Automatica 89 (2018) 117–124

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Worst-case stealthy innovation-based linear attack on remote state estimation☆

Ziyang Guo [a], Dawei Shi [b,*], Karl Henrik Johansson [c], Ling Shi [a]

[a] Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong
[b] Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[c] ACCESS Linnaeus Center, School of Electrical Engineering, Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 17 August 2016
Received in revised form 5 July 2017
Accepted 18 October 2017
Available online 23 December 2017

Keywords:
Integrity attack
Kullback–Leibler divergence
Remote state estimation
Cyber–Physical system security

ABSTRACT

In this work, a security problem in cyber–physical systems is studied. We consider a remote state estimation scenario where a sensor transmits its measurement to a remote estimator through a wireless communication network. The Kullback–Leibler divergence is adopted as a stealthiness metric to detect system anomalies. We propose an innovation-based linear attack strategy and derive the remote estimation error covariance recursion in the presence of attack, based on which a two-stage optimization problem is formulated to investigate the worst-case attack policy. It is proved that the worst-case attack policy is zero-mean Gaussian distributed and the numerical solution is obtained by semi-definite programming. Moreover, an explicit algorithm is provided to calculate the compromised measurement. The trade-off between attack stealthiness and system performance degradation is evaluated via simulation examples.

1. Introduction

Cyber–physical systems (CPS) are the next generation of engineered systems that tightly integrate computation, communication, control and physical processes (Kim & Kumar, 2012). Due to the interconnection of different technologies and components, CPS are vulnerable to adversarial intrusion which may cause severe consequences on national economy, social security or even loss of human lives (Poovendran et al., 2012). Recently reported cyber attacks, e.g., the StuxNet malware (Karnouskos, 2011), the Maroochy water bleach (Slay & Miller, 2007), evidently indicate that security is of fundamental importance to ensure safe operation of CPS.

With the increasing adoption of CPS, attack strategy and defense mechanism design have received considerable attention during past decade. Based on the available resources, malicious agents aim at crippling the system functionality and simultaneously remaining undetected (Teixeira, Sou, Sandberg, & Johansson, 2015). Denial-of-service (DoS) attacks attempt to block the communication channel and prevent the legitimate access to system components. Since jamming is a power-intensive activity and the available energy of an jammer might be limited, continuous action is often impossible. Thus, jamming attack models for resource-constrained attackers were studied (Gupta, Langbort, and Basar, 2010; Zhang, Cheng, Shi, and Chen 2015). Moreover, the optimal transmission scheduling against remote state estimation were analyzed under game-theoretic framework in Li, Quevedo, Dey, and Shi (2016) and Li, Shi, Cheng, Chen, and Quevedo (2015). Replay attacks attempt to inject fake control signal and simultaneously replay the past sensory data to keep stealthiness. The feasible conditions and countermeasures of such an attack were studied for linear Gaussian systems in Mo, Chabukswar, and Sinopoli (2014) and Mo and Sinopoli (2009). The trade-off between control performance and system security was investigated under a stochastic game framework in Miao, Pajic, and Pappas (2013). False data injection attacks were studied for electric power grids against remote state estimation in Liu, Ning, and Reiter (2011). The reachable consequence of such an attack on estimation performance was investigated in Mo, Garone, Casavola, and Sinopoli (2010). The explicit trade-off between attack stealthiness and system performance degradation was analyzed for control signal injection attack in Bai, Pasqualetti, and Gupta (2015, 2017). Further results and developments of integrity attack and secure state estimation problems were studied in Shi, Chen, and Darouach (2016) and Shi, Elliott, and Chen (2017).

The results on detectability of attacks discussed above have not been presented in a unified framework and there seems not to

☆ The work by Z. Guo and L. Shi is supported by an HKUST KTH Partnership FP804. The work by D. Shi was supported by Natural Science Foundation of China (61503027). The work by K. H. Johansson is supported by the Knut and Alice Wallenberg Foundation and the Swedish Research Council. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Alessandro Chiuso under the direction of Editor Torsten Söderström.

* Corresponding author.
E-mail addresses: zguoae@ust.hk (Z. Guo), dawei.shi@outlook.com (D. Shi), kallej@kth.se (K.H. Johansson), eesling@ust.hk (L. Shi).

https://doi.org/10.1016/j.automatica.2017.11.018
0005-1098/

exist an agreement on the measure of attack stealthiness. Existing works investigated the detectability of a certain attack with respect to a specific detection scheme, such as the classic χ^2 false-data detector (Liu et al., 2011; Miao et al., 2013; Mo et al., 2014, 2010; Mo & Sinopoli, 2009). Similarly, we considered such a detector in our previous work (Guo, Shi, Johansson, and Shi, 2017) and obtained a closed-form expression of the worst-case innovation-based linear attack. However, Guo et al. (2017) did not consider the situations when the attack strategies are not Gaussian distributed or not strictly stealthy to the false-data detector. Some recent works introduced the Kullback–Leibler (K–L) divergence as a measure of attack stealthiness (Bai & Gupta, 2014; Bai et al., 2015), which is independent of any specific detection scheme. Motivated by above observations, we adopt K–L divergence as a stealthiness measure and consider arbitrarily distributed innovation-based linear attacks for general dynamic systems in this work.

The contributions of this paper are threefold. First, we extend the innovation-based linear attack to an arbitrary distribution and use the K–L divergence between the modified and the legitimate dynamics of the measurement innovation as a measure of attack stealthiness, which results in more general models for both the malicious attacker and the false-data detector. Second, we prove that the worst-case linear attack strategy which maximizes the estimation error covariance is zero-mean Gaussian distributed and provide a closed-form expression of the resulting covariance matrix. Finally, we provide an algorithm to numerically calculate the worst possible action of the attacker.

A related work Bai and Gupta (2014) considered arbitrary measurement attacks for first-order systems under ϵ-weakly marginal stealthiness metric. Its advanced version Bai, Gupta, and Pasqualetti (2017) considered the same attack scenario while adopting ϵ-stealthiness metric. The optimal ϵ-stealthy attack policy was obtained to achieve the upper bound of the estimation error covariance. Different from above works focusing on arbitrary attacks but for first-order systems, our work considers higher-order dynamic systems but with linear attack strategies. Moreover, the stealthiness metrics used in Bai and Gupta (2014) and Bai, Gupta et al. (2017) and our work are slightly different. Simulation examples are also provided for result comparison.

The remainder of the paper is organized as follows. Section 2 introduces the system architecture. Section 3 presents the innovation-based linear attack strategy and its stealthiness constraint. Section 4 derives the worst-case attack policy. Numerical examples are provided in Section 5. Some concluding remarks are given in the end.

Notations: $\mathbb{N}$ and $\mathbb{R}$ denote the sets of natural and real numbers, respectively. $\mathbb{R}^n$ is the n-dimensional Euclidean space. For a matrix X, X', $\mathrm{Tr}(X)$ and $|X|$ stand for the transpose, trace and determinant of X. $X > 0$ ($X \geq 0$) means X is a positive definite (semi-definite) matrix. For $x \in \mathbb{R}$ and $x > 0$, $\log x$ denotes the natural logarithm of x.

2. System architecture

The system architecture is shown in Fig. 1. A sensor measures a physical process and transmits the data to a remote estimator through a wireless communication network. The attacker attempts to intercept and modify measurement data, which may degrade the estimation performance without triggering an alarm. The detailed model of each component is introduced as follows.

2.1. Process model

We consider a discrete-time linear time-invariant (LTI) process described by

$$x_{k+1} = Ax_k + w_k, \tag{1}$$
$$y_k = Cx_k + v_k, \tag{2}$$

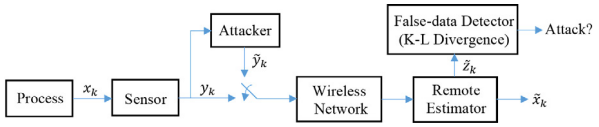


Fig. 1. System block diagram.

where $k \in \mathbb{N}$ is the time index, $x_k \in \mathbb{R}^n$ is the process state, $y_k \in \mathbb{R}^m$ is the sensor measurement, $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are zero-mean i.i.d. Gaussian noises with covariances $Q \geq 0$ and $R > 0$, respectively. The initial state x_0 is zero-mean Gaussian with covariance matrix $\Pi_0 \geq 0$ and independent of w_k and v_k for all $k \geq 0$. The pair (A, C) is detectable and $(A, \sqrt{Q})$ is stabilizable.

2.2. Remote estimator

At each time step, the sensor sends its measurement to a remote estimator through a wireless network. To estimate the system state, a Kalman filter is adopted by the remote estimator to process the received data:

$$\hat{x}_k^- = A\hat{x}_{k-1}, \tag{3}$$
$$P_k^- = AP_{k-1}A' + Q, \tag{4}$$
$$K_k = P_k^- C'(CP_k^- C' + R)^{-1}, \tag{5}$$
$$\hat{x}_k = \hat{x}_k^- + K_k(y_k - C\hat{x}_k^-), \tag{6}$$
$$P_k = (I - K_k C)P_k^-, \tag{7}$$

where $\hat{x}_k^-$ and $\hat{x}_k$ are the *a priori* and the *a posteriori* minimum mean squared error (MMSE) estimates of the state x_k at the remote estimator, P_k^- and P_k the corresponding error covariances. The recursion starts from $\hat{x}_0^- = 0$ and $P_0^- = \Pi_0 \geq 0$.

It is well known that the Kalman filter converges exponentially fast from any initial condition (Anderson & Moore, 2012). We define the steady-state error covariance

$$\overline{P} \triangleq \lim_{k \to +\infty} P_k^-, \tag{8}$$

where $\overline{P}$ is the unique positive semi-definite solution of $X = AXA' + Q - AXC'(CXC' + R)^{-1}CXA'$. Without loss of generality, we assume that the system starts from steady state with $P_0^- = \overline{P}$, which results in a fixed-gain Kalman filter, i.e.,

$$K \triangleq \overline{P}C'(C\overline{P}C' + R)^{-1}. \tag{9}$$

2.3. False-data detector

Note that the stochastic nature of the system provides the degree of freedom for the potential malicious attacker since the process noise and the measurement noise induce some uncertainty in system variables. Hence, a false-data detector is necessary at the remote side to monitor system behavior and detect the existence of cyber attacks. The K–L divergence (Cover & Thomas, 2012), a non-negative measure of the distance between two probability distributions, is well known in detection theory (Poor, 2013) and defined as follows.

Definition 1 (*Kullback–Leibler Divergence*)**.** Let x_k and y_k be two random sequences with joint probability density functions f_{x_k} and f_{y_k}, respectively. The Kullback–Leibler divergence between x_k and y_k is defined as

$$D(x_k \| y_k) = \int_{\{\xi_k | f_{x_k}(\xi_k) > 0\}} \log \frac{f_{x_k}(\xi_k)}{f_{y_k}(\xi_k)} f_{x_k}(\xi_k) d\xi_k.$$

It can be observed that $D(x_k\|y_k) = 0$ if and only if $f_{x_k} = f_{y_k}$. Also, the K–L divergence is generally not symmetric, i.e., $D(x_k\|y_k) \neq D(y_k\|x_k)$.

According to Anderson and Moore (2012), the innovation sequence $z_k = y_k - C\hat{x}_k^-$ has a steady-state Gaussian distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma = C\bar{P}C' + R$ and $\mathbb{E}[z_i z_j'] = 0$ for all $i \neq j$. Hence, the K–L divergence between the attacker induced innovation and the nominal innovation is adopted as a measure of attack stealthiness. When the K–L divergence exceeds a certain threshold, an alarm will be trigger, which indicates the existence of attacks.

3. Innovation-based linear attack strategy

In this section, we propose an innovation-based linear attack strategy and define the feasible attack space, based on which the problem we concerned is introduced.

3.1. Attack model

Similar to the attack models used in man-in-the-middle attacks (Callegati, Cerroni, & Ramilli, 2009; Meyer & Wetzel, 2004), we assume that the attacker has full knowledge of the process model and is capable of intercepting and modifying the transmitted measurement. The goal of the attacker is to degrade the system performance in the sense of maximizing the estimation error covariance, and simultaneously remaining stealthy to the false-data detector. It is worth noticing that the attacker can work equivalently with the measurement and the innovation under above assumptions. Specifically, based on the system knowledge, the attacker is able to implement a filter to first calculate the innovation z_k according to $z_k = y_k - C\hat{x}_k^-$, then generate the compromised innovation $\tilde{z}_k$, and finally go back to the manipulated measurement $\tilde{y}_k$ according to $\tilde{y}_k = \tilde{z}_k + C\tilde{x}_k^-$, where $\tilde{x}_k^-$ is the *a priori* MMSE estimate of x_k when the system is under attack. This procedure $y_k \to z_k \to \tilde{z}_k \to \tilde{y}_k$ means that generating attack signal $\tilde{y}_k$ is equivalent to generating $\tilde{z}_k$. Thus, we design the attack strategy with respect to the innovation sequence z_k in the subsequent discussion.

At each time k, a general attack strategy is defined as

$$\tilde{z}_k = f_k(z_k), \tag{10}$$

where $z_k \in \mathbb{R}^m$ is the currently intercepted innovation, $\tilde{z}_k \in \mathbb{R}^m$ the innovation modified by the attacker, and $f_k : \mathbb{R}^m \to \mathbb{R}^m$ an arbitrary function. However, for a nonlinear function f_k, it is difficult to analytically determine the statistical characteristics of $\tilde{z}_k$, not to mention the analysis of the corresponding system performance and the attack effect. Hence, we focus in this initial study on the subset of all linear attack strategies where f_k is an affine function of the innovation z_k:

$$\tilde{z}_k = T_k z_k + b_k, \tag{11}$$

where $T_k \in \mathbb{R}^{m\times m}$ is an arbitrary attack matrix, and $b_k \in \mathbb{R}^m$ is an arbitrary i.i.d. random variable independent of z_k with zero mean and covariance $\mathbb{E}[b_k b_k'] = \Gamma_k$. In this case, $\tilde{z}_k$ can follow an arbitrary distribution with covariance $\tilde{\Sigma}_k = T_k \Sigma T_k' + \Gamma_k$.

To avoid being detected by the false-data detector, the attacker needs to carefully design the attack signal $\tilde{y}_k$ at each time instant such that the K–L divergence between the compromised innovation and the steady-state innovation does not exceed the threshold, i.e.,

$$D(\tilde{z}_k\|z_k) \leq \delta, \tag{12}$$

where $\delta \geq 0$ is the threshold. Consequently, the feasible set of attacks is defined as follows.

Definition 2 (*Attack Space*)**.** For any given threshold δ, the feasible set of attacks, which contains all possible linear attack strategies (T_k, b_k) such that $\tilde{z}_k = T_k z_k + b_k$ satisfies $D(\tilde{z}_k\|z_k) \leq \delta$, is denoted as $\Phi_\delta \subset \mathbb{R}^{m\times m} \times \mathbb{R}^m$.

Let $\tilde{x}_k^-$ and $\tilde{x}_k$ be the *a priori* and the *a posteriori* MMSE estimates at the remote estimator in the presence of the attack, which can be obtained from the recursion

$$\tilde{x}_k^- = A\tilde{x}_{k-1}, \tag{13}$$

$$\tilde{x}_k = \tilde{x}_k^- + K\tilde{z}_k. \tag{14}$$

Since the remote estimator is unaware of the attack, the state estimate produced by (13)–(14) will deviate from the true system state. To quantify the system performance, we define $\tilde{P}_k^-$ and $\tilde{P}_k$ as the *a priori* and the *a posteriori* MMSE estimation error covariance matrices at the remote estimator under attack in the subsequent discussions.

3.2. Problem of interest

For the considered system (1)–(2) under the linear attack (11) with detection criterion (12), the attacker aims at maximizing the remote estimation error covariance without exceeding the upper bound of the K–L divergence at each time instant, i.e.,

$$\mathbf{P_1}: \max_{(T_k, b_k)} \ \mathrm{Tr}(\tilde{P}_k) \quad \text{s.t.} \ \ D(\tilde{z}_k\|z_k) \leq \delta, \quad \forall k.$$

The problem $\mathbf{P_1}$ we are thus interested in is to find the worst-case innovation-based linear attack strategy (T_k, b_k) in Φ_δ which leads to the largest degradation of the remote estimation performance. The detailed mathematical formulation and solution to this problem are introduced in the following sections.

4. Worst-case attack strategy

In this section, we first derive the iteration of the remote estimation error covariance in the presence of the proposed attack, based on which a two-stage optimization problem is formulated. Then, the worst-case linear attack $\tilde{z}_k$ is proved to be Gaussian distributed and the corresponding T_k is obtained by semi-definite programming. Finally, we provide an explicit algorithm to characterize the worst possible action of the attacker on the system measurement.

4.1. Iteration of estimation error covariance

Before solving the optimization problem $\mathbf{P_1}$, we first investigate the estimation error covariance recursion under the proposed attack strategy, which is summarized in the following lemma.

Lemma 1. *For the system (1)–(2) under the linear attack (11), the estimation error covariance at the remote estimator follows the recursion*

$$\tilde{P}_k = A\tilde{P}_{k-1}A' + Q + K\tilde{\Sigma}_k K' - \bar{P}C'T_k'K' - KT_kC\bar{P}, \tag{15}$$

where $\tilde{\Sigma}_k = \mathbb{E}[\tilde{z}_k\tilde{z}_k']$ is the covariance of $\tilde{z}_k$, $\bar{P}$ and K are given in (8) and (9).

Proof. According to the dynamics of the state estimate under linear attack (13)–(14), one has

$$x_k - \tilde{x}_k = A(x_{k-1} - \tilde{x}_{k-1}) + w_{k-1} - K\tilde{z}_k,$$

based on which the error covariance at the remote estimator can be represented as

$$\tilde{P}_k = A\tilde{P}_{k-1}A' + Q + K\tilde{\Sigma}_k K' - \mathbb{E}[(x_k - \tilde{x}_k^-)\tilde{z}_k' K'] - \mathbb{E}[K\tilde{z}_k(x_k - \tilde{x}_k^-)']. \tag{16}$$

To calculate the last two terms of (16), we first evaluate

$$x_k - \tilde{x}_k^- = A^k(x_0 - \hat{x}_0^-) + \sum_{i=0}^{k-1} A^i w_{k-1-i} - \sum_{i=0}^{k-1} A^{i+1} K\tilde{z}_{k-1-i},$$

where the equality follows from the assumption $\hat{x}_0^- = \tilde{x}_0^-$. Note that

$$\tilde{z}_k = T_k C[A(I-KC)]^k(x_0 - \hat{x}_0^-) + \sum_{i=0}^{k-1} T_k C[A(I-KC)]^i w_{k-1-i} + V,$$

where $V = T_k v_k + b_k - \sum_{i=0}^{k-1} T_k C[A(I-KC)]^i AK v_{k-1-i}$ is independent of $x_k - \tilde{x}_k^-$. Due to the fact that $\mathbb{E}[\tilde{z}_i\tilde{z}_j^T] = 0$ for any $i \neq j$, and $\overline{P}$ is the unique positive semi-definite fixed point of $h \circ \tilde{g}$, i.e., $(h \circ \tilde{g})^n(\overline{P}) = \overline{P}$, the second last term of (16) can be further evaluated as

$$\begin{aligned}
&\mathbb{E}[(x_k - \tilde{x}_k^-)\tilde{z}_k' K'] \\
&= \mathbb{E}\left[\left\{A^k(x_0 - \hat{x}_0^-) + \sum_{i=0}^{k-1} A^i w_{k-1-i}\right\}\left\{T_k C[A(I-KC)]^k \times (x_0 - \hat{x}_0^-) + \sum_{i=0}^{k-1} T_k C[A(I-KC)]^i w_{k-1-i}\right\}' K'\right] \\
&= \{A^k\overline{P}[(I-KC)'A']^k + \sum_{i=0}^{k-1} A^i Q[(I-KC)'A']^i\} C' T_k' K' \\
&= \overline{P}C' T_k' K'.
\end{aligned}$$

Similarly, we obtain

$$\mathbb{E}[K\tilde{z}_k(x_k - \tilde{x}_k^-)'] = KT_k C\overline{P},$$

which completes the proof. ■

4.2. Worst-case distribution of $\tilde{z}_k$

Based on the estimation error covariance iteration obtained in Lemma 1, we now focus on deriving the solution to the optimization problem $\mathbf{P_1}$. As a first and important step, we analyze the statistical characteristics of $\tilde{z}_k$ under which the error covariance at the remote estimator is maximized.

Note that maximizing the trace of the estimation error covariance at each time step is equivalent to maximizing the trace of the last three terms in (15). Thus, the goal of the malicious attacker can be stated as

$$\mathbf{P_2}: \max_{(T_k, b_k)} \ \mathrm{Tr}(K\tilde{\Sigma}_k K' - \overline{P}C'T_k'K' - KT_kC\overline{P}) \quad \text{s.t.} \ D(\tilde{z}_k\|z_k) \leq \delta, \ \forall k.$$

The worst-case distribution of the modified innovation $\tilde{z}_k$ is obtained in the following theorem, in which the associated covariance matrix is also provided.

Theorem 1. *Let $\tilde{z}_k^*$ be the optimal solution of problem $\mathbf{P_2}$ for any given T_k, then $\tilde{z}_k^*$ is zero-mean Gaussian distributed with covariance matrix*

$$\tilde{\Sigma}_k = \left(\Sigma^{-1} - \frac{2}{\mu}K'K\right)^{-1}, \tag{17}$$

where μ is the unique scalar such that $\mu > 2\min_{1\leq i\leq m}\lambda_i$ and

$$\sum_{i=1}^{m}\left[\frac{1}{1-\frac{2}{\mu}\lambda_i} + \log\left(1 - \frac{2}{\mu}\lambda_i\right)\right] = 2\delta + m, \tag{18}$$

with δ being the threshold of K–L divergence, m being the dimension of $\tilde{z}_k^*$, and $\lambda_1, \lambda_2, \ldots, \lambda_m$ being the eigenvalues of $K'K\Sigma$.

Proof. For any given T_k in the attack space Φ_δ, the worst-case distribution of $\tilde{z}_k$ obtained by solving problem $\mathbf{P_2}$ is equivalent to solving problem

$$\mathbf{P_3}: \max_{b_k} \ \mathrm{Tr}(K\tilde{\Sigma}_k K') \quad \text{s.t.} \ D(\tilde{z}_k\|z_k) \leq \delta.$$

Let $f_{\tilde{z}_k}(x)$ and $f_{z_k}(x)$ be the probability density functions of $\tilde{z}_k$ and z_k, respectively. Then, one has

$$\begin{aligned}
&D(\tilde{z}_k\|z_k) \\
&= \int f_{\tilde{z}_k}(x)\log\frac{f_{\tilde{z}_k}(x)}{f_{z_k}(x)}dx \\
&= -h(\tilde{z}_k) - \int f_{\tilde{z}_k}(x)\log[\frac{1}{\sqrt{(2\pi)^m|\Sigma|}}\exp(-\frac{1}{2}x'\Sigma^{-1}x)]dx \\
&= -h(\tilde{z}_k) + \frac{1}{2}\log\left[(2\pi)^m|\Sigma|\right] + \frac{1}{2}\mathbb{E}[\tilde{z}_k'\Sigma^{-1}\tilde{z}_k],
\end{aligned} \tag{19}$$

where $h(\tilde{z}_k) = -\int f_{\tilde{z}_k}(x)\log f_{\tilde{z}_k}(x)dx$ is the differential entropy of $\tilde{z}_k$.

Consider a Gaussian random variable $\eta \sim \mathcal{N}(0, \tilde{\Sigma}_k)$, where $\tilde{\Sigma}_k = \mathbb{E}[\tilde{z}_k\tilde{z}_k'] = T_k\Sigma T_k' + \Gamma_k$, such that

$$\begin{aligned}
&\mathbb{E}[\eta'\eta] = \mathrm{Tr}(\tilde{\Sigma}_k) = \mathbb{E}[\tilde{z}_k'\tilde{z}_k], \\
&\mathbb{E}[\eta'\Sigma^{-1}\eta] = \mathrm{Tr}(\Sigma^{-1}\tilde{\Sigma}_k) = \mathbb{E}[\tilde{z}_k'\Sigma^{-1}\tilde{z}_k].
\end{aligned} \tag{20}$$

Since the normal distribution has the maximum entropy among all real-valued distribution with a specified variance (Cover & Thomas, 2012), we obtain

$$h(\eta) \geq h(\tilde{z}_k), \tag{21}$$

with the equality holds if and only if $\tilde{z}_k$ is also a zero-mean Gaussian random variable.

According to (19), (20) and (21), it is easy to obtain

$$\begin{aligned}
D(\eta\|z_k) &= -h(\eta) + \frac{1}{2}\log\left[(2\pi)^m|\Sigma|\right] + \frac{1}{2}\mathbb{E}[\eta'\Sigma^{-1}\eta] \\
&\leq -h(\tilde{z}_k) + \frac{1}{2}\log\left[(2\pi)^m|\Sigma|\right] + \frac{1}{2}\mathbb{E}[\tilde{z}_k'\Sigma^{-1}\tilde{z}_k] \\
&\leq D(\tilde{z}_k\|z_k).
\end{aligned}$$

Hence, for any given T_k in the attack space, the remote estimation error covariance is maximized when $\tilde{z}_k$ is zero-mean Gaussian distributed, i.e., b_k is zero-mean Gaussian.

It now suffices to find the corresponding covariance matrix of b_k, which is equivalent to find the worst-case $\tilde{\Sigma}_k$. Since $z_k \sim \mathcal{N}(0, \Sigma)$ and $\tilde{z}_k \sim \mathcal{N}(0, \tilde{\Sigma}_k)$, the K–L divergence is given as

$$D(\tilde{z}_k\|z_k) = \frac{1}{2}\mathrm{Tr}(\Sigma^{-1}\tilde{\Sigma}_k) - \frac{m}{2} + \frac{1}{2}\log\frac{|\Sigma|}{|\tilde{\Sigma}_k|}.$$

We now consider the following problem:

$$\mathbf{P_4}: \max_{\tilde{\Sigma}_k} \ \mathrm{Tr}(K\tilde{\Sigma}_k K') \quad \text{s.t.} \ \frac{1}{2}\mathrm{Tr}(\Sigma^{-1}\tilde{\Sigma}_k) - \frac{m}{2} + \frac{1}{2}\log\frac{|\Sigma|}{|\tilde{\Sigma}_k|} \leq \delta.$$

It can be observed that the objective function is affine in $\tilde{\Sigma}_k$. Further, the feasible domain is convex in $\tilde{\Sigma}_k$ since determinant is

a log-concave function. Hence, we rewrite problem $\mathbf{P_4}$ in the form of a standard convex optimization problem

$$\mathbf{P_5}: \quad \min_{\tilde{\Sigma}_k} \quad -\operatorname{Tr}(K'K\tilde{\Sigma}_k)$$
$$\text{s.t.} \quad \frac{1}{2}\operatorname{Tr}(\Sigma^{-1}\tilde{\Sigma}_k) - \frac{m}{2} + \frac{1}{2}\log\frac{|\Sigma|}{|\tilde{\Sigma}_k|} - \delta \le 0 \tag{22}$$

and define the Lagrangian

$$\mathcal{L}_p(\tilde{\Sigma}_k, \mu) = -\operatorname{Tr}(K'K\tilde{\Sigma}_k) + \mu\left[\frac{1}{2}\operatorname{Tr}(\Sigma^{-1}\tilde{\Sigma}_k) - \frac{m}{2} + \frac{1}{2}\log\frac{|\Sigma|}{|\tilde{\Sigma}_k|} - \delta\right],$$

where $\mu \ge 0$ is the Lagrangian multiplier.

Let the derivative of $\mathcal{L}_p(\tilde{\Sigma}_k, \mu)$ with respect to $\tilde{\Sigma}_k$ be equal to zero. Due to the fact that $\frac{\partial \operatorname{Tr}(AX')}{\partial X} = A$ and $\frac{\partial \log|X|}{\partial X} = (X^{-1})' = (X')^{-1}$, we obtain

$$\frac{\partial \mathcal{L}_p(\tilde{\Sigma}_k, \mu)}{\partial \tilde{\Sigma}_k} = -K'K + \frac{\mu}{2}\Sigma^{-1} - \frac{\mu}{2}\tilde{\Sigma}_k^{-1} = 0, \tag{23}$$

which is equivalent to

$$\tilde{\Sigma}_k = \left(\Sigma^{-1} - \frac{2}{\mu}K'K\right)^{-1}.$$

According to the Karush Kuhn Tucker (KKT) conditions, besides (22), (23) and $\mu \ge 0$, we also have

$$\mu\left(\frac{1}{2}\operatorname{Tr}(\Sigma^{-1}\tilde{\Sigma}_k) - \frac{m}{2} + \frac{1}{2}\log\frac{|\Sigma|}{|\tilde{\Sigma}_k|} - \delta\right) = 0. \tag{24}$$

It is also worth noticing that $K'K = \frac{\mu}{2}(\Sigma^{-1} - \tilde{\Sigma}_k^{-1})$ cannot always be zero, i.e., the Lagrangian multiplier $\mu > 0$, from which we obtain that

$$2\delta + m = \operatorname{Tr}(\Sigma^{-1}\tilde{\Sigma}_k) + \log\frac{|\Sigma|}{|\tilde{\Sigma}_k|} = \sum_{i=1}^{m}\left[\frac{1}{1-\frac{2}{\mu}\lambda_i} + \log\left(1-\frac{2}{\mu}\lambda_i\right)\right]$$

where $\lambda_i, \in \{1, 2, \ldots, m\}$ are the eigenvalues of $K'K\Sigma$.

Therefore, the solution to the optimization problem $\mathbf{P_4}$ is obtained in the form of (17) and (18). ■

Remark 1. The following comments are made on the results obtained in Theorem 1:

(1) It can be observed from (17) that the covariance matrix $\tilde{\Sigma}_k$ of the optimal attack $\tilde{z}_k^*$ is time-invariant since Σ and K are steady-state values which do not change with time. For notation brevity, we ignore the time index and use $\tilde{\Sigma}$ instead of $\tilde{\Sigma}_k$ in the subsequent discussion. However, it is worth noticing that the obtained results still hold even if the system has not entered steady state. In this case, Σ and K become $\Sigma_k = CP_k^-C' + R$ and $K_k = P_k^-C'(CP_k^-C' + R)$, where P_k^- can be calculated from (4) and (7).

(2) Note that the maximum estimation error covariance is obtained when equality in (22) holds, which implies a trade-off between the attack stealthiness and the attack consequence.

4.3. Worst-case linear attack

For any given T_k in the attack space Φ_δ, Theorem 1 limits the worst-case linear attack to a Gaussian random variable whose variance depends on the threshold of K–L divergence, i.e., $\tilde{\Sigma}$ is a function of δ. In the subsection, we aim at finding the worst-case T_k among all these Gaussian distributed attack policies which maximizes the degradation of system estimation performance. The results are summarized in the following theorem.

Theorem 2. *The worst-case linear attack strategy T_k is given by the solution of the convex optimization problem*

$$\mathbf{P_6}: \quad \min_{T_k} \quad \operatorname{Tr}(C\overline{P}\overline{P}C'\Sigma^{-1}T_k)$$
$$\text{s.t.} \quad \begin{bmatrix}\tilde{\Sigma} & T_k\\ T_k' & \Sigma^{-1}\end{bmatrix} \ge 0.$$

The corresponding b_k is a zero-mean Gaussian with covariance $\Gamma_k = \tilde{\Sigma} - T_k\Sigma T_k'$.

Proof. According to the linear attack strategy (11) and the worst-case Gaussian distribution of the corrupted innovation derived in Theorem 1, the feasibility constraint of the malicious attacker becomes

$$T_k\Sigma T_k' + \Gamma_k = \tilde{\Sigma}.$$

Consequently, it must hold that

$$\Gamma_k = \tilde{\Sigma} - T_k\Sigma T_k' \ge 0.$$

According to the iteration of the remote estimation error covariance (15), for any given threshold δ, to maximize $\operatorname{Tr}(\tilde{P}_k)$ is equivalent to solve the problem

$$\mathbf{P_7}: \quad \max_{T_k} \quad \operatorname{Tr}(-\overline{P}C'T_k'K' - KT_kC\overline{P})$$
$$\text{s.t.} \quad \tilde{\Sigma} - T_k\Sigma T_k' \ge 0.$$

We simplify the objective using $\operatorname{Tr}(A') = \operatorname{Tr}(A)$ and change the constraint to a linear matrix inequality using Schur complement, which gives the result. ■

Remark 2. Problem $\mathbf{P_6}$ is a semi-definite programming (SDP) problem which can be readily solved using the CVX toolbox in MATLAB. Note that the obtained worst-case attack matrix T_k^* and covariance matrix Γ_k^* are also time-invariant since $\tilde{\Sigma}$ is proved to be time-invariant.

Remark 3. For scalar system $m = 1$, the constraint of problem $\mathbf{P_4}$ becomes

$$\frac{\tilde{\Sigma}_k}{\Sigma} - 1 + \log\frac{\Sigma}{\tilde{\Sigma}_k} \le 2\delta, \tag{25}$$

where the maximum value of the objective function is obtained on the boundary, i.e., when equality holds. Then, the worst-case linear attack strategy which yields the largest estimation error covariance is achieved when T_k is maximized, i.e., when $\Gamma_k = \tilde{\Sigma}_k - T_k^2\Sigma = 0$. Consequently, (25) can be further simplified as

$$T_k^2 - 1 - \log T_k^2 = 2\delta,$$

Therefore, the closed-form expression of the worst-case linear attack strategy for scalar systems is given by $T_k = -\sqrt{X}$, where X is the largest solution of the equation $X = 2\delta + 1 + \log X$.

Remark 4. According to Theorems 1 and 2, the worst-case linear attack strategy when $D(\tilde{z}_k\|z_k) = 0$ is $T_k = -I$, $b_k = 0$, which demonstrates that the worst-case linear attack under χ^2 false-data detector obtained in Guo et al. (2017) is a special case of this work.

4.4. Worst-case attack signal

In this subsection, we characterize the worst possible action of the attacker on the system measurement through Algorithm 1 based on the worst-case linear attack (T_k, b_k) obtained in

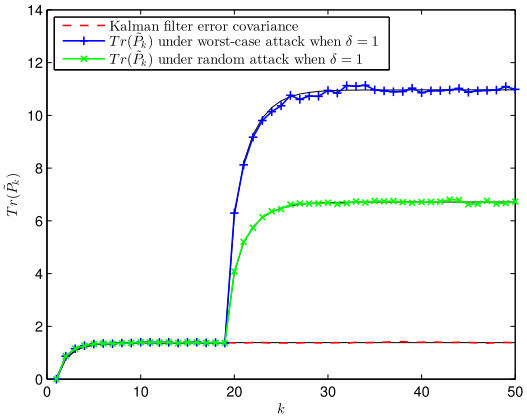


Fig. 2. Remote estimation error covariances when system is under worst-case linear attack and randomly generated attack for a given threshold $\delta = 1$.

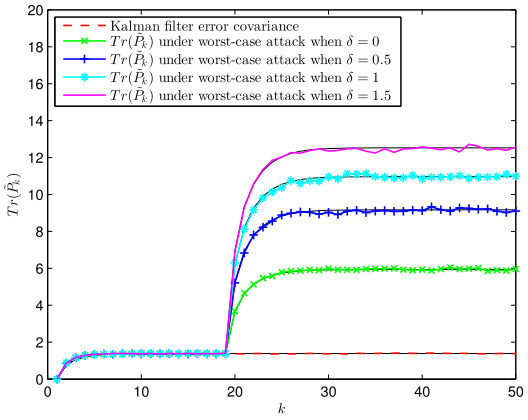


Fig. 3. Remote estimation error covariances when system is under worst-case linear attack for different thresholds δ.

Theorem 2. At each time instant k, the malicious attacker first solves the optimization problem $\mathbf{P}_6$ based on its knowledge of system parameters, from which the true innovation z_k and the corrupted innovation $\tilde{z}_k$ can be obtained. According to the relationship between the measurement and the innovation, the worst-case attack signal $\tilde{y}_k$ is obtained. Finally, the attacker updates the *priori* state estimates for the original and compromised processes, which it then use in its next iteration.

Algorithm 1 Calculation of the worst-case attack signal

1: **for** $k = 0 : 1 : \infty$ **do**
2: /*Find the optimal attack signal*/
3: Solve problem $\mathbf{P_6}$;
4: $z_k = y_k - C\hat{x}_k^-$, $\tilde{z}_k = T_k z_k + b_k$;
5: $\tilde{y}_k = \tilde{z}_k + C\tilde{x}_k^-$;
6: /*Update the prior state estimates*/
7: $\hat{x}_{k+1}^- = A(\hat{x}_k^- + Kz_k)$, $\tilde{x}_{k+1}^- = A(\tilde{x}_k^- + K\tilde{z}_k)$;
8: **end for**

5. Simulation example

We provide some numerical simulations to demonstrate the analytical results in this section. We consider a system with parameters

$$A = \begin{bmatrix} 0.7 & 0.2 \\ 0.05 & 0.64 \end{bmatrix}, \ C = \begin{bmatrix} 0.5 & -0.8 \\ 0 & 0.7 \end{bmatrix},$$
$$Q = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.7 \end{bmatrix}, \ R = \begin{bmatrix} 1 & 0 \\ 0 & 0.8 \end{bmatrix}.$$

Fig. 2 illustrates the worst-case linear attack strategy for a given threshold of K–L divergence $\delta = 1$. During time interval $[0, 20]$, the remote estimator runs a Kalman filter and enters steady state. The blue plus-mark line, the green x-mark line and the red dashed line represent the estimation error covariances when the system is under worst-case linear attack, randomly generated Gaussian noise, and no attack, respectively. It is shown that the worst-case linear attack strategy (T_k, b_k) obtained by solving the optimization problems in Theorems 1 and 2 leads to the largest estimation error covariance $\mathrm{Tr}(\tilde{P}_k)$.

We then analyze the potential degradation of the system estimation performance under different stealthiness measures. For different thresholds of K–L divergence, the evolutions of the remote estimation error covariance under the worst-case linear attack are shown in Fig. 3. Note that the attacks start from the steady state. Observed from the figure, the larger the threshold of K–L divergence δ, the larger the estimation error covariance $\mathrm{Tr}(\tilde{P}_k)$, which is consistent with the intuition that the attack stealthiness and consequence suggest a fundamental trade-off in the proposed framework. It is also worth noticing that even if $D(\tilde{z}_k \| z_k) = 0$, the attack space is not empty and the worst-case attack strategy $T_k = -I$ is still able to degrade the estimation performance.

To compare the attack consequence with the literature Bai and Gupta (2014) and Bai, Gupta et al. (2017), we consider a first-order system with parameters $A = 0.4, C = 1, Q = 0.2, R = 0.5$, which is the system model used in Bai, Gupta et al. (2017). The trade-off between worst-case estimation performance and attack stealthiness level is shown in Fig. 4, where the blue triangle line, the black cross line and the red circle line correspond to the result obtained in Bai and Gupta (2014) and Bai, Gupta et al. (2017) and our work, respectively. Observed from Fig. 4, we cannot conclude either the optimal attack obtained in Bai and Gupta (2014) or our work outperforms the other since different stealthiness metrics are adopted. However, due to a larger attack space considered in Bai, Gupta et al. (2017), the worst-case estimation error covariance under attack policy obtained in Bai, Gupta et al. (2017) is indeed larger than or equal to (when $\delta = 0$) that in our case for first-order systems. It is also worth noting that, although the stealthiness metrics adopted in these two works are not equivalent, the metric $D(\tilde{z}_k | \, |z_k) \leq \delta$ in our work implies $\lim_{k\to\infty} \frac{1}{k} D(\tilde{z}_1^k | \, |z_1^k) \leq \delta$ in Bai, Gupta et al. (2017), which leads to a reasonable comparison. Moreover, the correlation coefficient of the corrupted innovation sequence, $\rho_{\tilde{z}_k \tilde{z}_{k+1}} = \frac{1}{\sigma_{\tilde{z}_k} \sigma_{\tilde{z}_{k+1}}} \mathbb{E}[\tilde{z}_k \tilde{z}_{k+1}]$, is shown with respect to different thresholds δ in Fig. 5. This illustrates the underlying difference of the worst-case attack strategy obtained in our work and Bai and Gupta (2014) and Bai, Gupta et al. (2017). Specifically, in our case, the innovation approximately preserves the i.i.d. property even in the presence of attacks, while the innovations are relatively more correlated between different time instants under attack strategy obtained in Bai and Gupta (2014) and Bai, Gupta et al. (2017).

6. Conclusion

In this paper, we considered an arbitrarily distributed innovation-based linear attack in a remote state estimation scenario with K–L divergence as a stealthiness metric. To find the worst-case attack policy, the evolution of the remote estimation error covariance was derived, based on which a two-stage optimization problem was formulated. Furthermore, the worst-case linear

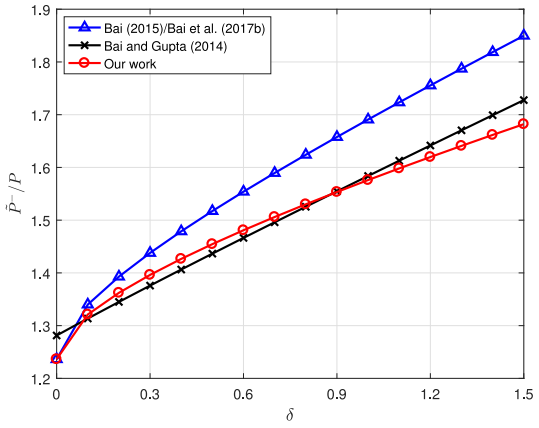


Fig. 4. Trade-off between system estimation performance and attack stealthiness.

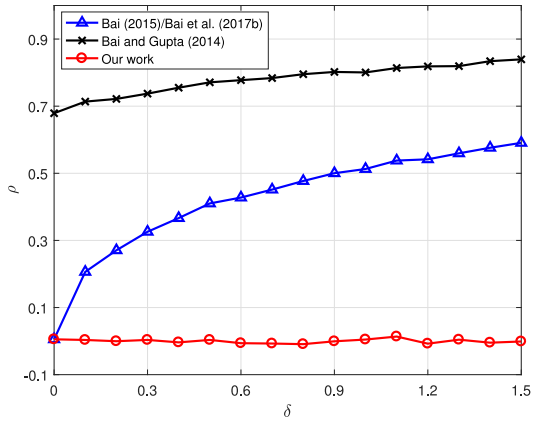


Fig. 5. Correlation coefficient of corrupted innovations with respect to different thresholds δ.

attack strategy was proved to be zero-mean Gaussian distributed and the numerical solution was obtained. An algorithm was provided to characterize the worst possible action of the attacker on the measurement data, which helps us better understand the attack consequence in designing control systems. Simulation and comparison were provided to demonstrate the analytical results. The worst-case analysis under arbitrary attack strategies and the design of associated protection scheme provide directions of the future work.

References

Anderson, B. D., & Moore, J. B. (2012). *Optimal filtering*. Courier Corporation.

Bai, C.-Z., & Gupta, V. (2014). On Kalman filtering in the presence of a compromised sensor: Fundamental performance bounds. In *2014 American control conference* (pp. 3029–3034). IEEE.

Bai, C.-Z., Gupta, V., & Pasqualetti, F. (2017). On Kalman filtering with compromised sensors: attack stealthiness and performance bounds. *IEEE Transactions on Automatic Control*. http://dx.doi.org/10.1109/TAC.2017.2714903.

Bai, C.-Z., Pasqualetti, F., & Gupta, V. (2015). Security in stochastic control systems: Fundamental limitations and performance bounds. In *American control conference* (pp. 195–200).

Bai, C.-Z., Pasqualetti, F., & Gupta, V. (2017). Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs. *Automatica*, *82*, 251–260.

Callegati, F., Cerroni, W., & Ramilli, M. (2009). Man-in-the-middle attack to the HTTPS protocol. *IEEE Security and Privacy Magazine*,(1), 78–81.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Guo, Z., Shi, D., Johansson, K. H., & Shi, L. (2017). Optimal linear cyber-attack on remote state estimation. *IEEE Transactions on Control of Network Systems*, *4*(1), 4–13.

Gupta, A., Langbort, C., & Basar, T. (2010). Optimal control in the presence of an intelligent jammer with limited actions. In *49th IEEE conference on decision and control* (pp. 1096–1101).

Karnouskos, S. (2011). Stuxnet worm impact on industrial cyber-physical system security. In *Proc. 37th annu. conf.* (pp. 4490–4494). IEEE Industrial Electronics Society.

Kim, K., & Kumar, P. R. (2012). Cyber–physical systems: A perspective at the centennial. *Proceedings of the IEEE*, *100*, 1287–1308. (Special Centennial Issue).

Li, Y., Quevedo, D., Dey, S., & Shi, L. (2016). SINR-based DoS attack on remote state estimation: A game-theoretic approach. *IEEE Transactions on Control of Network Systems*.

Li, Y., Shi, L., Cheng, P., Chen, J., & Quevedo, D. E. (2015). Jamming attacks on remote state estimation in cyber-physical systems: A game-theoretic approach. *IEEE Transactions on Automatic Control*, *60*(10), 2831–2836.

Liu, Y., Ning, P., & Reiter, M. K. (2011). False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, *14*(1), 13.

Meyer, U., & Wetzel, S. (2004). A man-in-the-middle attack on UMTS. In *Proceedings of the 3rd ACM workshop on wireless security* (pp. 90–97).

Miao, F., Pajic, M., & Pappas, G. J. (2013). Stochastic game approach for replay attack detection. In *52nd IEEE conference on decision and control* (pp. 1854–1859).

Mo, Y., Chabukswar, R., & Sinopoli, B. (2014). Detecting integrity attacks on SCADA systems. *IEEE Transactions on Control Systems Technology*, *22*(4), 1396–1407.

Mo, Y., Garone, E., Casavola, A., & Sinopoli, B. (2010). False data injection attacks against state estimation in wireless sensor networks. In *49th IEEE conference on decision and control* (pp. 5967–5972).

Mo, Y., & Sinopoli, B. (2009). Secure control against replay attacks. In *47th annual allerton conference on communication, control, and computing* (pp. 911–918).

Poor, H. V. (2013). *An introduction to signal detection and estimation*. Springer Science & Business Media.

Poovendran, R., Sampigethaya, K., Gupta, S. K. S., Lee, I., Prasad, K. V., Corman, D., et al. (2012). Special issue on cyber-physical systems. *Proceedings of the IEEE*, *100*(1), 1–12.

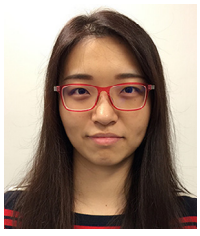
Shi, D., Chen, T., & Darouach, M. (2016). Event-based state estimation of linear dynamic systems with unknown exogenous inputs. *Automatica*, *69*, 275–288.

Shi, D., Elliott, R. J., & Chen, T. (2017). On finite-state stochastic modeling and secure estimation of cyber-physical systems. *IEEE Transactions on Automatic Control*, *62*(1), 65–80.

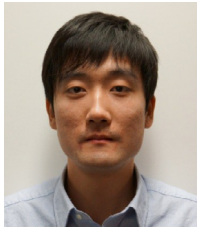
Slay, J., & Miller, M. (2007). *Lessons learned from the maroochy water breach*. Springer.

Teixeira, A., Sou, K. C., Sandberg, H., & Johansson, K. H. (2015). Secure control systems: A quantitative risk management approach. *IEEE Control Systems Magazine*, *35*(1), 24–45.

Zhang, H., Cheng, P., Shi, L., & Chen, J. (2015). Optimal Denial-of-Service attack scheduling with energy constraint. *IEEE Transactions on Automatic Control*, *60*(11), 3023–3028.



Ziyang Guo received her B.Eng. degree (Honors) in College of Control Science and Engineering from Zhejiang University, Hangzhou, China, in 2014. She is currently pursuing the Ph.D. degree in Electronic and Computer Engineering at the Hong Kong University of Science and Technology, Hong Kong. Her research interests include cyber–physical system security, state estimation and networked control systems.



Dawei Shi received his B.Eng. degree in Electrical Engineering and Automation from the Beijing Institute of Technology in 2008. He received his Ph.D. degree in Control Systems from the University of Alberta in 2014. In December 2014, he was appointed as an Associate Professor at the School of Automation, Beijing Institute of Technology, China. Since February 2017, he has been with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, as a postdoctoral fellow. His research interests include event-based control and estimation, robust model predictive control and tuning, and wireless sensor networks. He is a reviewer for a number of international journals, including IEEE Transactions on Automatic Control, Automatica, and Systems & Control Letters.

In 2009, he received the Best Student Paper Award in IEEE International Conference on Automation and Logistics.

Karl Henrik Johansson is Director of the Stockholm Strategic Research Area ICT The Next Generation and Professor at the School of Electrical Engineering, KTH Royal Institute of Technology. He received M.Sc. and Ph.D. degrees in Electrical Engineering from Lund University. He has held visiting positions at UC Berkeley, Caltech, NTU, HKUST Institute of Advanced Studies, and NTNU. His research interests are in networked control systems, cyber–physical systems, and applications in transportation, energy, and automation. He is a member of the IEEE Control Systems Society Board of Governors and the European Control Association Council. He has received several best paper awards and other distinctions, including a ten-year Wallenberg Scholar Grant, a Senior Researcher Position with the Swedish Research Council, the Future Research Leader Award from the Swedish Foundation for Strategic Research, and the triennial Young Author Prize from IFAC. He is member of the Royal Swedish Academy of Engineering Sciences, Fellow of the IEEE, and IEEE Distinguished Lecturer.

Ling Shi received the B.S. degree in electrical and electronic engineering from Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2002 and the Ph.D. degree in Control and Dynamical Systems from California Institute of Technology, Pasadena, CA, USA, in 2008. He is currently an associate professor at the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology. His research interests include cyber–physical systems security, networked control systems, sensor scheduling, and event-based state estimation. He is a senior member of IEEE. He served as an editorial board member for The European Control Conference 2013–2016. He has been serving as a subject editor for International Journal of Robust and Nonlinear Control from March 2015, an associate editor for IEEE Transactions on Control of Network Systems from July 2016, and an associate editor for IEEE Control Systems Letters from Feb 2017. He also served as an associate editor for a special issue on Secure Control of Cyber–physical Systems in the IEEE Transactions on Control of Network Systems in 2015–2017. He serves as the General Chair of the 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS 2018).