文章编号：1006-3080(2020)06-0839-05 DOI：10.14135/j.cnki.1006-3080.20191016001

# 针对信息物理系统线性欺诈攻击的水印加密策略

汪 迪， 李芳菲， 许思遥， 刘昌洪
（华东理工大学理学院，上海 200237）

**摘要：**研究了利用水印加密策略来防御信息物理系统(CPS)的线性欺诈攻击。线性欺诈攻击可以降低系统的远程估计性能，很难被检测器检测。采取KL散度检测器和水印加密的方式来保护系统传输的数据，并证明这一方法可以有效地帮助KL散度检测器检测攻击或降低攻击的效果。最后给出了数值模拟进一步说明水印加密策略的有效性。

**关键词：**信息物理系统；KL散度；水印加密；线性欺诈攻击；系统安全
**中图分类号：**TP3 **文献标志码：**A

随着计算机和动力学系统方面的研究长足发展，信息物理系统(CPS)在当今社会中起着越来越不可或缺的作用[1-2]。CPS作为一个集成计算、网络和物理环境的复杂系统，可通过3C(计算机、通信、控制)技术的有机集成和深度协作，实现大型工程系统的实时感知、动态控制和信息服务[3]。由于信号通过无线通信通道时系统的安全性不能得到有效的保障，对CPS的成功攻击可能会造成巨大的损失，包括信息泄漏、工业流程暂停和基础架构损坏等。因此，CPS的安全性问题逐渐引起研究人员的关注[4]。

为了保证CPS的安全运行，近十年来攻击策略和防御机制受到了广泛的关注。文献[5]中，从攻击者对系统开放资源破坏的角度出发，对CPS中的几种攻击模式进行了分类，其中拒绝服务(DoS)攻击是最常见的攻击类型之一，在这其中攻击者试图阻止通信通道，并使系统的估计误差变大[6]。与DoS攻击相比，欺诈攻击的重点是隐身性，隐身的攻击在攻击系统时不太可能被发现，从而使欺诈攻击所造成的危害更加严重。文献[7]研究了基于卡方检测器的最优线性欺诈攻击。文献[8]研究了基于KL散度的隐身攻击，该隐身攻击独立于任何特定的检测方案，并分析了此类攻击的隐身性和性能特征。文献[9]进一步证明了控制性能下降和攻击隐身性之间的权衡。最近的研究提供了针对某些攻击的防御策略。比如在文献[10]中，作者建议使用水印来达到检测重放攻击的目的，但同时牺牲了部分系统性能。

在信息物理系统的信息安全问题方面尽管已经取得了一些成果，但仍有一些问题需要进一步考虑。当提到基于KL散度检测器的CPS的安全性问题时，由于KL散度的复杂计算，一些代表性的结果仅考虑了系统为一维的情况[9-11]。另外，大多数研究仅考虑网络攻击问题，而没有给出针对这种攻击的防御或检测策略。与卡方检测器不同，KL散度检测器可以应用于非高斯分布的情况，因此具有广泛的应用[12]，国内学者则进一步针对KL散度检测器提出了新型的线下欺诈攻击[13]。然而，对基于KL散度检测器的CPS的安全性问题的研究还远远不够，尤其是在防御或检测问题上。出于上述原因，本文考虑使用水印加密来针对带有KL散度检测器的CPS中的线性欺诈攻击。

本文通过选择适当的水印参数，证明了KL散度检测器在攻击者不知道水印参数的统计特征时会触发警报，结果表明对于给出的特定欺诈攻击也有效，并且当攻击者知道水印的统计特征时，水印的存在

收稿日期：2019-10-16
基金项目：国家自然科学基金面上项目(61773161)；上海市探索基金(18ZR1409800)；华东理工大学优秀青年培育基金(50321111916010)
作者简介：汪 迪(1995—)，男，浙江人，硕士生，研究方向为信息安全。E-mail：1049767865@qq.com
通信联系人：李芳菲，E-mail：li_fangfei@163.com
引用本文：汪 迪，李芳菲，许思遥，等. 针对信息物理系统线性欺诈攻击的水印加密策略[J]. 华东理工大学学报(自然科学版)，2020，46(6)：839-843.
**Citation:** WANG Di, LI Fangfei, XU Siyao, *et al*. Watermark Encryption for Linear Deception Attacks in Cyber Physical Systems[J]. Journal of East China University of Science and Technology, 2020, 46(6): 839-843.

仍然可以削弱攻击所造成攻击效果，使得估计误差方差尽可能控制在能容忍的范围内。

## 1 问题设置

### 1.1 系统模型

本文考虑下述一维的线性不变的系统：

$$x_{k+1} = ax_k + \omega_k \tag{1}$$

$$y_k = cx_k + \upsilon_k \tag{2}$$

其中：$x_k$ 表示系统状态变量，$y_k$ 表示系统输出变量，$a$、$c$ 是系统参数，$\omega_k$、$\upsilon_k$ 则是均值为 0、方差为 $p$、$q$ 的正态分布噪声。本文假设 $a$、$c$ 是可测的且系统是稳定的。

下面给出卡尔曼滤波器的主要内容，用来在接收端估计系统运行变量[13]：

$$\begin{aligned}
&\hat{x}_k^- = a\hat{x}_{k-1}\\
&P_k^- = a^2P_{k-1} + p\\
&K_k = P_k^-c(c^2P_k^- + q)^{-1}\\
&\hat{x}_k = \hat{x}_k^- + K_k(y_k - c\hat{x}_k^-)\\
&P_k = (1 - K_kc)P_k^-
\end{aligned}$$

其中：$P_k^-$ 与 $P_k$ 分别表示先验与后验估计的估计误差方差，$K_k$ 表示卡尔曼增益。在下面的讨论中，我们假设系统已经运行到稳定状态，此时本文用 $\bar{P}$ 与 $K$ 分别表示稳态系统下的估计误差方差与卡尔曼增益，并定义新息序列为 $z_k = y_k - c\hat{x}_k^-$，它是均值为 0、方差为 $\sigma_z = c^2\bar{P} + q$ 的正态分布。

### 1.2 KL 散度检测器

**定义 1**：假设 $g_k$ 和 $h_k$ 是两个随机序列，$f(g_k)$ 和 $f(h_k)$ 则表示上述序列的概率密度函数，则这两个序列的 KL 散度可以表示为：

$$D(g_k \| h_k) = \int_{\{\xi | f_{g_k}(\xi_k) > 0\}} \lg \frac{f_{g_k}(\xi_k)}{f_{h_k}(\xi_k)} \, f_{g_k}(\xi_k)\mathrm{d}(\xi_k) \tag{3}$$

值得注意的是，当两个随机序列都满足均值为 0 的正态分布时，KL 散度的表达式可以化简为

$$D(g_k \| h_k) = \frac{1}{2}Tr\left(\sigma_h^{-1}\sigma_g\right) - \frac{m}{2} + \frac{1}{2}\lg\frac{|\sigma_h|}{|\sigma_g|} \tag{4}$$

其中：$Tr$ 为矩阵轨迹，$\sigma_g$ 与 $\sigma_h$ 为对应序列的方差。进一步，当维数为 1 维时，它关于 $\sigma_g$ 的二阶导数为 $\frac{1}{2\sigma_g^2} > 0$，即它关于 $\sigma_g$ 来说是一个凸函数，且当 $\sigma_g$ 大于 $\sigma_h$ 时，KL 散度值单调递增[14]。

### 1.3 攻击模型

假设攻击者能够拦截和修改传输的新息序列，它的目标是使估计误差最大化，并避免被错误数据检测器检测到。基于文献 [5]，我们主要考虑下述攻击形式：

$$\tilde{z}_k = t_kz_k + b_k \tag{5}$$

其中：$t_k$ 为任意的攻击系数，$b_k$ 为服从 0 均值，方差为 $\sigma_b$ 的正态分布序列，且 $\tilde{z}_k$ 的方差为

$$\sigma_{\tilde{z}} = t_k^2\sigma_z + \sigma_b \tag{6}$$

同时对于系统预先给定的检测器阈值 $\delta$，攻击必须满足以下条件：

$$D(\tilde{z}_k \| z_k) \le \delta \tag{7}$$

我们称满足这一条件的攻击形式为线性欺诈攻击，即它不会被检测器所发现。

## 2 水印加密以及性能分析

### 2.1 水印加密模型

基于上述内容的讨论，我们了解到 KL 散度的大小取决于两个正态分布的方差的差距，因此为了检测上述线性欺诈攻击，我们期望扩大攻击存在时的新息序列的方差以帮助检测器检测攻击的存在。我们给出水印加密的模块，具体模式如图 1 所示，在系统发送新息序列之前，我们对数据进行预处理：

$$r(z_k) = uz_k + m_k \tag{8}$$

其中：$u$ 为大于 0 的实数，$m_k$ 为满足均值为 0，方差为 $\sigma_m$ 的正态分布的随机序列。

并在接收端对数据进行还原：

$$z_k^f = \frac{\tilde{r}(z_k) - m_k}{u} \tag{9}$$

其中：$\tilde{r}(z_k)$ 是接收端实际收到的数据。可以看出，若数据并未被攻击，则数据可以完全被还原，从而对系统并不产生影响。而当攻击存在时，有

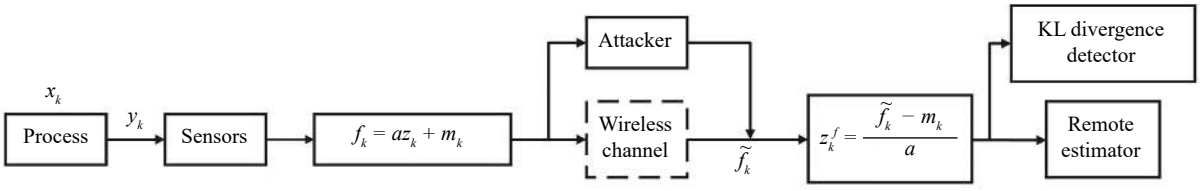


图 1 系统运行阶段的水印加密以及解密

Fig. 1 Watermark encryption and decryption during system operation

$$z_k^f = t_k z_k + \frac{1}{u} b_k + \frac{1}{u}(t_k - 1) m_k \tag{10}$$

容易得到, 它的均值为 0, 方差为:

$$\sigma_{z_f} = t_k^2 \sigma_z + \frac{1}{u^2}\sigma_b + \frac{1}{u^2}(t_k - 1)^2 \sigma_m \tag{11}$$

### 2.2　不同情况下的性能分析

2.2.1　缺少水印信息　若攻击者不知道水印的存在, 仍采取水印不存在时的最优攻击策略, 由文献 [7] 可知, 其最优攻击策略为 $t_k = -\sqrt{x}$, 其中 $x$ 为方程 $x=2\delta+1+\lg x$ 的最大解, $b_k=0$, 此时 $D(\tilde{z}_k \| z_k) = \delta$。下面的定理用来说明水印加密存在时, 攻击会触发 KL 散度检测器的警报。

**定理 1**: 当水印存在时, 原最优攻击会使得 KL 散度超过限定的阈值, 从而使检测器发出警报。

**证明**: 当攻击者使用参数 $t_k = -\sqrt{x}$, $b_k=0$ 时, $\sigma_{z_f} = x\sigma_z + \frac{1}{u^2}\left(\sqrt{x}+1\right)^2 \sigma_m$, $\sigma_{\tilde{z}} = x\sigma_z$, 由于 $x$ 为方程 $x=2\delta+1+\lg x$ 的最大解, 容易得知 $x \geqslant 1$, 即可以得到 $\sigma_{z_f} \geqslant \sigma_{\tilde{z}}$, 由 KL 散度在一侧的单调性, 我们可以得到:

$$D\left(z_k^f \| z_k\right) > D(\tilde{z}_k \| z_k) = \delta \tag{12}$$

故 KL 散度检测器会触发警报, 从而发现攻击的存在, 证毕。

进一步我们可以看到由于 KL 散度在一侧单调递增, 若水印参数较大, 则可以使 KL 散度值进一步变大, 使得攻击的暴露速度加快, 但同时较大的参数会使得容易被攻击者发现水印的存在。同时基于上述定理, 可以得知水印的存在可以有效地使文献 [13] 中的最优攻击被检测器发现。

2.2.2　完全了解水印信息　因较大的水印参数可能使得水印被攻击者发现, 若攻击者知道了水印参数 $u$ 以及 $m_k$ 的方差 $\sigma_m$, 此时攻击者会重新构造攻击, 从而避开 KL 散度检测器[15]。即使攻击者重构攻击, 水印的存在会使得攻击的效果变差, 即估计误差方差仍被控制在一定范围内。在此之前, 我们需要得到估计误差方差关于攻击参数 $t_k$ 的表达式:

**引理 1:** 当有水印存在时, 在如式(5)的攻击形式下, 估计误差方差的表达式为:

$$\tilde{P}_k = a^2 \tilde{P}_{k-1} + p + K^2 \sigma_z^* - 2\bar{P}\, c\, t_k\, K \tag{13}$$

其中: $\sigma_z^*$是使得 $D(\tilde{z}_k \| z_k) = \delta$ 时的 $\tilde{z}_k$ 的方差。

**证明:** 证明过程与 [7] 类似, 这里不再赘述, 证毕。

由引理 1 容易得知, 当攻击不存在时, 估计误差方差为:

$$\tilde{P}_k^* = a^2 \tilde{P}_{k-1} + p + K^2 \sigma_z - 2\bar{P}\, c\, K \tag{14}$$

下面给出水印能有效控制攻击效果的详细说明:

**定理 2**: 当水印参数 $u>0$, $\sigma_m \to +\infty$时, 估计误差方差

$$\tilde{P}_k \to a^2 \tilde{P}_{k-1} + p + K^2 \sigma_z^* - 2\bar{P}\, c\, K \tag{15}$$

**证明:** 要证明上述结论, 只需要证明当水印参数 $u>0$, $\sigma_m \to +\infty$ 时, $t_k \to 1$。固定参数 $u$, 重新构造的攻击需要满足下面的条件:

$$t_k^2 \sigma_z + \frac{1}{u^2}\sigma_b + \frac{1}{u^2}(t_k - 1)^2 \sigma_m = \sigma_z^* \tag{16}$$

可以得到

$$t_k = \frac{\dfrac{2\sigma_m}{u^2} \pm \sqrt{4\sigma_z\sigma_z^* + \dfrac{4\sigma_z^*}{u^2}\sigma_m - \dfrac{4\sigma_z\sigma_b}{u^2} - \dfrac{4\sigma_z}{u^2}\sigma_m - \dfrac{4\sigma_b}{u^4}\sigma_m}}{2\sigma_z + \dfrac{2\sigma_m}{u^2}} \tag{17}$$

由于 $\sigma_z^* > \frac{1}{u^2}\sigma_b$, 可以得到, 当 $\sigma_m \to +\infty$ 时, $t_k \to 1$, 故 $\tilde{P}_k \to a^2 \tilde{P}_{k-1} + p + K^2 \sigma_z^* - 2\bar{P}\, c\, K$。因此当水印参数较大时, 可以将估计误差限制在较小的范围, 可以有效的限制攻击者的决策范围, 证毕。

通过引理 1 以及定理 2 可知, 水印的存在可以降低估计误差方差, 从而减小对系统的影响, 在实际运用中, 适当大小的水印参数即可限制攻击者的策略, 并非要求趋向于无穷。

## 3　数值模拟

我们比较了不同水印情况下对于线性欺骗攻击的影响, 考虑下面这样的系统, 系统参数为 $a$=0.7, $c$=0.3, $p$=0.5 和 $q$=0.1, KL 散度检测器的阈值 $\delta$=1。

图 2 示出了不同的水印参数对于 KL 散度检测器的影响, 其中黑线为攻击不存在时的 KL 散度值, 蓝线, 红线和绿线分别表示原最优攻击存在时参数分别为 $u$=0.5, $\sigma_m$=1; $u$=0.5, $\sigma_m$=4; $u$=0.5, $\sigma_m$=8 时, KL 散度检测器所得到的 KL 散的值的变化曲线。结果表明, 在存在最佳攻击的情况下, 水印的存在可以增加 KL 散度的值, 从而触发警报。

当攻击者完全了解水印的存在, 即知道水印的协方差矩阵并重构最佳攻击策略时, 我们分析水印参数对攻击者造成的估计误差协方差的影响。对于相同的 KL 散度阈值, 图 3 给出了在重构的最优线性欺诈攻击下, 不同水印参数的影响下, 远程估计误差协方差的演变。其中红线表示任意水印存在、攻击不存在时的估计误差方差, 黄线, 黑线, 蓝线和绿线分别表示当存在攻击而不加水印时的远程估计误差协方差。从图 3 可以看出, 水印参数越大, 估计的误差协方差越小。根据定理 2 的结论, 参数的值变大,

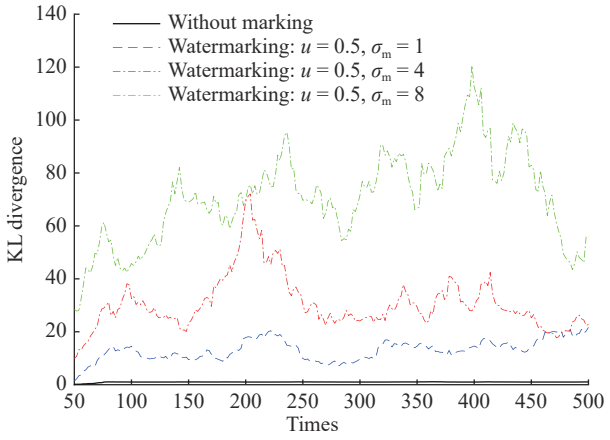


图 2 不同程度的水印参数对 KL 散度值的影响

Fig. 2 Influence of Different Degrees of Watermark Parameters on KL Divergence Value

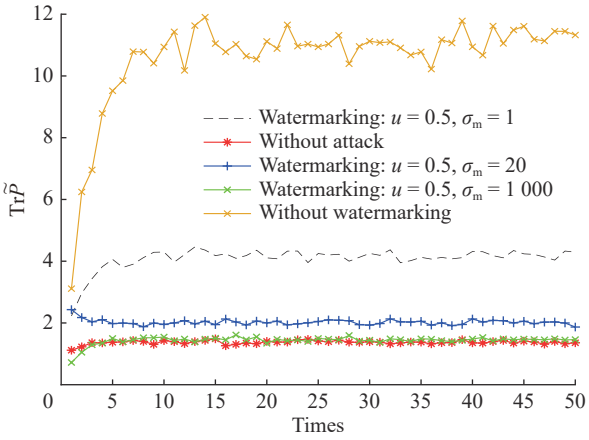


图 3 不同水印参数下,重构的最优欺诈攻击所能造成的估计误差方差的大小比较

Fig. 3 Comparison of estimated error variances caused by reconstructed optimal deception attacks under different watermark parameters

重构攻击可能引起的估计误差协方差变小。根据仿真结果,当 $u$=0.5, $\sigma_m$=1 000 时,估计误差方差的值趋向于攻击不存在时的估计误差方差的值。

## 4 结 论

本文使用水印加密和解密来帮助 KL 散度检测器在远程状态估计方案中识别线性欺骗攻击。与以前的结果不同,当不存在攻击时,我们使用加水印加密和解密的过程来恢复传输的数据,以避免影响系统的估计性能。针对不同的线性攻击场景,进一步分析了该方法的可靠性及其对估计误差的影响。当攻击者采取最佳线性欺骗攻击时,可以证明该攻击将触发警报,事实上在适当的参数选择下,KL 散度的值趋于正无穷大。当攻击者重构针对水印的最佳线性欺诈攻击时,通过选择适当的参数,这种攻击只能对系统性能产生有限的影响。总结上述情况并获得最佳参数选择,可以削弱攻击者的影响或使 KL 散度检测器尽快检测到攻击。同时本文提供了一个仿真示例,以验证本文的方法对线性欺骗攻击的影响。

## 参考文献:

[1] JIN X, HE W L, TANG Y, *et al*. Twisting-based finite-time consensus for euler-lagrange systems with an event-triggered strategy[J]. IEEE Transactions on Network Science and Engineering, 2020, 7(3): 1007-1018.

[2] TANG Y, ZHANG D D, HO D W C, *et al*. Event-based tracking control of mobile robot with denial-of-service attacks[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(9): 3300-3310.

[3] SHEN L J, WU Q D, SUN J T. Approximate controllability of fractional order semilinear systems with bounded delay[J]. Journal of Differential Equations, 2012, 252(11): 6163-6174.

[4] GHADIMI E, TEIXEIRA A, SHAMES I, *et al*. Optimal parameter selection for the alternating direction method of multipliers (ADMM): Quadratic problems[J]. IEEE Transaction on Automatic Control, 2014, 60(3): 644-658.

[5] LIANG G Q, WELLER S R, ZHAO J H, *et al*. The 2015 ukraine blackout: Implications for false data injection attacks[J]. IEEE Transaction on Power System, 2016, 32(4): 3317-3318.

[6] MO Y L, WEERAKKODY S, SINOPOLI B. Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs[J]. IEEE Control Systems Magazine, 2015, 35(1): 93-109.

[7] GUO Z Y, SHI D W, JOHANSSON K H, *et al*. Worst-case stealthy innovation-based linear attack on remote state estimation[J]. Automatica, 2018, 89(1): 117-124.

[8] BAI C Z, PASQUALETTI F, GUPTA V. Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs[J]. IEEE Transactions on Automatic Control, 2017, 62(12): 6641-6648.

[9] BAI C Z, PASQUALETTI F, GUPTA V. Security in stochastic control systems: Fundamental limitations and performance bounds[C]// American Control Conference, Chicago ZL USA: [s.n.], 2015, 195-200.

[10] MO L Y, CHABUKSWAR R, SINOPOLI B. Detecting integrity attacks on SCADA systems[J]. IEEE Transactions on Control Systems Technology, 2013, 22(4): 1396-1407.

[11] BAI C Z, PASQUALETTI F, GUPTA V. Data-injection attacks in stochastic control systems: detectability and performance tradeoffs[J]. Automatica, 2017, 82(1): 251-260.

[12] HUG G, GIAMPAPA J A. Vulnerability assessment of AC

state estimation with respect to false data injection cyber-attacks[J]. Automatica, 2017, 3(3): 1362-1370.

[13] GUO Z Y, SHI D W, JOHANSSON K H, *et al*. Optimal linear cyber-attack on remote state estimation[J]. IEEE Transactions on Control of Network Systems, 2016, 4(1): 4-13.

[14] ZHANG H, CHENG P, SHI L, *et al*. Optimal DoS attack scheduling in wireless networked control system[J]. IEEE Transactions on Control Systems Technology, 2015, 24(3): 843-852.

[15] ANDERSON B D, MOORE J B. Optimal filtering[M]. [S.L.]: Courier Corporation, 2012.

# Watermark Encryption for Linear Deception Attacks in Cyber Physical Systems

WANG Di, LI Fangfei, XU Siyao, LIU Changhong

（School of Science, East China University of Science and Technology, Shanghai 200237, China）

**Abstract:** A defense method for linear deception attacks using KL divergence detector to detect watermarks in cyber physics system (CPS) is proposed. It is well known that linear deception attacks can degrade the performance of the remote estimator without being detected by the KL divergence detector. In order to detect such attacks, computer-based encryption methods use watermarks to encrypt and decrypt data transmitted over a wireless network to protect the system. In the absence of attack, the decryption section can recover the transmitted data to ensure remote estimation performance. In the case of linear deception attacks, these data will be watermarked to expand the variance of the transmitted data, so that the data at the receiving end will no longer obey the original normal distribution. Due to the convex nature of KL divergence, the KL divergence value can be proved. It is enlarged to exceed the thresholds that have been tested by the detector, so that they can help the KL divergence detector to detect the attack. The watermark encryption method has been proved to help the KL divergence detector to detect attacks or to mitigate the effects of attacks in different situations, and further discusses how to select the appropriate value of the watermark parameters, which makes the system's protection performance more perfect. Finally, numerical simulations are given to further illustrate the limitations of the presence of watermarks on attackers, further highlighting the practical effects of parameter selection.

**Key words:** Cyber physical system; KL divergence; watermarking encryption; linear deception attack; system security