Secure Estimation under Causative Attacks

Saurabh Sihag and Ali Tajer
Rensselaer Polytechnic Institute

***Abstract*—This paper considers the problem of secure parameter estimation when the estimation algorithm is prone to *causative* attacks. Causative attacks, in principle, target decision-making algorithms (e.g., inference and learning algorithms) to alter their decisions by making them oblivious to specific attacks. Such attacks influence inference algorithms by tampering with the mechanism through which the algorithm is provided with the statistical model of the population about which an inferential decision is made. Causative attacks are viable, for instance, by contaminating the historical or training data, or by compromising an expert who provides the model. In the presence of causative attacks, the inference algorithms operate under a distorted statistical model for the population from which they collect data samples. This paper introduces specific notions of secure estimation and provides a framework under which secure estimation under causative attacks can be formulated. Closed-form decision rules, and the fundamental tradeoffs between security guarantee and decision qualities are characterized. To circumvent the computational complexity associated with growing parameter dimension or attack complexity, a scalable estimation algorithm and its attendant optimality guarantees are provided.**

I. Introduction

A. Motivation

Anomaly detection, which has diverse applications in intrusion detection, fraud detection, fault detection, system health monitoring, and event detection, constitutes a major class of inference problems in which the objective is raising alarms when the data pattern (e.g., statistical model) deviates significantly from the expected patterns. Effective detection of anomalies in the data strongly hinges on the known rules for distinguishing normal and abnormal data segments. These rules, for instance, can be specified by an expert or by leveraging the historical data, depending on the context of the application.

While anomaly detection, which in essence copes with the vulnerability of the sampled data to being contaminated or compromised, has been studied extensively (c.f. [1]–[3]), the vulnerability of the *inference algorithms* to being compromised is far less-investigated. The nature of security vulnerabilities that inference algorithms are exposed to is fundamentally distinct from that of data. Specifically, in the case of compromised sample data, the information of the decision algorithm about the model remains intact, while the data fed to the algorithm is anomalous. In contrast, attacks on the algorithms can be exerted by providing the algorithm with an incorrect statistical model for the data. This is viable by, for instance, contaminating the historical data or by confusing the expert that produces a model, which are critical for furnishing the true model for the statistical model of the data. Therefore, when the sampled data is compromised, an inference algorithm produces decisions based on an un-compromised known model for the data, while the data that it receives and processes are compromised. On the other hand, when the historical data or the expert are compromised, an inference algorithm functions based on an incorrect model for the data, in which case even un-compromised sampled data produces unreliable decisions.

The aforementioned security vulnerabilities for the inference algorithms can be capitalized on by adversaries in order to force an inference algorithm to deviate from its optimal structure and produce decisions in ways that serve an adversary's purposes. Such attacks on decision algorithms are often referred to as **causative attacks**, through which an adversary aims to (i) make the inference algorithms oblivious to specific attacks, or (ii) degrade the performance of the inference algorithm in the presence of such an attack [4].

While secure decision-making in adjacent domains (e.g., machine learning) has been heavily investigated in recent years, the fundamental limits of secure statistical inference are not well-investigated, and all the limited existing studies remain rather ad-hoc. In this paper, we provide a framework for secure parameter estimation under the potential presence of *causative* attacks. We establish the fundamental tradeoffs involved in decision-making under causative attacks and characterize the optimal decision rules for securely estimating the parameters and concurrently detecting the presence of the attackers.

B. Overview and Contributions

To lay the context for discussing the problem investigated, consider the canonical parameter estimation problem in which we have a collection of probability distributions $\{P_X : X \in \mathcal{X}\}$ defined over a common measurable space. The objective is to estimate X, which lies in a known set $\mathcal{X} \subseteq \mathbb{R}^p$, from data samples $\mathbf{Y} \triangleq [Y_1, \ldots, Y_n]$, where the sample Y_r is distributed according to P_X and lies in a known set $\mathcal{Y} \subseteq \mathbb{R}^m$. We denote the probability density functions (pdfs) that the statistician *assumes* about the underlying distributions of X and Y_r by π and $f(\cdot \mid X)$, respectively, i.e.,

$$Y_r \sim f(\cdot \mid X), \qquad \text{with} \quad X \sim \pi . \tag{1}$$

For convenience, we will assume that the pdfs do not have any non-zero probability masses over lower-dimensional manifolds. The objective of the statistician is formalizing a reliable estimator

$$\hat{X}(\mathbf{Y}) : \mathcal{Y}^n \mapsto \mathcal{X} . \tag{2}$$

Causative Attacks: In an adversarial environment, a malicious attacker might launch a ***causative*** attack to influence (degrade) the quality of $\hat{X}(\mathbf{Y})$. The purpose of such an attack is to compromise the *process that underlies acquiring the statistical models*. We emphasize that such an attack is different from those that aim to compromise the *data*, e.g., false data injection attacks, which aim to distort the data samples $\mathbf{Y}$. Consequently, the effect of a causative attack is misleading the statistician about the true model $f(\cdot \mid X)$ that it assumes about the data. Such attacks are possible by compromising the historical (or training) data that is used for defining a model for the data. Depending on the specificity and the extent of a causative attack, e.g., the fraction of the historical or training data that is compromised, the true model $f(\cdot \mid X)$ can deviate to alternative forms, the space of which we denote by $\mathcal{F}$. The attack can affect the statistical distribution of any number of the m coordinates of $\mathbf{Y}$. There are two major aspects to selecting $\mathcal{F}$ as viable model space.

- An attack is effective if the compromised model is sufficiently distinct from the model assumed by the statistician. Hence, even though in general $\mathcal{F}$ can be any representation of possible

This research was supported in part by the U. S. National Science Foundation under the CAREER Award ECCS-1554482 and the grant DMS-1737976.

kernels $f(\cdot \mid X)$ mapping $\mathcal{Y}$ to $\mathbb{R}^m$, only a subset of such mappings suffices to describe the set of effective attacks.
- There exists a tradeoff between the complexity of the model space and its expressiveness. If it is overly expressive, it can represent the possible compromised models with a more refined accuracy at the expense of more complex inferential rules.

(q,β)***-Security:*** The potential presence of an adversary introduces a new dimension to the estimation problem in (2). Specifically, on the one hand, the stochastic model of the data can be altered by an attack and detecting whether the data model is compromised, itself being an inference task, is never perfect. On the other hand, designing an optimal estimation rule strongly hinges on successfully isolating the true model. Hence, there exists an inherent coupling between the original estimation problem of interest and the introduced auxiliary problem (i.e., detecting the presence of an attacker and isolating the true model). Based on this observation, in an adversarial setting, there exists uncertainty about the true model, based on which the quality of the estimator is expected to degrade with respect to an attack-free setting. We are interested in establishing the fundamental interplay between the quality of discerning the true model and the degradation level in the estimation quality. To establish this interplay, we say that an estimator is (q,β)***-secure*** if its estimation cost is weaker than that of the attack-free setting by a factor $q \in [1,+\infty)$, while missing at most $\beta \in (0,1]$ fraction of the attacks.

C. Related Studies

The problem of secure inference is studied primarily in the context of sensor networks. The study in [5], in particular, considers a two-sensor network in which one sensor is known to be secured, and one sensor is vulnerable to attacks. The objective is forming an estimate based on the mean-squared error criterion, for which a heuristic detection-driven estimator is designed. The adversarial setting defined in this paper is also similar to the widely-studied Byzantine attack models in sensor networks, in which the data generated by the compromised sensors are modified arbitrarily by the adversaries in order to degrade or the inference quality (c.f. [2], [3], [6], [7]). Strategies for isolating the compromised nodes in sensor networks are investigated in [8]–[10]. The emphasis of these studies is primarily focused on detecting attacks, or isolating the attacked sensors, which is different from the scope of our paper, which is focused on parameter estimation. All the aforementioned studies that involve secure estimation, irrespective of their focus or objective, conform in their design principle, which decouples the estimation decisions from all other decisions involved (e.g., attack detection or attacked sensor isolation), and produces either detection-driven estimators or estimation-driven detection routines. Such approaches implicitly assume that the detection decision has been perfect. The premise that decoupling such intertwined estimation and detection problems into independent estimation and detection routines is sub-optimal is well-investigated (e.g., in [11]–[14]).

II. Data Model and Definitions

A. Attack Model

Our focus is on the canonical estimation problem in (2). The objective is to form an optimal estimate $\hat{X}(\mathbf{Y})$ (under the general cost functions specified later) in the potential presence of a causative attack. Under the attack-free setting, the data is assumed to be generated according to the known distribution

$$Y_r \;\sim\; f(\cdot \mid X)\;, \qquad \text{with} \quad X \;\sim\; \pi\;, \qquad \text{for } r \in \{1,\dots n\}\;. \tag{3}$$

In an adversarial setting, an adversary, depending on its strength and preference, can launch an attack that can compromise the underlying process that the statistician uses for acquiring $f(\cdot \mid X)$. An attack will be carried out for the ultimate purpose of degrading the estimation quality of X. We assume that the adversary can corrupt the data model of *up to* $K \in \{1,\dots,m\}$ coordinates of $\mathbf{Y}$. Hence, for a given K, there exist $T = \sum_{i=1}^{K} \binom{m}{i}$ number of attack scenarios under which the compromised data models are distinct. Define $\mathcal{S} \triangleq \{S_1,\dots,S_T\}$ as the set of all possible combinations of attack scenarios, where $S_i \subseteq \{1,\dots,m\}$ describes the set of coordinates the models of which are compromised under scenario $i \in \{1,\dots,T\}$.

Under the attack scenario $i \in \{1,\dots,T\}$, the distribution of Y_r deviates from f and changes to a model in the space $\mathcal{F}_i$. As discussed earlier, there exists a tradeoff between the expressiveness of this space and the complexity of the ensuing inferential rules. Specifically, a larger space $\mathcal{F}_i$ can distinguish different attack strategies with a more accurate resolution at the expense of high complexity in the analysis and the resulting decision rules. Also, the model can be effective if it encompasses sufficiently distinct models. Throughout the analysis of the paper, we assume that $\mathcal{F}_i \triangleq \{f_i(\cdot \mid X)\}$, i.e., $\mathcal{F}_i$ consists of one alternative distribution. This is primarily for the convenience in notations, and all the results presented can be generalized to any arbitrary space with countable elements. Based on this model, when the data models in the coordinates contained in S_i are compromised, the joint distribution changes from $f(\cdot \mid X)$ to $f_i(\cdot \mid X)$.

Different attack scenarios might occur with different likelihoods, e.g., compromising one coordinate is easier than compromising two, and it might turn out to be more likely. To distinguish such likelihoods we adopt a Bayesian framework in which we define ϵ_0 as the prior probability of having an attack-free scenario and define ϵ_i as the prior probability of the event that the attacker compromises the model under the coordinates specified by S_i. A block diagram of the attack model and the inferential goals to be characterized, which are discussed in the remainder of this section, is depicted in Fig. 1. Finally, we define the marginal pdf of the data at coordinate $l \in \{1,\dots,m\}$ under the attack-free setting and when the coordinate is compromised by g_l^0 and g_l^1, respectively.

B. Decision Cost Functions

1) Attack Detection Costs: The possibility of having multiple alternatives to the attack-free model renders the model detection problem as the following composite hypothesis testing problem.

$$\begin{array}{lll} \mathsf{H}_0 \;:\; & \mathbf{Y} \sim f(\mathbf{Y} \mid X), & \text{with } X \sim \pi(X) \\ \mathsf{H}_i \;:\; & \mathbf{Y} \sim f_i(\mathbf{Y} \mid X), & \text{with } X \sim \pi(X)\;, \quad \text{for } \in \{1,\dots,T\} \end{array}\;. \tag{4}$$

The likelihood of deciding in favor of H_j under the true model H_i is

$$\mathbb{P}(\mathsf{D}=\mathsf{H}_j \,|\, \mathsf{T}=\mathsf{H}_i) = \int_{\mathbf{Y}} \delta_j(\mathbf{Y}) f_i(\mathbf{Y})\, \mathrm{d}\mathbf{Y}\;. \tag{5}$$

We define P_{md} as the aggregate probability of incorrectly identifying the true model under the presence of compromised coordinates, i.e.,

$$\mathsf{P}_{\mathsf{md}}(\boldsymbol{\delta}) \triangleq \mathbb{P}(\mathsf{D} \neq \mathsf{T} \mid \mathsf{T} \neq \mathsf{H}_0) = \sum_{i=1}^{T} \frac{\epsilon_i \cdot \mathbb{P}(\mathsf{D} \neq \mathsf{H}_i \mid \mathsf{T} = \mathsf{H}_i)}{1-\epsilon_0}\;. \tag{6}$$

Furthermore, we define P_{fa} as the aggregate probability of erroneously declaring that a set of coordinates are compromised, while operating in an attack-free scenario. We have

$$\mathsf{P}_{\mathsf{fa}}(\boldsymbol{\delta}) \triangleq \mathbb{P}(\mathsf{D} \neq \mathsf{H}_0 \mid \mathsf{T} = \mathsf{H}_0) = \sum_{i=1}^{T} \mathbb{P}(\mathsf{D}=\mathsf{H}_i \,|\, \mathsf{T}=\mathsf{H}_0)\;. \tag{7}$$

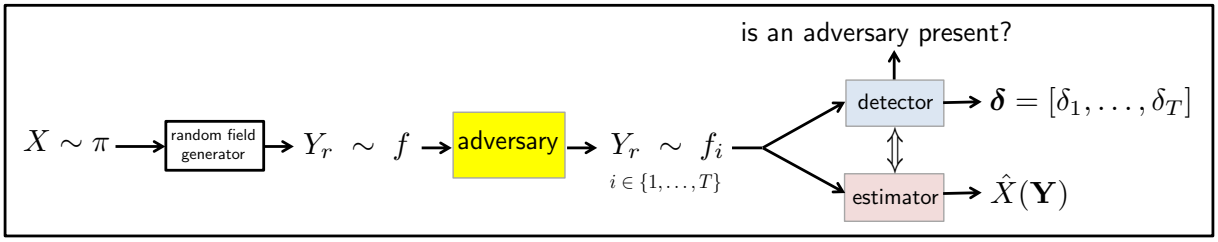

Fig. 1: The effect of the adversary on the data model, and the inferential decisions involved.

2) Secure Estimation Costs: In this subsection, we define two estimation cost functions for capturing the fidelity of the estimate $\hat{X}(\mathbf{Y})$ we aim to form for X. For this purpose, we adopt a generic and *non-negative* cost function $\mathsf{C}(X, U(\mathbf{Y}))$ to quantify the discrepancy between the ground truth X and a generic estimator $U(\mathbf{Y})$. Based on this, corresponding to each model H_i and given data $\mathbf{Y}$ we first define the *average posterior cost function* as

$$\mathsf{C}_{\mathrm{p},i}(U(\mathbf{Y}) \mid \mathbf{Y}) \triangleq \mathbb{E}_i \left[\mathsf{C}(X, U(\mathbf{Y})) \mid \mathbf{Y}\right] \ \forall i \in \{0, \ldots, T\} \ , \quad (8)$$

where the conditional expectation is with respect to X when the true model is H_i. Besides this, due to having distinct data models under different attack models, we consider having possibly distinct estimators under different models. Driven by this, we consider the design for an estimate for X under each model. We denote the estimate of X under model H_i by $\hat{X}_i(\mathbf{Y})$, and accordingly, we define

$$\hat{\mathbf{X}}(\mathbf{Y}) \triangleq [\hat{X}_0(\mathbf{Y}), \ldots, \hat{X}_T(\mathbf{Y})] \ . \quad (9)$$

Considering such distinct estimators, the estimation cost $\mathsf{C}(X, \hat{X}_i(\mathbf{Y}))$ is relevant only if the decision is H_i. Hence, for any generic estimator $U_i(\mathbf{Y})$ of X under model H_i, we define the *decision-specific average cost function* as

$$J_i(\delta_i, U_i(\mathbf{Y})) \triangleq \mathbb{E}_i[\mathsf{C}(X, U_i(\mathbf{Y})) \mid \mathsf{D} = \mathsf{H}_i] \ , \forall i \ , \quad (10)$$

where the conditional expectation is with respect to X and $\mathbf{Y}$. Accordingly, we define an aggregate average estimation as

$$J(\boldsymbol{\delta}, \mathbf{U}) \triangleq \max_{i \in \{0, \ldots, T\}} J_i(\delta_i, U_i(\mathbf{Y})) \ , \quad (11)$$

where we have defined $\mathbf{U} \triangleq [U_0(\mathbf{Y}), \ldots, U_T(\mathbf{Y})]$. Finally, corresponding to the attack-free scenario, in which the only possible data model is the assumed model f, corresponding to any generic estimator $V(\mathbf{Y})$ we define the average estimation according to

$$J_0(V) = \mathbb{E}[\mathsf{C}(X, V(\mathbf{Y}))] \ , \quad (12)$$

where the expectation is with respect to X and $\mathbf{Y}$ under model f. It is noteworthy that J_0 defined in (12) is fundamentally different from $J(\boldsymbol{\delta}, \mathbf{U})$ defined in (11), since the former is the estimation cost when there is no alternative to f (i.e., the attack-free scenario), while the latter is the estimation cost in an adversarial setting in which we have decided that the attacker has not compromised the data, which being a detection decision is never perfect and can be inaccurate with a non-zero probability. The role of $J_0(V)$ in our analysis is furnishing a baseline for the estimation quality in order to assess the impact of the potential presence of an adversary on the estimation quality.

III. Secure Parameter Estimation

The core premise underlying the notion of secure estimation presented is that there exists an inherent interplay between the quality of estimating X and the quality of isolating the true model governing the data. Specifically, perfect detection of an adversary's attack model is impossible. At the same time, the estimation quality strongly relies on the successful isolation of the true data model. Lack of a perfect decision about the data model is expected to degrade the estimation quality compared to the attack-free scenario. To quantify such an interplay as well as the degradation in estimation quality with respect to the attack-free scenario, we provide the following definition.

Definition 1: For a given estimator V in the attack-free scenario, and a secure estimation procedure specified by rules $(\boldsymbol{\delta}, \mathbf{U})$ in the adversarial scenario, we define the estimation degradation factor as

$$q(\boldsymbol{\delta}, \mathbf{U}, V) \triangleq \frac{J(\boldsymbol{\delta}, \mathbf{U})}{J_0(V)} \ . \quad (13)$$

Based on this definition, next we define the performance region, which encompasses all the pairs of decision qualities $q(\boldsymbol{\delta}, \mathbf{U}, V)$ and $\mathsf{P}_{\mathsf{md}}(\boldsymbol{\delta})$ over the space of all possible decision rules $(\boldsymbol{\delta}, \mathbf{U}, V)$.

Definition 2 (Performance Region): We define the performance region as the region of all simultaneously achievable estimation quality $q(\boldsymbol{\delta}, \mathbf{U}, V)$ and detection performance $\mathsf{P}_{\mathsf{md}}(\boldsymbol{\delta})$.

By leveraging the characteristics of the performance region, next we define the notion of (q, β)-security, which is instrumental in defining the secure estimation problem of interest. For this purpose, note that a defined in (13) normalizes the estimation cost in the adversarial setting by that of the attack-free scenario. The two estimation cost functions involved in $q(\boldsymbol{\delta}, \mathbf{U}, V)$ can be computed independently, and as a result, determining their attendant decision rules can be carried out independently. For this purpose, we define V^* as the optimal decision rule under the attack-free setting, and J_0^* as the corresponding estimation cost, i.e.,

$$V^* \triangleq \arg\min_V J_0(V) \ , \qquad \text{and} \qquad J_0^* \triangleq \min_V J_0(V) \ . \quad (14)$$

Definition 3 ((q, β)-security): An estimation procedure specified by $(\boldsymbol{\delta}, \mathbf{U}, V^*)$ for the adversarial scenario is said to be (q, β)-secure if the decision rules $(\boldsymbol{\delta}, \mathbf{U})$ yield the minimal EDF among all the decision rules corresponding to which the average rate of missing the attacks does not exceed $\beta \in (0, 1]$, i.e.,

$$q \triangleq \min_{\boldsymbol{\delta}, \mathbf{U}} q(\boldsymbol{\delta}, \mathbf{U}, V^*) \ , \qquad \text{s.t.} \qquad \mathsf{P}_{\mathsf{md}}(\boldsymbol{\delta}) \leq \beta \ . \quad (15)$$

The performance region, and its boundary that specifies the interplay between q and β are illustrated in Fig. 2. Based on these definitions, we aim to characterize:

1) The region of all simultaneously achievable values of $q(\boldsymbol{\delta}, \mathbf{U}, V^*)$ and $\mathsf{P}_{\mathsf{md}}(\boldsymbol{\delta})$, which is illustrated by the dashed region in Fig. 2.
2) The (q, β)-secure decision rules $(\boldsymbol{\delta}, \mathbf{U}, V^*)$ that solve (15), and specify the boundary of the performance region, which is illustrated by a solid line as the boundary of the performance region in Fig. 2.

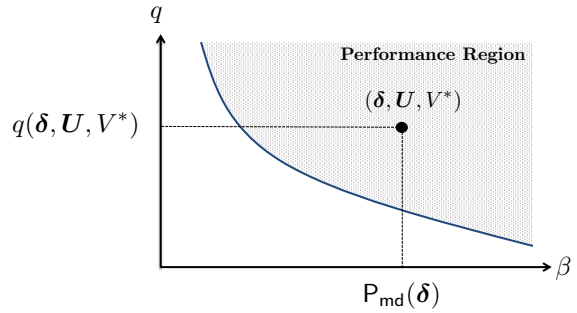

Fig. 2: Performance region.

Since$q(\boldsymbol{\delta}, \mathbf{U}, V^*) = \frac{J(\boldsymbol{\delta}, \mathbf{U})}{J_0^*}$, where J_0^* is a constant, the performance region and the (q, β)-secure decision rules are found by solving

$$\mathcal{Q}(\beta) \triangleq \begin{cases} \min_{\boldsymbol{\delta}, \mathbf{U}} & J(\boldsymbol{\delta}, \mathbf{U}) \\ \text{s.t.} & \mathsf{P}_{\mathsf{md}}(\boldsymbol{\delta}) \leq \beta \end{cases} . \tag{16}$$

IV. Secure Estimation: Optimal Decision Rules

We characterize an optimal solution to the more general problem $\mathcal{Q}(\beta)$, i.e., the estimators $\{\hat{X}_i(\mathbf{Y}) : i \in \{0, \ldots, T\}\}$ and the detectors $\{\delta_i(\mathbf{Y}) : i \in \{0, \ldots, T\}\}$. By noting (5) and (7), we obtain

$$\mathcal{Q}(\beta) = \begin{cases} \min_{(\boldsymbol{\delta}, \boldsymbol{U})} & J(\boldsymbol{\delta}, \mathbf{U}) \\ \text{s.t.} & \sum\limits_{i=1}^{T} \frac{\epsilon_i}{1-\epsilon_0} \sum\limits_{\substack{j=0 \\ j \neq i}}^{T} \int_{\mathbf{Y}} \delta_j(\mathbf{Y}) f_i(\mathbf{Y}) \, \mathrm{d}\mathbf{Y} \leq \beta \end{cases} . \tag{17}$$

The roles of the estimators $\{U_i(\mathbf{Y}) : i \in \{0, \ldots, T\}\}$ appear only in the utility function $J(\boldsymbol{\delta}, \mathbf{U})$. This allows for decoupling the optimization problem $\mathcal{Q}(\beta)$ into two sub-problems, as formalized in Theorem 1.

Theorem 1: The optimal secure estimators of X under different models, i.e., $\hat{\mathbf{X}} = [\hat{X}_0, \ldots, \hat{X}_T]$ is the solution to

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{U}} J(\boldsymbol{\delta}, \mathbf{U}) . \tag{18}$$

Furthermore, the solution of $\mathcal{Q}(\beta)$, and subsequently the design of the attack detectors, can be found by equivalently solving

$$\mathcal{Q}(\beta) = \begin{cases} \min_{\boldsymbol{\delta}} & J(\boldsymbol{\delta}, \hat{\mathbf{X}}) \\ \text{s.t.} & \sum\limits_{i=1}^{T} \frac{\epsilon_i}{1-\epsilon_0} \sum\limits_{\substack{j=0 \\ j \neq i}}^{T} \int_{\mathbf{Y}} \delta_j(\mathbf{Y}) f_i(\mathbf{Y}) \, \mathrm{d}\mathbf{Y} \leq \beta \end{cases} . \tag{19}$$

This theorem establishes a property for the optimal estimator in (18), by leveraging which, in the following theorem we provide optimal designs for the secure estimators. Interestingly, it is shown that the optimal estimator under each model can be specified by optimizing a relevant cost function defined exclusively for that model.

Theorem 2 ((q, β)-secure Estimators): For the optimal secure estimators $\hat{\mathbf{X}}$ we have

1) The minimizer of the estimation cost $J_i(\delta_i, U_i(\mathbf{Y}))$, i.e., the estimation cost function under model H_i, is given by

$$U_i^*(\mathbf{Y}) \triangleq \arg \inf_{U_i(\mathbf{Y})} \mathsf{C}_{\mathrm{p},i}(U_i(\mathbf{Y}) \mid \mathbf{Y}) . \tag{20}$$

2) The optimal estimator $\hat{\mathbf{X}} = [\hat{X}_0, \ldots, \hat{X}_T]$is given by

$$\hat{X}_i(\mathbf{Y}) = U_i^*(\mathbf{Y}) . \tag{21}$$

3) The cost function $J(\boldsymbol{\delta}, \hat{\mathbf{X}})$ is given by

$$J(\boldsymbol{\delta}, \hat{\mathbf{X}}) = \max_i \left\{ \frac{\int_{\mathbf{Y}} \delta_i(\mathbf{Y}) \mathsf{C}_{\mathrm{p},i}^*(\mathbf{Y}) f_i(\mathbf{Y}) d\mathbf{Y}}{\int_{\mathbf{Y}} \delta_i(\mathbf{Y}) f_i(\mathbf{Y}) d\mathbf{Y}} \right\} , \tag{22}$$

where we have defined

$$\mathsf{C}_{\mathrm{p},i}^*(\mathbf{Y}) \triangleq \inf_{U_i(\mathbf{Y})} \mathsf{C}_{\mathrm{p},i}(U_i(\mathbf{Y}) \mid \mathbf{Y}) . \tag{23}$$

Next, given the optimal estimators $\hat{\mathbf{X}}$, we characterize the optimal detection rules in the next theorem. The main observation is that even though we started by considering general randomized decision rules, these rules in their optimal forms reduce to deterministic ones. Furthermore, the decisions rules depend on the estimation costs that are computed based on the optimal estimation costs. These estimation costs make the decisions coupled. In order to proceed, we first show that $\mathcal{Q}(\beta)$ in (19) can be solved by leveraging the result of the following theorem, which specifies an auxiliary convex problem.

Theorem 3: For any arbitrary $u \in \mathbb{R}_+$, we have $\mathcal{Q}(\beta) \leq u$ if and only if $\mathcal{R}(\beta, u) \leq 0$, where we have defined

$$\mathcal{R}(\beta, u) = \begin{cases} \min_{\boldsymbol{\delta}} & \gamma \\ \text{s.t.} & \int_{\mathbf{Y}} \delta_i(\mathbf{Y}) f_i(\mathbf{Y}) [\mathsf{C}_{\mathrm{p},i}^*(\mathbf{Y}) - u] \, \mathrm{d}\mathbf{Y} \leq \gamma, \\ & \sum\limits_{i=1}^{T} \frac{\epsilon_i}{1-\epsilon_0} \sum\limits_{\substack{j=0 \\ j \neq i}}^{T} \int_{\mathbf{Y}} \delta_j(\mathbf{Y}) f_i(\mathbf{Y}) \, \mathrm{d}\mathbf{Y} \leq \beta + \gamma \end{cases} . \tag{24}$$

Furthermore, $\mathcal{R}(\beta, u)$ is convex, and $\mathcal{R}(\beta, u) = 0$ has a unique solution in u, which we denote by u^*.

The point u^* has a pivotal role in specifying the optimal detection decision rules. We define the constants $\{\ell_i : i \in \{0, \ldots, T+2\}\}$ as the dual variables in the Lagrange function associated with the convex problem $\mathcal{R}(\beta, u^*)$. Based on these parameters, the optimal detection rules can be characterized in closed-forms, as specified in the following theorem.

Theorem 4 ((q, β)-secure Detection Rules): The optimal decision rules for isolating the compromised coordinates are given by

$$\delta_i(\mathbf{Y}) = \begin{cases} 1, & \text{if} \quad i = i^* \\ 0, & \text{if} \quad i \neq i^* \end{cases} , \tag{25}$$

where we have defined $i^* \triangleq \operatorname{argmin}_{i \in \{0, \ldots, T\}} A_i$. Constants $\{A_0, \ldots, A_T\}$ are specified by the models, u^*, and its associated Langrangian multipliers $\{\ell_i : i \in \{0, \ldots, T+2\}\}$. Specifically,

$$A_0 \triangleq \ell_0 f_0(\mathbf{Y}) [\mathsf{C}_{\mathrm{p},0}^*(\mathbf{Y}) - u^*] + \ell_{T+1} \sum_{i=1}^{T} \frac{\epsilon_i}{1-\epsilon_0} f_i(\mathbf{Y}) , \tag{26}$$

$$A_i \triangleq \ell_i f_i(\mathbf{Y}) [\mathsf{C}_{\mathrm{p},i}^*(\mathbf{Y}) - u^*] + \ell_{T+1} \sum_{j=1, j \neq i}^{T} \frac{\epsilon_j}{1-\epsilon_0} f_j(\mathbf{Y}) + \ell_{T+2} f_0(\mathbf{Y}) . \tag{27}$$

Algorithm 1 summarizes all the steps involved for solving $\mathcal{Q}(\beta)$.

V. Case Study: Secure Estimation in Sensor Networks

We consider a network of two sensors and a fusion center (FC) to evaluate the estimation frameworks presented in this paper. Each sensor is collecting a stream of data. Sensor $i \in \{1, 2\}$ collects n measurements, denoted by $\mathbf{Y}_i = [Y_1^i, \ldots, Y_n^i]$, where

$$Y_j^i = h^i X + N_j^i . \tag{28}$$

Algorithm 1 – Solving $\mathcal{Q}(\beta)$

Initialize $u_0 = 0, u_1$
Evaluate optimal posterior estimation costs
repeat
  $\hat{u} \leftarrow (u_0 + u_1)/2$
  for every $\hat{\boldsymbol{\ell}} \succcurlyeq 0$ in the discretized space $\|\hat{\boldsymbol{\ell}}\|_1 = 1$ **do**
    Compute $\boldsymbol{\delta}$ from Theorem 4
    Compute $M(\hat{\boldsymbol{\ell}}) \triangleq \mathcal{R}(\beta, \hat{u})$
  end for
  if $\min_{\hat{\boldsymbol{\ell}}} M(\hat{\boldsymbol{\ell}}) \leq 0$ **then**
    $u_1 \leftarrow \hat{u}$
    $\boldsymbol{\ell} \leftarrow \hat{\boldsymbol{\ell}}$
  else
    $u_0 \leftarrow \hat{u}$
  end if
until $u_1 - u_0 \leq \epsilon$, for ϵ sufficiently small
$\mathcal{Q}(\beta) \leftarrow u^* = u_1$

h^i models the channel connecting sensor i to the FC and N_j^i accounts for the additive channel noise. Different noise terms are assumed to be independent and identically distributed (i.i.d.) generated according to a known distribution.

We consider an adversarial setting in which only sensor 1 is vulnerable. Hence, we have only one attack scenario. Accordingly, we have $\epsilon_0 + \epsilon_1 = 1$. Under the attack-free scenario, we assume that the noise terms N_j^i are i.i.d. with distribution $\mathcal{N}(0, \sigma_n^2)$. When data from sensor 1 is compromised, the actual conditional distribution of $Y_j^1|X$ is distinct from the distribution assumed by the statistician. The inference objective, in principle, becomes similar to the adversarial setting of [5], which focuses on data injection. Hence, in order to be able to compare the performance of the optimal framework with that of [5], we assume that the conditional distribution of $Y_j^1|X$ when sensor 1 is under a causative attack is $\mathcal{N}(h^i X, \sigma_n^2) * \mathsf{Unif}[a, b]$, where $a, b \in \mathbb{R}$ are fixed constants and $*$ denotes convolution.

Figure 3 depicts the variations of q, versus the tolerable miss-detection rate β, where it is observed that the estimation quality improves monotonically as β increases, and it reaches its maximum quality when $\beta = 1$. A similar setting is studied in [5], where the attack is induced additively into the data of sensor 1 and can be any real number. This setting can be studied in the context of causative attacks where the attacker's mode of compromising the data is adding a disturbance which has uniform distribution. Figure 3 also compares the estimation quality of the methodology developed in this paper, with that obtained by applying the methodology of [5], which characterizes a single point in the (q, β) plane. Specifically, in [5], an estimator is designed to obtain the most robust estimate by exploring the dependence of the estimation quality on the false alarm probability, using which an optimal false alarm probability is obtained, which in turn, fixes the miss-detection error probability, and does not provide the flexibility to change the miss-detection rate β. The results presented in Fig. 3 correspond to $\sigma = 3$, $\sigma_n = 1$, $h^1 = 1$, $h^2 = 4$, $a = -40, b = 40$. The upper bound on $\mathsf{P_{fa}}$ is set to $\alpha^* = 0.1$, where α^* is obtained using the methodology in [5].

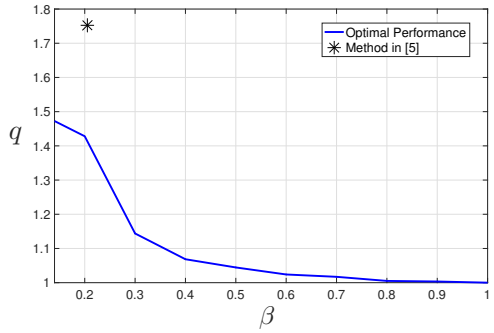

Fig. 3: q versus β.

VI. Conclusion

We have formalized and analyzed the problem of secure parameter estimation problem under the potential presence of causative attacks on the estimation algorithm. Under causative attacks, the information of the estimation algorithm about the statistical model of the sampled data is compromised. This leads the estimation algorithm exhibit degraded performance compared to the attack-free setting. We have provided closed-form optimal decision rules that ensure the best estimation quality (minimum estimation cost) while controlling the error in detecting the attacks and isolating the true model of the data. We have designed the optimal decision rules, which combine both estimation performance and detection power.

References

[1] A. Tajer, V. V. Veeravalli, and H. V. Poor, "Outlying Sequence Detection in Large Datasets - A Data-Drive Approach," *IEEE Signal Processing Magazine - Special Issue on Signal Processing for Big Data*, vol. 31, no. 5, pp. 44–56, September 2014.

[2] A. Vempaty, L. Tong, and P. K. Varshney, "Distributed inference with Byzantine data: State-of-the-art review on data falsification attacks," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 65–75, Sep. 2013.

[3] A. Vempaty, Y. S. Han, and P. K. Varshney, "Target localization in wireless sensor networks using error correcting codes," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 697–712, Jan. 2014.

[4] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proc. ACM Symposium on Information, computer and communications security*, Taipei, Taiwan, Mar. 2006, pp. 16–25.

[5] C. Wilson and V. V. Veeravalli, "MMSE estimation in a sensor network in the presence of an adversary," in *Proc. IEEE International Symposium on Information Theory*, Barcelona, Spain, Jul. 2016, pp. 2479–2483.

[6] A. Vempaty, O. Ozdemir, K. Agrawal, H. Chen, and P. K. Varshney, "Localization in wireless sensor networks: Byzantines and mitigation techniques," *IEEE Transactions on Signal Processing*, vol. 61, no. 6, pp. 1495–1508, Mar. 2013.

[7] J. Zhang, R. S. Blum, X. Lu, and D. Conus, "Asymptotically optimum distributed estimation in the presence of attacks," *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1086–1101, Mar. 2015.

[8] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Countering Byzantine attacks in cognitive radio networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, Mar. 2010, pp. 3098–3101.

[9] E. Soltanmohammadi, M. Orooji, and M. Naraghi-Pour, "Decentralized hypothesis testing in wireless sensor networks in the presence of misbehaving nodes," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 205–215, Jan. 2013.

[10] A. Vempaty, K. Agrawal, P. Varshney, and H. Chen, "Adaptive learning of Byzantines' behavior in cooperative spectrum sensing," in *Proc. IEEE Wireless Communications and Networking Conference*, Cancun, Mexico, Mar. 2011, pp. 1310–1315.

[11] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 434–444, May 1968.

[12] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1597–1602, Sep. 1992.

[13] G. V. Moustakides, G. H. Jajamovich, A. Tajer, and X. Wang, "Joint detection and estimation: Optimum tests and applications," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4215–4229, Jul. 2012.

[14] G. H. Jajamovich, A. Tajer, and X. Wang, "Minimax-optimal hypothesis testing with estimation-dependent costs," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6151–6165, Dec. 2012.