DESIGN OF DATA-INJECTION ADVERSARIAL ATTACKS AGAINST SPATIAL FIELD DETECTORS

Roberto López-Valcarce[*]

University of Vigo, Spain
valcarce@gts.uvigo.es

Daniel Romero

University of Minnesota, USA
romer155@umn.edu

ABSTRACT

Data-injection attacks on spatial field detection corrupt a subset of measurements to cause erroneous decisions. We consider a centralized decision scheme exploiting spatial field smoothness to overcome lack of knowledge on system parameters such as noise variance. We obtain closed-form expressions for system performance and investigate strategies for an intruder injecting false data in a fraction of the sensors in order to reduce the probability of detection. The problem of determining the most vulnerable subset of sensors is also analyzed.

Index Terms— Adversarial detection, Byzantine sensors, cyber security, spatial field detection, sensor networks.

1. INTRODUCTION

Advances in embedded sensors have renewed interest in multisensor signal detection for surveillance and environmental and safety monitoring [1, 2]. Typical sensor networks comprise a large number of low-cost nodes measuring some physical phenomena and reporting readings to a fusion center (FC). Being deployed over large areas with unattended nodes, they remain susceptible to external and internal attacks [3]. Sensor readings may be compromised before they reach the FC if an intruder either alters the contents of data packets after capturing a node, or directly modifies the environmental parameters around some sensors. Cryptographic measures are ineffective against such data-injection attacks.

We focus on typical dense deployments, for which the measurements of the monitored physical phenomenon will exhibit spatial smoothness [4]. This allows for parsimonious parametric modeling of the corresponding spatial field, which can be exploited for inference purposes [5–7]. The setting is based on a linear model in spatially independent Gaussian noise with unknown variance, with the FC detecting the presence of a spatially smooth field (toxic chemical spill, bacterial activity, electromagnetic radiation, etc.) Since the distributions have unknown parameters, a Generalized Likelihood Ratio (GLR) approach is adopted. We then investigate the design of attacks by an adversary injecting false data in a number of sensors, with the goal of decreasing the probability of detecting the phenomenon of interest.

If successful, such adversarial actions may have catastrophic consequences, and therefore it is of great importance to understand their potential and limitations in order to devise adequate defense mechanisms. Our work constitutes a first step in that direction: we quantify the damage that the adversary can inflict to the system, and provide guidelines to determine an appropriate subset of sensors for the adversary to capture in order to maximize system degradation, thus revealing network vulnerabilities.

Although detection in adversarial environments has been considered in previous works [8–12], in all of these it is assumed that the sensors (malicious or not) are *binary,* i.e., they report local 1-bit decisions to the FC. However, in our setting, and due to the presence of unknown parameters in the distributions, it is not possible for an individual sensor to make a local decision by itself, because the GLR with a single measurement does not exist. Therefore, in contrast with the aforementioned works, sensors must send their measurements (rather than local 1-bit decisions) to the FC, which in turn makes the corresponding decision. Our model is similar to that from [13, 14], which studied the effects of data attacks for state *estimation* (rather than *detection*) for cyber-physical systems.

Notation: For a matrix $\boldsymbol{A}$, $\boldsymbol{A}^\dagger$ denotes its pseudoinverse, and $\mathcal{R}(\boldsymbol{A})$ and $\mathcal{N}(\boldsymbol{A})$ its column and null spaces, respectively. For $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ symmetric, its ordered eigenvalues are denoted by $\lambda_1(\boldsymbol{A}) \geq \ldots \geq \lambda_n(\boldsymbol{A})$. The Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{C}$ is denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$, whereas χ^2_ν and $\chi'^2_\nu(\lambda)$ denote the central and, respectively, noncentral χ^2-distributions with ν degrees of freedom (d.o.f.) and centrality parameter λ. We denote by $\mathcal{F}_{\nu_1,\nu_2}$, $\mathcal{F}'_{\nu_1,\nu_2}(\lambda_1)$ and $\mathcal{F}''_{\nu_1,\nu_2}(\lambda_1,\lambda_2)$ respectively the central, noncentral, and doubly noncentral F-distributions with ν_1 and ν_2 d.o.f. and centrality parameters λ_1, λ_2. The corresponding cdfs are denoted by $F_{\nu_1,\nu_2}(x)$, $F'_{\nu_1,\nu_2}(x;\lambda)$, and $F''_{\nu_1,\nu_2}(x;\lambda_1,\lambda_2)$.

[*]Supported by the Spanish Government and the European Regional Development Fund (ERDF) (projects TEC2013-47020-C2-1-R COMPASS, TEC2015-69648-REDC Red COMONSENS), and by the Galician Regional Government and ERDF (projects GRC2013/009 "Consolidation of Research Units", R2014/037 REdTEIC and AtlantTIC).

2. SYSTEM MODEL

We first describe the operation of an attack-unaware monitoring system. Consider n sensor nodes deployed to detect some physical phenomenon of interest. A scalar measurement at the i-th sensor is modeled as

$$y_i = \boldsymbol{h}_i^T \boldsymbol{x} + w_i, \qquad i = 1, \ldots, n, \tag{1}$$

with $\boldsymbol{x} \in \mathbb{R}^m$ an unknown vector related to the monitored physical process, $\boldsymbol{h}_i \in \mathbb{R}^m$ known, and i.i.d. measurement noise $\{w_i\}$ with $w_i \sim \mathcal{N}(0, \sigma^2)$; the variance σ^2 is unknown. The spatial field is assumed sufficiently smooth, so that it can be parameterized by a low-dimensional $\boldsymbol{x}$ with $m \ll n$. This model is fairly general and accommodates a wide range of signal repesentations based on Fourier series, polynomial basis functions, wavelets, splines, etc. [5, 7]. Thus, $\boldsymbol{h}_i$ is a function of the basis expansion model chosen as well as of the location of the i-th sensor. The sensors report their observations to an FC, which then builds the $n \times 1$ vector

$$\boldsymbol{y} \triangleq [\ y_1 \quad y_2 \quad \cdots \quad y_n\ ]^T = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{w}, \tag{2}$$

where $\boldsymbol{w} \triangleq [\, w_1 \ \cdots \ w_n \,]^T \in \mathbb{R}^n$, and $\boldsymbol{H} \in \mathbb{R}^{n \times m}$ has $\boldsymbol{h}_i^T$ as its i-th row. We assume w.l.o.g. that $\boldsymbol{H}$ is full-column rank.

2.1. GLR Detection

The goal of the system is to decide whether the phenomenon of interest (e.g., a toxic chemical spill) is present, i.e., whether $\boldsymbol{x} \neq \boldsymbol{0}$. Modeling $\boldsymbol{x}$ as unknown deterministic, vector $\boldsymbol{y}$ in (2) is Gaussian distributed. The corresponding hypothesis test is

$$\mathcal{H}_0 : \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \qquad \mathcal{H}_1 : \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{H}\boldsymbol{x}, \sigma^2 \boldsymbol{I}_n). \tag{3}$$

Since the pdf $p(\boldsymbol{y})$ contains unknown parameters under both hypotheses, a sensible approach is to adopt a GLR test [15]:

$$L_G(\boldsymbol{y}) \triangleq \frac{\max_{\boldsymbol{x}, \sigma^2} p(\boldsymbol{y}\,;\, \boldsymbol{x}, \sigma^2)}{\max_{\sigma^2} p(\boldsymbol{y}\,;\, \boldsymbol{0}, \sigma^2)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{4}$$

with γ a suitable threshold. This results in

$$L_G(\boldsymbol{y}) = \left(\frac{\|\boldsymbol{y}\|^2}{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{H}^\dagger \boldsymbol{y}\|^2)} \right)^{\frac{n}{2}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \tag{5}$$

Let $\boldsymbol{P}_{||} \triangleq \boldsymbol{H}\boldsymbol{H}^\dagger$ and $\boldsymbol{P}_\perp \triangleq \boldsymbol{I}_n - \boldsymbol{H}\boldsymbol{H}^\dagger$ be respectively the orthogonal projectors onto the *signal subspace* $\mathcal{R}(\boldsymbol{H})$ and the *noise subspace* $\mathcal{N}(\boldsymbol{H}^T)$. Hence $\|\boldsymbol{y}\|^2 = \|\boldsymbol{P}_\perp \boldsymbol{y}\|^2 + \|\boldsymbol{P}_{||} \boldsymbol{y}\|^2$, and, with $c \triangleq \frac{n-m}{m}$, the GLR test (5) is equivalent to

$$T \triangleq c\, \frac{\|\boldsymbol{P}_{||} \boldsymbol{y}\|^2}{\|\boldsymbol{P}_\perp \boldsymbol{y}\|^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma', \quad \text{with} \quad \gamma' \triangleq c\,(\gamma^{\frac{2}{n}} - 1). \tag{6}$$

Note that if $n = 1$, T in (6) becomes a data-independent constant. Thus, it is not possible for the sensors to individually make local decisions: their data has to be gathered at the FC.

2.2. Detection Performance

Let the columns of $\boldsymbol{U}_{||} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{U}_\perp \in \mathbb{R}^{n \times (n-m)}$ constitute orthonormal bases for $\mathcal{R}(\boldsymbol{H})$ and $\mathcal{N}(\boldsymbol{H}^T)$, respectively. Then $\boldsymbol{P}_{||} = \boldsymbol{U}_{||} \boldsymbol{U}_{||}^T$ and $\boldsymbol{P}_\perp = \boldsymbol{U}_\perp \boldsymbol{U}_\perp^T$, so that T in (6) can be written as $T = \frac{\|\boldsymbol{U}_{||}^T \boldsymbol{y}\|^2/m}{\|\boldsymbol{U}_\perp^T \boldsymbol{y}\|^2/(n-m)}$. Since $\boldsymbol{U}_\perp^T \boldsymbol{H} = \boldsymbol{0}$, one has $\boldsymbol{U}_\perp^T \boldsymbol{y} = \boldsymbol{U}_\perp^T \boldsymbol{w} \triangleq \boldsymbol{w}_\perp$ under both $\mathcal{H}_0$ and $\mathcal{H}_1$, and

$$\begin{array}{rcl} \boldsymbol{U}_{||}^T \boldsymbol{y} & = & \boldsymbol{U}_{||}^T (\boldsymbol{H}\boldsymbol{x} + \boldsymbol{w}) \\ & = & \boldsymbol{x}_{||} + \boldsymbol{w}_{||}, \end{array} \quad \text{with} \quad \left\{ \begin{array}{l} \boldsymbol{x}_{||} \triangleq \boldsymbol{U}_{||}^T \boldsymbol{H}\boldsymbol{x}, \\ \boldsymbol{w}_{||} \triangleq \boldsymbol{U}_{||}^T \boldsymbol{w}. \end{array} \right. \tag{7}$$

Note that $\boldsymbol{w}_\perp, \boldsymbol{w}_{||}$ are zero-mean Gaussian with $E\{\boldsymbol{w}_\perp \boldsymbol{w}_\perp^T\} = \sigma^2 \boldsymbol{I}_{n-m}$ and $E\{\boldsymbol{w}_{||} \boldsymbol{w}_{||}^T\} = \sigma^2 \boldsymbol{I}_m$. In addition, $E\{\boldsymbol{w}_\perp \boldsymbol{w}_{||}^T\} = \boldsymbol{0}$, so that $\boldsymbol{w}_\perp, \boldsymbol{w}_{||}$ are statistically independent. Therefore,

$$\frac{1}{\sigma^2} \|\boldsymbol{U}_{||}^T \boldsymbol{y}\|^2 \sim \left\{ \begin{array}{ll} \chi^2_m & \text{under } \mathcal{H}_0, \\ \chi'^2_m(\rho) & \text{under } \mathcal{H}_1, \end{array} \right. \tag{8}$$

where we have introduced the Signal-to-Noise Ratio (SNR)

$$\rho \triangleq \frac{\|\boldsymbol{x}_{||}\|^2}{\sigma^2} = \frac{\|\boldsymbol{H}\boldsymbol{x}\|^2}{\sigma^2}, \tag{9}$$

since $\boldsymbol{H}^T \boldsymbol{H} = \boldsymbol{H}^T \boldsymbol{U}_{||} \boldsymbol{U}_{||}^T \boldsymbol{H}$. On the other hand,

$$\frac{1}{\sigma^2} \|\boldsymbol{U}_\perp^T \boldsymbol{y}\|^2 \sim \chi^2_{n-m} \quad \text{under both } \mathcal{H}_0 \text{ and } \mathcal{H}_1. \tag{10}$$

From (8) and (10), it follows that T in (6) is distributed as

$$T \,|\, \mathcal{H}_0 \sim \mathcal{F}_{m,n-m}, \qquad T \,|\, \mathcal{H}_1 \sim \mathcal{F}'_{m,n-m}(\rho). \tag{11}$$

Hence, the probabilities of false alarm and detection are respectively given by

$$P_{\text{FA}} = \Pr\{T > \gamma' \,|\, \mathcal{H}_0\} = 1 - F_{m,n-m}(\gamma'), \tag{12}$$

$$P_{\text{D}} = \Pr\{T > \gamma' \,|\, \mathcal{H}_1\} = 1 - F'_{m,n-m}(\gamma'; \rho). \tag{13}$$

3. THREAT MODEL

The adversary is interested in instigating some event (fire, chemical spill, etc.) while evading detection by the sensor network. He can modify the data from a subset of sensors $\mathcal{S}_{\text{A}}$, with $|\mathcal{S}_{\text{A}}| = k \ll n$. Instead of (2), the FC observes

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{a} \qquad \text{with} \quad \boldsymbol{a} \in \mathcal{A}, \tag{14}$$

with $\mathcal{A} = \{\boldsymbol{a} \in \mathbb{R}^n \,|\, a_i = 0 \text{ if } i \notin \mathcal{S}_{\text{A}}\}$ the set of feasible attack vectors. The adversary's goal is to decrease P_{D} as much as possible; to this aim, he may select the attack vector $\boldsymbol{a}$, and (possibly) the set of adversary sensors $\mathcal{S}_{\text{A}}$. It is assumed that the adversary has knowledge of the signal space $\mathcal{R}(\boldsymbol{H})$, but not of the parameter $\boldsymbol{x}$ (this is reasonable since, even though the event was triggered by the adversary, he is unlikely to be able to accurately estimate the corresponding field from the

few sensors in $\mathcal{S}_{\rm A}$ at his disposal). This will be the focus in the sequel.

With corrupted observations, the test statistic T becomes

$$T = c\,\frac{\|\boldsymbol{U}_{||}^T\tilde{\boldsymbol{y}}\|^2}{\|\boldsymbol{U}_{\perp}^T\tilde{\boldsymbol{y}}\|^2} = c\,\frac{\|\boldsymbol{x}_{||}+\boldsymbol{U}_{||}^T\boldsymbol{a}+\boldsymbol{w}_{||}\|^2}{\|\boldsymbol{U}_{\perp}^T\boldsymbol{a}+\boldsymbol{w}_{\perp}\|^2}, \qquad (15)$$

so that, under $\mathcal{H}_1$ and in the presence of an attack,

$$T \sim \mathcal{F}''_{m,n-m}(\rho_{||},\rho_{\perp}) \text{ with } \begin{cases} \rho_{||} & \triangleq & \frac{\|\boldsymbol{x}_{||}+\boldsymbol{U}_{||}^T\boldsymbol{a}\|^2}{\sigma^2}, \\ \rho_{\perp} & \triangleq & \frac{\|\boldsymbol{U}_{\perp}^T\boldsymbol{a}\|^2}{\sigma^2}. \end{cases} \qquad (16)$$

It is seen from (15)-(16) that the attack vector component $\boldsymbol{U}_{\perp}^T\boldsymbol{a}$ in $\mathcal{N}(\boldsymbol{H}^T)$ always contributes to reducing $P_{\rm D}$, which is monotonically decreasing in $\rho_{\perp}$ (and increasing in $\rho_{||}$). On the other hand, the component $\boldsymbol{U}_{||}^T\boldsymbol{a}$ in $\mathcal{R}(\boldsymbol{H})$ may or may not do so, depending on the value of $\boldsymbol{x}_{||} = \boldsymbol{U}_{||}^T\boldsymbol{H}\boldsymbol{x}$.

4. ATTACK DESIGN

Let $\boldsymbol{S}_{\rm A} \in \mathbb{R}^{n\times k}$ comprise the k columns of $\boldsymbol{I}_n$ with indices in $\mathcal{S}_{\rm A}$. Then any $\boldsymbol{a} \in \mathcal{A}$ can be written as $\boldsymbol{a} = \alpha\boldsymbol{S}_{\rm A}\boldsymbol{v}$, where $\boldsymbol{v} \in \mathbb{R}^k$ has unit norm ($\|\boldsymbol{v}\| = 1$) and $\alpha = \|\boldsymbol{a}\|$, so that

$$\rho_{||} = \frac{\|\boldsymbol{x}_{||}+\alpha\boldsymbol{U}_{||}^T\boldsymbol{S}_{\rm A}\boldsymbol{v}\|^2}{\sigma^2}, \quad \rho_{\perp} = \frac{\alpha^2}{\sigma^2}\|\boldsymbol{U}_{\perp}^T\boldsymbol{S}_{\rm A}\boldsymbol{v}\|^2. \quad (17)$$

Attack design involves the selection of α, $\boldsymbol{S}_{\rm A}$ and $\boldsymbol{v}$ in order to minimize $P_{\rm D}$ at the FC. Since $\boldsymbol{x}_{||} = \boldsymbol{U}_{||}^T\boldsymbol{H}\boldsymbol{x}$ is unknown to the adversary, a sensible attack design approach is to focus on maximizing the noise subspace contribution $\rho_{\perp}$. Nevertheless, the attack effect on the signal subspace and its impact on $P_{\rm D}$ (which is the ultimate goal) must be kept in sight.

4.1. Selection of $\boldsymbol{v}$

Suppose the set of adversary vectors $\mathcal{S}_{\rm A}$ (and therefore the matrix $\boldsymbol{S}_{\rm A}$) is given. Since $\boldsymbol{U}_{\perp}\boldsymbol{U}_{\perp}^T = \boldsymbol{P}_{\perp}$, it is clear that $\rho_{\perp}$ in (17) is maximized w.r.t. the spherical component $\boldsymbol{v}$ if

$$\boldsymbol{v}_{\star} = \text{principal eigenvector of } \boldsymbol{S}_{\rm A}^T\boldsymbol{P}_{\perp}\boldsymbol{S}_{\rm A}. \qquad (18)$$

Note that in order to compute (18) it suffices to have knowledge of $\mathcal{R}(\boldsymbol{S}_{\rm A}^T\boldsymbol{H})$, which could be estimated from measurements from the adversary sensors alone, as in [14].

With the choice (18), one has $\rho_{\perp} = \frac{\alpha^2}{\sigma^2}\lambda_{\star} = \eta\lambda_{\star}$, where $\lambda_{\star} \le 1$ is the largest eigenvalue of $\boldsymbol{S}_{\rm A}^T\boldsymbol{P}_{\perp}\boldsymbol{S}_{\rm A}$, and where

$$\eta \triangleq \frac{\alpha^2}{\sigma^2} \qquad (19)$$

is the *Distortion-to-Noise Ratio* (DNR). Observe now that

$$\begin{aligned} \|\boldsymbol{x}_{||}+\alpha\boldsymbol{U}_{||}^T\boldsymbol{S}_{\rm A}\boldsymbol{v}_{\star}\| &\le \|\boldsymbol{x}_{||}\| + \alpha\|\boldsymbol{U}_{||}^T\boldsymbol{S}_{\rm A}\boldsymbol{v}_{\star}\| \quad (20)\\ &= \|\boldsymbol{x}_{||}\| + \alpha\sqrt{1-\lambda_{\star}}, \quad (21)\end{aligned}$$

where the last step follows from the fact that $\|\boldsymbol{U}_{\perp}^T\boldsymbol{S}_{\rm A}\boldsymbol{v}_{\star}\|^2 + \|\boldsymbol{U}_{||}^T\boldsymbol{S}_{\rm A}\boldsymbol{v}_{\star}\|^2 = \|\boldsymbol{S}_{\rm A}\boldsymbol{v}_{\star}\|^2 = 1$. Since $F''_{m,m-n}$ is decreasing in its second argument, this results in the following upper bound for the probability of detection:

$$P_{\rm D} \le 1 - F''_{m,n-m}\left(\gamma'; \left(\sqrt{\rho}+\sqrt{\eta(1-\lambda_{\star})}\right)^2, \eta\lambda_{\star}\right) \quad (22)$$

Note that (22) is a *guaranteed* or *worst-case* performance metric for the adversary; equality holds in (20)-(22) iff $\boldsymbol{x} = \kappa\boldsymbol{H}^{\dagger}\boldsymbol{S}_{\rm A}\boldsymbol{v}_{\star}$ with $\kappa > 0$. On the other hand, since

$$\|\boldsymbol{x}_{||}+\alpha\boldsymbol{U}_{||}^T\boldsymbol{S}_{\rm A}\boldsymbol{v}_{\star}\| \ge \left|\|\boldsymbol{x}_{||}\| - \alpha\sqrt{1-\lambda_{\star}}\right|, \qquad (23)$$

the following lower bound is obtained:

$$P_{\rm D} \ge 1 - F''_{m,n-m}\left(\gamma'; \left(\sqrt{\rho}-\sqrt{\eta(1-\lambda_{\star})}\right)^2, \eta\lambda_{\star}\right) \quad (24)$$

Note that $1-\lambda_{\star}$ can be seen as the fraction of attack power leaking into the signal subspace. The bounds (22), (24) approach each other (and thus become tight) as $\lambda_{\star} \to 1$.

4.2. Selection of α

For $\lambda_{\star} = 1$ (no attack power leakage in the signal subspace), one has

$$P_{\rm D} = 1 - F''_{m,n-m}(\gamma'; \rho, \eta) \qquad (25)$$

which is monotonically decreasing in the DNR η. In fact, this property holds as long as $\lambda_{\star}$ is sufficiently close to 1 (although it need not hold otherwise), so the adversary should select α as large as possible. The following example illustrates this fact: consider a network with $n = 50$ nodes, $m = 6$, and the threshold set via (12) to achieve $P_{\rm FA} = 10^{-2}$. Fig. 1 depicts the bounds (22) and (24) *vs.* DNR for leakage values of $1-\lambda_{\star} = 0.01$ and 0.001. Clearly, both bounds decrease monotonically as the attack power is increased. Of course, with an attack-aware FC, a large value of α will make it likely for the adversary to be discovered, triggering in turn a data-cleansing stage which will considerably reduce the impact of the attack. Thus, the choice of α is dictated by a distortion-detectability tradeoff, which is the subject of ongoing work.

4.3. Selection of $\mathcal{S}_{\rm A}$

The considerations above have made clear that the adversary's achievable reduction in $P_{\rm D}$ is strongly influenced by the eigenvalue $\lambda_{\star}$. This fact can be observed in Fig. 1, and is further illustrated in Fig. 2, which shows the upper bound (22) on $P_{\rm D}$ as a function of $\lambda_{\star}$, assuming an SNR of 15 dB and an attack with DNR = 20 dB, for networks with $n = 50$ and 100 nodes. It can be seen that it is essential for the adversary to achieve a sufficiently small leakage $1-\lambda_{\star}$ in order for the attack to be effective. Note that $\lambda_{\star}$ depends not only on the number of sensors available to the adversary, but also on the specific set $\mathcal{S}_{\rm A}$ of such sensors. Nevertheless, the following result holds.

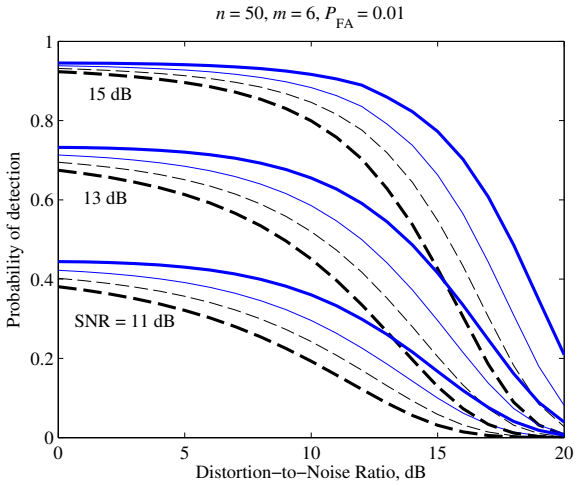

Fig. 1. Upper (solid) and lower bounds (dashed) on $P_{\rm D}$, for $n = 50$ sensors, parameter dimension $m = 6$, and $P_{\rm FA} = 0.01$: $1 - \lambda_\star = 0.01$ (thick lines), 0.001 (thin lines).

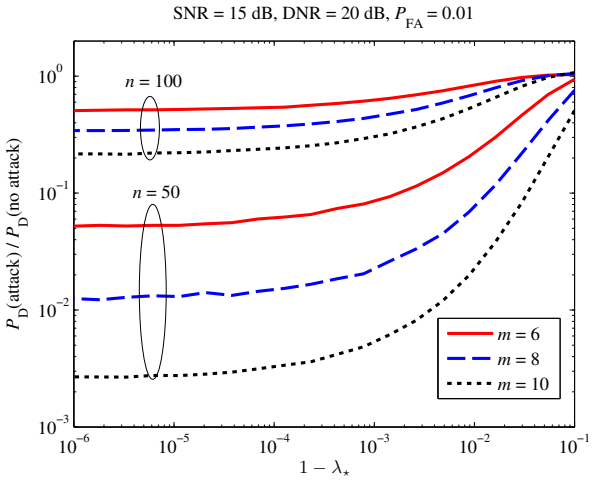

Fig. 2. Upper bound on $P_{\rm D}$ (relative to the $P_{\rm D}$ value in the absence of attack) for SNR $= 15$ dB and DNR $= 20$ dB. Threshold set for $P_{\rm FA} = 0.01$.

Theorem 1. *If the number of adversary sensors k is strictly larger than the dimension m of the unknown parameter (i.e., if $k > m$), then $\lambda_\star = 1$ regardless of the size n of the network, and of the particular set of adversary sensors $\mathcal{S}_{\rm A}$.*

Proof. Since $\lambda_\star = \lambda_1(\boldsymbol{S}_{\rm A}^T \boldsymbol{P}_\perp \boldsymbol{S}_{\rm A})$, and $\boldsymbol{S}_{\rm A} \in \mathbb{R}^{n\times k}$ has orthonormal columns, by Poincaré separation theorem [16],

$$\lambda_1(\boldsymbol{P}_\perp) \geq \lambda_\star \geq \lambda_{n-k+1}(\boldsymbol{P}_\perp). \tag{26}$$

Now $\boldsymbol{P}_\perp \in \mathbb{R}^{n\times n}$ is a projection matrix onto a subspace of dimension $n - m$, so that its eigenvalues are $\lambda_i(\boldsymbol{P}_\perp) = 1$ for $1 \leq i \leq n - m$ and $\lambda_i(\boldsymbol{P}_\perp) = 0$ otherwise. Hence, one has $\lambda_1(\boldsymbol{P}_\perp) = 1$ and, if $k > m$, then $\lambda_{n-k+1}(\boldsymbol{P}_\perp) = 1$ and the result follows from (26). □

Thus, the adversary can completely avoid leakage into the signal subspace by capturing a sufficiently large number of sensors $k > m$, in which case the particular set of adversary sensors is not relevant. This is not the case, however, if $k \leq m$, and the selection of the k sensors to capture becomes crucial. Finding $\boldsymbol{S}_{\rm A}$ in order to maximize $\lambda_\star = \lambda_1(\boldsymbol{S}_{\rm A}^T \boldsymbol{P}_\perp \boldsymbol{S}_{\rm A})$ is a combinatorial problem requiring the computation of the largest eigenvalue of a total of $\binom{n}{k}$ matrices of size $k \times k$. This quickly becomes infeasible as the network size n increases.

The problem can be rephrased as follows, in order to get a better understanding. Since the nonzero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$ are the same as those of $\boldsymbol{A}^T\boldsymbol{A}$, one has

$$\lambda_1(\boldsymbol{S}_{\rm A}^T \boldsymbol{P}_\perp \boldsymbol{S}_{\rm A}) = \lambda_1(\boldsymbol{S}_{\rm A}^T \boldsymbol{U}_\perp \boldsymbol{U}_\perp^T \boldsymbol{S}_{\rm A}) = \lambda_1(\boldsymbol{U}_\perp^T \boldsymbol{D}_{\rm A} \boldsymbol{U}_\perp),$$

where $\boldsymbol{D}_{\rm A} = \boldsymbol{S}_{\rm A}\boldsymbol{S}_{\rm A}^T$ is a diagonal matrix with k diagonal elements equal to 1 at the positions of the adversary sensor indices, and zeros elsewhere. Writing $\boldsymbol{U}_\perp$ row-wise as $\boldsymbol{U}_\perp^T = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_n \end{bmatrix}$, and with $d_i \triangleq (\boldsymbol{D}_{\rm A})_{ii}$, one has

$$\boldsymbol{U}_\perp^T \boldsymbol{D}_{\rm A} \boldsymbol{U}_\perp = \sum_{i=1}^n d_i \boldsymbol{u}_i \boldsymbol{u}_i^T. \tag{27}$$

Letting $\boldsymbol{d} = [\, d_1 \; d_2 \; \cdots \; d_n \,]^T$, the problem can be written as

$$\max_{\boldsymbol{d}} \lambda_1 \left(\sum_{i=1}^n d_i \boldsymbol{u}_i \boldsymbol{u}_i^T \right) \quad \text{s. to} \quad \begin{cases} \boldsymbol{d} \in \{0,1\}^n, \\ \sum_{i=1}^n d_i = k. \end{cases} \tag{28}$$

A natural way to deal with sensor selection problems similar in structure to (28) is to replace the nonconvex constraint $\boldsymbol{d} \in \{0,1\}^n$ with the convex one $\boldsymbol{d} \in [0,1]^n$ [17, 18]. However, in this case even such relaxation remains nonconvex, as it involves the *maximization* of a *convex* objective[1], in contrast with [17, 18]. Deriving efficient means to approximately solve (28) is the object of ongoing work.

5. CONCLUSIONS

Spatial field detection by sensor networks can be severely degraded by injection of false data in a few sensors. We have characterized the power of the attack analytically, providing strategies for attack design. We showed that if the number of captured sensors is sufficiently large, selecting which sensors to capture is not important to the adversary. Our results can be used in network design in order to evaluate the impact of, and make the system more robust to this class of attacks.

In practice it is of interest to provide defense mechanisms against these data-injection attacks. Designing suitable attack detectors and analyzing distortion-detectability tradeoffs is the focus of future work.

[1] If $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n\}$ is any set of symmetric matrices, the function $f(\boldsymbol{x}) = \lambda_1(x_1\boldsymbol{A}_1 + \cdots + x_n\boldsymbol{A}_n)$ is convex in $\boldsymbol{x} = [\, x_1 \; \cdots \; x_n \,]^T$.

6. REFERENCES

[1] P. Corke, T. Wark, R. Jurdak, Wen Hu, P. Valencia, and D. Moore, "Environmental wireless sensor networks," *Proc. IEEE*, vol. 98, no. 11, pp. 1903–1916, Nov. 2010.

[2] J.-F. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 16–25, May 2007.

[3] Xiangqian Chen, K. Makki, Kang Yen, and N. Pissinou, "Sensor network security: a survey," *IEEE Commun. Surveys & Tutorials*, vol. 11, no. 2, pp. 52–73, 2nd quarter 2009.

[4] M. C. Vuran, Ö. B. Akan, and I. F. Akyildiz, "Spatio-temporal correlation: theory and applications for wireless sensor networks," *Computer Networks*, vol. 45, pp. 245–259, 2004.

[5] C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, and S. Madden, "Distributed regression: an efficient framework for modeling sensor network data," in *Int. Symp. Info. Process. Sensor Networks (IPSN)*, 2004, pp. 2394–2398.

[6] A. Ribeiro and G. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks–Part II: Unknown probability density function," *IEEE Trans. Signal Process.*, vol. 54, no. 7, pp. 2784–2796, Jul. 2006.

[7] A. Dogandžić and K. Qiu, "Decentralized random-field estimation for sensor networks using quantized spatially correlated data and fusion-center feedback," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 6069–6085, Dec. 2008.

[8] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of Byzantine attacks," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 16–29, Jan. 2009.

[9] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of Byzantine attacks in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 59, no. 2, Feb. 2011.

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

[11] K. G. Vamvoudakis, J. P. Hespanha, B. Sinopoli, and Y. Mo, "Detection in adversarial environments," *IEEE Trans. Autom. Control*, vol. 59, no. 12, pp. 3209–3223, Dec. 2014.

[12] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Asymptotic analysis of distributed Bayesian detection with Byzantine data," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 608–612, May. 2015.

[13] S. Cui, Z. Han, S. Kar, T. T. Kim, H. V. Poor, and A Tajer, "Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 106–115, Sep. 2012.

[14] J. Kim, L. Tong, and R. J. Thomas, "Subspace methods for data attack on state estimation: a data driven approach," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1102–1114, Mar. 2015.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing, vol. 2: Detection Theory*, Prentice-Hall, 1998.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 2nd edition, 2013.

[17] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.

[18] S. P. Chepuri and G. Leus, "Sparsity-promoting sensor selection for non-linear measurement models," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 684–698, Feb. 2015.