




Letter

Side Information-Based Stealthy False Data Injection Attacks Against Multi-Sensor Remote Estimation

Haibin Guo , Zhong-Hua Pang , Senior Member, IEEE, and Chao Li 

Dear Editor,

This letter investigates a novel stealthy false data injection (FDI) attack scheme based on side information to deteriorate the multi-sensor estimation performance of cyber-physical systems (CPSs). Compared with most existing works depending on the full system knowledge, this attack scheme is only related to attackers' sensor and physical process model. The design principle of the attack signal is derived to diverge the system estimation performance. Next, it is proven that the proposed attack scheme can successfully bypass the residual-based detector. Finally, all theoretical results are verified by numerical simulation.

CPSs integrate computation, communication and control, breaking through the geographical restrictions in traditional control modes, which have been widely applied to many fields such as power grid, smart factory, and multi-agent systems [1], [2]. However, communication networks incidentally increase the operation risk of CPSs, e.g., external cyber intrusions [3], [4]. Thus, the security of CPSs is of great importance and has attracted more attention in recent years.

As one of typical cyber attacks, stealthy FDI attacks hold the greater damage on the system performance meanwhile evading the anomaly detector [5]. In [6], a residual-based linear attack scheme was proposed to maximally degrade the performance of remote estimation without causing the alarm of χ^2 detector. Then, a historical and current residuals-based attack scheme was developed in [7] to further improve the attack impact. And, in [8], an extra sensor was adopted by the attacker to obtain side information that was utilized to design the attack signal. Both side information and historical residual were employed in [9] to construct the attack scheme. In [10], the security protection of multi-sensor remote estimation was investigated, and it was analyzed and proven that such a system can effectively expose the above single-sensor attack schemes. However, few works investigate the stealthy FDI attacks against the multi-sensor remote estimation except [11], [12]. The stealthy FDI attack schemes proposed in [11], [12] required the full system knowledge as well as depending on the measurement information of all sensors. Compared with such a case, it is relatively easy and freedom for malicious attackers to measure the output of the target system by their own sensor [8], [9]. Furthermore, the remote center is carefully protected in general, and it is very difficult to illegally disclose its knowledge [13].

This letter considers a more reasonable and practical scenario that malicious attackers only know the physical system model and the side information measured by the extra sensor. The specific contributions of this letter are detailed as follows. 1) For the multi-sensor remote estimation, a novel stealthy FDI attack scheme, which is only

Corresponding author: Zhong-Hua Pang.

Citation: H. Guo, Z.-H. Pang, and C. Li, "Side information-based stealthy false data injection attacks against multi-sensor remote estimation," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 4, pp. 1054–1056, Apr. 2024.

The authors are with the Key Laboratory of Fieldbus Technology and Automation of Beijing, North China University of Technology, Beijing 100144, China (e-mail: haibin.guo@bit.edu.cn; zhonghua.pang@ia.ac.cn; lichao@ncut.edu.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JAS.2023.124086

related to the side information, is proposed to maximally damage the system estimation performance meanwhile evading the residual-based detector. 2) The design principle of attack signal is derived to cause the estimation error to diverge. And, it is proved that the statistical distribution of the compromised residual is the same as the normal case, which illustrates the stealthiness of the proposed attack scheme.

Notations: \mathbb{R}^n denotes n -dimensional Euclidean space. $\Phi > 0$ ($\Phi \geq 0$) denotes Φ is a positive definite (semi-definite) matrix. $\mathcal{N}(q, Q)$ denotes the Gaussian distribution with mean q and covariance Q . $\mathbb{E}[\cdot]$ denotes the mathematical expectation. I_p denotes p -dimensional identity matrix, and 0 denotes the zero matrix with appropriate dimensions.

Multi-sensor remote estimation: Consider a linear time-invariant discrete-time system with N sensors

$$x_{k+1} = Ax_k + w_k \quad (1)$$

$$y_{i,k} = C_i x_k + v_{i,k} \quad (2)$$

where $x_k \in \mathbb{R}^n$ is the system state, and $y_{i,k} \in \mathbb{R}^{m_i}$ is the measurement of sensor i ($i = 1, 2, \dots, N$). The noises $w_k \in \mathbb{R}^n$ and $v_{i,k} \in \mathbb{R}^{m_i}$ are independent of each other, which satisfy $w_k \sim \mathcal{N}(0, \Phi)$ with $\Phi \geq 0$ and $v_{i,k} \sim \mathcal{N}(0, \Psi_i)$ with $\Psi_i > 0$. It is assumed that system matrices $(A, \Phi^{\frac{1}{2}})$ is stabilizable and (A, C_i) is detectable.

The entire measurement $Y_k \triangleq [y_{1,k}^T, \dots, y_{N,k}^T]^T$ is described as

$$Y_k = Cx_k + V_k \quad (3)$$

where $C \triangleq [C_1^T, C_2^T, \dots, C_N^T]^T$, and $V_k \triangleq [v_{1,k}^T, v_{2,k}^T, \dots, v_{N,k}^T]^T$ with the covariance $\Psi \triangleq \text{blkdiag}\{\Psi_1, \Psi_2, \dots, \Psi_N\}$.

The measurement residual of sensor i is defined as

$$z_{i,k} = y_{i,k} - C_i \hat{x}_{k|k-1} \quad (4)$$

where $\hat{x}_{k|k-1}$ is the priori estimate of the state x_k , which is broadcasted from the remote center to each sensor [10]–[12]. With (3), the measurement residuals of all sensors $Z_k = [z_{1,k}^T, \dots, z_{N,k}^T]^T$ can be expressed as

$$Z_k = Y_k - C \hat{x}_{k|k-1} \quad (5)$$

based on which, the following centralized Kalman filter is employed in the remote center:

$$\hat{x}_{k+1|k} = A \hat{x}_k \quad (6a)$$

$$\hat{x}_k = \hat{x}_{k|k-1} + K Z_k \quad (6b)$$

where \hat{x}_k is the posteriori estimate of x_k . The filter gain K is

$$K \triangleq PC^T (CPC^T + \Psi)^{-1} \quad (7)$$

where P is the priori estimation error covariance in the steady state satisfying $P = APA^T + \Phi - APC^T (CPC^T + \Psi)^{-1} CPA^T$.

It is well known that in the steady state, the measurement residual (5) obeys

$$Z_k \sim \mathcal{N}(0, S) \quad (8)$$

where $S = CPC^T + \Psi$. Based on this property, a χ^2 detector is adopted to find the system anomaly, which is defined as

$$g_k = \sum_{i=k-\tau+1}^k Z_i^T S^{-1} Z_i \stackrel{H_0}{\leq} \vartheta \quad (9)$$

where τ is the detection window size, and H_0 and H_1 denote the system without and with the attack, respectively. Once g_k exceeds the detection threshold ϑ , the detector would trigger an alarm.

FDI attack scheme: As shown in Fig. 1, malicious attackers utilize an extra sensor different from the system sensors as follows:

$$y_{e,k} = C_e x_k + v_{e,k} \quad (10)$$

where $y_{e,k} \in \mathbb{R}^{m_e}$ denotes the measurement output, and $v_{e,k} \in \mathbb{R}^{m_e}$ is the measurement noise satisfying $v_{e,k} \sim \mathcal{N}(0, \Psi_e)$ with $\Psi_e > 0$, which is independent of the noises w_k and V_k . It is assumed that (A, C_e) is detectable. Then, the following auxiliary filter is adopted by mali-

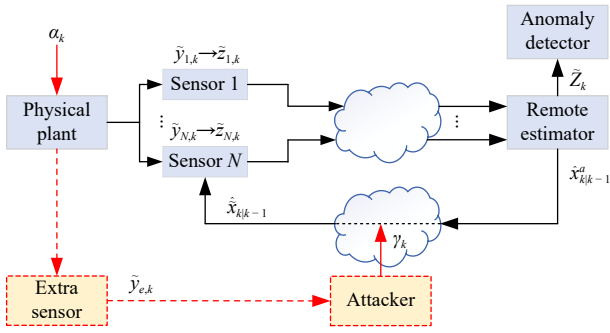


Fig. 1. A multi-sensor remote estimation system under attack.

cious attackers:

$$\hat{x}_{k+1|k}^e = A\hat{x}_k^e \quad (11a)$$

$$\hat{x}_k^e = \hat{x}_{k|k-1}^e + K^a(v_{e,k} - C_e\hat{x}_{k|k-1}^e) \quad (11b)$$

where $\hat{x}_{k|k-1}^e$ and \hat{x}_k^e are the priori and posteriori estimates of the state x_k , respectively. Define $\delta_k^e \triangleq x_k - \hat{x}_k^e$, and with (1), (10) and (11), we can obtain

$$\delta_{k+1}^e = (A - K^a C_e A)\delta_k^e + (I_n - K^a C_e)w_k - K^a v_{e,k+1} \quad (12)$$

which converges if and only if the filter gain K^a is designed to ensure the matrix $A - K^a C_e A$ stable.

It is assumed that malicious attackers can directly insert false data into the physical system (1) meanwhile modifying the information $\hat{x}_{k|k-1}$ broadcasted by the remote center. Under the attack, the system (1) is rewritten as

$$\tilde{x}_{k+1} = A\tilde{x}_k + \alpha_k + w_k \quad (13)$$

where \tilde{x}_k denotes the compromised system state, and α_k is the attack signal to be designed. The compromised priori estimate received by each sensor is

$$\hat{\tilde{x}}_{k|k-1} = \hat{x}_{k|k-1}^a + \gamma_k \quad (14)$$

where $\hat{x}_{k|k-1}^a$ is the system estimate under the attack, and γ_k is the injection signal to be designed.

For the compromised system (13), the measurement (10) of the extra sensor is represented as

$$\tilde{y}_{e,k} = C_e \tilde{x}_k + v_{e,k}. \quad (15)$$

The filter (11) is rewritten as

$$\hat{\tilde{x}}_{k+1|k}^e = A\hat{\tilde{x}}_{k|k-1}^e + \alpha_k + AK^a(\tilde{y}_{e,k} - C_e\hat{\tilde{x}}_{k|k-1}^e) \quad (16)$$

where $\hat{\tilde{x}}_{k|k-1}^e$ is the priori estimate of the state \tilde{x}_k under the attack.

The injection signal γ_k is designed as

$$\gamma_k = \hat{\tilde{x}}_{k|k-1}^e - \hat{x}_{k|k-1}^{ea} \quad (17)$$

where $\hat{x}_{k|k-1}^{ea}$ satisfies

$$\hat{\tilde{x}}_{k+1|k}^{ea} = A\hat{\tilde{x}}_{k|k-1}^{ea} + AK^a(\tilde{y}_{e,k} - C_e\hat{\tilde{x}}_{k|k-1}^{ea}) \quad (18)$$

with $\tilde{y}_{e,k}^a = \tilde{y}_{e,k} - C_e\gamma_k$.

Remark 1: As shown in (3) and (10), both the system sensors and the extra sensor are employed to measure the information of the physical system. Compared with utilizing the information of all system sensors and the full system knowledge to design the attack signal in [11], [12], it is more freedom for the proposed attack scheme to merely require the partial physical system knowledge A and the side information measured by the extra sensor. And, it can be found from (16)–(18) that the injection signal γ_k is unrelated to the real-time measurement information. Furthermore, the filter gain K^a in (11) needs to guarantee $A - K^a C_e A$ stable, which is unrelated with the Kalman filter (6). These break through the strict assumptions of attack design and reduce the attack complexity.

Attack destructiveness: Under the attack, the entire measurement (3) is rewritten as

$$\tilde{Y}_k = C\tilde{x}_k + V_k. \quad (19)$$

Correspondingly, the residual received by the remote center becomes

$$\tilde{Z}_k = \tilde{Y}_k - C\hat{\tilde{x}}_{k|k-1} \quad (20)$$

based on which, the Kalman filter (6) is redescribed as

$$\hat{\tilde{x}}_{k+1|k}^a = A\hat{\tilde{x}}_{k|k-1}^a + AK\tilde{Z}_k \quad (21)$$

where $\hat{\tilde{x}}_{k|k-1}^a$ denotes the priori estimate under the attack. Define the priori estimation error under the attack as

$$e_{k+1|k}^a \triangleq \tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1|k}^a$$

which is utilized to quantify the estimation performance of the compromised system.

Theorem 1: Under the attack, the estimation error $e_{k+1|k}^a$ diverges, if and only if the system matrix A is unstable or the attack signal α_k is unbounded.

Proof: Combining with (13), (14), (20) and (21) yields

$$e_{k+1|k}^a = (A - AKC)e_{k|k-1}^a + AKC\gamma_k + \alpha_k + w_k - AKV_k. \quad (22)$$

Then, with (16) and (18), the injection signal (17) satisfies

$$\gamma_{k+1} = A\gamma_k + \alpha_k. \quad (23)$$

Let $\xi_k \triangleq [e_{k|k-1}^{aT} \ \gamma_k^T]^T$, and with (22) and (23), we can get

$$\xi_{k+1} = \begin{bmatrix} \mathcal{A} & AKC \\ 0 & A \end{bmatrix} \xi_k + \begin{bmatrix} I_n \\ 0 \end{bmatrix} \alpha_k + \begin{bmatrix} I_n & -AK \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_k \\ V_k \end{bmatrix} \quad (24)$$

where $\mathcal{A} = A - AKC$.

It can be found that the system estimation error with unstable matrix A under the attack always diverges no matter how to design the attack signal α_k . On the other hand, when A is stable, an unbounded attack signal is designed as

$$\alpha_{k+1} = \mathcal{F}\alpha_k \quad (25)$$

where \mathcal{F} is an unstable matrix. \blacksquare

Attack stealthiness: With (14) and (19), the residual (20) is further derived as

$$\tilde{Z}_k = C(e_{k|k-1}^a - \gamma_k) + V_k. \quad (26)$$

Let $\delta_{k|k-1}^a \triangleq e_{k|k-1}^a - \gamma_k$, and then subtracting (23) from (22) yields

$$\delta_{k+1|k}^a = \mathcal{A}\delta_{k|k-1}^a + w_k - AKV_k \quad (27)$$

whose covariance $P_{k+1|k}^a \triangleq \mathbb{E}[\delta_{k+1|k}^a \delta_{k+1|k}^{aT}]$ is derived as

$$P_{k+1|k}^a = \mathcal{A}P_{k|k-1}^a \mathcal{A}^T + \Phi + AK\Psi K^T A^T. \quad (28)$$

It is assumed that the attack starts at time $k = k_a$. Before the attack occurs, with (16) and (18), $\hat{\tilde{x}}_{k_a|k_a-1}^e = \hat{\tilde{x}}_{k_a|k_a-1}^{ea}$ holds, that is, $\gamma_{k_a} = 0$. And further, $\delta_{k_a|k_a-1}^a = e_{k_a|k_a-1}^a$ holds, which leads to $P_{k_a|k_a-1}^a = P$.

In the normal situation, with (1)–(6), the priori estimation error $e_{k+1|k} \triangleq x_{k+1} - \hat{x}_{k+1|k}$ is derived as

$$e_{k+1|k} = \mathcal{A}e_{k|k-1} + w_k - AKV_k \quad (29)$$

whose steady-state covariance $P \triangleq \lim_{k \rightarrow +\infty} \mathbb{E}[e_{k+1|k} e_{k+1|k}^T]$ is

$$P = \mathcal{A}P\mathcal{A}^T + \Phi + AK\Psi K^T A^T. \quad (30)$$

With (28) and (30), it can be observed that since $P_{k_a|k_a-1}^a = P$, $P_{k+1|k}^a$ is time invariant and satisfies

$$P_{k+1|k}^a = P \quad (31)$$

which leads to the following theorem.

Theorem 2: The compromised residual \tilde{Z}_k obeys

$$\tilde{Z}_k \sim \mathcal{N}(0, S) \quad (32)$$

which has the same distribution as normal residual (8), i.e., the proposed attack scheme successfully evades the residual-based detector.

Proof: Compromised residual (26) can be rewritten as

$$\tilde{Z}_k = C\delta_{k|k-1}^a + V_k.$$

From (29), it is obtained that

$$\mathbb{E}[e_{k+1|k}] = \mathcal{A}^{k+1}\mathbb{E}[e_{0|-1}]$$

which converges to zero in few steps due to stable \mathcal{A} . At the attack start time k_a , we can get $\mathbb{E}[e_{k_a|k_a-1}] = 0$. And $\gamma_{k_a} = 0$ leads to $\mathbb{E}[\delta_{k_a|k_a-1}^a] = \mathbb{E}[e_{k_a|k_a-1}] = 0$. Then, the mathematical expectation of

(27) is derived as

$$\mathbb{E}[\delta_{k+1|k}^a] = \mathcal{A}^{k+1-k_a} \mathbb{E}[\delta_{k_a|k_a-1}^a] = 0$$

which leads to $\mathbb{E}[\tilde{Z}_k] = 0$. And, the covariance $\tilde{S}_k \triangleq \mathbb{E}[\tilde{Z}_k \tilde{Z}_k^T]$ is calculated as

$$\tilde{S}_k = C P_{k|k-1}^a C^T + \Psi.$$

With (31), $\tilde{S}_k = S$ holds. Hence, the distribution of the compromised residual is

$$\tilde{Z}_k \sim \mathcal{N}(0, S)$$

which keeps the same distribution as (8). This means that the proposed attack scheme can bypass the residual-based detector. ■

Numerical simulation: A stable system is considered to verify the effectiveness of the proposed attack scheme, whose parameters are

$$A = \begin{bmatrix} 0.3201 & 0.4712 & 0.2605 \\ 0.1420 & 0.1027 & 0.0141 \\ 0.0231 & 0.2526 & 0.4357 \end{bmatrix}, \Phi = 0.01I_3$$

$$C_1 = \begin{bmatrix} 0.1501 & 0.6423 & 0.5712 \\ 0.0231 & 0.3201 & 0.2013 \end{bmatrix}, \Psi_1 = 0.03I_2$$

$$C_2 = \begin{bmatrix} 0.1012 & 0.2701 & 0.6235 \\ 0.0724 & 0.5413 & 0.2514 \end{bmatrix}, \Psi_2 = 0.02I_2.$$

With (7), the Kalman filter gain is designed as

$$K = \begin{bmatrix} 0.0684 & 0.0113 & 0.0737 & 0.0482 \\ 0.1338 & 0.0720 & 0.0483 & 0.1908 \\ 0.1272 & 0.0386 & 0.2476 & 0.0599 \end{bmatrix}.$$

For the attacker, the following parameters are chosen:

$$C_e = \begin{bmatrix} 0.4205 & 0.1314 & 0.6425 \\ 0.3425 & 0.8301 & 0.2624 \end{bmatrix}, \Psi_e = 0.04I_2.$$

The unstable attack matrix \mathcal{F} and the filter gain K^a are respectively designed as

$$\mathcal{F} = \begin{bmatrix} 0.7021 & 0.0642 & 0.7021 \\ 0.2301 & 0.8501 & 0.6142 \\ 0.0032 & 0.1824 & 0.7024 \end{bmatrix}, K^a = \begin{bmatrix} 0.1362 & 0.1035 \\ 0.0172 & 0.1754 \\ 0.1747 & 0.0625 \end{bmatrix}$$

where K^a ensures $A - K^a C_e A$ is stable. The attack occurs at [101, 200] with the initial condition $\alpha_{101} = [1 \ 1 \ 1]^T$. The simulation results are shown in Fig. 2. It is clear from Fig. 2(a) that the detection index under the attack keeps the same distribution as that under no attack. And, the estimation error of the system diverges when the attack occurs as shown in Fig. 2(b). These results directly verify Theorems 1 and 2.

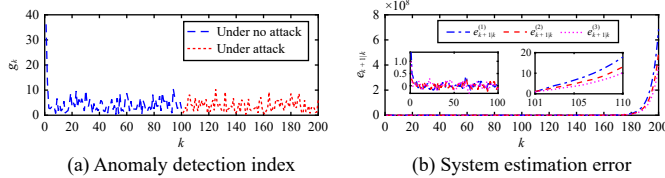


Fig. 2. The system performance under attack.

Conclusion: This letter has explored the security issue of multi-sensor remote estimation of cyber-physical systems from the adverse standpoint. A stealthy FDI attack scheme, which is based on the side information measured by an extra sensor, has been proposed. The design principle of the attack signal has been derived for stable and

unstable systems aiming at extremely degrading the estimation performance. Then, the distribution of the polluted residual has been derived to illustrate the attack stealthiness. Finally, the numerical simulation has been carried out to verify the effectiveness of the proposed attack scheme. In our future work, the proposed scheme will be further investigated for the cooperative control of networked multi-agent systems [14].

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (62173002) and the Beijing Natural Science Foundation (4222045).

References

- [1] D. Ding, Q.-L. Han, X. Ge, and J. Wang, "Secure state estimation and control of cyber-physical systems: A survey," *IEEE Trans. Syst. Man Cybern.: Syst.*, vol. 51, no. 1, pp. 176–190, Jan. 2021.
- [2] C.-B. Zheng, Z.-H. Pang, J.-X. Wang, J. Sun, G.-P. Liu, and Q.-L. Han, "Null-space-based time-varying formation control of uncertain nonlinear second-order multi-agent systems with collision avoidance," *IEEE Trans. Ind. Electron.*, vol. 70, no. 10, pp. 10476–10485, Oct. 2023.
- [3] H. Guo, J. Sun, Z.-H. Pang, and G.-P. Liu, "Event-based optimal stealthy false data injection attacks against remote state estimation systems," *IEEE Trans. Cybern.*, vol. 53, no. 10, pp. 6714–6724, Oct. 2023.
- [4] M. Xie, D. Ding, X. Ge, Q.-L. Han, H. Dong, and Y. Song, "Distributed platooning control of automated vehicles subject to replay attacks based on proportional integral observers," *IEEE/CAA J. Autom. Sinica*, 2022. DOI: 10.1109/JAS.2022.105941
- [5] Z.-H. Pang, L.-Z. Fan, H. Guo, Y. Shi, R. Chai, J. Sun, and G.-P. Liu, "Security of networked control systems subject to deception attacks: A survey," *Int. J. Syst. Sci.*, vol. 53, no. 16, pp. 3577–3598, Dec. 2022.
- [6] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyber-attack on remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 4–13, Mar. 2017.
- [7] H. Liu, Y. Ni, L. Xie, and K. H. Johansson, "How vulnerable is innovation-based remote state estimation: Fundamental limits under linear attacks," *Automatica*, vol. 136, p. 110079, Feb. 2022.
- [8] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Worst-case innovation-based integrity attacks with side information on remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 1, pp. 48–59, Mar. 2019.
- [9] D. Ye, B. Yang, and T.-Y. Zhang, "Optimal stealthy linear attack on remote state estimation with side information," *IEEE Syst. J.*, vol. 16, no. 1, pp. 1499–1507, Mar. 2022.
- [10] Y. Li, L. Shi, and T. Chen, "Detection against linear deception attacks on multi-sensor remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 846–856, 2018.
- [11] H. Guo, J. Sun, and Z.-H. Pang, "Stealthy false data injection attacks with resource constraints against multi-sensor estimation systems," *ISA Trans.*, vol. 127, pp. 32–40, Aug. 2022.
- [12] H. Guo, J. Sun, and Z.-H. Pang, "Residual-based false data injection attacks against multi-sensor estimation systems," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1181–1191, May 2023.
- [13] Q. Zhang, K. Liu, D. Han, G. Su, and Y. Xia, "Design of stealthy deception attacks with partial system knowledge," *IEEE Trans. Autom. Control*, vol. 68, no. 2, pp. 1069–1076, Feb. 2023.
- [14] Z.-H. Pang, W.-C. Luo, G.-P. Liu, and Q.-L. Han, "Observer-based incremental predictive control of networked multi-agent systems with random delays and packet dropouts," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 68, no. 1, pp. 426–430, Jan. 2021.