Detecting Feedback-path Delay Injection Attacks Using Interacting Multiple Model Filtering

Lovisa Eriksson, Torbjörn Wigren, Dave Zachariah and André Teixeira

I. INTRODUCTION

Time-delays are known to have a detrimental effect on feedback systems. In the context of networked cyber-physical systems, delays can be injected by malicious adversaries and detecting them early is an important challenge. Recently, it was proposed to conceal the delay attack in the feedback loop, for systems switching between open and closed loop settings [1], and detecting such instances is the aim of this paper. This paper proposes using Interacting Multiple Model (IMM) filtering to detect delay injection attacks in feedback control systems in an open loop setting. The detection scheme is formalised, and a theoretical analysis of the stationary distribution informs the choices of parameters. The method is applied to a cruise control application, and shows fast detection and a low false alarm probability. It is compared to previous work on delay detection [1], [2].

The contributions of the paper are as follows:

1) The IMM-based scheme for delay attack detection is formulated under a quickest change detection setting, leveraging the posterior probabilities as the test statistic for the detection rule;
2) Based on the underlying Hidden Markov Model (HMM), the theoretical analysis provides guidelines for tuning the hyperparameters of the HMM and the detection threshold;
3) The numerical example on a safety-critical application, cruise control, illustrates the applicability of the method and shows good detection performance.

II. PROBLEM FORMULATION

In this paper, the linear state space model

$$\begin{cases} x_{t+1} &= Ax_t + Bu_t + w_t \\ y_t &= Cx_{t-\delta(t)} + Du_{t-\delta(t)} + \nu_t, \end{cases} \tag{1}$$

is considered with independent white noise processes $w_t \sim \mathcal{N}(0, Q)$, $\nu_t \sim \mathcal{N}(0, R)$. The system model parameters A, B, C, D are known. However, the *observed* output y_t is subject to a possibly time-varying delay $\delta(t) \in \mathbb{N}$ that is *unknown*. It should be noted that the delay is only applied in the feedback path and thus only affects performance when the feedback loop is closed. This can be seen in Fig. 1, where the placement of the IMM detector is also visible.

The performance is evaluated by Average Detection Delay (ADD), Probability of False Alarm (PFA) and Average Run Length (ARL), which are all commonly used metrics in quickest change detection [3].

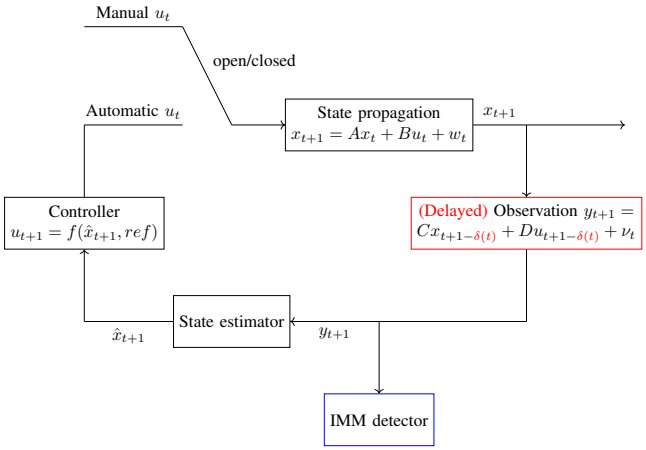

Fig. 1: Block diagram of the model. The attack only affects the observation.

III. DELAY DETECTION BY ESTIMATION

The delay detection problem is approached by considering the possible integer delays $\delta(t)$ as *modes* $s_t = 0, 1, 2, ..., D$ of the system. An attack is then understood as a change from the nominal mode to one of the delay modes. By estimating the probabilities of each mode at each time, a detection scheme is constructed. An attack detector is then defined by thresholding the posterior probability of the system being in the nominal mode, with the threshold a design parameter. To obtain the required probabilities, IMM filtering is applied.

A. Interacting Multiple Model (IMM)

The general idea of IMM is to apply Kalman filters for simultaneous estimation under several different hypotheses, or modes, of the system dynamics, with the system potentially switching between the modes. The potential of switching is handled by combining previous estimates based on the probabilities of having been in the mode. IMM is described in more details in [4] and in the full paper.

B. Construction of the Markov Model

The transition probabilities model how the delay modes may change. By focusing on the transitions between the nominal (zero delay) and attack (non-zero delay) modes, $s = 0$ and $s > 0$, respectively, $\mathbb{P}(s_t = i | s_{t-1} = j)$ is modelled by p_0, the probability of remaining in the nominal mode, and p_1, the probability of transitioning from an attacked to the nominal mode. Transitions between delayed modes are modeled by a uniform distribution. This formulation

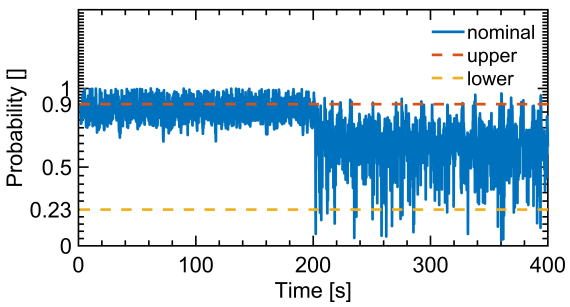

Fig. 2: A single trajectory of how the nominal probability change over time, using $\theta = 0.99$ and a delay attack of 0.2 seconds being introduced after 200 seconds. The upper and lower bound for reasonable threshold derived in the full paper are shown for reference.

circumvents excessive modelling requirements, as only two parameters are used.

The transition probabilities form a Markov chain. It can be shown that to achieve a desired stationary distribution $\begin{pmatrix}\bar{\pi} & 1-\bar{\pi}\end{pmatrix}^T$ the parameters should be chosen as

$$0 \leq p_1 = (1-p_0)\frac{\bar{\pi}^0}{1-\bar{\pi}^0} \leq 1 \tag{2}$$

where the probability of remaining in the nominal mode can be selected freely as

$$p_0 = (1-\theta)\cdot 1 + \theta \cdot \left(1 - \frac{\bar{\pi}^0}{1-\bar{\pi}^0}\right), \tag{3}$$

where $\theta \in [0,1]$. A small θ corresponds to slow movement between modes and vice versa, and $(1-\theta)$ can be interpreted as a temperature of the system, similar to that of the Boltzmann distribution.

Further, by analysing the stationary distribution under attack, a range for reasonable thresholds of the detector can be obtained. This range is visualised in Fig. 2.

IV. Delay Detection in Cruise Control

A. Simulation Results

The simulations are for a linearised cruise control system, for which the derivation can be found in [1]. See the full paper for details on model parameters and hyperparameter tuning.

All simulation results consider a delay attack defined as a step function, with $\delta(t) = 0$ until some attack time t^*, when a fixed delay is inserted.

In Figure 2, a single trajectory of the IMM filter estimate is shown. It is clear from the figure that the filter is quick to adjust the probabilities when the attack is introduced. A threshold of 0.4 is supported by this trajectory as well.

In Table I, the estimated ADD and PFA are shown after running the model for 500 runs for various injected delays. It should be noted that the two largest delays in this simulation, 1.1 and 1.2 seconds, is higher than any of the hypothesises used by the model, and the detector performs slightly worse for these.

TABLE I: Estimated ADD and PFA over 500 runs, for varied delays attacks inserted at a uniformly random time. In parenthesis, the standard deviation is reported.

Delay [s]	ADD [s]	PFA []
0.1	108($\pm$81.7)	0.0027($\pm$0.052)
0.2	3.67($\pm$2.25)	0.0036($\pm$0.060)
0.3	3.50($\pm$2.17)	0.0028($\pm$0.051)
0.4	3.47($\pm$2.08)	0.0026($\pm$0.051)
0.5	3.51($\pm$2.11)	0.0028($\pm$0.053)
0.6	3.47($\pm$2.11)	0.0024($\pm$0.049)
0.7	3.52($\pm$2.21)	0.0024($\pm$0.049)
0.8	3.48($\pm$2.20)	0.0024($\pm$0.049)
0.9	3.45($\pm$2.10)	0.0028($\pm$0.053)
1.0	3.42($\pm$2.06)	0.0016($\pm$0.040)
1.1	4.07($\pm$2.71)	0.0038($\pm$0.062)
1.2	4.79($\pm$3.45)	0.0024($\pm$0.049)
Average	12.38	0.0027
Average excl. 0.1	3.67	0.0027

B. Comparison with Existing Detection Schemes

Previous work on delay detection builds on identification of the system parameters. In [1], identification is applied to the same vehicle model as in this paper, but assuming unknown system dynamics. This gave a detection time exceeding 100 seconds on an example trajectory, which is significantly worse than the results for the IMM framework. They are not fully comparable, since the IMM approach assumes known model dynamics, but it shows that much better performance is possible when the dynamics are known.

It is harder to compare the current work to that of [2] since those experiments are not performed on the same application and do not use a fixed-threshold detector. However, using that method, it is possible to note a difference in the distribution after around 10 seconds, only slightly worse than the IMM method, but only if the delay is inserted momentarily. A gradually deployed attack is not noticeable since the method assumes a nominal mode for the previous time step so any uncertainty of previous modes are discarded. This issue should not be present in IMM, since the uncertainties of the mode from past time steps are carried over to future steps.

V. Future Work

Interesting directions for future work are to find theoretical bounds on the false alarm probability given the threshold, or to detect delay attacks in closed loop settings using the IMM approach.

References

[1] T. Wigren and A. Teixeira, "Feedback path delay attacks and detection" in *IEEE conf. on Decision and Control*, 2023, pp. 3864-3871.

[2] E. Korkmaz, M. Davis, A. Dolgikh, and V. Skormin, "Detection and mitigation of time delay injection attacks on industrial control systems with PLCs" in *Computer Network Security*, Cham.: Springer, 2017, pp. 62-74.

[3] V. V. Veeravalli and T. Banerjee, "Quickest change detection" in *Academic Press Library in Signal Processing: Volume 3*, Elsevier, 2014, pp. 209-255.

[4] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, "IMM estimator versus optimal estimator for hybrid systems", *Trans. Aerosp. Electron. Syst.*, vol. 41, no. 3, pp. 986-991, Jul. 2003.