution or insertion of extra packets). The random set paradigm is adopted in order to model, via *random finite sets* (RFSs), the switching nature of both system attacks and the injection of fake measurements. The problem of detecting an attack on the system and jointly estimating its state, possibly in the presence of fake measurements, is then formulated and solved in the Bayesian framework for systems with and without direct feedthrough of the attack input to the output. This leads to the analytical derivation of a *hybrid Bernoulli filter* (HBF) that updates in real time the joint posterior density of a Bernoulli attack RFS and of the state vector. A closed-form Gaussian mixture implementation of the proposed HBF is fully derived in the case of invertible direct feedthrough. Finally, the effectiveness of the developed tools for joint attack detection and secure state estimation is tested on two case studies concerning a benchmark system for unknown input estimation and a standard IEEE power network application.

KEYWORDS

Bayesian state estimation, Bernoulli filter, cyber-physical systems, extra packet injection, random finite sets, secure state estimation

1 | INTRODUCTION

Cyber-physical systems (CPSs) are complex engineered systems arising from the integration of computational resources and physical processes, tightly connected through a communication infrastructure. Typical examples of CPSs include next-generation systems in building and environmental monitoring/control, health care, electric power grids, transportation and mobility, and industrial process control. While, on one hand, advances in CPS technology will enable enhanced autonomy, efficiency, seamless interoperability, and cooperation, on the other hand, the increased interaction between cyber and physical realms is unavoidably providing novel security vulnerabilities, which make CPSs subject to nonstandard malicious threats. Recent real-world attacks, such as the Maroochy Shire sewage spill, the Stuxnet worm sabotaging an industrial control system, and the lately reported massive power outage against Ukrainian electric grid,[1] have brought into particularly sharp focus the urgency of designing secure CPSs. It is worth pointing out that, in presence of malicious threats against CPSs, standard approaches extensively used for control systems subject to benign faults and failures are no longer suitable. Moreover, the design and implementation of defense mechanisms usually employed for cybersecurity can only guarantee limited layers of protection, since they do not take into account vulnerabilities like the ones on physical components. This is why recent research efforts on the design of secure systems have explored different routes.

Preliminary work addressed the issues of attack detection/identification and proposed attack monitors for deterministic control systems.[2] Secure strategies have been studied for *replay* attacks[3,4] where the adversary first records and then replays the observed data, as well as for *denial-of-service* (DoS) attacks[5,6] disrupting the flow of data. Moreover, active detection methods have been designed in order to detect *stealthy* attacks via manipulation of, eg, control inputs[7] or dynamics.[8] Over the last few years, the problem of secure state estimation, ie, capable of reconstructing the state even when the CPS of interest is under attack, has gained considerable attention.[9-17] Initial work considered a worst-case approach for the special class of SISO systems.[9] Under the assumption of linear systems subject to an unknown but bounded number of *false-data injection* attacks on sensor outputs, the problem for a noise-free system has been cast into an ℓ_0-optimization problem, which can be relaxed as a more efficient convex problem[10] and, in turn, adapted to systems with bounded noise.[11] Further advances tried to tackle the combinatorial complexity of the problem by resorting to satisfiability modulo theories[12] and investigated, in the same context, the case of Gaussian measurement noise[13] and the concept of observability under attacks.[14] Most recently, deterministic models of the most popular attack policies have been presented based on adversary's resources and system knowledge,[15] and secure state estimation of CPSs has been addressed[16] by modeling in a stochastic framework the attacker's decision-making by assuming Markov (possibly uninformative) decision processes instead of unknown or worst-case models.

Although the literature on attack-resilient state estimation is quite abundant, most of the existing contributions have adopted a deterministic (worst-case) approach and/or have been restricted to linear systems. In practice, the system monitor (defender) might have some (even no) probabilistic prior knowledge on the attacker's strategy and the CPS of interest might easily be affected by nonlinearities. In this respect, a Bayesian approach where prior knowledge on the attacks is characterized in terms of probability distributions and nonlinearities are possibly handled by particle filtering or Gaussian mixture (GM) methods seems well suited and will be pursued in this paper. This allows great flexibility in that knowledge available to the attack monitor can range from complete knowledge to no prior knowledge (uninformative prior) depending on the assumed distributions.

Specifically, in this paper, three different types of adversarial attacks on CPSs are considered: (i) *signal* attack, ie, signal of arbitrary magnitude and location injected (with known structure) to corrupt sensor/actuator data, (ii) *packet substitution* attack, describing an intruder that possibly intercepts and then replaces the system-generated measurement with a fake (unstructured) one, and (iii) *extra packet injection*, a new type of attack against state estimation, already introduced in information security,[18,19] in which multiple counterfeit observations (junk packets) are possibly added to the system-generated measurement. Note that the key feature distinguishing signal attacks on sensors from packet substitution relies on the fact that the former are assumed to alter the measurement through a given structure (ie, known measurement function), whereas the latter mechanism captures integrity attacks that spoof sensor data packets with no care of the model structure. By considering both structured and unstructured injections, we do not restrict the type of attack the adversary can enforce on the sensor measurements. Please notice that, as a further by-product, the Bayesian approach with uninformative prior can also deal with the situation in which the attacker has the ability to choose arbitrarily large attack and/or fake measurements, while the worst-case attack paradigm in such a case is not viable.

The present paper aims to address the problem of simultaneously detecting a signal attack while estimating the state of the monitored system, possibly in presence of fake measurements independently injected into the system's monitor by cyberattackers. A random set attack modeling approach is undertaken by representing the signal attack presence/absence by means of a Bernoulli random set (BRS) (ie, a set that, with some probability, can be either empty or a singleton depending on the presence or not of the attack) and by taking into account possible fake measurements by means of a random measurement set. We follow the approach of Forti et al[17] and formulate the joint attack detection–state estimation problem within the Bayesian framework as the recursive determination of the joint posterior density of the signal attack Bernoulli set and of the state vector at each time given all the measurement sets available up to that time. Strictly speaking, the posed Bayesian estimation problem is neither standard[20] nor *Bernoulli* filtering[21-24] but is rather a *hybrid* Bayesian filtering problem that aims to jointly estimate a BRS for the signal attack and a random vector for the system state. An analytical solution of the hybrid filtering problem has been found in terms of integral equations that generalize the Bayes and Chapman-Kolmogorov equations of the Bernoulli filter. In particular, the proposed *hybrid Bernoulli Bayesian filter* for joint attack detection–state estimation propagates in time, via a two-step prediction-correction procedure, a joint posterior density completely characterized by a triplet consisting of (1) a signal attack probability, (2) a *probability density function* (PDF) in the state space for the system under no signal attack, and (3) a PDF in the joint attack input-state space for the system under signal attack.

The adopted approach enjoys the following positive features: (1) it encompasses in a unique framework different types of attacks (signal attacks, packet substitution, extra packet injection, temporary DoS, etc); (2) it takes into account the presence of disturbances and noise and deals with general nonlinear systems; (3) it propagates probability distributions of the system state, attack signal, and attack existence, which can be useful for, respectively, real-time dynamic state estimation, attack reconstruction, and security decision-making. Notice that, unlike most previous work cited above, in the present paper, we address the problem from the estimator's perspective and, hence, we cannot assume any specific strategy for the attacker. This motivates the modeling of the signal attack as a switching unknown input affecting the system.

Preliminary work on Bayesian state estimation against switching unknown inputs and extra packet injection was carried out by Forti et al.[17] The present paper extends this preliminary work in the following directions.

1. It also considers the *packet substitution* attack (in addition to the already considered *extra packet injection* attack). This novel type of attack refers to the practically relevant situation wherein the attacker has the ability to intercept and manipulate packets sent to the system monitor so as to replace system-originated measurements by fake ones but, unlike the extra packet injection attack, cannot send additional indistinguishable packets containing fake measurements to confuse the system monitor.
2. It provides the full derivation of a closed-form solution of the posed Bayesian filtering problem for linear Gaussian models based on a GM approach. This allows a computationally efficient implementation of the proposed joint *attack detector-state estimator* also generalizable to nonlinear models via extended or unscented (instead of standard) Kalman filtering techniques.
3. It considers also the case of no direct feedthrough of the attack input into the observed output.

The main challenge has been to provide an unifying framework to deal with different types of attacks/faults, possible nonlinearities, presence of noise and disturbances, and, at the same time, produce implementable algorithms, via GM techniques, that in real time are able to detect attacks and safely monitor the plant state. It is the authors' opinion that Bayesian random set filtering can promote significant advances in the research on security of CPSs providing theoretically principled and flexible tools that can be used in many practical scenarios.

The rest of this paper is organized as follows. Section 2 introduces the considered attack models and provides the necessary background on joint input and state estimation (JISE) as well as on random set estimation. Sections 3 and 4 formulate and solve the joint *attack detection–state estimation* problem of interest in the Bayesian framework. Section 5 provides detailed derivations of the GM hybrid Bernoulli filter (HBF) for linear Gaussian models. Then, Section 6 demonstrates the effectiveness of the proposed approach via numerical simulations on a benchmark example taken from the literature on resilient state estimation for CPSs[25] as well as on a practical application pertaining to the monitoring of power electrical networks. Finally, Section 7 ends this paper with concluding remarks and perspectives for future work.

2 | PROBLEM SETUP AND PRELIMINARIES

2.1 | System description and attack model

Let the discrete-time CPS of interest be modeled by

$$x_{k+1} = \begin{cases} f_k^0(x_k) + w_k, & \text{under no attack} \\ f_k^1(x_k, a_k) + w_k, & \text{under attack,} \end{cases} \tag{1}$$

where k is the time index; $x_k \in \mathbb{R}^n$ is the state vector to be estimated; $a_k \in \mathbb{R}^m$, called attack vector, is an unknown input affecting the system only when it is under attack; $f_k^0(\cdot)$ and $f_k^1(\cdot,\cdot)$ are known state transition functions that describe the system evolution in the *no-attack* and, respectively, *attack* cases; w_k is a random process disturbance also affecting the system. For monitoring purposes, the state of the above system is observed through the measurement model

$$y_k = \begin{cases} h_k^0(x_k) + v_k, & \text{under no attack} \\ h_k^1(x_k, a_k) + v_k, & \text{under attack,} \end{cases} \tag{2}$$

where $h_k^0(\cdot)$ and $h_k^1(\cdot,\cdot)$ are known measurement functions that refer to the *no-attack* and, respectively, *attack* cases; v_k is a random measurement noise. It is assumed that the measurement y_k is delivered with probability $p_d \in (0, 1]$, where the nonunit probability might be due to a number of reasons (eg, temporary denial of service, packet loss, and sensor inability to detect or sense the system). The attack modeled in (1)-(2) via the attack vector a_k is usually referred to as *signal* attack. While for ease of presentation, only the case of a single attack model is taken into account, multiple attack models[26] could be accommodated in the considered framework by letting (1)-(2) depend on a discrete variable, say, v_k, which specifies the particular attack model and has to be estimated together with a_k. Besides the system-originated measurement y_k in (2), it is assumed that the system monitor might receive *fake* measurements from some cyberattacker. In this respect, the following two cases will be considered.

1. *Packet substitution*: With some probability $p_f \in [0, 1)$, the attacker replaces the system-originated measurement y_k with a fake one y_k^f (see Figure 1).
2. *Extra packet injection*: The attacker sends to the monitor one or multiple fake measurements indistinguishable from the system-originated one (see Figure 2).

For the subsequent developments, it is convenient to introduce the *attack set* at time k, $\mathcal{A}_k$, which is either equal to the empty set if the system is not under signal attack at time k or to the singleton $\{a_k\}$ otherwise, ie,

$$\mathcal{A}_k = \begin{cases} \emptyset, & \text{if the system is not under signal attack} \\ \{a_k\}, & \text{otherwise.} \end{cases}$$

It is also convenient to define the *measurement set* at time k, $\mathcal{Z}_k$. For the *packet substitution* attack (Figure 1),

$$\mathcal{Z}_k = \begin{cases} \emptyset, & \text{with probability } \ 1 - p_d \\ \{y_k\}, & \text{with probability } p_d(1 - p_f) \\ \{y_k^f\}, & \text{with probability } p_d p_f, \end{cases} \tag{3}$$

where y_k is given by (2) and y_k^f is a fake measurement provided by the attacker in place of y_k. Conversely, for the *extra packet injection* attack (Figure 2) the definition (3) is replaced by

$$\mathcal{Z}_k = \mathcal{Y}_k \cup \mathcal{F}_k, \tag{4}$$

where

$$\mathcal{Y}_k = \begin{cases} \emptyset, & \text{with probability } 1 - p_d \\ \{y_k\}, & \text{with probability } p_d \end{cases} \tag{5}$$

is the set of system-originated measurements and $\mathcal{F}_k$ the finite set of fake measurements.

The aim of this paper is to address the problem of joint attack detection and state estimation, which amounts to jointly estimating, at each time k, the state x_k and signal attack set $\mathcal{A}_k$ given the set of measurements $\mathcal{Z}^k \triangleq \cup_{i=1}^{k} \mathcal{Z}_i$ up to time k.

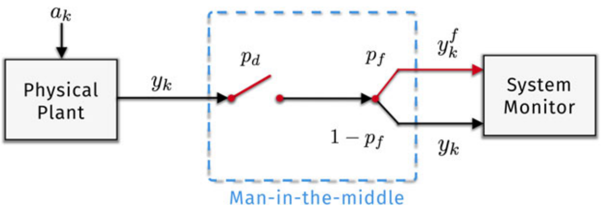

FIGURE 1 *Packet substitution* attack [Colour figure can be viewed at wileyonlinelibrary.com]

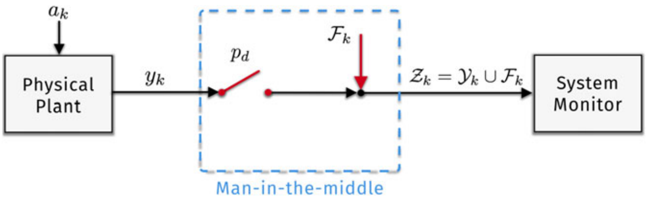

FIGURE 2 *Extra packet injection* attack [Colour figure can be viewed at wileyonlinelibrary.com]

2.2 | Joint input and state estimation

This section recalls the Bayesian approach to JISE[27] that is exploited, in this work, in order to estimate, besides the plant state, also the signal attack input to which the CPS may be subjected. Consider a system affected by an unknown input a_k

$$\begin{cases} x_{k+1} = f(x_k, a_k) + w_k \\ y_k = h(x_k, a_k) + v_k. \end{cases} \tag{6}$$

In JISE,[27-29] it is customary to distinguish the case in which there is a direct feedthrough of the unknown input a_k to the output y_k from the case of no direct feedthrough.

Direct feedthrough: Suppose that there is an invertible direct feedthrough[27,28] of the unknown input a_k to the output y_k, which amounts to assuming that the function $h(x, a)$ is injective with respect to a for any x. In this case, the Bayesian approach is based on the recursive computation of the joint PDF $p(a_k, x_k \,|\, y^k)$ of the unknown input a_k and state x_k conditioned on all the information available up to the current time. Given the conditional PDF, optimal estimates of a_k and x_k can be computed according to any given criterion, the most typical ones being maximum a posteriori probability (MAP) and minimum mean square error (MMSE). In this respect, it is worthy to point out that the two methods (MAP and MMSE) are actually both reasonable choices corresponding to extract either the mean (MMSE) or the mode (MAP) from the conditional PDF. If such a conditional PDF is highly multimodal, eg, due to the presence of multiple counterfeit measurements (extra packet injection attack), the MAP approach, ie, extracting from the posterior density the Gaussian mean with the highest weight, seems preferable. The joint conditional PDF can be computed by means of a two-step procedure of correction and prediction. Suppose that, at time $k-1$, the predicted posterior $p(a_k, x_k \,|\, y^{k-1})$ has been computed. Then, at time k, when the new measurement y_k is collected, in the correction step, the new conditional PDF $p(a_k, x_k \,|\, y^k)$ can be obtained by means of the Bayes rule

$$p\left(a_k, x_k \,|\, y^k\right) = \frac{p(y_k \,|\, a_k, x_k)\, p\left(a_k, x_k \,|\, y^{k-1}\right)}{p\left(y_k \,|\, y^{k-1}\right)}. \tag{7}$$

Conversely, the prediction step concerns the propagation of the conditional PDF from time k to time $k+1$. In the literature on unknown input estimation, it is usually supposed that the values a_k and x_k of unknown input and, respectively, of state at time k do not provide any information on the value a_{k+1} taken by the unknown input at time $k+1$. Accordingly, $p(a_{k+1}, x_{k+1} \,|\, y^k)$ takes the form

$$p\left(a_{k+1}, x_{k+1} \,|\, y^k\right) = p\left(x_{k+1} \,|\, y^k\right) p(a_{k+1}), \tag{8}$$

where the conditional PDF $p(x_{k+1} \,|\, y^k)$ is computed via the Chapman-Kolmogorov equation

$$p\left(x_{k+1} \,|\, y^k\right) = \iint p(x_{k+1} \,|\, a_k, x_k)\, p\left(a_k, x_k \,|\, y^k\right) da_k dx_k. \tag{9}$$

With this respect, when no information on the unknown input a_{k+1} is supposed to be available, it is customary[27] to resort to the so-called *principle of indifference* and take $p(a_{k+1})$ as an uninformative (flat) prior. It is easy to check that, in this case, the conditional PDF $p(a_k, x_k \,|\, y^k)$ resulting from the correction step can be rewritten as

$$p\left(a_k, x_k \,|\, y^k\right) = \frac{p(y_k \,|\, a_k, x_k)\, p\left(x_k \,|\, y^{k-1}\right)}{\int\int p(y_k \,|\, a, x) p(x \,|\, y^{k-1})\, dx\, da}. \tag{10}$$

Then, maximization of (10) with respect to x_k and a_k provides a MAP estimate of x_k and a Maximum Likelihood (ML) estimate of the unknown input a_k. This is the approach followed by Fang et al[27] that allows to generalize the traditional techniques for linear systems[28,29] to general nonlinear systems (see Theorems 1 and 2 in the work of Fang et al[27]).

No direct feedthrough: Suppose that there is no direct feedthrough[27,29] of the unknown input a_k to the output y_k so that $y_k = h(x_k) + v_k$. In this case, the unknown input must be estimated with one step delay, since y_{k+1} is the first measurement containing information on a_k. Hence, the Bayesian approach is based on the recursive computation of the joint PDF $p(a_{k-1}, x_k \,|\, y^k)$ of the unknown input a_{k-1} and state x_k conditioned on all the information available up to time k. Suppose that, at time $k-1$, the predicted posterior $p(a_{k-1}, x_k \,|\, y^{k-1})$ has been computed. Then, at time k, when the new measurement

y_k is collected, in the correction step, the new conditional PDF $p(a_{k-1}, x_k \,|\, y^k)$ can be obtained by means of the Bayes rule

$$p\left(a_{k-1}, x_k \,|\, y^k\right) = \frac{p(y_k \,|\, x_k) p\left(a_{k-1}, x_k \,|\, y^{k-1}\right)}{p\left(y_k \,|\, y^{k-1}\right)}, \tag{11}$$

while, in the prediction step, $p(a_k, x_{k+1} \,|\, y^k)$ takes the form

$$p\left(a_k, x_{k+1} \,|\, y^k\right) = p\left(x_{k+1} \,|\, a_k, y^k\right) p(a_k), \tag{12}$$

where the conditional PDF $p(x_{k+1} \,|\, a_k, y^k)$ is computed via the Chapman-Kolmogorov equation

$$p\left(x_{k+1} \,|\, a_k, y^k\right) = \int p(x_{k+1} \,|\, a_k, x_k) p\left(x_k \,|\, y^k\right) dx_k. \tag{13}$$

When no information on the unknown input a_k is supposed to be available so that $p(a_k)$ is taken as an uninformative (flat) prior, the conditional PDF $p(a_{k-1}, x_k \,|\, y^k)$ resulting from the correction step can be rewritten as

$$p\left(a_{k-1}, x_k \,|\, y^k\right) = \frac{p(y_k \,|\, x_k) p\left(x_k \,|\, a_{k-1}, y^{k-1}\right)}{\int \int p(y_k \,|\, x) p(x \,|\, a, y^{k-1}) dx da}. \tag{14}$$

2.3 | Random set estimation

A *random finite set* (RFS) $\mathcal{X}$ over $\mathbb{X}$ is a random variable taking values in $\mathcal{F}(\mathbb{X})$, the collection of all finite subsets of $\mathbb{X}$. The mathematical background needed for Bayesian random set estimation can be found in Mahler's book;[22] here, the basic concepts needed for the subsequent developments are briefly reviewed. From a probabilistic viewpoint, an RFS $\mathcal{X}$ is completely characterized by its *set density* $f(\mathcal{X})$, also called FISST (*FInite Set STatistics*) density. In fact, given $f(\mathcal{X})$, the cardinality *probability mass function* $\rho(n)$ that $\mathcal{X}$ have $n \geq 0$ elements and the joint PDFs $f(x_1, x_2, \ldots, x_n \,|\, n)$ over $\mathbb{X}^n$ given that $\mathcal{X}$ have n elements are obtained as follows:

$$\begin{aligned} \rho(n) &= \frac{1}{n!} \int\limits_{\mathbb{X}^n} f(\{x_1, \ldots, x_n\}) dx_1 \ldots dx_n \\ f(x_1, x_2, \ldots, x_n \,|\, n) &= \frac{1}{n!\,\rho(n)} f(\{x_1, \ldots, x_n\}). \end{aligned}$$

In order to measure probability over subsets of $\mathbb{X}$ or compute expectations of random set variables, Mahler[22] introduced the notion of *set integral* for a generic real-valued function $g(\mathcal{X})$ of an RFS $\mathcal{X}$ as

$$\int g(\mathcal{X}) \delta \mathcal{X} = g(\emptyset) + \sum_{n=1}^{\infty} \frac{1}{n!} \int g(\{x_1, \ldots, x_n\}) dx_1 \ldots dx_n. \tag{15}$$

In particular, in this work, we will consider the Bernoulli RFS, ie, a random set that can be either empty or, with some probability $r \in [0, 1]$, a singleton $\{x\}$ whose element is distributed over $\mathbb{X}$ according to the PDF $p(x)$. Accordingly, its set density is defined as follows:

$$f(\mathcal{X}) = \begin{cases} 1 - r, & \text{if } \mathcal{X} = \emptyset \\ r \cdot p(x), & \text{if } \mathcal{X} = \{x\}. \end{cases} \tag{16}$$

Please notice that the above equation as well as all subsequent definitions of probability distributions involving a Bernoulli set argument have two branches on the right-hand side depending on whether the Bernoulli argument is empty or a singleton.

3 | BAYESIAN RANDOM SET FILTER FOR JOINT ATTACK DETECTION AND STATE ESTIMATION: DIRECT FEEDTHROUGH CASE

Let us suppose that, when the attack input is present, there is a direct feedthrough from the attack a_k to the output y_k. More specifically, in accordance with the considerations of Section 2.2, it is assumed that, when the attack input is present, the mapping from a_k to y_k is full rank, ie, invertible. Let the attack input at time k be modeled as a BRS $\mathcal{A}_k \in \mathscr{B}(\mathbb{A})$, where $\mathscr{B}(\mathbb{A}) = \emptyset \cup \mathcal{S}(\mathbb{A})$ is a set of all finite subsets of the attack space $\mathbb{A} \subseteq \mathbb{R}^m$, and $\mathcal{S}(\mathbb{A})$ denotes the set of all singletons (ie, sets with cardinality 1) $\{a\}$ such that $a \in \mathbb{A}$. Further, let $\mathbb{X} \subseteq \mathbb{R}^n$ denote the state space for the system state vector, then we can define the *hybrid BRS* (HBRS) $(\mathcal{A}, x)$ as a new state variable that incorporates the Bernoulli attack random set $\mathcal{A}$ and the random state vector x, taking values in the hybrid space $\mathscr{B}(\mathbb{A}) \times \mathbb{X}$. An HBRS is fully specified by the (signal attack) probability r of $\mathcal{A}$ being a singleton, the PDF $p^0(x)$ defined on the state space $\mathbb{X}$, and the joint PDF $p^1(a, x)$ defined on the joint attack input state space $\mathbb{A} \times \mathbb{X}$, ie,

$$p(\mathcal{A}, x) = \begin{cases} (1-r)p^0(x), & \text{if } \mathcal{A} = \emptyset \\ r \cdot p^1(a, x), & \text{if } \mathcal{A} = \{a\}. \end{cases} \tag{17}$$

Moreover, since integration over $\mathscr{B}(\mathbb{A}) \times \mathbb{X}$ takes the form

$$\int_{\mathscr{B}(\mathbb{A})\times\mathbb{X}} p(\mathcal{A}, x)\delta\mathcal{A}\, dx = \int p(\emptyset, x)\, dx + \iint p(\{a\}, x)\, da\, dx, \tag{18}$$

where the set integration with respect to $\mathcal{A}$ is defined according to (15) while the integration with respect to x is an ordinary one, it is easy to see that $p(\mathcal{A}, x)$ integrates to one by substituting (17) in (18), and noting that $p^0(x)$ and $p^1(a, x)$ are conventional PDFs on $\mathbb{X}$ and $\mathbb{A} \times \mathbb{X}$, respectively. This, in turn, guarantees that (17) is a FISST probability density for the HBRS $(\mathcal{A}, x)$. The notion of *attack existence*, embodied by parameter r in (17), is introduced so as to detect the presence (existence) of a signal attack and hence initiate its estimation. Because of this concept, as shown later on, the probability of attack existence is directly computed by the filter.

In this paper, the attack input is modeled as a BRS to account for the fact that the attack can switch (from *off* to *on* or vice-versa) at any time with no prior knowledge on the attack onset/termination from the system monitor side. The switching nature of the attack could be tackled in different ways, eg, with multiple models (one for the attack and another for the no-attack cases), but the random set approach undertaken in this work turns out to be advantageous also to include other type of attacks, specifically packet substitution and extra packet injection to be considered in the next section.

3.1 | Measurement models and correction

3.1.1 | Packet substitution

Let us consider the *packet substitution* attack model introduced in Section 2.1 and denote by $\lambda(\mathcal{Z}_k \mid \mathcal{A}_k, x_k)$ the likelihood function of the measurement set defined in (3), which has obviously two possible forms, $\mathcal{A}_k$ being a BRS. In particular, for $\mathcal{A}_k = \emptyset$,

$$\lambda(\mathcal{Z}_k \mid \emptyset, x_k) = \begin{cases} 1 - p_d, & \text{if } \mathcal{Z}_k = \emptyset \\ p_d[(1-p_f)\ell(y_k \mid x_k) + p_f\, \kappa(y_k)], & \text{if } \mathcal{Z}_k = \{y_k\}, \end{cases} \tag{19}$$

where $\{y_k\}$ denotes the singleton whose element represents a delivered measurement, ie, $\lambda(\{y_k\} \mid \mathcal{A}_k, x_k)$ is the likelihood that a single measurement y_k will be collected. Furthermore, $\ell(y_k \mid x_k)$ is the standard likelihood function of the system-generated measurement y_k when no signal attack is present, whereas $\kappa(\cdot)$ is a PDF modeling the fake measurement y_k^f, assumed to be independent of the system state. Conversely, for $\mathcal{A}_k = \{a_k\}$,

$$\lambda(\mathcal{Z}_k \mid \{a_k\}, x_k) = \begin{cases} 1 - p_d, & \text{if } \mathcal{Z}_k = \emptyset \\ p_d[(1-p_f)\ell(y_k \mid a_k, x_k) + p_f\, \kappa(y_k)], & \text{if } \mathcal{Z}_k = \{y_k\}, \end{cases} \tag{20}$$

where $\ell(y_k \mid a_k, x_k)$ denotes the conventional likelihood of measurement y_k, due to the system under attack a_k in state x_k. Notice that, by using the definition of set integral (15), it is easy to check that both forms (19) and (20) of the likelihood function $\lambda(\mathcal{Z}_k \,|\, \mathcal{A}_k, x_k)$ integrate to one. Using the aforementioned measurement model, it is possible to derive the exact correction equations of the Bayesian random set filter for joint attack detection and state estimation, in case of substitution attack.

Theorem 1. (Correction under packet substitution attack)**.** *Suppose that the prior density at time k is hybrid Bernoulli of the form*

$$p(\mathcal{A}_k, x_k \,|\, \mathcal{Z}^{k-1}) = \begin{cases} (1 - r_{k\,|\,k-1}) p^0_{k\,|\,k-1}(x_k), & \text{if } \mathcal{A}_k = \emptyset \\ r_{k\,|\,k-1} \cdot p^1_{k\,|\,k-1}(a_k, x_k), & \text{if } \mathcal{A}_k = \{a_k\}. \end{cases} \tag{21}$$

Then, given the measurement random set $\mathcal{Z}_k$ defined in (3), also the posterior density at time k turns out to be hybrid Bernoulli of the form

$$p(\mathcal{A}_k, x_k \,|\, \mathcal{Z}^{k}) = \begin{cases} (1 - r_{k\,|\,k}) p^0_{k\,|\,k}(x_k), & \text{if } \mathcal{A}_k = \emptyset \\ r_{k\,|\,k} \cdot p^1_{k\,|\,k}(a_k, x_k), & \text{if } \mathcal{A}_k = \{a_k\}, \end{cases} \tag{22}$$

completely specified by the triplet

$$\left(r_{k\,|\,k}, p^0_{k\,|\,k}(x_k), p^1_{k\,|\,k}(a_k, x_k) \right) = \left(r_{k\,|\,k-1}, p^0_{k\,|\,k-1}(x_k), p^1_{k\,|\,k-1}(a_k, x_k) \right),$$

if $\mathcal{Z}_k = \emptyset$ or, if $\mathcal{Z}_k = \{y_k\}$, by

$$r_{k\,|\,k} = \frac{(1 - p_f)\Psi_1 + p_f \kappa(y_k)}{(1 - p_f)(\Psi_0 - r_{k\,|\,k-1}\Psi) + p_f \kappa(y_k)} r_{k\,|\,k-1} \tag{23}$$

$$p^0_{k\,|\,k}(x_k) = \frac{(1 - p_f)\ell(y_k \,|\, x_k) + p_f \kappa(y_k)}{(1 - p_f)\Psi_0 + p_f \kappa(y_k)} p^0_{k\,|\,k-1}(x_k) \tag{24}$$

$$p^1_{k\,|\,k}(a_k, x_k) = \frac{(1 - p_f)\ell(y_k \,|\, a_k, x_k) + p_f \kappa(y_k)}{(1 - p_f)\Psi_1 + p_f \kappa(y_k)} p^1_{k\,|\,k-1}(a_k, x_k), \tag{25}$$

where

$$\Psi_0 \triangleq \int \ell(y_k \,|\, x_k) p^0_{k\,|\,k-1}(x_k) \, \mathrm{d}x_k \tag{26}$$

$$\Psi_1 \triangleq \iint \ell(y_k \,|\, a_k, x_k) p^1_{k\,|\,k-1}(a_k, x_k) \, \mathrm{d}a_k \mathrm{d}x_k \tag{27}$$

$$\Psi \triangleq \Psi_0 - \Psi_1. \tag{28}$$

Proof. The correction equation of the Bayes random set filter for joint attack detection and state estimation follows from a generalization of (7), which yields

$$p(\mathcal{A}_k, x_k \,|\, \mathcal{Z}^{k}) = \frac{\lambda(\mathcal{Z}_k \,|\, \mathcal{A}_k, x_k) p\left(\mathcal{A}_k, x_k \,|\, \mathcal{Z}^{k-1}\right)}{p\left(\mathcal{Z}_k \,|\, \mathcal{Z}^{k-1}\right)}, \tag{29}$$

where $\lambda(\mathcal{Z}_k \,|\, \mathcal{A}_k, x_k)$ is given by (19) and (20), while

$$\begin{aligned} p\left(\mathcal{Z}_k \,|\, \mathcal{Z}^{k-1}\right) &= \iint \lambda(\mathcal{Z}_k \,|\, \mathcal{A}_k, x_k) p\left(\mathcal{A}_k, x_k \,|\, \mathcal{Z}^{k-1}\right) \delta\mathcal{A}_k \mathrm{d}x_k \\ &= \int \lambda(\mathcal{Z}_k \,|\, \emptyset, x_k) p\left(\emptyset, x_k \,|\, \mathcal{Z}^{k-1}\right) \mathrm{d}x_k + \iint \lambda(\mathcal{Z}_k \,|\, \{a_k\}, x_k) p\left(\{a_k\}, x_k \,|\, \mathcal{Z}^{k-1}\right) \mathrm{d}a_k \mathrm{d}x_k. \end{aligned} \tag{30}$$

For the case $\mathcal{Z}_k = \emptyset$, the above reduces to

$$p(\emptyset \,|\, \mathcal{Z}^{k-1}) = 1 - p_d, \tag{31}$$

by substituting (19)-(20) and (21) in (30) and simply noting that $\int p^0_{k\,|\,k-1}(x_k) \mathrm{d}x_k = 1$ and $\iint p^1_{k\,|\,k-1}(a_k, x_k) \mathrm{d}a_k \mathrm{d}x_k = 1$. The posterior probability of attack existence $r_{k|k}$ can be obtained from the posterior density (29) with $\mathcal{A}_k = \emptyset$ via

$$r_{k\,|\,k} = 1 - \int p\left(\emptyset, x_k \,|\, \mathcal{Z}^{k}\right) \mathrm{d}x_k, \tag{32}$$

where, using (19), (21), and (31) in (29), we have

$$p\left(\emptyset, x_k \,|\, \mathcal{Z}^k\right) = (1 - r_{k\,|\,k-1}) p^0_{k\,|\,k-1}(x_k). \tag{33}$$

Moreover, $p^0_{k\,|\,k}(x_k) = p(\emptyset, x_k \,|\, \mathcal{Z}^k)/(1 - r_{k\,|\,k})$, and the joint density for the system under attack can be easily derived from the posterior density with $\mathcal{A}_k = \{a_k\}$ by recalling that $p^1_{k\,|\,k}(a_k, x_k) = p(\{a_k\}, x_k \,|\, \mathcal{Z}^k)/r_{k\,|\,k}$, where

$$p\left(\{a_k\}, x_k \,|\, \mathcal{Z}^k\right) = r_{k\,|\,k-1} \cdot p^1_{k\,|\,k-1}(a_k, x_k) \tag{34}$$

results from replacing (20), (21), and (31) in (29). Notice that, from the set integral definition (15) and densities (33)-(34), it holds that $\int p(\emptyset, x_k \,|\, \mathcal{Z}^k) \mathrm{d}x_k + \iint p(\{a_k\}, x_k \,|\, \mathcal{Z}^k) \mathrm{d}a_k \mathrm{d}x_k = 1$. Hence, as stated, the Bayes correction (22) provides a hybrid Bernoulli density. Next, for the case $\mathcal{Z}_k = \{y_k\}$, (30) leads to

$$p(\{y_k\} \,|\, \mathcal{Z}^{k-1}) = p_d[(1 - p_f)(\Psi_1 - r_{k\,|\,k-1}\Psi) + p_f \kappa(y_k)], \tag{35}$$

so that, from (29), one gets

$$p\left(\emptyset, x_k \,|\, \mathcal{Z}^k\right) = \frac{[(1 - p_f)\ell(y_k \,|\, x_k) + p_f \kappa(y_k)]}{(1 - p_f)(\Psi_1 - r_{k\,|\,k-1}\Psi) + p_f \kappa(y_k)} (1 - r_{k\,|\,k-1}) p^0_{k\,|\,k-1}(x_k), \tag{36}$$

which, in turn, is used to obtain (23) through (32). Once $r_{k|k}$ is known, (24) immediately follows as previously shown for the case $\mathcal{Z}_k = \emptyset$, while (25) comes from dividing the posterior

$$p\left(\{a_k\}, x_k \,|\, \mathcal{Z}^k\right) = \frac{[(1 - p_f)\ell(y_k \,|\, x_k) + p_f \kappa(y_k)]}{(1 - p_f)(\Psi_1 - r_{k\,|\,k-1}\Psi) + p_f \kappa(y_k)} r_{k\,|\,k-1} p^1_{k\,|\,k-1}(a_k, x_k) \tag{37}$$

by $r_{k|k}$ in (23). □

3.1.2 | Extra packet injection

A complete derivation of the correction step for the *extra packet injection* model introduced in Section 2.1 can be found in the work of Forti et al.[30] We summarize below the main results, since they are the basis for the derivation of the GM filter of Section 4. First recall that, in this case, the measurement set $\mathcal{Z}_k$ is given by the union of the two independent random sets $\mathcal{Y}_k$ and $\mathcal{F}_k$. Clearly, in view of (5), $\mathcal{Y}_k$ is a BRS whose cardinality is either 0 or 1 depending on whether the system-originated measurement y_k is delivered or not. Conversely, it is supposed that no prior knowledge on the number of fake measurements, ie, the cardinality of $\mathcal{F}_k$, is available. Accordingly, $\rho(n)$ is taken as an uninformative distribution, and hence, the FISST PDF of fake-only measurements turns out to be

$$\gamma(\mathcal{F}_k) \propto |\mathcal{F}_k|! \prod_{y_k \in \mathcal{F}_k} \kappa(y_k), \tag{38}$$

where $\kappa(y_k)$ is a PDF describing the distribution of fake measurements on the measurement space $\mathbb{Y}$. Clearly, if no prior knowledge on such a distribution can be assumed, the same approach of Section 2.1 can be followed by taking $\kappa(y_k)$ as an uninformative (ie, uniform) PDF over $\mathbb{Y}$. The following result holds.

Theorem 2 (Correction under extra packet injection attack, see the work of Forti et al[30] for the proof). *Suppose that the prior density at time k is hybrid Bernoulli of the form*

$$p\left(\mathcal{A}_k, x_k \,|\, \mathcal{Z}^{k-1}\right) = \begin{cases} (1 - r_{k\,|\,k-1}) p^0_{k\,|\,k-1}(x_k), & \text{if } \mathcal{A}_k = \emptyset \\ r_{k\,|\,k-1} \cdot p^1_{k\,|\,k-1}(a_k, x_k), & \text{if } \mathcal{A}_k = \{a_k\}. \end{cases} \tag{39}$$

Then, given the measurement random set $\mathcal{Z}_k$ defined in (4), also the posterior density at time k turns out to be hybrid Bernoulli of the form

$$p(\mathcal{A}_k, x_k \,|\, \mathcal{Z}^k) = \begin{cases} (1 - r_{k|k}) p^0_{k|k}(x_k), & \text{if } \mathcal{A}_k = \emptyset \\ r_{k|k} \cdot p^1_{k|k}(a_k, x_k), & \text{if } \mathcal{A}_k = \{a_k\}, \end{cases} \tag{40}$$

completely specified by the triplet

$$r_{k|k} = \frac{1 - p_d(1 - \Gamma_1)}{1 - p_d[1 - (\Gamma_0 - r_{k|k-1}\Gamma)]} r_{k|k-1} \tag{41}$$

$$p^0_{k|k}(x_k) = \frac{1 - p_d + p_d \sum\limits_{y_k \in \mathcal{Z}_k} \frac{\ell(y_k | x_k)}{n\,\kappa(y_k)}}{1 - p_d(1 - \Gamma_0)} p^0_{k|k-1}(x_k) \tag{42}$$

$$p^1_{k|k}(a_k, x_k) = \frac{1 - p_d + p_d \sum\limits_{y_k \in \mathcal{Z}_k} \frac{\ell(y_k | a_k, x_k)}{n\,\kappa(y_k)}}{1 - p_d(1 - \Gamma_1)} p^1_{k|k-1}(a_k, x_k), \tag{43}$$

where

$$\Gamma_0 \triangleq \sum_{y_k \in \mathcal{Z}_k} \frac{\int \ell(y_k | x_k) p^0_{k|k-1}(x_k) \mathrm{d}x_k}{n\,\kappa(y_k)} \tag{44}$$

$$\Gamma_1 \triangleq \sum_{y_k \in \mathcal{Z}_k} \frac{\iint \ell(y_k | a_k, x_k) p^1_{k|k-1}(a_k, x_k) \mathrm{d}a_k \mathrm{d}x_k}{n\,\kappa(y_k)} \tag{45}$$

and $\Gamma \triangleq \Gamma_0 - \Gamma_1$.

3.2 | Dynamic model and prediction

Let us now focus on the prediction step of the Bayesian HBF. Concerning the propagation of the signal attack from time k to time $k+1$, we consider the most general model for signal attacks where any value can be injected and, accordingly, we model a_{k+1} as a completely unknown input whose value does not depend on the values a_k and x_k of attack and, respectively, state at time k. However, concerning the existence of the attack at time $k+1$, we introduce two parameters p_s and p_b to model the fact that the presence of an attack at time $k+1$ is more probable when an attack is already present at time k: p_b denotes the probability that an attack a_{k+1} is launched to the system at time $k+1$ when the system is under normal operation at time k; p_s denotes the probability that an adversarial action affecting the system at time k will endure to time $k+1$. Notice that the probabilities p_b and p_s have to be regarded as design parameters for the filter that can be tuned depending on the desired properties: the lower is p_b, the more cautious will be the filter in declaring the presence of an attack; the higher is p_s, the more cautious will be the filter in declaring that the attack has disappeared. According to this model, the transition density $\pi(\mathcal{A}_{k+1} \,|\, \mathcal{A}_k)$ of the attack BRS takes the form

$$\pi(\mathcal{A}_{k+1} \,|\, \emptyset) = \begin{cases} 1 - p_b, & \text{if } \mathcal{A}_{k+1} = \emptyset \\ p_b\, p(a_{k+1}), & \text{if } \mathcal{A}_{k+1} = \{a_{k+1}\} \end{cases}$$

$$\pi(\mathcal{A}_{k+1} \,|\, \{a_k\}) = \begin{cases} 1 - p_s, & \text{if } \mathcal{A}_{k+1} = \emptyset \\ p_s\, p(a_{k+1}), & \text{if } \mathcal{A}_{k+1} = \{a_{k+1}\}. \end{cases}$$

Like in Section 2.2, $p(a_{k+1})$ is the PDF summarizing the available knowledge on a_{k+1}, which can be taken equal to an uninformative PDF (eg, uniform over the attack space) when the attack vector is completely unknown.

Then, the joint transition density of $(\mathcal{A}, x)$ at time $k+1$ takes the form

$$\pi(\mathcal{A}_{k+1}, x_{k+1} \,|\, \mathcal{A}_k, x_k) = \pi(x_{k+1} \,|\, \mathcal{A}_k, x_k)\, \pi(\mathcal{A}_{k+1} \,|\, \mathcal{A}_k), \tag{46}$$

where, in accordance with (1), we have

$$\pi(x_{k+1} \,|\, \mathcal{A}_k, x_k) = \begin{cases} \pi(x_{k+1} \,|\, x_k), & \text{if } \mathcal{A}_k = \emptyset \\ \pi(x_{k+1} \,|\, a_k, x_k), & \text{if } \mathcal{A}_k = \{a_k\}, \end{cases} \tag{47}$$

with $\pi(x_{k+1} \mid x_k)$ and $\pi(x_{k+1} \mid a_k, x_k)$ known Markov transition PDFs.

Under the above assumptions, an exact recursion for the prior density can be obtained.

Theorem 3 (See the work of Forti et al[30] for the proof). *Given the posterior hybrid Bernoulli density $p(\mathcal{A}_k, x_k \mid \mathcal{Z}^k)$ at time k of the form (22), fully characterized by the triplet $(r_{k\mid k}, p^0_{k\mid k}(x_k), p^1_{k\mid k}(a_k, x_k))$, also the predicted density turns out to be hybrid Bernoulli of the form*

$$p\left(\mathcal{A}_{k+1}, x_{k+1} \mid \mathcal{Z}^k\right) = \begin{cases} (1 - r_{k+1\mid k})p^0_{k+1\mid k}(x_{k+1}), & \text{if } \mathcal{A}_{k+1} = \emptyset \\ r_{k+1\mid k} \cdot p^1_{k+1\mid k}(a_{k+1}, x_{k+1}), & \text{if } \mathcal{A}_{k+1} = \{a_{k+1}\}, \end{cases} \tag{48}$$

with

$$r_{k+1\mid k} = (1 - r_{k\mid k})p_b + r_{k\mid k}p_s \tag{49}$$

$$p^0_{k+1\mid k}(x_{k+1}) = \frac{(1 - r_{k\mid k})(1 - p_b)p_{k+1\mid k}(x_{k+1} \mid \emptyset)}{1 - r_{k+1\mid k}} + \frac{r_{k\mid k}(1 - p_s)p_{k+1\mid k}(x_{k+1} \mid \{a_k\})}{1 - r_{k+1\mid k}} \tag{50}$$

$$p^1_{k+1\mid k}(a_{k+1}, x_{k+1}) = \frac{(1 - r_{k\mid k})p_b p_{k+1\mid k}(x_{k+1} \mid \emptyset)p(a_{k+1})}{r_{k+1\mid k}} + \frac{r_{k\mid k}p_s p_{k+1\mid k}(x_{k+1} \mid \{a_k\})p(a_{k+1})}{r_{k+1\mid k}}, \tag{51}$$

where

$$p_{k+1\mid k}(x_{k+1} \mid \emptyset) = \int \pi(x_{k+1} \mid x_k)p^0_{k\mid k}(x_k)\,\mathrm{d}x_k \tag{52}$$

$$p_{k+1\mid k}(x_{k+1} \mid \{a_k\}) = \iint \pi(x_{k+1} \mid a_k, x_k)p^1_{k\mid k}(a_k, x_k)\,\mathrm{d}a_k\mathrm{d}x_k. \tag{53}$$

Notice that, if $p_b = 0$, $p_s = 1$, and $r_{k\mid k} = 1$, it follows that $r_{k+1\mid k} = 1$ and $p^1_{k+1\mid k}(a_{k+1}, x_{k+1}) = p_{k+1\mid k}(x_{k+1} \mid \{a_k\})p(a_{k+1})$. Hence, in this case, we recover the standard Chapman-Kolmogorov Equation (9) for the system under attack.

Remark 1. Given the conditional density $p(\mathcal{A}_k, x_k \mid \mathcal{Z}^k)$, characterized by the triplet $(r_{k\mid k}, p^0_{k\mid k}(\cdot), p^1_{k\mid k}(\cdot,\cdot))$, the joint attack detection and state estimation problem can be solved as follows. First of all, we perform attack detection using $r_{k|k}$ from the available current hybrid Bernoulli density $p(\mathcal{A}_k, x_k \mid \mathcal{Z}^k)$. By using a MAP decision rule, given $\mathcal{Z}_k$, the detector will assign $\hat{\mathcal{A}}_k \neq \emptyset$ (the system is under attack) if and only if $\text{Prob}(\mathcal{A}_k \neq \emptyset \mid \mathcal{Z}^k) > \text{Prob}(\mathcal{A}_k = \emptyset \mid \mathcal{Z}^k)$, ie, if and only if $r_{k\mid k} > 1/2$. Then, if the signal attack has been detected, one can maximize $p(\mathcal{A}_k, x_k \mid \mathcal{Z}^k)$ with respect to x_k and a_k. In this way, it is possible to obtain a MAP estimate of x_k and an ML estimate of the unknown attack input a_k. The above joint attack detection and state estimation process is illustrated in Figure 3.

Remark 2. The Bayesian formulation of this section has allowed to generalize the standard *joint input and state* filtering process to take into account several practically relevant issues like the switching nature of the attack input, the injection of fake measurements or replacement of system-originated by fake measurements, and the possible lack of system-originated measurements. Please notice that all such phenomena are not contemplated in the standard filtering process.

Remark 3. The hybrid Bernoulli filtering recursions derived in this section are rarely solvable in explicit form but, as it will be shown in the next section, this is possible in the linear Gaussian case. In such a case, in fact, the propagated PDFs $p^0_{k\mid k}(\cdot)$ and $p^1_{k\mid k}(\cdot,\cdot)$ turn out to be GMs at any time k, even if with a number of Gaussian components growing with time and hence to be reduced via suitable pruning and merging procedures.

Remark 4. It is clear from the previous derivations that the defense method against signal attacks is embedded in the proposed HBF and can be coordinated with any of the defense methods against the two considered data attacks, either packet substitution or extra packet injection. In fact, it suffices to perform the correction step of the HBF according to either Theorem 1 or Theorem 2 while the prediction step is clearly unaffected by the choice of the data attack model. Please notice that packet substitution and extra packet injection attacks are clearly alternative and that the HBF can switch from counteracting one or the other at any time, just by choosing the appropriate correction step, depending on whether the system monitor receives a single or multiple data packets during the sampling interval. The

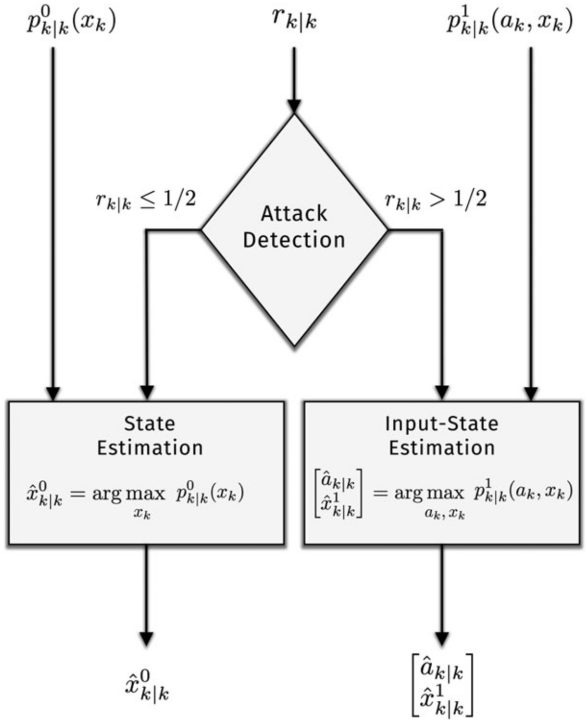
$p^0_{k|k}(x_k)$
$r_{k|k}$
$p^1_{k|k}(a_k, x_k)$
$r_{k|k} \leq 1/2$
Attack Detection
$r_{k|k} > 1/2$
State Estimation
$\hat{x}^0_{k|k} = \arg\max_{x_k} p^0_{k|k}(x_k)$
Input-State Estimation
$\begin{bmatrix}\hat{a}_{k|k}\\ \hat{x}^1_{k|k}\end{bmatrix} = \arg\max_{a_k, x_k} p^1_{k|k}(a_k, x_k)$
$\hat{x}^0_{k|k}$
$\begin{bmatrix}\hat{a}_{k|k}\\ \hat{x}^1_{k|k}\end{bmatrix}$

FIGURE 3 Block diagram of the joint attack detection and state estimation process

above described strategy could, therefore, provide a sensible way to coordinate the defense methods against packet substitution and extra packet injection cyberattacks.

4 | BAYESIAN RANDOM SET FILTER FOR JOINT ATTACK DETECTION AND STATE ESTIMATION: NO DIRECT FEEDTHROUGH CASE

Suppose now that, even when the attack input is present, there is no direct feedthrough from the attack a_k to the output y_k, so that the measurement model is

$$y_k = h(x_k) + v_k, \tag{54}$$

irrespectively of the presence of the attack. In this case, clearly, the attack set $\mathcal{A}_k$ must be estimated with one step delay, since $\mathcal{Z}_{k+1}$ is the first measurement set containing information on $\mathcal{A}_k$. In the following sections, a detailed derivation of the correction and prediction steps of the Bayes recursion in the case of no direct feedthrough is provided.

4.1 | Measurement models and correction

In the case of packet substitution with no direct feedthrough, the likelihood function $\lambda(\mathcal{Z}_k | x_k)$ takes the following form:

$$\lambda(\mathcal{Z}_k | x_k) = \begin{cases} 1 - p_d, & \text{if } \mathcal{Z}_k = \emptyset \\ p_d[(1-p_f)\ell(y_k | x_k) + p_f \, \kappa(y_k)], & \text{if } \mathcal{Z}_k = \{y_k\}, \end{cases} \tag{55}$$

where $\ell(y_k \,|\, x_k)$ is the standard likelihood function of the system-generated measurement y_k. It is easy to check that the likelihood function $\lambda(\mathcal{Z}_k | x_k)$ integrates to one.

Instead, in the case of extra packet injection attack with no direct feedthrough, it can be shown that the likelihood function $\lambda(\mathcal{Z}_k | x_k)$ can be written as

$$\lambda(\mathcal{Z}_k | x_k) = \gamma(\mathcal{Z}_k) \left[1 - p_d + p_d \sum_{y_k \in \mathcal{Z}_k} \frac{\ell(y_k | x_k)}{n \, \kappa(y_k)} \right], \tag{56}$$

where n denotes the cardinality of $\mathcal{Z}_k$, ie, the number of received measurements.

Hence, the following result holds (the proof is omitted since it follows along the same lines as the proofs of Theorems 1 and 2).

Theorem 4 (Correction without direct feedthrough). *Suppose that the prior density at time k is hybrid Bernoulli of the form*

$$p\left(\mathcal{A}_{k-1}, x_k \,|\, \mathcal{Z}^{k-1}\right) = \begin{cases} (1 - r_{k\,|\,k-1}) p^0_{k\,|\,k-1}(x_k), & \text{if } \mathcal{A}_{k-1} = \emptyset \\ r_{k\,|\,k-1} \cdot p^1_{k\,|\,k-1}(a_{k-1}, x_k), & \text{if } \mathcal{A}_{k-1} = \{a_{k-1}\}. \end{cases} \tag{57}$$

Then, given the measurement random set $\mathcal{Z}_k$ for packet substitution attack, also the posterior density at time k turns out to be hybrid Bernoulli of the form

$$p\left(\mathcal{A}_{k-1}, x_k \,|\, \mathcal{Z}^{k}\right) = \begin{cases} (1 - r_{k\,|\,k}) p^0_{k\,|\,k}(x_k), & \text{if } \mathcal{A}_{k-1} = \emptyset \\ r_{k\,|\,k} \cdot p^1_{k\,|\,k}(a_{k-1}, x_k), & \text{if } \mathcal{A}_{k-1} = \{a_{k-1}\}. \end{cases} \tag{58}$$

The triplet $(r_{k\,|\,k}, p^0_{k\,|\,k}(x_k), p^1_{k\,|\,k}(a_{k-1}, x_k))$ completely specifying the posterior density can be computed as in Theorem 1 for the case of packet substitution and as in Theorem 2 for the case of extra packet injection attack, provided that a_k, $\mathcal{A}_k$, and $\ell(y_k \,|\, a_k, x_k)$ are replaced by a_{k-1}, $\mathcal{A}_{k-1}$, and $\ell(y_k \,|\, x_k)$, respectively.

4.2 | Dynamic model and prediction

The joint transition density takes the form

$$\pi(\mathcal{A}_k, x_{k+1} \,|\, \mathcal{A}_{k-1}, x_k) = \pi(x_{k+1} \,|\, \mathcal{A}_k, x_k)\, \pi(\mathcal{A}_k \,|\, \mathcal{A}_{k-1}), \tag{59}$$

where

$$\pi(x_{k+1} \,|\, \mathcal{A}_k, x_k) = \begin{cases} \pi(x_{k+1} \,|\, x_k), & \text{if } \mathcal{A}_k = \emptyset \\ \pi(x_{k+1} \,|\, a_k, x_k), & \text{if } \mathcal{A}_k = \{a_k\}, \end{cases} \tag{60}$$

with $\pi(x_{k+1} \,|\, x_k)$ and $\pi(x_{k+1} \,|\, a_k, x_k)$ known Markov transition PDFs.

The transition density $\pi(\mathcal{A}_k \,|\, \mathcal{A}_{k-1})$ of the attack BRS takes the form

$$\pi(\mathcal{A}_k \,|\, \emptyset) = \begin{cases} 1 - p_b, & \text{if } \mathcal{A}_k = \emptyset \\ p_b\, p(a_k), & \text{if } \mathcal{A}_k = \{a_k\} \end{cases}$$

$$\pi(\mathcal{A}_k | \{a_{k-1}\}) = \begin{cases} 1 - p_s, & \text{if } \mathcal{A}_k = \emptyset \\ p_s\, p(a_k), & \text{if } \mathcal{A}_k = \{a_k\}, \end{cases}$$

where $p(a_k)$ is the PDF summarizing the available knowledge on a_k, which can be taken equal to an uninformative PDF (eg, uniform over the attack space) when the attack vector is completely unknown.

Theorem 5 (Prediction without direct feedthrough). *Given the posterior hybrid Bernoulli density $p(\mathcal{A}_{k-1}, x_k \,|\, \mathcal{Z}^k)$ at time k of the form (58), fully characterized by the triplet $(r_{k\,|\,k}, p^0_{k\,|\,k}(x_k), p^1_{k\,|\,k}(a_{k-1}, x_k))$, also the predicted density turns out to be hybrid Bernoulli of the form*

$$p\left(\mathcal{A}_k, x_{k+1} \,|\, \mathcal{Z}^{k}\right) = \begin{cases} (1 - r_{k+1\,|\,k}) p^0_{k+1\,|\,k}(x_{k+1}), & \text{if } \mathcal{A}_k = \emptyset \\ r_{k+1\,|\,k} \cdot p^1_{k+1\,|\,k}(a_k, x_{k+1}), & \text{if } \mathcal{A}_k = \{a_k\}, \end{cases} \tag{61}$$

with

$$r_{k+1\,|\,k} = (1 - r_{k\,|\,k}) p_b + r_{k\,|\,k} p_s \tag{62}$$

$$p^0_{k+1\,|\,k}(x_{k+1}) = \frac{(1 - r_{k\,|\,k})(1 - p_b) p_{k+1\,|\,k}(x_{k+1} \,|\, \emptyset)}{1 - r_{k+1\,|\,k}} + \frac{r_{k\,|\,k}(1 - p_s) p_{k+1\,|\,k}(x_{k+1} | \{a_{k-1}\})}{1 - r_{k+1\,|\,k}} \tag{63}$$

$$p^1_{k+1\,|\,k}(a_k, x_{k+1}) = \frac{(1 - r_{k\,|\,k}) p_b\, p_{k+1\,|\,k}(x_{k+1} | \{a_k\}, \emptyset) p(a_k)}{r_{k+1\,|\,k}} + \frac{r_{k\,|\,k} p_s\, p_{k+1\,|\,k}(x_{k+1} | \{a_k\}, \{a_{k-1}\}) p(a_k)}{r_{k+1\,|\,k}}, \tag{64}$$

where

$$p_{k+1\,|\,k}(x_{k+1}\,|\,\emptyset) \triangleq \int \pi(x_{k+1}\,|\,x_k)\,p^0_{k\,|\,k}(x_k)\,\mathrm{d}x_k \tag{65}$$

$$p_{k+1\,|\,k}(x_{k+1}|\{a_{k-1}\}) \triangleq \iint \pi(x_{k+1}\,|\,x_k)\,p^1_{k\,|\,k}(a_{k-1},x_k)\,\mathrm{d}a_{k-1}\mathrm{d}x_k \tag{66}$$

$$p_{k+1\,|\,k}(x_{k+1}|\{a_k\},\emptyset) \triangleq \int \pi(x_{k+1}\,|\,a_k,x_k)\,p^0_{k\,|\,k}(x_k)\,\mathrm{d}x_k \tag{67}$$

$$p_{k+1\,|\,k}(x_{k+1}|\{a_k\},\{a_{k-1}\}) \triangleq \iint \pi(x_{k+1}\,|\,a_k,x_k)\,p^1_{k\,|\,k}(a_{k-1},x_k)\,\mathrm{d}a_{k-1}\mathrm{d}x_k. \tag{68}$$

Proof. The prediction equation is given by the following generalization of (12):

$$\begin{aligned} p\left(\mathcal{A}_k,x_{k+1}\,|\,\mathcal{Z}^k\right) &= \iint \pi(\mathcal{A}_k,x_{k+1}\,|\,\mathcal{A}_{k-1},x_k)\,p\left(\mathcal{A}_{k-1},x_k\,|\,\mathcal{Z}^k\right)\,\delta\mathcal{A}_{k-1}\mathrm{d}x_k \\ &= (1-r_{k\,|\,k})\int \pi(\mathcal{A}_k,x_{k+1}\,|\,\emptyset,x_k)\,p^0_{k\,|\,k}(x_k)\,\mathrm{d}x_k + r_{k\,|\,k}\iint \pi(\mathcal{A}_k,x_{k+1}|\{a_{k-1}\},x_k)\,p^1_{k\,|\,k}(a_{k-1},x_k)\,\mathrm{d}a_{k-1}\mathrm{d}x_k. \end{aligned}$$

Then, for $\mathcal{A}_k = \emptyset$, one has

$$\begin{aligned} p\left(\emptyset,x_{k+1}\,|\,\mathcal{Z}^k\right) &= (1-r_{k\,|\,k})(1-p_b)\int \pi(x_{k+1}\,|\,x_k)\,p^0_{k\,|\,k}(x_k)\,\mathrm{d}x_k + r_{k\,|\,k}(1-p_s)\iint \pi(x_{k+1}\,|\,x_k)\,p^1_{k\,|\,k}(a_{k-1},x_k)\,\mathrm{d}a_{k-1}\mathrm{d}x_k \\ p\left(\emptyset,x_{k+1}\,|\,\mathcal{Z}^k\right) &= (1-r_{k\,|\,k})(1-p_b)\,p_{k+1\,|\,k}(x_{k+1}\,|\,\emptyset) + r_{k\,|\,k}(1-p_s)\,p_{k+1\,|\,k}(x_{k+1}|\{a_{k-1}\}). \end{aligned}$$

Analogously, for $\mathcal{A}_k = \{a_k\}$, we obtain

$$p(\{a_k\},x_{k+1}\,|\,\mathcal{Z}^k) = [(1-r_{k\,|\,k})\,p_b\,p_{k+1\,|\,k}(x_{k+1}|\{a_k\},\emptyset) + r_{k\,|\,k}\,p_s\,p_{k+1\,|\,k}(x_{k+1}|\{a_k\},\{a_{k-1}\})]\,p(a_k).$$

Thus, the output of the prediction step is of the form (61), fully specified by (62)-(64). □

5 | GAUSSIAN MIXTURE HYBRID BERNOULLI FILTER

While, in general, no exact closed-form solution to the proposed HBF is admitted, for the special class of linear Gaussian models, this problem can be effectively mitigated by parameterizing the posterior densities $p^0_{k\,|\,k}(\cdot)$ and $p^1_{k\,|\,k}(\cdot,\cdot)$ via GMs so as to derive a GM-HBF. This approach can be generalized to nonlinear models and/or non-Gaussian noises via nonlinear extensions of the GM approximation based on nonlinear filtering techniques such as the extended Kalman filter (EKF) or the unscented Kalman filter (UKF). In what follows, a detailed derivation of the GM-HBF for linear Gaussian models is provided. For the sake of brevity, only the direct feedthrough case (Section 3) is considered. The GM implementation in the case of no direct feedthrough (Section 4) can be derived in a similar way.

Denoting by $\mathcal{N}(x;m,P)$ a Gaussian PDF in the variable x, with mean m and covariance P, the closed-form GM-HBF assumes linear Gaussian observation, transition, and (a priori) attack models, ie,

$$\ell(y_k\,|\,x_k) = \mathcal{N}(y_k;Cx_k,R) \tag{69}$$

$$\ell(y_k\,|\,a_k,x_k) = \mathcal{N}(y_k;Cx_k+Ha_k,R) \tag{70}$$

$$\pi(x_{k+1}\,|\,x_k) = \mathcal{N}(x_{k+1};Ax_k,Q) \tag{71}$$

$$\pi(x_{k+1}\,|\,a_k,x_k) = \mathcal{N}(x_{k+1};Ax_k+Ga_k,Q) \tag{72}$$

$$p(a) = \sum_{j=1}^{J^a} \tilde{\varpi}^{a,j}\,\mathcal{N}(a;\tilde{a}^j,\tilde{P}^{a,j}). \tag{73}$$

Note that (73) uses given model parameters $J^a, \tilde{\omega}^{a,j}, \tilde{a}^j, \tilde{P}^{a,j}, j = 1, \dots, J^a$, to define the a priori PDF of the signal attack, here expressed as a GM and supposed time independent.

In the GM implementation, each probability density at time k is represented by the following set of parameters:

$$\left(r_{k|k}, p^0_{k|k}(x_k), p^1_{k|k}(a_k, x_k)\right) = \left(r_{k|k}, \left\{\omega^{0,j}_{k|k}, m^{0,j}_{k|k}, P^{0,j}_{k|k}\right\}_{j=1}^{J^0_{k|k}}, \left\{\omega^{1,j}_{k|k}, m^{1,j}_{k|k}, P^{1,j}_{k|k}\right\}_{j=1}^{J^1_{k|k}}\right), \tag{74}$$

where ω and J indicate, respectively, weights and number of mixture components, such that

$$p^0_{k|k}(x_k) = \sum_{j=1}^{J^0_{k|k}} \omega^{0,j}_{k|k} \mathcal{N}\left(m^{0,j}_{k|k}, P^{0,j}_{k|k}\right) \tag{75}$$

$$p^1_{k|k}(a_k, x_k) = \sum_{j=1}^{J^1_{k|k}} \omega^{1,j}_{k|k} \mathcal{N}\left(m^{1,j}_{k|k}, P^{1,j}_{k|k}\right), \tag{76}$$

with $m^0_{k|k} = \hat{x}^0_{k|k}$, $m^1_{k|k} = [\hat{x}^{1^T}_{k|k}, \hat{a}^T_k]^T$, $P^0_{k|k} \triangleq \mathbb{E}[(x_k - \hat{x}^0_{k|k})(x_k - \hat{x}^0_{k|k})^T]$, $P^1_{k|k} = \begin{bmatrix} P^{1x}_{k|k} & P^{xa}_k \\ P^{ax}_k & P^a_k \end{bmatrix}$, and $P^{1x}_{k|k} \triangleq \mathbb{E}[(x_k - \hat{x}^1_{k|k})(x_k - \hat{x}^1_{k|k})^T]$, $(P^{xa}_k)^T = P^{ax}_k \triangleq \mathbb{E}[(a_k - \hat{a}_k)(x_k - \hat{x}^1_{k|k})^T]$, $P^a_k \triangleq \mathbb{E}[(a_k - \hat{a}_k)(a_k - \hat{a}_k)^T]$. The weights are such that $\sum_{j=1}^{J^0_{k|k}} \omega^{0,j}_{k|k} = 1$ and $\sum_{j=1}^{J^1_{k|k}} \omega^{1,j}_{k|k} = 1$.

The GM implementation of the HBF (GM-HBF) is described as follows.

5.1 | GM-HBF correction for packet substitution

Proposition 1. *Suppose that assumptions (69)-(73) hold, the measurement set $\mathcal{Z}_k$ is defined by (3), the predicted FISST density at time k is fully specified by the triplet $(r_{k|k-1}, p^0_{k|k-1}(x_k), p^1_{k|k-1}(a_k, x_k))$, and $p^0_{k|k-1}(\cdot)$, $p^1_{k|k-1}(\cdot,\cdot)$ are GMs of the form*

$$p^0_{k|k-1}(x_k) = \sum_{j=1}^{J^0_{k|k-1}} \omega^{0,j}_{k|k-1} \mathcal{N}\left(m^{0,j}_{k|k-1}, P^{0,j}_{k|k-1}\right) \tag{77}$$

$$p^1_{k|k-1}(a_k, x_k) = \sum_{j=1}^{J^1_{k|k-1}} \omega^{1,j}_{k|k-1} \mathcal{N}\left(m^{1,j}_{k|k-1}, P^{1,j}_{k|k-1}\right). \tag{78}$$

Then, the posterior FISST density $(r_{k|k}, p^0_{k|k}(x_k), p^1_{k|k}(a_k, x_k))$ is given by

$$r_{k|k} = \frac{(1 - p_f)\Psi_1 + p_f \kappa(y_k)}{(1 - p_f)(\Psi_0 - r_{k|k-1}\Psi) + p_f \kappa(y_k)} r_{k|k-1} \tag{79}$$

$$p^0_{k|k}(x_k) = \sum_{j=1}^{J^0_{k|k}} \omega^{0,j}_{k|k} \mathcal{N}\left(m^{0,j}_{k|k}, P^{0,j}_{k|k}\right) = \sum_{j=1}^{J^0_{k|k-1}} \omega^{0,j}_{F,k|k} \mathcal{N}\left(m^{0,j}_{k|k-1}, P^{0,j}_{k|k-1}\right) + \sum_{j=1}^{J^0_{k|k-1}} \omega^{0,j}_{\bar{F},k|k} \mathcal{N}\left(m^{0,j}_{k|k}, P^{0,j}_{k|k}\right) \tag{80}$$

$$p^1_{k|k}(a_k, x_k) = \sum_{j=1}^{J^1_{k|k}} \omega^{1,j}_{k|k} \mathcal{N}\left(m^{1,j}_{k|k}, P^{1,j}_{k|k}\right) = \sum_{j=1}^{J^1_{k|k-1}} \omega^{1,j}_{F,k|k} \mathcal{N}\left(m^{1,j}_{k|k-1}, P^{1,j}_{k|k-1}\right) + \sum_{j=1}^{J^1_{k|k-1}} \omega^{1,j}_{\bar{F},k|k} \mathcal{N}\left(m^{1,j}_{k|k}, P^{1,j}_{k|k}\right), \tag{81}$$

where

$$\omega^{i,j}_{F,k|k} = \frac{p_f \kappa(y_k) \omega^{i,j}_{k|k-1}}{(1 - p_f)\Psi_i + p_f \kappa(y_k)}, \tag{82}$$

$$\omega^{i,j}_{\bar{F},k|k} = \frac{(1 - p_f) \omega^{i,j}_{k|k-1}}{(1 - p_f)\Psi_i + p_f \kappa(y_k)} q^{i,j}_k(y_k), \tag{83}$$

for i $= 0, 1$, *while*

$$q^{0,j}_k(y_k) = \mathcal{N}\left(y; C m^{0,j}_{k|k-1}, C P^{0,j}_{k|k-1} C^T + R\right) \tag{84}$$

$$q_k^{1,j}(y_k) = \mathcal{N}\left(y; \tilde{C}m_{k|k-1}^{1,j}, \tilde{C}P_{k|k-1}^{1,j}\tilde{C}^T + R\right), \tag{85}$$

with $\tilde{C} \overset{\triangle}{=} [C, H]$, $\Psi_0 = \sum_{j=1}^{J_{k|k-1}^0} \omega_{k|k-1}^{0,j} q_k^{0,j}(y_k)$, *and* $\Psi_1 = \sum_{j=1}^{J_{k|k-1}^1} \omega_{k|k-1}^{1,j} q_k^{1,j}(y_k)$.

Proof. From Theorem 1, the corrected probability of signal attack existence is provided by (23) where Ψ_0 is obtained by substituting (69) and (77) into (26), so that

$$\Psi_0 = \int \mathcal{N}(y; Cx_k, R) \sum_{j=1}^{J_{k|k-1}^0} \omega_{k|k-1}^{0,j} \mathcal{N}\left(m_{k|k-1}^{0,j}, P_{k|k-1}^{0,j}\right) \mathrm{d}x_k. \tag{86}$$

Then, by applying a standard result for Gaussian functions,[31, Lemma 1] we can write

$$\int \mathcal{N}(y; Cx_k, R) \mathcal{N}\left(m_{k|k-1}^{0,j}, P_{k|k-1}^{0,j}\right) \mathrm{d}x_k = q_k^{0,j}(y_k), \tag{87}$$

where $q_k^{0,j}(y_k)$ is given by (84), and hence, (86) takes the form

$$\Psi_0 = \sum_{j=1}^{J_{k|k-1}^0} \omega_{k|k-1}^{0,j} q_k^{0,j}(y_k). \tag{88}$$

Moreover, Ψ_1 in (79) can be analogously obtained by substituting (70) and (78) into (27), and by applying lemma 1 in the work of Vo and Ma[31] to the (double) integral $\iint \mathcal{N}(y; Cx_k + Ha_k, R)\mathcal{N}(m_{k|k-1}^{1,j}, P_{k|k-1}^{1,j}) \mathrm{d}a_k \mathrm{d}x_k$, so as to obtain

$$\Psi_1 = \sum_{j=1}^{J_{k|k-1}^1} \omega_{k|k-1}^{1,j} q_k^{1,j}(y_k), \tag{89}$$

where $q^{1,j}(y_k)$ is given by (85) and $m_{k|k-1}^{1,j} = [(\hat{x}_{k|k-1}^1)^T, (\hat{a}_k^j)^T]^T$.

Next, the posterior density $p_{k|k}^0(\cdot)$ can be derived from (24) in Theorem 1 as

$$p_{k|k}^0(x_k) = \frac{p_f\, \kappa(y_k)}{(1-p_f)\Psi_0 + p_f\, \kappa(y_k)} p_{k|k-1}^0(x_k) + \frac{(1-p_f)\ell(y_k|x_k)}{(1-p_f)\Psi_0 + p_f\, \kappa(y_k)} p_{k|k-1}^0(x_k). \tag{90}$$

By substituting (69) and (77) into (90), we obtain

$$p_{k|k}^0(x_k) = \sum_{j=1}^{J_{k|k-1}^0} \frac{p_f\, \kappa(y_k)\omega_{k|k-1}^{0,j}}{(1-p_f)\Psi_0 + p_f\, \kappa(y_k)} \mathcal{N}\left(m_{k|k-1}^{0,j}, P_{k|k-1}^{0,j}\right) + \sum_{j=1}^{J_{k|k-1}^0} \frac{(1-p_f)\omega_{k|k-1}^{0,j} \mathcal{N}(y; Cx_k, R)}{(1-p_f)\Psi_0 + p_f\, \kappa(y_k)} \mathcal{N}\left(m_{k|k-1}^{0,j}, P_{k|k-1}^{0,j}\right). \tag{91}$$

Then, by applying lemma 2 in the work of Vo and Ma,[31] we can write

$$\mathcal{N}(y; Cx_k, R)\mathcal{N}\left(m_{k|k-1}^{0,j}, P_{k|k-1}^{0,j}\right) = q_k^{0,j}(y_k)\mathcal{N}\left(m_{k|k}^{0,j}, P_{k|k}^{0,j}\right), \tag{92}$$

where $q_k^{0,j}(y_k)$ has been defined in (84), while $m_{k|k}^{0,j}, P_{k|k}^{0,j}$ have been introduced in (75).

In the special case of linear Gaussian models, $m_{k|k}^{0,j}$ and $P_{k|k}^{0,j}$ can be easily calculated following the standard Bayes filter correction step, which, in this case, boils down to the standard Kalman filter (KF) for linear discrete-time systems[28]:

$$m_{k|k}^{0,j} = m_{k|k-1}^{0,j} + L_k^{0,j}\left(y_k - Cm_{k|k-1}^{0,j}\right) \tag{93}$$

$$P_{k|k}^{0,j} = \left(I - L_k^{0,j}C\right)P_{k|k-1}^{0,j}, \tag{94}$$

where

$$L^{0,j}_k = P^{0,j}_{k\mid k-1} C^T \left(S^{0,j}_k\right)^{-1} \tag{95}$$

$$S^{0,j}_k = C P^{0,j}_{k\mid k-1} C^T + R. \tag{96}$$

Thus, by substituting (92) into (91) with means and covariances given by (93)-(94), we can write

$$p^0_{k\mid k}(x_k) = \sum_{j=1}^{J^0_{k\mid k}} \omega^{0,j}_{k\mid k} \mathcal{N}\left(m^{0,j}_{k\mid k}, P^{0,j}_{k\mid k}\right), \tag{97}$$

which consists of $2J^0_{k\mid k-1}$ Gaussian components, ie,

$$p^0_{k\mid k}(x_k) = \sum_{j=1}^{J^0_{k\mid k-1}} \omega^{0,j}_{F,k\mid k} \mathcal{N}\left(m^{0,j}_{k\mid k-1}, P^{0,j}_{k\mid k-1}\right) + \sum_{j=1}^{J^0_{k\mid k-1}} \omega^{0,j}_{\bar{F},k\mid k} \mathcal{N}\left(m^{0,j}_{k\mid k}, P^{0,j}_{k\mid k}\right), \tag{98}$$

with weights $\omega^{0,j}_{F,k\mid k}, \omega^{0,j}_{\bar{F},k\mid k}$ given by (82)-(83) for $i = 0$. Note that, as it can be seen from (98), it turns out that $J^0_{k\mid k} = 2J^0_{k\mid k-1}$, where the first *legacy* (not corrected) components correspond to the hypothesis of the system-originated measurement being replaced by a fake one y^f_k, while the remaining components are the ones corrected under the hypothesis of receiving y_k with probability $1 - p_f$.

Following the same rationale, analogous results can be obtained for $p^1_{k\mid k}(\cdot,\cdot)$, with the exception that also signal attack estimation has to be performed. By substituting (70) and (78) into (25) in Theorem 1, we obtain

$$\begin{aligned} p^1_{k\mid k}(a_k, x_k) = &\sum_{j=1}^{J^1_{k\mid k-1}} \frac{p_f\, \kappa(y_k)\, \omega^{1,j}_{k\mid k-1}}{(1-p_f)\Psi_1 + p_f\, \kappa(y_k)} \mathcal{N}\left(m^{1,j}_{k\mid k-1}, P^{1,j}_{k\mid k-1}\right) \\ &+ \sum_{j=1}^{J^1_{k\mid k-1}} \frac{(1-p_f)\, \omega^{1,j}_{k\mid k-1} \mathcal{N}(y; Cx_k + Ha_k, R)}{(1-p_f)\Psi_1 + p_f\, \kappa(y_k)} \mathcal{N}\left(m^{1,j}_{k\mid k-1}, P^{1,j}_{k\mid k-1}\right). \end{aligned} \tag{99}$$

Then, by applying lemma 2 in the work of Vo and Ma,[31] we can write

$$\mathcal{N}(y; Cx_k + Ha_k, R)\, \mathcal{N}\left(m^{1,j}_{k\mid k-1}, P^{1,j}_{k\mid k-1}\right) = q^{1,j}_k(y_k)\, \mathcal{N}\left(m^{1,j}_{k\mid k}, P^{1,j}_{k\mid k}\right), \tag{100}$$

where $q^{1,j}_k(y_k)$ has been defined in (85), while $m^{1,j}_{k\mid k}, P^{1,j}_{k\mid k}$ have been introduced in (76). For linear Gaussian models, $m^{1,j}_{k\mid k}$ and $P^{1,j}_{k\mid k}$ can be calculated following the correction step of the filter for JISE of linear discrete-time systems,[28] introduced in Section 2.2. In particular, $m^{1,j}_{k\mid k}$ consists of

$$\hat{x}^{1,j}_{k\mid k} = \hat{x}^{1,j}_{k\mid k-1} + \tilde{L}^{1,j}_k \left(y_k - C\hat{x}^{1,j}_{k\mid k-1} - H\hat{a}^j_k\right) = \hat{x}^{1,j}_{k\mid k-1} + L^{1,j}_k \left(y_k - C\hat{x}^{1,j}_{k\mid k-1}\right) \tag{101}$$

$$\hat{a}^j_k = M^j_k \left(y_k - C\hat{x}^{1,j}_{k\mid k-1}\right), \tag{102}$$

where

$$L^{1,j}_k = \tilde{L}^{1,j}_k \left(I - HM^j_k\right) \tag{103}$$

$$\tilde{L}^{1,j}_k = P^{1x,j}_{k\mid k-1} C^T \left(S^{1,j}_k\right)^{-1} \tag{104}$$

$$S^{1,j}_k = C P^{1x,j}_{k\mid k-1} C^T + R \tag{105}$$

$$M^j_k = \left[H^T \left(S^{1,j}_k\right)^{-1} H\right]^{-1} H^T \left(S^{1,j}_k\right)^{-1}. \tag{106}$$

The elements composing $P^{1,j}_{k\mid k}$ can be computed as

$$P^{1x,j}_{k\mid k} = \left(I - L^{1,j}_k C\right) P^{1x,j}_{k\mid k-1} \tag{107}$$

$$P_k^{a,j} = \left[H^T \left(S_k^{1,j}\right)^{-1} H\right]^{-1} \tag{108}$$

$$P_k^{xa,j} = \left(P_k^{ax,j}\right)^T = -\tilde{L}_k^{1,j} H P_k^{a,j}. \tag{109}$$

Thus, by substituting (100) into (99) with means and covariances given by (101)-(102) and (107)-(109), we can write

$$p_{k\,|\,k}^1(a_k, x_k) = \sum_{j=1}^{J_{k\,|\,k}^1} \omega_{k\,|\,k}^{1,j} \mathcal{N}\left(m_{k\,|\,k}^{1,j}, P_{k\,|\,k}^{1,j}\right), \tag{110}$$

which comprises $2J_{k\,|\,k-1}^1$ components, ie,

$$p_{k\,|\,k}^1(a_k, x_k) = \sum_{j=1}^{J_{k\,|\,k-1}^1} \omega_{F,k\,|\,k}^{1,j} \mathcal{N}\left(m_{k\,|\,k-1}^{1,j}, P_{k\,|\,k-1}^{1,j}\right) + \sum_{j=1}^{J_{k\,|\,k-1}^1} \omega_{\bar{F},k\,|\,k}^{1,j} \mathcal{N}\left(m_{k\,|\,k}^{1,j}, P_{k\,|\,k}^{1,j}\right), \tag{111}$$

with weights $\omega_{F,k\,|\,k}^{1,j}, \omega_{\bar{F},k\,|\,k}^{1,j}$ given by (82)-(83) for $i = 1$. □

5.2 | GM-HBF correction for extra packet injection

Proposition 2. *Suppose that assumptions (69)-(73) hold, the measurement set $\mathcal{Z}_k$ is defined by (4), the predicted FISST density at time k is fully specified by the triplet $(r_{k\,|\,k-1}, p_{k\,|\,k-1}^0(x_k), p_{k\,|\,k-1}^1(a_k, x_k))$, and $p_{k\,|\,k-1}^0(\cdot)$, $p_{k\,|\,k-1}^1(\cdot,\cdot)$ are GMs of the form (77) and (78), respectively. Then, the posterior FISST density $(r_{k\,|\,k}, p_{k\,|\,k}^0(x_k), p_{k\,|\,k}^1(a_k, x_k))$ is given by*

$$r_{k\,|\,k} = \frac{1 - p_d + p_d \Gamma_1}{1 - p_d + p_d(1 - r_{k\,|\,k-1})\Gamma_0 + p_d r_{k\,|\,k-1}\Gamma_1} r_{k\,|\,k-1} \tag{112}$$

$$p_{k\,|\,k}^0(x_k) = \sum_{j=1}^{J_{k\,|\,k}^0} \omega_{k\,|\,k}^{0,j} \mathcal{N}\left(m_{k\,|\,k}^{0,j}, P_{k\,|\,k}^{0,j}\right) = \sum_{j=1}^{J_{k\,|\,k-1}^0} \omega_{\bar{D},k\,|\,k}^{0,j} \mathcal{N}\left(m_{k\,|\,k-1}^{0,j}, P_{k\,|\,k-1}^{0,j}\right) + \sum_{y_k \in \mathcal{Z}_k} \sum_{j=1}^{J_{k\,|\,k-1}^0} \omega_{D,k\,|\,k}^{0,j} \mathcal{N}\left(m_{k\,|\,k}^{0,j}, P_{k\,|\,k}^{0,j}\right) \tag{113}$$

$$p_{k\,|\,k}^1(a_k, x_k) = \sum_{j=1}^{J_{k\,|\,k}^1} \omega_{k\,|\,k}^{1,j} \mathcal{N}\left(m_{k\,|\,k}^{1,j}, P_{k\,|\,k}^{1,j}\right) = \sum_{j=1}^{J_{k\,|\,k-1}^1} \omega_{\bar{D},k\,|\,k}^{1,j} \mathcal{N}\left(m_{k\,|\,k-1}^{1,j}, P_{k\,|\,k-1}^{1,j}\right) + \sum_{y_k \in \mathcal{Z}_k} \sum_{j=1}^{J_{k\,|\,k-1}^1} \omega_{D,k\,|\,k}^{1,j} \mathcal{N}\left(m_{k\,|\,k}^{1,j}, P_{k\,|\,k}^{1,j}\right), \tag{114}$$

where, for i $= 0, 1,$

$$\omega_{\bar{D},k\,|\,k}^{i,j} = \frac{(1 - p_d)\omega_{k\,|\,k-1}^{i,j}}{1 - p_d + p_d \Gamma_i}, \tag{115}$$

$$\omega_{D,k\,|\,k}^{i,j} = \frac{p_d \omega_{k\,|\,k-1}^{i,j} q_k^{i,j}(y_k)}{(1 - p_d + p_d \Gamma_i) n \kappa(y_k)} \tag{116}$$

and

$$\Gamma_0 = \sum_{y_k \in \mathcal{Z}_k} \sum_{j=1}^{J_{k\,|\,k-1}^0} \frac{\omega_{k\,|\,k-1}^{0,j}}{n \kappa(y_k)} q_k^{0,j}(y_k) \tag{117}$$

$$\Gamma_1 = \sum_{y_k \in \mathcal{Z}_k} \sum_{j=1}^{J_{k\,|\,k-1}^1} \frac{\omega_{k\,|\,k-1}^{1,j}}{n \kappa(y_k)} q_k^{1,j}(y_k). \tag{118}$$

Proof. We first derive the corrected probability of signal attack existence, which can be directly written from (41) as

$$r_{k\,|\,k} = \frac{1 - p_d + p_d \Gamma_1}{1 - p_d + p_d(1 - r_{k\,|\,k-1})\Gamma_0 + p_d r_{k\,|\,k-1}\Gamma_1} r_{k\,|\,k-1}, \tag{119}$$

where Γ_0 is obtained by substituting (69) and (77) into (44), so that

$$\Gamma_0 = \sum_{y_k \in \mathcal{Z}_k} \frac{\int \mathcal{N}(y; Cx_k, R) \sum_{j=1}^{J^0_{k|k-1}} \omega^{0,j}_{k|k-1} \mathcal{N}\left(m^{0,j}_{k|k-1}, P^{0,j}_{k|k-1}\right) \mathrm{d}x_k}{n\kappa(y_k)}. \tag{120}$$

Then, by applying (87), (120) takes the form (117). Moreover, Γ_1 in (119) can be analogously obtained by substituting (70) and (78) into (45) and by applying (100) that leads to (118).

Next, the posterior density $p^0_{k|k}(\cdot)$ can be derived from (42) in Theorem 2 as

$$p^0_{k|k}(x_k) = \frac{1-p_d}{1-p_d+p_d\Gamma_0} p^0_{k|k-1}(x_k) + \frac{p_d}{1-p_d+p_d\Gamma_0} \sum_{y_k \in \mathcal{Z}_k} \frac{\ell(y_k|x_k)}{n\kappa(y_k)} p^0_{k|k-1}(x_k). \tag{121}$$

By substituting (69) and (77) into (121), we obtain

$$\begin{aligned} p^0_{k|k}(x_k) = &\sum_{j=1}^{J^0_{k|k-1}} \frac{1-p_d}{1-p_d+p_d\Gamma_0} \omega^{0,j}_{k|k-1} \mathcal{N}\left(m^{0,j}_{k|k-1}, P^{0,j}_{k|k-1}\right) \\ &+ \sum_{y_k \in \mathcal{Z}_k} \sum_{j=1}^{J^0_{k|k-1}} \omega^{0,j}_{k|k-1} \frac{p_d}{1-p_d+p_d\Gamma_0} \frac{\mathcal{N}(y; Cx_k, R)}{n\kappa(y_k)} \mathcal{N}\left(m^{0,j}_{k|k-1}, P^{0,j}_{k|k-1}\right). \end{aligned} \tag{122}$$

Thus, by substituting (87) into (122), with means and covariances given by (93)-(94), we can write

$$p^0_{k|k}(x_k) = \sum_{j=1}^{J^0_{k|k}} \omega^{0,j}_{k|k} \mathcal{N}\left(m^{0,j}_{k|k}, P^{0,j}_{k|k}\right), \tag{123}$$

which comprises $J^0_{k|k-1}(1+|\mathcal{Z}_k|)$ components, where $|\mathcal{Z}_k|$ denotes the cardinality of the measurement set $\mathcal{Z}$ at time k, ie,

$$p^0_{k|k}(x_k) = \sum_{j=1}^{J^0_{k|k-1}} \omega^{0,j}_{\bar{D},k|k} \mathcal{N}\left(m^{0,j}_{k|k-1}, P^{0,j}_{k|k-1}\right) + \sum_{y_k \in \mathcal{Z}_k} \sum_{j=1}^{J^0_{k|k-1}} \omega^{0,j}_{D,k|k} \mathcal{N}\left(m^{0,j}_{k|k}, P^{0,j}_{k|k}\right), \tag{124}$$

with weights

$$\begin{aligned} \omega^{0,j}_{\bar{D},k|k} &= \frac{(1-p_d)\omega^{0,j}_{k|k-1}}{1-p_d+p_d \sum_{y_k \in \mathcal{Z}_k} \sum_{h=1}^{J^0_{k|k-1}} \frac{\omega^{0,h}_{k|k-1}}{n\kappa(y_k)} q^{0,h}_k(y_k)} \\ \omega^{0,j}_{D,k|k} &= \frac{p_d \omega^{0,j}_{k|k-1} q^{0,j}_k(y_k)}{\left[1-p_d+p_d \sum_{y_k \in \mathcal{Z}_k} \sum_{h=1}^{J^0_{k|k-1}} \frac{\omega^{0,h}_{k|k-1}}{n\kappa(y_k)} q^{0,h}_k(y_k)\right] n\kappa(y_k)}. \end{aligned}$$

Note that, as it can be seen from (124), it turns out that $J^0_{k|k} = J^0_{k|k-1} + |\mathcal{Z}_k| J^0_{k|k-1} = J^0_{k|k-1}(1+|\mathcal{Z}_k|)$, where the first *legacy* components correspond to the fact that no measurement has been delivered, and hence, no update is carried out, while the remaining components are the ones corrected when one or multiple measurements are received.

Following the same rationale, analogous results can be obtained for $p^1_{k\,|\,k}(\cdot,\cdot)$. From (43) in Theorem 2,

$$p^1_{k\,|\,k}(a_k,x_k) = \frac{1-p_d}{1-p_d+p_d\Gamma_1}\,p^1_{k\,|\,k-1}(a_k,x_k) + \frac{p_d}{1-p_d+p_d\Gamma_1}\sum_{y_k\in\mathcal{Z}_k}\frac{\ell(y_k\,|\,a_k,x_k)}{n\,\kappa(y_k)}p^1_{k\,|\,k-1}(a_k,x_k). \tag{125}$$

By substituting (70) and (78) into (125), we obtain

$$\begin{aligned} p^1_{k\,|\,k}(a_k,x_k) = {} & \sum_{j=1}^{J^1_{k\,|\,k-1}} \frac{1-p_d}{1-p_d+p_d\Gamma_1}\,\omega^{1,j}_{k\,|\,k-1}\mathcal{N}\left(m^{1,j}_{k\,|\,k-1},P^{1,j}_{k\,|\,k-1}\right) \\ & + \sum_{y_k\in\mathcal{Z}_k}\sum_{j=1}^{J^1_{k\,|\,k-1}}\omega^{1,j}_{k\,|\,k-1}\frac{p_d}{1-p_d+p_d\Gamma_1}\,\frac{\mathcal{N}(y;Cx_k+Ha_k,R)}{n\,\kappa(y_k)}\,\mathcal{N}\left(m^{1,j}_{k\,|\,k-1},P^{1,j}_{k\,|\,k-1}\right). \end{aligned} \tag{126}$$

Thus, by substituting (100) into (126), with means and covariances given by (101)-(102) and (107)-(109), we can write

$$p^1_{k\,|\,k}(a_k,x_k) = \sum_{j=1}^{J^1_{k\,|\,k}}\omega^{1,j}_{k\,|\,k}\mathcal{N}\left(m^{1,j}_{k\,|\,k},P^{1,j}_{k\,|\,k}\right), \tag{127}$$

which comprises $J^1_{k\,|\,k-1}(1+|\mathcal{Z}_k|)$ components, ie,

$$p^1_{k\,|\,k}(a_k,x_k) = \sum_{j=1}^{J^1_{k\,|\,k-1}}\omega^{1,j}_{\bar{D},k\,|\,k}\mathcal{N}\left(m^{1,j}_{k\,|\,k-1},P^{1,j}_{k\,|\,k-1}\right) + \sum_{y_k\in\mathcal{Z}_k}\sum_{j=1}^{J^1_{k\,|\,k-1}}\omega^{1,j}_{D,k\,|\,k}\mathcal{N}\left(m^{1,j}_{k\,|\,k},P^{1,j}_{k\,|\,k}\right), \tag{128}$$

with weights

$$\begin{aligned} \omega^{1,j}_{\bar{D},k\,|\,k} &= \frac{(1-p_d)\,\omega^{1,j}_{k\,|\,k-1}}{1-p_d+p_d\displaystyle\sum_{y_k\in\mathcal{Z}_k}\sum_{h=1}^{J^1_{k\,|\,k-1}}\frac{\omega^{1,h}_{k\,|\,k-1}}{n\,\kappa(y_k)}q^{1,h}_k(y_k)} \\ \omega^{1,j}_{D,k\,|\,k} &= \frac{p_d\,\omega^{1,j}_{k\,|\,k-1}q^{1,j}_k(y_k)}{\left[1-p_d+p_d\displaystyle\sum_{y_k\in\mathcal{Z}_k}\sum_{h=1}^{J^1_{k\,|\,k-1}}\frac{\omega^{1,h}_{k\,|\,k-1}}{n\,\kappa(y_k)}q^{1,h}_k(y_k)\right]n\,\kappa(y_k)}. \end{aligned}$$

□

5.3 | GM-HBF prediction

Proposition 3. *Suppose that assumptions (69)-(73) hold, the posterior FISST density at time k is fully specified by the triplet* $(r_{k\,|\,k},p^0_{k\,|\,k}(x_k),p^1_{k\,|\,k}(a_k,x_k))$, *and* $p^0_{k\,|\,k}(\cdot)$, $p^1_{k\,|\,k}(\cdot,\cdot)$ *are GMs of the form (75)-(76). Then, the predicted FISST density* $(r_{k+1\,|\,k},p^0_{k+1\,|\,k}(x_{k+1}),p^1_{k+1\,|\,k}(a_{k+1},x_{k+1}))$ *is given by*

$$r_{k+1\,|\,k} = (1-r_{k\,|\,k})p_b + r_{k\,|\,k}p_s \tag{129}$$

$$p^0_{k+1\,|\,k}(x_{k+1}) = \sum_{j=1}^{J^0_{k+1\,|\,k}}\omega^{0,j}_{k+1\,|\,k}\mathcal{N}\left(m^{0,j}_{k+1\,|\,k},P^{0,j}_{k+1\,|\,k}\right) \tag{130}$$

$$p^1_{k+1\,|\,k}(a_{k+1},x_{k+1}) = \sum_{j=1}^{J^1_{k+1\,|\,k}}\omega^{1,j}_{k+1\,|\,k}\mathcal{N}\left(m^{1,j}_{k+1\,|\,k},P^{1,j}_{k+1\,|\,k}\right), \tag{131}$$

where (130) comprises $J^0_{k+1\,|\,k} = J^0_{k\,|\,k} + J^1_{k\,|\,k}$ *components, ie,*

$$p^0_{k+1\,|\,k}(x_{k+1}) = \underbrace{\sum_{j=1}^{J^0_{k\,|\,k}} \omega^{0,j}_{\bar{B},k+1\,|\,k} \mathcal{N}\left(m^{0,j}_{\bar{B},k+1\,|\,k}, P^{0,j}_{\bar{B},k+1\,|\,k}\right)}_{\text{no attack-birth}} + \underbrace{\sum_{j=1}^{J^1_{k\,|\,k}} \omega^{0,j}_{\bar{S},k+1\,|\,k} \mathcal{N}\left(m^{0,j}_{\bar{S},k+1\,|\,k}, P^{0,j}_{\bar{S},k+1\,|\,k}\right)}_{\text{no attack-survival}}, \tag{132}$$

with

$$m^{0,j}_{\bar{B},k+1\,|\,k} = A\,m^{0,j}_{k\,|\,k} \tag{133}$$

$$P^{0,j}_{\bar{B},k+1\,|\,k} = AP^{0,j}_{k\,|\,k}A^T + Q \tag{134}$$

$$\omega^{0,j}_{\bar{B},k+1\,|\,k} = \frac{(1 - r_{k\,|\,k})(1 - p_b)}{1 - r_{k+1\,|\,k}}\,\omega^{0,j}_{k\,|\,k} \tag{135}$$

and

$$m^{0,j}_{\bar{S},k+1\,|\,k} = \tilde{A}\,m^{1,j}_{k\,|\,k} \tag{136}$$

$$P^{0,j}_{\bar{S},k+1\,|\,k} = \tilde{A}P^{1,j}_{k\,|\,k}\tilde{A}^T + Q \tag{137}$$

$$\omega^{0,j}_{\bar{S},k+1\,|\,k} = \frac{r_{k\,|\,k}(1 - p_s)}{1 - r_{k+1\,|\,k}}\,\omega^{1,j}_{k\,|\,k}, \tag{138}$$

where $\tilde{A} \overset{\triangle}{=} [A, G]$. *Moreover, (131) comprises* $J^1_{k+1\,|\,k} = J^a(J^0_{k\,|\,k} + J^1_{k\,|\,k})$ *components, ie,*

$$p^1_{k+1\,|\,k}(a_{k+1}, x_{k+1}) = \underbrace{\sum_{j=1}^{J^0_{k\,|\,k}} \sum_{h=1}^{J^a} \omega^{1,jh}_{B,k+1\,|\,k} \mathcal{N}\left(m^{1,jh}_{B,k+1\,|\,k}, P^{1,jh}_{B,k+1\,|\,k}\right)}_{\text{attack-birth}} + \underbrace{\sum_{j=1}^{J^1_{k\,|\,k}} \sum_{h=1}^{J^a} \omega^{1,jh}_{S,k+1\,|\,k} \mathcal{N}\left(m^{1,jh}_{S,k+1\,|\,k}, P^{1,jh}_{S,k+1\,|\,k}\right)}_{\text{attack-survival}}, \tag{139}$$

where

$$m^{1,jh}_{B,k+1\,|\,k} = \begin{bmatrix} A\,m^{0,j}_{k\,|\,k} \\ \tilde{a}^h \end{bmatrix} \tag{140}$$

$$P^{1,jh}_{B,k+1\,|\,k} = \begin{bmatrix} AP^{0,j}_{k\,|\,k}A^T + Q & 0 \\ 0 & \tilde{P}^{a,h} \end{bmatrix} \tag{141}$$

$$\omega^{1,jh}_{B,k+1\,|\,k} = \frac{(1 - r_{k\,|\,k})p_b}{r_{k+1\,|\,k}}\,\omega^{0,j}_{k\,|\,k}\tilde{\omega}^{a,h} \tag{142}$$

and

$$m^{1,jh}_{S,k+1\,|\,k} = \begin{bmatrix} \tilde{A}\,m^{1,j}_{k\,|\,k} \\ \tilde{a}^h \end{bmatrix} \tag{143}$$

$$P^{1,jh}_{S,k+1\,|\,k} = \begin{bmatrix} \tilde{A}P^{1,j}_{k\,|\,k}\tilde{A}^T + Q & 0 \\ 0 & \tilde{P}^{a,h} \end{bmatrix} \tag{144}$$

$$\omega^{1,jh}_{S,k+1\,|\,k} = \frac{r_{k\,|\,k}p_s}{r_{k+1\,|\,k}}\,\omega^{1,j}_{k\,|\,k}\tilde{\omega}^{a,h}. \tag{145}$$

Proof. The predicted signal attack probability comes directly from (49). Let us now derive the predicted density $p^0_{k+1\,|\,k}(\cdot)$. From (50) in Theorem 3,

$$p^0_{k+1\,|\,k}(x_{k+1}) = \frac{(1 - r_{k\,|\,k})(1 - p_b)}{1 - r_{k+1\,|\,k}} \int \pi(x_{k+1}\,|\,x_k), p^0_{k\,|\,k}(x_k)\,\mathrm{d}x_k + \frac{r_{k\,|\,k}(1 - p_s)}{1 - r_{k+1\,|\,k}} \iint \pi(x_{k+1}\,|\,a_k, x_k), p^1_{k\,|\,k}(a_k, x_k)\,\mathrm{d}a_k\mathrm{d}x_k. \tag{146}$$

Using (71), (75) in the first term and (72), (76) in the second term, we can rewrite

$$
\begin{aligned}
p^0_{k+1\,|\,k}(x_{k+1}) = {} & \frac{(1-r_{k\,|\,k})(1-p_b)}{1-r_{k+1\,|\,k}} \int \mathcal{N}(x;Ax_k,Q)\sum_{j=1}^{J^0_{k\,|\,k}}\omega^{0,j}_{k\,|\,k}\mathcal{N}\left(m^{0,j}_{k\,|\,k},P^{0,j}_{k\,|\,k}\right)\mathrm{d}x_k \\
& + \frac{r_{k\,|\,k}(1-p_s)}{1-r_{k+1\,|\,k}} \iint \mathcal{N}(x;Ax_k+Ga_k,Q)\sum_{j=1}^{J^1_{k\,|\,k}}\omega^{1,j}_{k\,|\,k}\mathcal{N}\left(m^{1,j}_{k\,|\,k},P^{1,j}_{k\,|\,k}\right)\mathrm{d}a_k\mathrm{d}x_k.
\end{aligned}
\tag{147}
$$

Hence, using lemma 1 in the work of Vo and Ma[31] in both the above terms, we finally derive (132)

$$
\begin{aligned}
p^0_{k+1\,|\,k}(x_{k+1}) = {} & \sum_{j=1}^{J^0_{k\,|\,k}}\frac{(1-r_{k\,|\,k})(1-p_b)}{1-r_{k+1\,|\,k}}\,\omega^{0,j}_{k\,|\,k}\mathcal{N}\left(x;Am^{0,j}_{k\,|\,k},AP^{0,j}_{k\,|\,k}A^T+Q\right) \\
& + \sum_{j=1}^{J^1_{k\,|\,k}}\frac{r_{k\,|\,k}(1-p_s)}{1-r_{k+1\,|\,k}}\,\omega^{1,j}_{k\,|\,k}\mathcal{N}\left(x;\tilde{A}m^{1,j}_{k\,|\,k},\tilde{A}P^{1,j}_{k\,|\,k}\tilde{A}^T+Q\right).
\end{aligned}
$$

In a similar fashion, we can obtain $p^1_{k+1\,|\,k}(\cdot,\cdot)$. From (51) in Theorem 3,

$$
p^1_{k+1\,|\,k}(a_{k+1},x_{k+1}) = \frac{(1-r_{k\,|\,k})p_b}{r_{k+1\,|\,k}}\int \pi(x_{k+1}\,|\,x_k),p^0_{k\,|\,k}(x_k)\,\mathrm{d}x_k\,p(a) + \frac{r_{k\,|\,k}p_s}{r_{k+1\,|\,k}}\iint \pi(x_{k+1}\,|\,a_k,x_k)p^1_{k\,|\,k}(a_k,x_k)\,\mathrm{d}a_k\mathrm{d}x_k\,p(a),
$$

which, using (71), (72), (73), (75), and (76), leads to

$$
\begin{aligned}
p^1_{k+1\,|\,k}(a_{k+1},x_{k+1}) = {} & \frac{(1-r_{k\,|\,k})p_b}{r_{k+1\,|\,k}}\int \mathcal{N}(x;Ax_k,Q)\sum_{j=1}^{J^0_{k\,|\,k}}\omega^{0,j}_{k\,|\,k}\mathcal{N}\left(m^{0,j}_{k\,|\,k},P^{0,j}_{k\,|\,k}\right)\mathrm{d}x_k\sum_{h=1}^{J^a}\tilde{\varpi}^{a,h}\mathcal{N}(a;\tilde{a}^h,\tilde{P}^{a,h}) \\
& + \frac{r_{k\,|\,k}p_s}{r_{k+1\,|\,k}}\iint \mathcal{N}(x;Ax_k+Ga_k,Q)\sum_{j=1}^{J^1_{k\,|\,k}}\omega^{1,j}_{k\,|\,k}\mathcal{N}\left(m^{1,j}_{k\,|\,k},P^{1,j}_{k\,|\,k}\right)\mathrm{d}a_k\mathrm{d}x_k\sum_{h=1}^{J^a}\tilde{\varpi}^{a,h}\mathcal{N}(a;\tilde{a}^h,\tilde{P}^{a,h}).
\end{aligned}
\tag{148}
$$

Finally, by applying the same result on integrals of Gaussians used above, we have

$$
\begin{aligned}
p^1_{k+1\,|\,k}(a_{k+1},x_{k+1}) = {} & \sum_{j=1}^{J^0_{k\,|\,k}}\sum_{h=1}^{J^a}\frac{(1-r_{k\,|\,k})p_b}{r_{k+1\,|\,k}}\,\omega^{0,j}_{k\,|\,k}\tilde{\varpi}^{a,h}\mathcal{N}\left(x;Am^{0,j}_{k\,|\,k},AP^{0,j}_{k\,|\,k}A^T+Q\right)\mathcal{N}(a;\tilde{a}^h,\tilde{P}^{a,h}) \\
& + \sum_{j=1}^{J^1_{k\,|\,k}}\sum_{h=1}^{J^a}\frac{r_{k\,|\,k}p_s}{r_{k+1\,|\,k}}\,\omega^{1,j}_{k\,|\,k}\tilde{\varpi}^{a,h}\mathcal{N}\left(x;\tilde{A}m^{1,j}_{k\,|\,k},\tilde{A}P^{1,j}_{k\,|\,k}\tilde{A}^T+Q\right)\mathcal{N}(a;\tilde{a}^h,\tilde{P}^{a,h}),
\end{aligned}
\tag{149}
$$

from which (139) is obtained. □

It is worth pointing out that, likewise other GM filters, also the proposed *GM-HBF* is characterized by a number of Gaussian components that increases with no bound over time. As already noticed in the above derivation, at time k, the GM-HBF requires

$$
J^0_{k\,|\,k} = \begin{cases} 2J^0_{k\,|\,k-1}, & \text{packet substitution} \\ J^0_{k\,|\,k-1}(1+|\mathcal{Z}_k|), & \text{extra packet injection} \end{cases}, \quad
J^1_{k\,|\,k} = \begin{cases} 2J^1_{k\,|\,k-1}, & \text{packet substitution} \\ J^1_{k\,|\,k-1}(1+|\mathcal{Z}_k|), & \text{extra packet injection} \end{cases}
$$

components to exactly represent the posterior densities $p^0_{k|k}(\cdot)$ and $p^1_{k|k}(\cdot,\cdot)$, respectively. Here,

$$J^0_{k|k-1} = J^0_{k-1|k-1} + J^1_{k-1|k-1},$$
$$J^1_{k|k-1} = J^a \left(J^0_{k-1|k-1} + J^1_{k-1|k-1} \right)$$

denote the number of components generated in the prediction step. Heuristic *pruning* and *merging* procedures[31] can be performed at each time step so as to remove low-weight components and combine statistically close components and, hence, ensure that the total number of GM components is always less than a pre-specified maximum value, say, $J_{\max}$. In this way, each step of the HBF has a bounded complexity in the order of $J_{\max}$ KFs (or EKFs/UKFs in the nonlinear case).

Remark 5. The GM implementation of this section has actually revealed a connection between the proposed HBF and the KF in that the former uses multiple KFs (or EKFs/UKFs) to propagate in time means and covariances of the various components of the GM (see Equations (93)-(96), (101)-(109), (133)-(134) and (136)-(137)). It is also worth to point out that, due to the switching nature of the attack input and the presence of fake measurements, the HBF turns out to be nonlinear even if the CPS (1)-(2) is linear and Gaussian. However, the handling of nonlinearities in the state-space model is actually a minor issue in this Bayesian setting, only requiring to replace the KFs (adopted for propagating means and covariances of the Gaussian components of the conditional PDFs) with either EKF or UKF that can cope with nonlinear functions $f^0_k(\cdot), f^1_k(\cdot,\cdot), h^0_k(\cdot), f^1_k(\cdot,\cdot)$ in (1)-(2).

6 | NUMERICAL EXAMPLES

The effectiveness of the developed tools, based on Bayesian random set theory, for joint attack detection and secure state estimation of CPSs has been tested on two numerical examples concerning a benchmark linear dynamical system and a standard IEEE power network case study. Simulations have been carried out in the presence of both signal and extra packet injection attacks as well as uncertainty on measurement delivery. Results on the performance of the GM-HBF under packet substitution attack are shown in Section 5.2.

6.1 | Benchmark linear system

Let us first consider the following benchmark linear system, already used in the JISE literature[32]:

$$\begin{aligned} x_{k+1} &= Ax_k + Ga_k + w_k \\ y_k &= Cx_k + Ha_k + v_k, \end{aligned} \tag{150}$$

where A, C, R, and Q are the same as in the work of Yong et al,[25] while $G = [e_1, e_2]$ and $H = [e_3, e_1]$, where $e_1, \ldots, e_5$ denote the canonical basis vectors. For this numerical study, the probabilities of attack-birth and attack-survival are fixed, respectively, at $p_b = 0.2$ and $p_s = 0.8$. The system-generated measurement is supposed to be delivered at the monitor/control center with probability $p_d = 0.98$, while the initial signal attack probability is set to $r_{1|0} = 0.1$. The initial state has been set equal to $x_0 = 0$, whereas both densities $p^0(\cdot)$ and $p^1(\cdot,\cdot)$ have been initialized as single Gaussian components with first guess mean $\hat{x}^0_{1|0} = [10, 10, 0, 0, 0]^T$ and covariance $P^0_{1|0} = 10^4 I_5$. Moreover, the first estimate of the attack vector has been randomly initialized as $\hat{a}_{1|0} = [15.1, 25.53]^T$, with associated initial covariance matrix $P^a_{1|0} = 50 I_2$. The extra fake measurements are modeled as uniformly distributed over the interval $[-0.3, 140.3]$. Finally, a pruning threshold $\gamma_p = 10^{-3}$ and a merging threshold $\gamma_m = 3$ have been chosen. As shown in Figure 4, at time $k = 150$, a signal attack vector $a = [10, 20]^T$ is injected into the system, persisting for 200 time steps. The proposed GM-HBF promptly detects the unknown signal attack, by simply comparing the attack probability $r_{k|k}$ obtained in (41) with the threshold 0.5. Figure 5 provides a comparison between the true and the estimated values of states x_1 and x_2 (clearly the only state components affected by the signal attack). Note that the state estimate is obtained by means of a MAP estimator, ie, by extracting the Gaussian mean with the highest weight from the posterior density $p^0(\cdot)$ (42) or $p^1(\cdot,\cdot)$ (43), according to the current value of the attack probability. Finally, Figure 6 shows how the attack estimates extracted from $p(a)$ of the two components of the attack vector, coincide with the actual values inside the attack time interval $[150, 350]$. Note that, outside that interval, the estimates of the attack vector are not meaningful because the attack probability $r_{k|k}$ is almost 0.

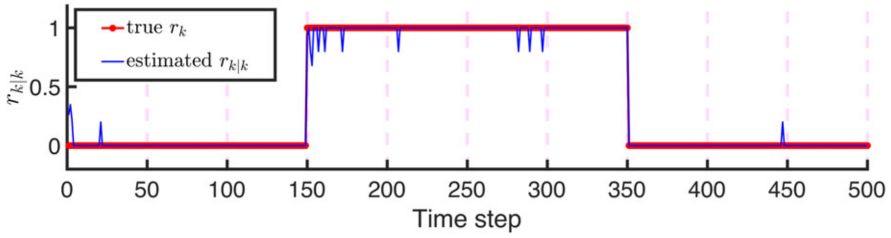

FIGURE 4 True and estimated attack probability [Colour figure can be viewed at wileyonlinelibrary.com]

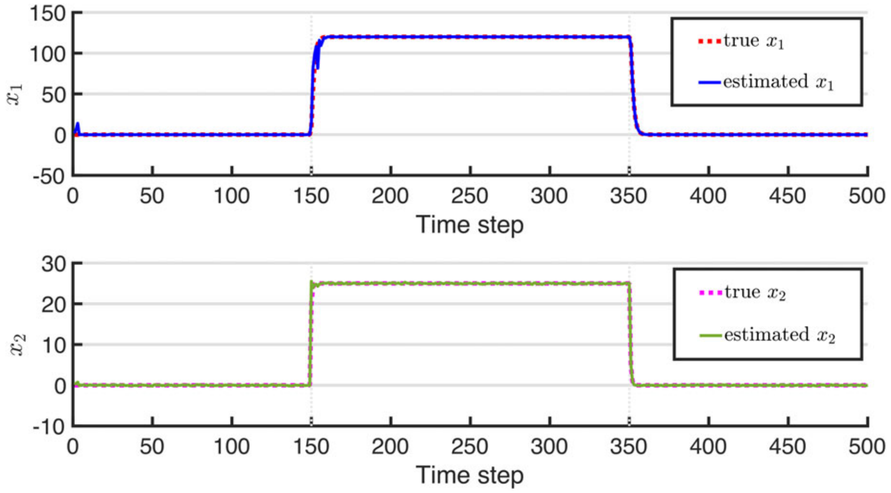

FIGURE 5 True and estimated state components x_1 and x_2 [Colour figure can be viewed at wileyonlinelibrary.com]

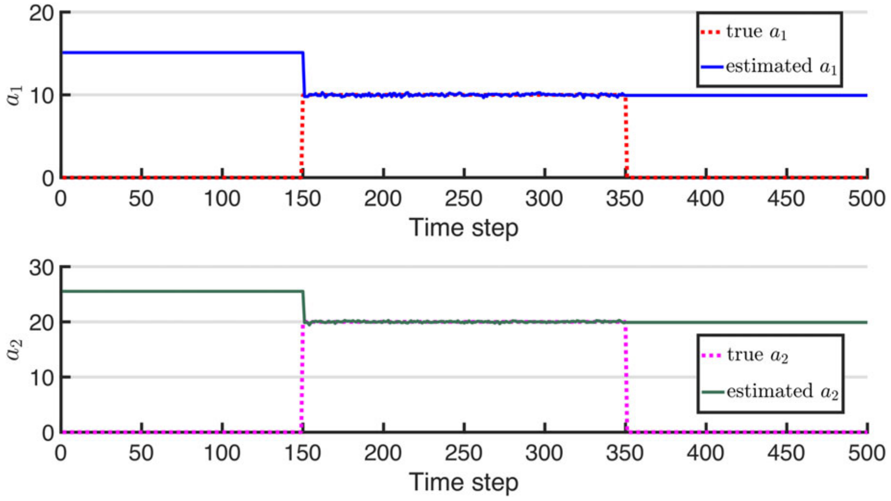

FIGURE 6 True and estimated attack components a_1 and a_2 [Colour figure can be viewed at wileyonlinelibrary.com]

6.2 | IEEE 14-bus power network

State estimation is of paramount importance to ensure the reliable operation of energy delivery systems since it provides estimates of the power grid state by processing meter measurements and exploiting power system models. Cyberattacks on power systems can alter available information at the control center and generate fake meter and input data, potentially causing power outage and forcing the energy management system to make erroneous decisions, eg, on contingency analysis and economic dispatch. The proposed GM-HBF was tested on the IEEE 14-bus system (Figure 7) consisting of 5 synchronous generators and 11 load buses, with parameters taken from MATPOWER.[33] The dynamics of the system can be described by the linearized swing equation[34] derived through Kron reduction[35] of the linear small-signal power network model. The DC state estimation model assumes 1 p.u. (per unit) voltage magnitudes in all buses and $j1$ p.u. branch impedance, with j denoting imaginary unit. The system dynamics is represented by the evolution of $n = 10$ states comprising both the rotor angles δ_j and the frequencies ω_j of each generator j in the network. After discretization (with sampling interval $T = 0.01$ s), the model of the system takes the form (1)-(2), where the whole state is measured by a

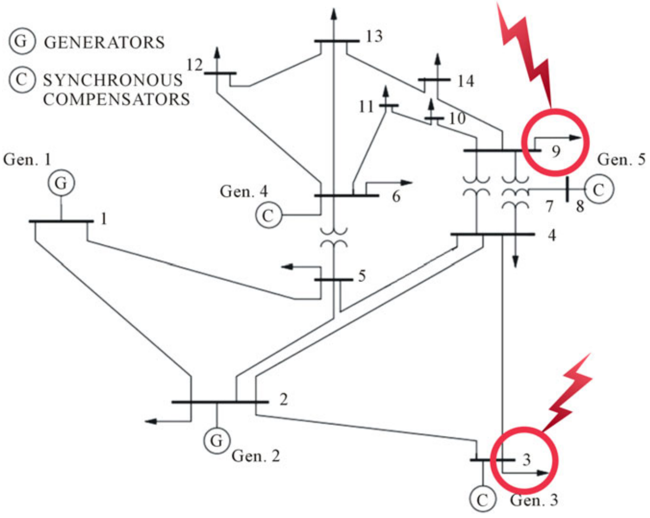

FIGURE 7 Single-line model of the IEEE 14-bus system. The *true* victim load buses 3 and 9 are circled in red [Colour figure can be viewed at wileyonlinelibrary.com]

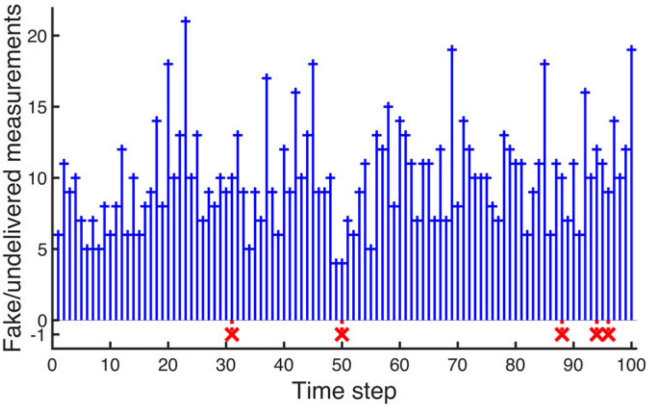

FIGURE 8 Number of extra fake measurements injected (blue circles) and undelivered ($p_d = 0.95$) system-originated observations (red cross in −1) vs time. The proposed Gaussian mixture hybrid Bernoulli filter turns out to be particularly robust to *extra packet injection* attacks [Colour figure can be viewed at wileyonlinelibrary.com]

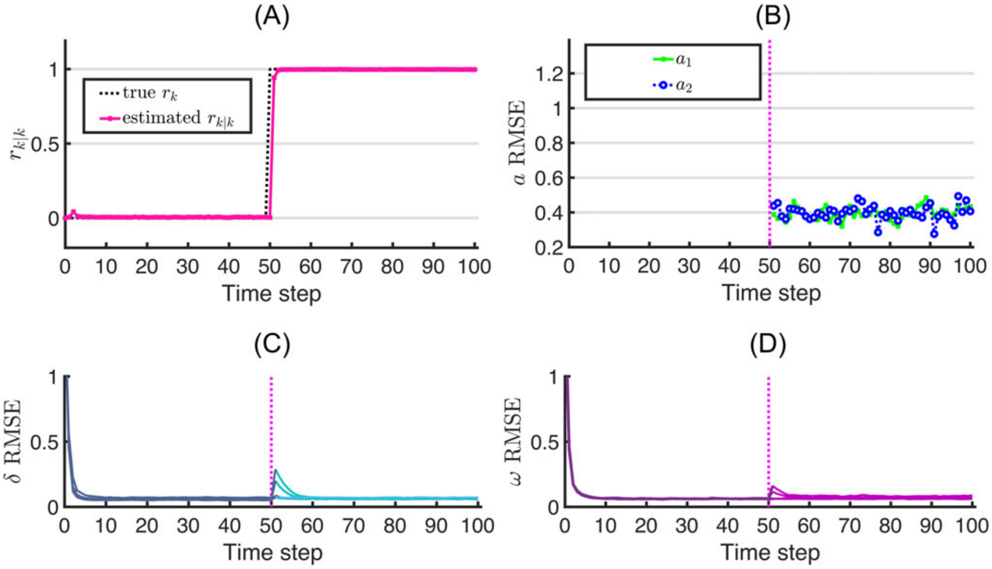

FIGURE 9 Performance of the Gaussian mixture hybrid Bernoulli filter in terms of (A) joint attack detection and (B) estimation of attack signal, (C) rotor angles δ_i, $i = 1, \ldots, 5$, and (D) frequencies ω_i, $i = 1, \ldots, 5$. RMSE, root mean square error [Colour figure can be viewed at wileyonlinelibrary.com]

network of sensors $\mathcal{S}_i$. The system is assumed to be corrupted by additive zero mean Gaussian white process and measurement noises with variances $\sigma_w^2 = 0.01$ and $\sigma_v^2 = 0.01$. At time $k = 50$, a signal attack vector $a = [0.2, 0.1]^T$ p.u. is injected into the system to abruptly increase the real power demand of the two victim load buses 3 and 9 with an additional loading of 21.23% and, respectively, 33.9%. This type of attack, referred to as *load altering attack*,[36] can provoke a loss of synchrony of the rotor angles and hence a deviation of the rotor speeds of all generators from their nominal value. In addition, we fixed the following parameters: $p_b = 0.05$, $p_s = 0.95$, $p_d = 0.95$, pruning and merging thresholds $\gamma_p = 10^{-2}$ and

$\gamma_m = 3$ for the GM implementation. Let us first consider the system under extra packet injection attack. The additional fake measurements injected into the sensor channels are modeled as uniformly distributed over the interval $[-10, 5]$, suitably chosen to emulate system-originated observations. Fake and missed packets are shown in Figure 8 for a specific run. The joint attack detection and state estimation performance of the GM-HBF algorithm has been analyzed by Monte Carlo simulations. Figure 9 shows the *true* and estimated probability of attack existence (Figure 9A) and the root mean square error (RMSE), averaged over 1000 Monte Carlo runs, relative to the rotor angle (Figure 9B) and frequency (Figure 9C) estimates. Figure 9D shows the RMSE of the estimated components of the signal attack, extracted from $p^1_{k|k}(a, x)$. As shown in the results (Figures 9A-D), the proposed secure state estimator succeeds in promptly detecting a signal attack altering the nominal energy delivery system behavior and hence in being simultaneously resilient to integrity attacks on power demand and robust to extra fake packets and undelivered measurements. Figure 10 provides, for a single Monte Carlo trial, a comparison between the *true* and the estimated values of the two rotor angles mainly affected by the victim

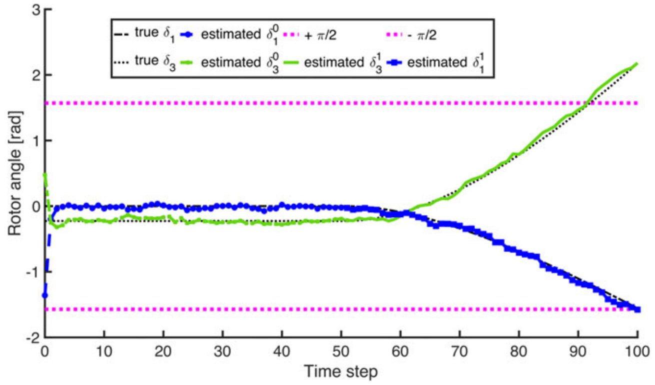

FIGURE 10 Estimated vs *true* trajectory of rotor angles δ_j, $j = 1, 3$. Note that, if $|\delta_j|$ is sufficiently large (values close to $\pi/2$), the linear small signal approximation significantly deviates from the nonlinear dynamics of the system and, hence, the assumed dynamic model becomes inaccurate [Colour figure can be viewed at wileyonlinelibrary.com]

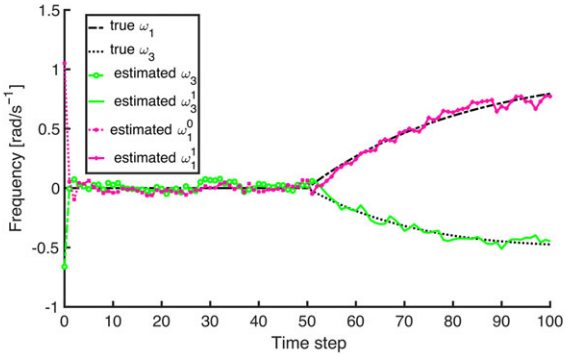

FIGURE 11 Estimated vs *true* trajectory of frequencies ω_1 and ω_3 [Colour figure can be viewed at wileyonlinelibrary.com]

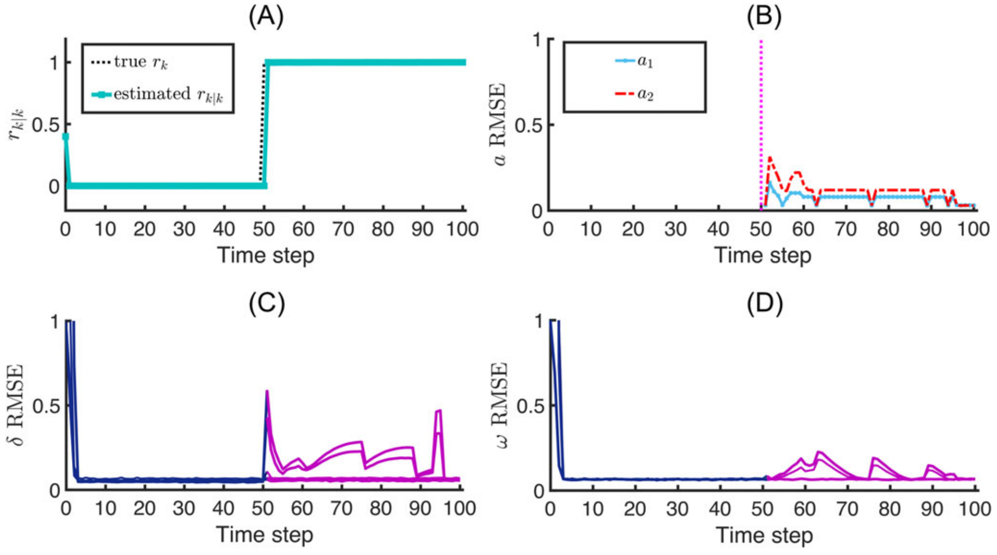

FIGURE 12 Performance of the Gaussian mixture hybrid Bernoulli filter under packet substitution attack ($p_f = 0.3$) in terms of (A) attack detection, (B) attack reconstruction, and (C)-(D) state estimation. RMSE, root mean square error [Colour figure can be viewed at wileyonlinelibrary.com]

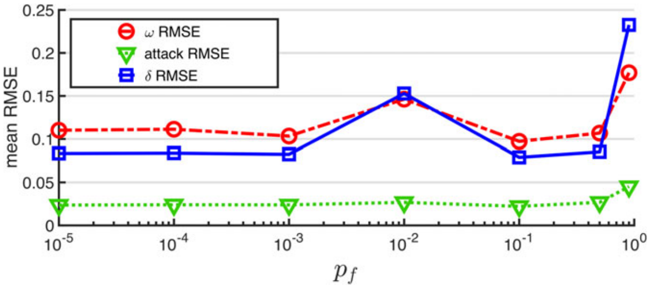

FIGURE 13 Mean RMSE on state (generator rotor angles and frequencies) and attack estimation under packet substitution as a function of filter parameter p_f. Simulated packet substitutions occur with probability $\bar{p}_f = 0.1$. The choice of p_f can improve estimation performance (the best results are obtained when $p_f = \bar{p}_f$), which, however, turns out to be comparable for most parameter values in the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.5, 0.9\}$. RMSE, root mean square error [Colour figure can be viewed at wileyonlinelibrary.com]

load buses and clearly shows how δ_1 and δ_3 lose synchrony once the load altering attack enters into action. Nevertheless, the proposed secure filter keeps tracking the state evolution with high accuracy even after time $k = 50$, once recognized that the system is under attack.

Finally, Figure 11 shows the performance of the GM-HBF in estimating the generator frequencies ω_1 and ω_3, before and after the appearance of the signal attack on the victim loads. The performance of the proposed GM-HBF under packet substitution attack, ie, the filter adopting the correction step described in Section 3.1, is shown in Figure 12 for $p_f = 0.3$ and $p_d = 1$. It is worth noting that the probability of packet substitution p_f can be seen as a design parameter that can be suitably tuned so as to enhance estimation performance. This is illustrated in Figure 13 where the mean (over time, components, and Monte Carlo runs) RMSE on state/attack estimation is shown as a function of parameter p_f. By contrast, simulation results indicated that the choice on p_f does not significantly affect the overall attack detection performance.

7 | CONCLUSIONS

This paper proposed a general framework to solve resilient state estimation for (linear/nonlinear) CPSs considering switching signal attacks, fake measurement injection, and packet substitution. RFSs have been exploited in order to model the switching nature of the signal attack as well as the possible presence of fake measurements, and a Bayesian random set estimation problem has been formulated for jointly detecting a signal attack and estimating the system state. In this way, a HBF for the Bayes-optimal solution of the posed problem has been derived and implemented as a Gaussian-sum filter. Numerical examples concerning both a benchmark system with direct feedthrough and a realistic energy delivery system have been presented so as to demonstrate the potentials and the real-world applicability of the proposed approach. Future work will concern worst-case performance degradation analysis for the developed filter and its application to resilient state estimation in distributed settings with nonsecure communication links.

ORCID

Nicola Forti https://orcid.org/0000-0001-5510-1616
Giorgio Battistelli https://orcid.org/0000-0002-0124-4715
Luigi Chisci https://orcid.org/0000-0001-5049-3577

REFERENCES

1. The Industrial Control Systems Cyber Emergency Response Team (ICS-CERT). https://ics-cert.us-cert.gov/
2. Pasqualetti F, Dörfler F, Bullo F. Attack detection and identification in cyber-physical systems. *IEEE Trans Autom Control*. 2013;58(11):2715-2729.
3. Mo Y, Sinopoli B. Secure control against replay attacks. Paper presented at: 2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton); 2009; Monticello, IL.
4. Miao F, Pajic M, Pappas GJ. Stochastic game approach for replay attack detection. Paper presented at: 52nd IEEE Conference on Decision and Control; 2013; Florence, Italy.
5. De Persis C, Tesi P. Input-to-state stabilizing control under denial-of-service. *IEEE Trans Autom Control*. 2015;60(11):2930-2944.
6. Zhang H, Cheng P, Shi L, Chen J. Optimal denial-of-service attack scheduling with energy constraint. *IEEE Trans Autom Control*. 2015;60(11):3023-3028.
7. Mo Y, Weerakkody S, Sinopoli B. Physical authentication of control systems: designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Syst Mag*. 2015;35(1):93-109.
8. Weerakkody S, Sinopoli B. Detecting integrity attacks on control systems using a moving target approach. Paper presented at: 2015 54th IEEE Conference on Decision and Control (CDC); 2015; Osaka, Japan.

9. Mo Y, Sinopoli B. Secure estimation in the presence of integrity attacks. *IEEE Trans Autom Control*. 2015;60(4):1145-1151.
10. Fawzi H, Tabuada P, Diggavi S. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Trans Autom Control*. 2014;59(6):1454-1467.
11. Pajic M, Lee I, Pappas GJ. Attack-resilient state estimation for noisy dynamical systems. *IEEE Trans Control Netw Syst*. 2017;4(1):82-92.
12. Shoukry Y, Puggelli A, Nuzzo P, Sangiovanni-Vincentelli AL, Seshia SA, Tabuada P. Sound and complete state estimation for linear dynamical systems under sensor attacks using satisfiability modulo theory solving. Paper presented at: 2015 American Control Conference (ACC); 2015; Chicago, IL.
13. Mishra S, Shoukry Y, Karamchandani N, Diggavi SN, Tabuada P. Secure state estimation against sensor attacks in the presence of noise. *IEEE Trans Control Netw Syst*. 2017;4(1):49-59.
14. Chong MS, Wakaiki M, Hespanha JP. Observability of linear systems under adversarial attacks. Paper presented at: 2015 American Control Conference (ACC); 2015; Chicago, IL.
15. Teixeira A, Shames I, Sandberg H, Johansson KH. A secure control framework for resource-limited adversaries. *Automatica*. 2015;51(1):135-148.
16. Shi D, Elliott RJ, Chen T. On finite-state stochastic modeling and secure estimation of cyber-physical systems. *IEEE Trans Autom Control*. 2017;62(1):65-80.
17. Forti N, Battistelli G, Chisci L, Sinopoli B. A Bayesian approach to joint attack detection and resilient state estimation. Paper presented at: 2016 IEEE 55th Conference on Decision and Control (CDC); 2016; Las Vegas, NV.
18. Gu Q, Liu P, Zhu S, Chu C-H. Defending against packet injection attacks in unreliable ad hoc networks. Paper presented at: GLOBECOM '05. IEEE Global Telecommunications Conference; 2005; St. Louis, MO.
19. Zhang X, Chan H, Jain A, Perrig A. *Bounding Packet Dropping and Injection Attacks in Sensor Networks*. Technical Report 07-019. Pittsburgh, PA: CMU-CyLab; 2007. https://www.cylab.cmu.edu/files/pdfs/tech_reports/cmucylab07019.pdf.
20. Ho Y, Lee R. A Bayesian approach to problems in stochastic estimation and control. *IEEE Trans Autom Control*. 1964;9(4):333-339.
21. Ristic B, Vo B-T, Vo B-N, Farina A. Tutorial on Bernoulli filters: theory, implementation and applications. *IEEE Trans Signal Process*. 2013;61(13):3406-3430.
22. Mahler RPS. *Statistical Multisource Multitarget Information Fusion*. Norwood, MA: Artech House Inc; 2007.
23. Vo B-T, Clark D, Vo B-N, Ristic B. Bernoulli forward-backward smoothing for joint target detection and tracking. *IEEE Trans Signal Process*. 2011;59(9):4473-4477.
24. Vo B-T, See CM, Ma N, Ng WT. Multi-sensor joint detection and tracking with the Bernoulli filter. *IEEE Trans Aerosp Electron Syst*. 2012;48(2):1385-1402.
25. Yong SZ, Zhu M, Frazzoli E. Resilient state estimation against switching attacks on stochastic cyber-physical systems. Paper presented at: 2015 54th IEEE Conference on Decision and Control (CDC); 2015; Osaka, Japan.
26. Forti N, Battistelli G, Chisci L, Sinopoli B. Secure state estimation of cyber-physical systems under switching attacks. *IFAC Pap*. 2017;50(1):4979-4986.
27. Fang H, De Callafon RA, Cortés J. Simultaneous input and state estimation for nonlinear systems with applications to flow field estimation. *Automatica*. 2013;49(9):2805-2812.
28. Gillijns S, De Moor B. Unbiased minimum-variance input and state estimation for linear discrete-time systems with direct feedthrough. *Automatica*. 2007;43(5):934-937.
29. Gillijns S, De Moor B. Unbiased minimum-variance input and state estimation for linear discrete-time systems. *Automatica*. 2007;43(1):111-116.
30. Forti N, Battistelli G, Chisci L, Sinopoli B. Bayesian state estimation against unknown switching inputs and extra packet injections. 2019. http://www.nicolaforti.com/wp-content/uploads/2019/02/728.pdf
31. Vo B-N, Ma WK. The Gaussian mixture probability hypothesis density filter. *IEEE Trans Signal Process*. 2006;54(11):4091-4104.
32. Cheng Y, Ye H, Wang Y, Zhou D. Unbiased minimum-variance state estimation for linear systems with unknown input. *Automatica*. 2009;45(2):485-491.
33. Zimmerman RD, Murillo-Sanchez CE, Thomas RJ. MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Trans Power Syst*. 2011;26(1):12-19.
34. Kundur P, Balu NJ, Lauby MG. *Power System Stability and Control*. New York, NY: McGraw-Hill; 1994.
35. Pasqualetti F, Bicchi A, Bullo F. A graph-theoretical characterization of power network vulnerabilities. In: Proceedings of the 2011 American Control Conference; 2011; San Francisco, CA.
36. Amini S, Mohsenian-Rad H, Pasqualetti F. Dynamic load altering attacks in smart grid. Paper presented at: 2015 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference (ISGT); 2015; Washington, DC.

How to cite this article: Forti N, Battistelli G, Chisci L, Sinopoli B. Joint attack detection and secure state estimation of cyber-physical systems. *Int J Robust Nonlinear Control*. 2019;1–28. https://doi.org/10.1002/rnc.4724