Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Maestría en Ciencias en Electrónica y Telecomunicaciones con orientación en Instrumentación y Control

---

Detección de ataques en sistemas ciberfísicos

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Maestro en Ciencias

Presenta:

Jesus Eduardo Verdugo German

Ensenada, Baja California, México

2019

Tesis defendida por

Jesus Eduardo Verdugo German

y aprobada por el siguiente Comité

Dr. Jonatán Peña Ramírez
Codirector de tesis

Dr. Justin Ruths
Codirector de tesis

Dr. Joaquín Álvarez Gallegos

Dr. José Ricardo Cuesta García

Dr. Ubaldo Ruiz López

Dr. Daniel Sauceda Carvajal
Coordinador del Posgrado en Electrónica y Telecomunicaciones

Dra. Rufina Hernández Martínez
Directora de Estudios de Posgrado

Resumen de la tesis que presenta Jesus Eduardo Verdugo German como requisito parcial para la obtención del grado de Maestro en Ciencias en Electrónica y Telecomunicaciones con orientación en Instrumentación y Control.

Detección de ataques en sistemas ciberfísicos

Resumen aprobado por:

Dr. Jonatán Peña Ramírez	Dr. Justin Ruths
Codirector de tesis	Codirector de tesis

El término sistemas ciberfísicos se utiliza para referirse a sistemas cuyo funcionamiento depende de una estrecha integración de tecnologías de computación, comunicación y control para lograr cierto funcionamiento deseado. Ejemplos de sistemas ciberfísicos son una central nuclear, la red eléctrica, un vehículo autónomo, una línea de producción automatizada, por mencionar solo algunos. Los sistemas ciberfísicos (CPS por sus siglas en inglés) son sistemas dinámicos que se caracterizan por contener una parte física, por ejemplo un proceso industrial, un robot manipulador, un sistema mecánico, entre otros, y una parte virtual, la cual consiste en una serie de algoritmos de control, los cuales comúnmente son implementados mediante una computadora. Debido a que la interacción entre la capa física y la capa virtual en los CPS es generalmente indirecta, por ejemplo a través de una conexión alámbrica remota o inalámbrica, este tipo de sistemas son susceptibles de ser atacados. En esta tesis, se presenta una estrategia de detección de ataques o anomalías en sistemas lineales, utilizando conceptos y herramientas sobre Teoría del Control, la cual permite mejorar la seguridad del sistema. La estrategia propuesta está basada en el uso de un observador discontinuo y un filtro pasa bajas. Además, se presenta el diseño de un detector para sistemas conmutados el cual constituye un primer paso hacia la detección de ataques en sistemas con dinámica híbrida. Los resultados analíticos obtenidos se ilustran mediante simulaciones numéricas y se validan experimentalmente usando un sistema mecánico y un circuito electrónico.

Palabras clave: Sistema ciberfísico, ataque, algoritmo de detección

Abstract of the thesis presented by Jesus Eduardo Verdugo German as a partial requirement to obtain the Master of Science degree in Electronics and Telecomunications with orientation in Instrumentation and Control.

Attack detection for ciber-physical systems

Abstract approved by:

Dr. Jonatán Peña Ramírez	Dr. Justin Ruths
Thesis Co-Director	Thesis Co-Director

The concept of cyber-physical systems (CPS) refers to control systems that integrate different technologies. Examples of CPS are a nuclear plant, the power grid, an autonomous vehicle, an automatized production line, among other examples. A CPS system contains two parts: a physical part, for instance, an industrial process, a robot manipulator, a mechanical system, among others, and a virtual part, consisting of control algorithms which commonly are implemented in a computer. Since the interaction between the physical layer and the virtual layer in a CPS is generally indirect, for example via a remote wire or a wireless connection, the system is susceptible to external attacks.In this thesis, we presented an anomaly detection strategy for linear CPS, which is based on tools and techniques from Control Theory. Specifically, the proposed strategy is based on a discontinuous observer and a low pass filter. Furthermore, the thesis also presents the design of a detector for piece-wise linear systems, which constitutes a first step towards the detection of attacks or anomalies in hybrid CPS. The obtained analytic results are illustrated by means of computer simulations and experimentally validated by using a mechanical system and an electronic circuit.

Keywords: Cyber-physical systems, attack, detection algorithm

Dedicatoria

A mi familia, gracias por todo su amor, por siempre brindarme su apoyo y motivación.

Agradecimientos

Al Centro de Investigación Científica y de Educación Superior de Ensenada por aceptarme como estudiante y por las facilidades que otorga para realizar nuestros estudios.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de maestría. No. de becario:634277

A mis padres y mis hermanos, por su amor y apoyo incodicional, a pesar de estar en otra ciudad pero siempre estar al pendiente de mi.

Agradezco a los doctores: Jonatan Peña y Justin Ruths por aceptar ser mis codirectores de tesis. Gracias por sus consejos, observaciones la confianza para realizar este proyecto de tesis. Además agradezco que siempre buscaron la manera de brindarme la comodidad para desarrollarme en un gran ambiente de trabajo.

También agradezco a los doctores: Joaquín Alvarez, Ricardo Cuesta Y Ubaldo López, por aceptar ser parte del comité de tesis, y también por sus comentarios y sugerencias, las cuales fueron muy valiosas para nuestro trabajo.

De igual manera agradezco al ing. René Torres encargado del taller de electrónica, por haberme brindado el espacio y equipo del laboratorio. De la misma manera al doctor Ricardo Cuesta por su tiempo y su apoyo para llevar a cabo los experimentos de este proyecto de tesis.

A mis amigos más cercanos de CICESE a Hirata, Guillermo, Alberto y Willem (El holandés) que siempre me apoyaron en lo académico durante la maestría y sobre todo por brindarme su amistad sincera. También al grupo de musica donde me permitieron aprender de ustedes en lo personal y musical, en especial a mis buenos amigos Roilhi y Paola.

Tabla de contenido

Página

Tabla de contenido (continuación)

Lista de figuras

Figura Página

Lista de figuras (continuación)

Figura Página

Lista de figuras (continuación)

Figura Página

Lista de tablas

Tabla Página

Capítulo 1. Introducción

El término sistemas ciberfísicos se utiliza para referirse a sistemas cuyo funcionamiento depende de una estrecha integración de tecnologías de computación, comunicación y control para lograr ciertos niveles deseados de estabilidad, rendimiento, confiabilidad, robustez y eficiencia (Kim y Kumar, 2012). Ejemplos de sistemas ciberfísicos son una central nucleoeléctrica, la red eléctrica, un vehículo autónomo, una línea de producción automatizada, por mencionar solo algunos.

Los sistemas ciberfísicos (CPS por sus siglas en inglés) son sistemas dinámicos que se caracterizan por contener una parte física, por ejemplo un proceso industrial, un robot manipulador como el mostrado en la Figura 1, un sistema mecánico, entre otros, y una parte virtual, la cual consiste en una serie de algoritmos de control, los cuáles comúnmente son implementados mediante una computadora (Serpanos, 2018; Kim y Kumar, 2012). La Figura 2 muestra un centro virtual de control, el cual puede verse como la capa virtual de un CPS.

La capa física y la capa virtual intercambian información a través de sensores y actuadores. En particular, los sensores envían información del sistema físico (p.ej. posición, temperatura, presión, velocidad, etc.) a la capa virtual, en donde se encuentran implementados los algoritmos de control. Con la información recibida, se procesa la acción de control necesaria, la cual es enviada a la capa física del CPS en donde se encuentran los actuadores del sistema (Bangemann *et al.*, 2016; Sun *et al.*, 2014).

Figura 1. Un robot manipulador es un ejemplo del tipo de sistemas que se encuentran en la capa física de un CPS.

Figura 2. Un centro remoto de control es un ejemplo del tipo de sistemas que se encuentran en la capa virtual de un CPS.

Debido a que la interacción entre la capa física y la capa virtual en los CPS es generalmente indirecta, por ejemplo a través de una conexión alámbrica remota o inalámbrica, este tipo de sistemas son susceptibles de ser atacados (Cárdenas *et al.*, 2008; Koren, 2018; Ding *et al.*, 2018; Dibaji *et al.*, 2019).

De acuerdo con (Loukas, 2015) un ataque se puede definir como un asalto intencional a un sistema, aprovechando las vulnerabilidades y debilidades del mismo con el fin de afectar, de manera adversa, su operación y/o el rendimiento. Por otra parte, existen distintos tipos de ataques los cuales pueden clasificarse de la siguiente manera (Pasqualetti *et al.*, 2013):

- **Ataque sigiloso o engaño:** la función de este ataque es comprometer la integridad de los paquetes de datos o mediciones obtenidas de los sensores y/o de las señales de control enviadas a los actuadores sin provocar algún cambio significactivo en el sistema.
- **Negación de servicio:** los ataques de negación de servicio, en cambio comprometen la disponibilidad del recurso o de las variables medidas. Por ejemplo cuando se bloquea el canal de comunicación entre el sistema y la unidad de control (Cárdenas *et al.*, 2008).
- **Ataque de repetición:** los ataques de repetición se presentan en un sistema de control cuando los sensores son secuestrados, es decir, se registran las lecturas de las mediciones durante un cierto período de tiempo y estas mediciones se repiten mientras se inyecta una señal exógena provocando un daño en el sistema.

- **Dato falso:** De acuerdo con (Zhong *et al.*, 2018) este tipo de ataque puede manipular deliberadamente las mediciones para modificar y/o reemplazar las mediciones obtenidas de los sensores y es un ataque que puede ser difícil de identificar y puede crear un gran daño en el sistema. Una caracteristica de este ataque es que el atacante debe tener conocimiento absoluto del sistema y sus parámetros.

En la Figura 3 se muestra el diagrama a bloques de un sistema con ataque en la salida medida. La brecha que existe entre la etapa de medición de la variable deseada mediante un sensor y la etapa de control presenta cierta vulnerabilidad de ser atacada por algún intruso o adversario.

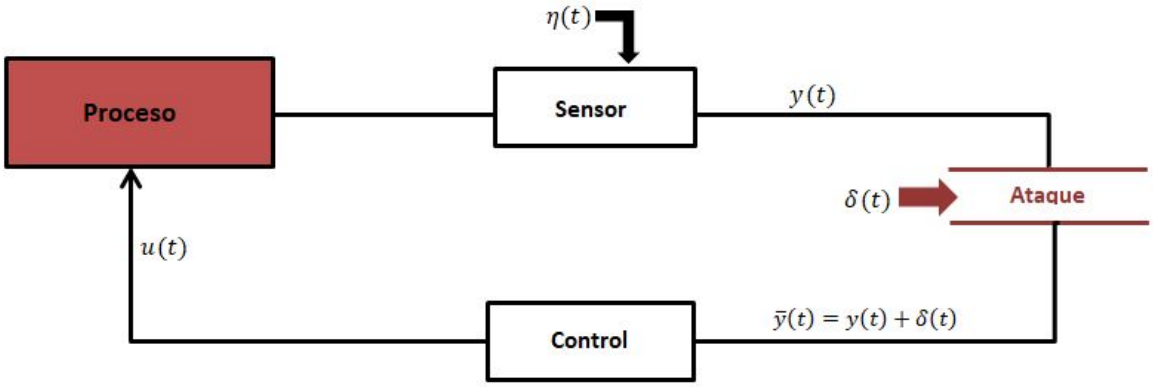

Figura 3. Diagrama a bloques de un sistema bajo ataque.

El creciente número de ciberataques a las industrias exige atención inmediata para proporcionar mecanismos más seguros para salvaguardar las empresas y minimizar los riesgos, ya que estos riesgos representan grandes pérdidas económicas e incluso pueden llegar a atentar contra la seguridad e integridad de las personas.

Un antecedente de este tipo ocurrió a mediados del 2010 cuando tuvo lugar uno de los ciberataques más conocidos, el del “gusano informático” Stuxnet, a los sistemas de control de las plantas nucleares iraníes (Shakarian *et al.*, 2013; Collins y McCombie, 2012; Wang *et al.*, 2012). Este ataque primero analizó y detectó redes Windows y sistemas informáticos vulnerables los cuales serían infectados. Una vez que se infiltró en los sistemas de computación, comenzó a replicarse continuamente y después dañó el software de los controladores lógicos programables (PLC), teniendo la capacidad de manipular las centrifugadoras de uranio hasta provocar el daño de ellas.

Un caso similar se presentó en Indonesia y en la India (Karnouskos, 2011). Estos virus permanecen inactivos y aprenden el comportamiento del sistema de tal manera que logran corromper la información crítica obtenida de los sensores, ocultándola y/o alterándola y finalmente provocando un fallo en la medición monitoreada (Dunaka y McMillin, 2017). Por otra parte, de acuerdo a las estadísticas del Consejo de Confiabilidad Eléctrica de América del Norte, en Estados Unidos han ocurrido 11 apagones en las redes eléctricas, los cuales han sido causados por anomalías en los sistemas cibernéticos (Zhang *et al.*, 2017).

Los ataques antes mencionados han mostrado que muchas de las medidas de seguridad sobre el entorno operativo, las capacidades tecnológias y los posibles análisis de riesgos, distan mucho de ser eficaces y no están a la altura de lo nuevos desafíos—en lo que a seguridad se refiere—que enfrentan los sistemas CPS.

En la actualidad existen varias propuestas para mejorar la seguridad en CPS, como lo son los métodos de detección de anomalías en las mediciones (Mo y Sinopoli, 2016),(Mo y Sinopoli, 2009) y (Fawzi *et al.*, 2011). Dentro de estos métodos se encuentra el método de detección Chi-cuadrado propuesto por (Murguia y Ruths, 2016b) y el método de suma acumulativa (CUSUM) propuesto por (Murguia y Ruths, 2016a), los cuales están basados en la estimación del estado del sistema mediante un filtro de Kalman y la construcción de una señal de residuo, la cual da una medida de la discrepancia entre la salida del sistema y la salida estimada. Posteriormente, esta señal de residuo se analiza con herramientas de probabilidad para definir un parámetro de umbral, de tal manera que si la norma de la señal de residuo excede dicho umbral, entonces se activa una alarma para indicar la presencia de un ataque. Otra técnica de detección de anomalías en CPS es la presentada en (Goh *et al.*, 2017), la cual se basa en el uso de redes neuronales recurrentes. Esta técnica, aplicable al caso en que se tienen múltiples sensores, no solo detecta la presencia de una anomalía sino que también es capaz de identificar el sensor que ha sido atacado.

Por otra parte, existen otras técnicas en las que se caracteriza a priori el comportamiento del sistema CPS sin ataque. En este caso, cualquier escenario en el que el comportamiento del sistema se desvía del comportamiento nominal se considera un ataque. Sin embargo, esta técnica presenta la desventaja de que es suceptible a una tasa alta de detección de falsos positivos (Mitchell y Chen, 2014).

1.1. Panorama de esta tesis

Esta tesis tiene como finalidad desarrollar una técnica de detección de ataques en sistemas CPS, usando conceptos y herramientas disponibles en la literatura sobre Teoría del Control, que permita reducir la vulnerabilidad del sistema. En particular, se diseñará e implementará un algoritmo de detección de ataques para sistemas ciberfísicos usando un observador discontinuo y un filtro pasa bajas. El diseño del observador está inspirado en un resultado previo (Alvarez *et al.*, 2009).

Además, este trabajo de tesis representa un primer paso hacia la detección de ataques en sistemas CPS con dinámica híbrida ya que, en la literatura, este problema aún no ha sido abordado.

1.1.1. Motivación

Como se mencionó anteriormente, los CPS son sistemas propensos a ser atacados debido a que la interacción entre la capa física y la capa virtual es indirecta. Los ataques a dichos sistemas no sólo pueden derivar en pérdidas económicas sino que pueden ocasionar catástrofes. Por consiguiente, es de suma importancia desarrollar métodos y estrategias de defensa que permitan blindar los sistemas CPS.

Una alternativa sería encriptar la información. Sin embargo, los ‘hackers’ han demostrado tener capacidades y habilidades para descifrar y manipular la información. Por lo tanto, es necesario utilizar métodos alternativos o complementarios que garanticen mayor confiabilidad.

Por ello, esta tesis propone el diseño de un detector que permite identificar ataques y/o anomalías en las mediciones de los sensores de un CPS y, en ciertos casos y bajo ciertas circunstancias, permite reconstruir un estimado del ataque o anomalía.

1.1.2. Objetivo general

El objetivo general de esta tesis es evidenciar la presencia de ataques externos en la salida medida de una clase particular de sistemas ciberfísicos por medio del desarrollo de una técnica de detección de ataques utilizando conceptos y herramientas de

la Teoría del Control.

1.1.3. Objetivos particulares

- Diseñar e implementar un algoritmo de detección de ataques para sistemas ciberfísicos lineales de segundo orden utilizando un observador robusto.
- Diseñar e implementar un algoritmo de detección de ataques en una clase particular de sistemas con dinámica híbrida, a saber sistemas conmutados.
- Validar experimentalmente los desarrollos teóricos y numéricos obtenidos. Dicha validación se hará en sistemas CPS, cuya capa física consiste en un sistema mecánico y un circuito eléctronico y la capa virtual es una PC conectada a una tarjeta de adquisición de datos.

1.1.4. Contribuciones

Las contribuciones de este trabajo se resumen a continuación.

1. Se desarrolló un nuevo detector de anomalías/ataques para sistemas lineales de segundo orden y se ha demostrado, de manera experimental, la efectividad y fiabilidad de este detector.
2. Se propuso la idea de filtrar la señal medida por sensor antes de aplicarla al detector de ataques y se mostró que haciendo esto, la sensibilidad del detector se incrementa. Esto es cierto tanto para el detector aquí desarrollado como para los detectores clásicos mencionados en el Capítulo 2 de esta tesis.
3. Se diseñó un nuevo detector para sistemas conmutados. Dado que, no se encontró en la literatura ningún detector para este tipo de sistemas, se cree que el detector aquí propuesto constituye un primer paso hacia la detección de ataques en sistemas con dinámica híbrida.

1.1.5. Productos generados

Se escribió con colaboración del Dr.Justin Ruths, Navid Hashemi y Dr.Jonatán Peña el artículo titulado:"Filtering Approaches for Dealing with Noise in Anomaly Detection", el cual será presentado en la conferencia internacional Conference on Decision and Control CDC 2019, a celebrarse en la ciudad de Niza, Francia.

Una copia de este artículo se encuentra en la sección de anexos de esta tesis. En el artículo se muestra la detección de ataques en sistemas lineales de segundo orden, usando un observador discontinuo en combinación con un filtro pasa bajas. El artículo está basado en el Capítulo 3 de esta tesis.

1.1.6. Estancias cortas de investigación

Se realizaron dos estancias cortas de investigación, una con duración una semana y otra de un mes, en la Universidad de Texas Campus Dallas, en el Departamento de Ingeniería Mecánica en el grupo de investigación del Dr. Justin Ruths. Durante dichas estancias se trabajó en el diseño de los experimentos reportados en el Capítulo 4 de esta tesis y se estudiaron algunos detectores clásicos mencionados en el Capítulo 2.

1.1.7. Estructura de esta tesis

Esta tesis está organizada de la siguiente manera.

En el Capítulo 2 está dedicado al estudio de detectores existentes en la literatura, tales como el detector CUSUM y Chi cuadrado, que son técnicas de deteccion clásicas.

Después, en el Capítulo 3 se presenta el diseño de un detector de ataques, basado en un observador discontinuo, aplicable a sistemas lineales de segundo orden. También, se discuten dos posibles modificaciones y/o extensiones del detector propuesto, las cuales son, filtración de la señal proveniente del sensor antes de ser aplicada al detector y uso de dos sensores (redundancia).

A continuación, en el Capítulo 4 se presenta un estudio experimental con el que se valida la efectividad del detector. Como casos de estudio se consideran un oscillador

mecánico y un circuito electrónico.

El Capítulo 5 está dedicado al diseño e implementación numérica de un detector de ataques para sistemas conmutados. Como caso de estudio se considera un sistema conmutado de dos tanques de agua.

Finalmente, el Capítulo 6 presenta las conclusiones generadas durante el desarrollo de esta tesis y en la sección de anexos se incluye el artículo de conferencia derivado de este trabajo.

Capítulo 2. Detectores clásicos de anomalías

En este capítulo se presentan algunos detectores disponibles en la literatura. En particular, se analiza el algoritmo Chi cuadrado y el algoritmo CUSUM, los cuales están basados en la estimación del estado mediante un filtro de Kalman, el cálculo de una señal de residuo y el uso de herramientas estocásticas.
Cabe notar que la información presentada en este capítulo no se utiliza en capítulos posteriores. La razón es porque los detectores aquí presentados son aplicables a sistemas estocásticos discretos, mientras que la estrategia de detección que se propondrá en esta tesis es para sistemas lineales continuos. Sin embargo, el propósito del capítulo es familiarizar al lector con las técnicas de detección disponibles en la literatura.

Además, algunos de los resultados obtenidos en el Capítulo 3 son aplicados a estos detectores para mejorar su efectividad.

2.1. Descripción del sistema

Considere el siguiente sistema discreto estocástico invariante en el tiempo descrito por

$$x(t_{k+1}) = Fx(t_k) + Gu(t_k) + \nu(t_k), \quad (1)$$

$$y(t_k) = Cx(t_k) + \eta(t_k), \quad (2)$$

donde $x \in \mathbb{R}^n$ es el estado, $y \in \mathbb{R}^m$ es la salida, y $u \in \mathbb{R}^l$ denota la entrada de control. Las matrices (F, C) son detectables y $G \in \mathbb{R}^{nxm}$ es la matriz de entrada. Por otra parte, $\nu \in \mathbb{R}^n$ representa el ruido Gaussiano de media cero y covarianza $R_1 \in \mathbb{R}^{nxn}$, y $\eta \in \mathbb{R}^m$ es el ruido en el sensor con matriz de covarianza $R_2 \in \mathbb{R}^{mxm}$. Durante los instantes de muestreo t_k, $k \in \mathbb{N}$, la salida $y(t_k)$ es muestreada y transmitida a un canal de comunicación. La salida $y(t_k)$ recibida se utiliza para calcular las acciones de control $u(t_k)$ que son retroalimentadas al sistema. Finalmente, se supone que el bucle de control completo se realiza instantáneamente, es decir el muestreo y la transmisión de la medición del sensor y la transmisión de llegada de la acción de control no experimentan retardos en el tiempo.

Cuando la señales medidas del sensor son reemplazadas o modificadas por un atacante, entonces la salida 'atacada' del sistema está dada por

$$\bar{y}(t_k) := y(t_k) + \delta(t_k) = Cx(t_k) + \eta(t_k) + \delta(t_k), \tag{3}$$

donde $\delta(t_k) \in \mathbb{R}^m$ representa el ataque al sensor.

En principio, la naturaleza o forma de $\delta(t_k)$ es desconocida y cualquier $\delta(t_k) \neq 0$ será considerado un ataque.

Con el propósito de simplificar la notación, se define lo siguiente: $x_k = x(t_k)$, $u_k = u(t_k)$, $\bar{y}_k = \bar{y}(t_k)$, $\eta_k = \eta(t_k)$,$\nu_k = \nu(t_k)$, y $\delta_k = \delta(t_k)$. Usando esta notación, es posible escribir el sistema (1)-(3) de la siguiente forma

$$\begin{aligned} x_{k+1} &= Fx_k + Gu_k + \nu_k, \\ \bar{y}_k &= Cx_k + \eta_k + \delta_k. \end{aligned} \tag{4}$$

Para la estimación del estado del sistema se utiliza un filtro de Kalman, el cual está dado por la ecuación

$$\hat{x}_{k+1} = F\hat{x}_k + Gu_k + L(\bar{y}_k - C\hat{x}_k), \tag{5}$$

donde $\hat{x}_k \in \mathbb{R}^n$ es el estado estimado y la ganancia L del estimador está dada por (Murguia y Ruths, 2016a)

$$L = (FPF^T)(R_2 + CPC^T)^{-1}, \tag{6}$$

donde la matríz $(R_2 + CPC^T)^{-1}$ es definida positiva y la matriz de covarianza P es la solución de la siguiente ecuación algebraica de Ricatti

$$FPF^T - P + R_1 = FPC^T(R_2 + CPC^T)^{-1}CPF^T. \tag{7}$$

A continuación, se define el residuo

$$r_k = \bar{y}_k - C\hat{x}_k = Ce_k + \eta_k + \delta_k, \tag{8}$$

donde $e_k = x_k - \hat{x}_k$. Nótese que el residuo r_k refleja la discrepancia entre la medición $\bar{y}_k$ proveniente del sensor, la cual potencialmente puede estar atacada, y la salida estimada por el filtro de Kalman. Un residuo de cero significa que ambas salidas son iguales. Cuando no hay ataques en el sistema, la media del residuo está dada

$$E[r_{k+1}] = CE[e_{k+1}] + E[\eta_{k+1}] = 0, \tag{9}$$

y la covarianza es

$$\Sigma = E[r_{k+1} r_{k+1}^T] = CPC^T + R_2. \tag{10}$$

2.2. Algoritmo de detección CUSUM

El metodo de detección Cusum está basado en la distancia de medición, la cual se define como la desviación que hay del estimador con el sistema físico. Dicha distancia de medición está definida por la siguiente ecuación

$$z_k = |r_k| = |Ce_k + \eta_k + \delta_k|. \tag{11}$$

Cuando el sistema se encuentra libre de ataque, es decir cuando $\delta = 0$, la distancia de medición z_k sigue una distribución media normal con

$$E[|r_k|] = \frac{\sqrt{2}}{\sqrt{\pi}}\sigma, \tag{12}$$

donde σ denota la desviación estandar.

A continuación se describe el procedimiento de CUSUM en las siguientes líneas.

Cusum: $S_1 = 0$,

$S_k = max(0, S_{k-1} + z_k - b)$, sí $S_{k-1} \leq \tau$

$S_k = 0$, y $\tilde{k} = k - 1$, sí $S_{k-1} > \tau$

Parámetros de diseño: umbral $\tau \in \mathbb{R}^+$ y valor de sesgo $b \in \mathbb{R}^+$.

Salida: Tiempo de alarmas $\tilde{k}$.

La idea de este detector es que la secuencia S_k acumula la distancia de medida z_k y las alarmas son activadas cuando S_k supera el valor del umbral τ. Sin embargo la secuencia S_k del CUSUM se reinicia a cero cada vez que S_k toma un valor negativo o supera el valor del umbral τ. El procedimiento para seleccionar el valor de τ se presenta en (Brook y Evans, 1972) y (Murguia y Ruths, 2016a). Por otra parte, la elección del valor de sesgo b se realiza en base al siguiente resultado.

Teorema 1 Considere el sistema discreto (4) y el filtro de Kalman (5)-(7). A continuación, suponga que el sistema esta libre de ataque, es decir, $\delta_k = 0$. Considere la suma acumulada S_k con sesgo $b \in \mathbb{R}^+$ y valor de umbral $\tau \in \mathbb{R}^+$ siendo determinada por la distancia de medición $z_k = r_k^T \Sigma^{-1} r_k$, $k \in \mathbb{N}$, con la secuencia de residuo $r_k \sim \mathcal{N}(0, \Sigma)$. Entonces, para $b > \bar{b} =: \sigma\sqrt{\frac{2}{\pi}}$ la secuencia S_k está acotada.

La prueba del Teorema 1 se presenta en (Murguia y Ruths, 2016b).

Nótese que si el valor del parámetro b es seleccionado por arriba de $\bar{b}$, entonces la secuencia S_k no diverge.

2.2.1. Controlador del sistema

Sea el control por retroalimentación de la salida

$$u_k = K\hat{x}_k, \tag{13}$$

donde $\hat{x}_k \in \mathbb{R}^n$ es el estado del filtro de Kalman y $K \in \mathbb{R}^{lxn}$ es la matriz de control, y el par (F, G) es estabilizable. Considerando el sistema en lazo cerrado y siendo el error de estimación $e_k = x_k - \hat{x}_k$, se tiene el sistema acoplado

$$\begin{cases} x_{k+1} = (F + GK)x_k - GKe_k + \nu_k, \\ e_{k+1} = (F - LC)e_k - L\delta_k - L\eta_k + \nu_k, \end{cases} \tag{14}$$

de tal manera que el ataque δ_k afecta directamente la dinámica del error de estimación. La presencia de los ataques se da por la interconexión del controlador al sistema.

2.2.2. Ataque sigiloso

En este tipo de ataque se considera que el atacante es capaz de inducir al sistema un ataque de manera muy sigilosa, para lo cual requiere tener conocimiento perfecto de la dinámica del sistema y también de las mediciones y del filtro de Kalman. De tal manera puede comprometer la medición del sensor en cualquier instante de tiempo mediante las secuencias δ_k, las cuales provocan cambios en la dinámica del sistema sin ser detectadas por el método de detección.

En este caso, la secuencia del CUSUM en términos del error de estimación está dada por

$$S_k = max(0, S_{k-1} + |Ce_k + \eta_k + \delta_k| - b), \tag{15}$$

y el ataque δ_k a implementar en el sistema está diseñado de la siguiente manera

$$\delta_k = \begin{cases} \tau + b - Ce_k - \eta_k - S_{k-1}, & k = k^*, \\ b - Ce_k - \eta_k, & k > k^*, \end{cases} \tag{16}$$

donde k^* es el instante de muestreo en que se activa el ataque. Cuando el sistema en lazo cerrado está bajo la secuencia de ataque (16) el esperado del estado del sistema, y el valor esperado del sistema del error de estimación estan dados por

$$\begin{aligned} E[x_{k+1}] &= (F + GK)E[x_k] - GKE[e_k], \\ E[e_{k+1}] &= FE[x_k] - Lb. \end{aligned} \tag{17}$$

Proposición 1

Considere el proceso (4), el filtro de Kalman (5)-(7), el controlador (13) y el algoritmo de CUSUM y suponga que el sensor es atacado por la secuencia de ataque sigiloso (16). Entonces, si se satisface que la densidad radial $\rho[F] < 1$, se tiene que $\lim_{k\to\infty} |E[x_k]| = \gamma$, donde

$$\gamma := |(I - F - GK)^{-1} GK(I - F)^{-1} Lb|. \tag{18}$$

La prueba de esta proposición se presenta en (Murguia y Ruths, 2016b).

2.2.3. Ejemplo numérico

A continuación se presenta un ejemplo numérico que ilustra el funcionamiento del detector CUSUM.

Considere el sistema (4) con matrices

$$F = \begin{bmatrix} 0.84 & 0.23 \\ -0.47 & 0.12 \end{bmatrix}, \quad G = \begin{bmatrix} 0.07 \\ 0.23 \end{bmatrix}, \tag{19}$$

y suponga que el vector de salida del sistema está dado por

$$C = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{20}$$

Además, considerese que el sensor está afectado por ruido Gaussiano η_k y la planta está afectada por el ruido Gaussinao ν_k ambos de media cero y con matrices de covarianza

$$R_1 = \begin{bmatrix} 0.45 & -0.11 \\ -0.11 & 0.20 \end{bmatrix}, \quad R_2 = 1, \tag{21}$$

respectivamente. Mediante la solución de la ecuación de Ricatti (7) se obtener la matriz de covarianza mínima P, resultando en

$$P = \begin{bmatrix} 0.70 & -0.27 \\ -0.27 & 0.32 \end{bmatrix}. \tag{22}$$

En consecuencia, la matriz de ganancia del estimador del filtro de Kalman L (6) es

$$L = \begin{bmatrix} 0.31 \\ -0.21 \end{bmatrix}, \tag{23}$$

y la covarianza del residuo sin ataque está dada por

$$\Sigma = CPC^T + R_2 = 1.70. \tag{24}$$

Finalmente, suponga que se implementa el control por retroalimentación dado en (13) con

$$K = (-1.85 \ -0.96). \tag{25}$$

Para la sintonización del detector Cusum se consideran los parámetros de umbral $\tau = 1.1801$ y de sesgo $b = 1.1985$.

Los resultados obtenidos se muestran en la Figura (4), permiten puede observar cómo el estado estimado por el filtro del Kalman presenta media cero cuando no está atacado. En la figura (5) se muestra cómo el residuo sigue una distribución normal al no ser afectado por un ataque.

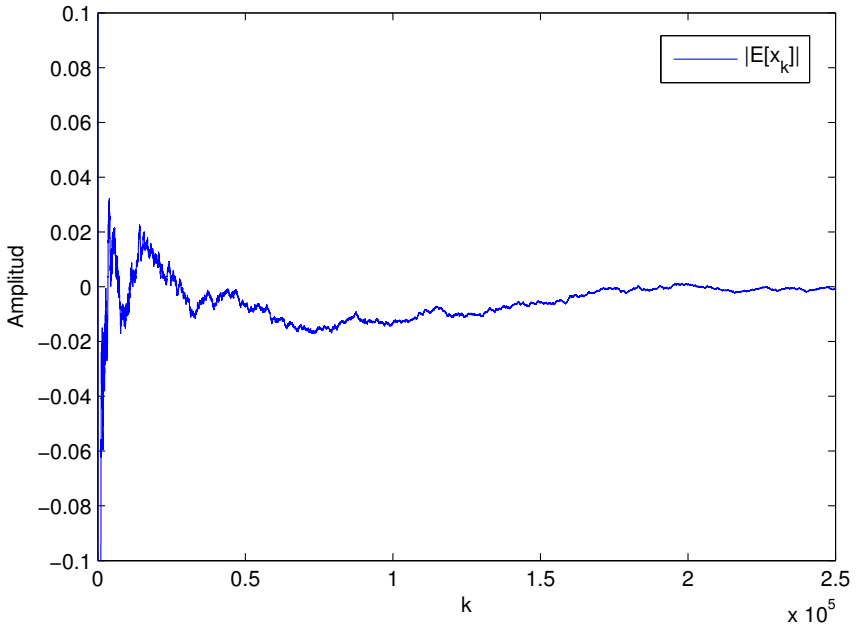

Figura 4. Evolución del valor esperado $|E[x_k]|$ con $S_k = 0$.

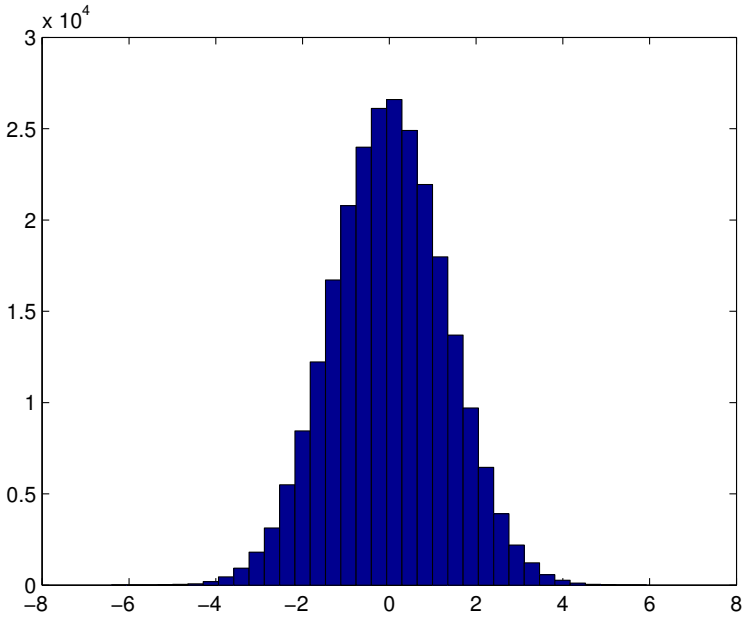

Figura 5. Distribución normal del residuo con $S_k = 0$.

Por otra parte, la Figura 6 muestra la evolución de la secuencia S_k del algoritmo CUSUM para diferentes valores del parámetro b. Como puede verse en la figura, si $b > \bar{b}$, la secuencia está acotada, tal y como se esperaba de acuerdo al Teorema 1. Por otro lado, los resultados presentados en la figura muestran que si $b < \bar{b}$ la secuencia S_k diverge.

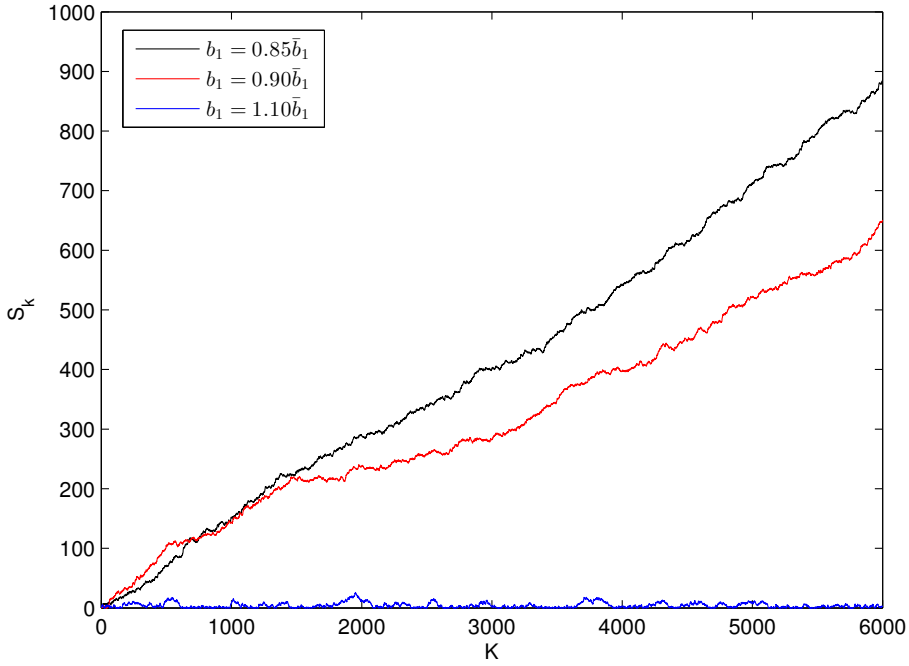

Figura 6. Evolución del CUSUM con distintos valores de b.

A continuación se considera el ataque de la forma (16) el cual es inducido en el instante de muestreo $k = 2 \times 10^4$, considerando los valores de τ y b dados anteriormente. En la Figura 7 se puede observar la degradación del estado estimado debido a que el valor de la media es distinto de cero a partir de que el sistema ha sido atacado. También se observa que la media del estado estimado converge al valor de γ predicho por la Proposición 1.

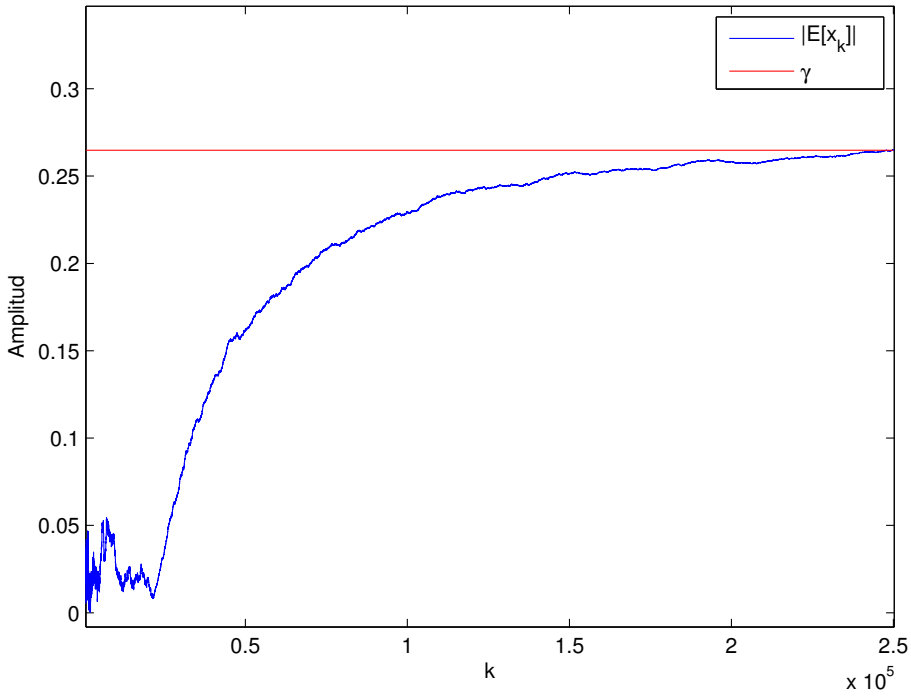

Figura 7. Degradación del valor esperado $|E[x_k]|$ cuando $S_k \neq 0$ (bajo ataque).

2.3. Algoritmo de detección Chi cuadrado

El detector Chi cuadrado es otro detector basado en la estimación del estado mediante el filtro de Kalman (4)-(7) descrito anteriormente. El rol de este detector es crear una variable no negativa aleatoria a partir del residuo r_k, ver (8), la cual se puede comparar fácilmente con un umbral α.

El detector Chi cuadrado introduce una medida de distancia definida de la siguiente manera

$$z_k = r_k^T \Sigma^{-1} r_k, \tag{26}$$

donde Σ es una matriz de covarianza.
El algoritmo de detección Chi cuadrado está definido de la siguiente manera .

Algoritmo Chi cuadrado:

if $z_k = r_k^T \Sigma^{-1} r_k > \alpha, \quad \tilde{k} = k.$

Parámetro de diseño: umbral $\alpha \in \mathbb{R}_{>0}$.

Salida: Tiempo alarma (s) $\tilde{k}$.

La idea de este algoritmo es que si z_k excede el valor del umbral α las alarmas son activadas. El procedimiento para seleccionar el valor de umbral α se presenta en (Murguia y Ruths, 2016b).

2.3.1. Ataque sigiloso

En el caso particular que el atacante tiene acceso a la salida $y_k = Cx_k + \eta_k$ y además tiene conocimiento perfecto del filtro de Kalman, el atacante puede inducir una secuencia δ_k la cual provoque un daño al sistema. En este caso, la distancia (26), escrita en términos del error de estimación y del ataque está dada por

$$z_k = (Ce_k + \eta_k + \delta_k)^T \Sigma^{-1} (Ce_k + \eta_k + \delta_k). \tag{27}$$

donde δ_k es un ataque sigiloso descrito por

$$\delta_k = -Ce_k - \eta_k + \Sigma^{\frac{1}{2}} \bar{\alpha} \longrightarrow z_k = \alpha, \tag{28}$$

con $\bar{\alpha}$

$$\bar{\alpha} := col(\sqrt{\alpha}, 0, \ldots, 0). \tag{29}$$

Esta secuencia de ataque δ_k mantiene z_k en el valor de umbral α, por lo que este tipo de ataque es indetectable.

Cuando el sistema en lazo cerrado está bajo la secuencia de ataque (28) el valor

esperado del estado del sistema y del error de estimación están dados por

$$E[x_{k+1}] = (F + GK)E[x_k] - GKE[e_k], \tag{30}$$
$$E[e_{k+1}] = FE[e_k] - L\Sigma^{\frac{1}{2}}\bar{\alpha},$$

respectivamente. El valor de $E[x_k]$ obtenido de (30) da una indicación de la degradación en el estado producida por el ataque.

2.3.2. Caso numérico

A continuación se presenta un ejemplo numérico para ilustrar el funcionamiento del detector Chi cuadrado. Sea el sistema (4) con matrices F y G dadas en (19). La salida del sistema está dada por

$$C = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{31}$$

El ruido en el sensor η_k y el ruido en la planta ν_k son ruidos Gaussianos de media cero con matrices de covarianza R_1 dada en (21) y

$$R_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{32}$$

respectivamente. Resolviendo la ecuación de Ricatti (7) para obtener la matriz de covarianza mínima P se obtiene

$$P = \begin{bmatrix} 0.70 & -0.27 \\ -0.27 & 0.32 \end{bmatrix}, \tag{33}$$

y la matriz L del filtro de Kalman es la siguiente

$$L = \begin{bmatrix} 0.25 & 0.17 \\ -0.18 & -0.07 \end{bmatrix}. \tag{34}$$

Por otra parte, la covarianza del residuo sin ataque está dada

$$\Sigma = CPC^T + R_2 = \begin{bmatrix} 1.65 & 0.40 \\ 0.40 & 1.46 \end{bmatrix}, \tag{35}$$

Y la matriz de ganancias del controlador (13) está dada por

$$K = (-1.85 \ \ -0.96). \tag{36}$$

Finalmente, en el algoritmo Chi cuadrado se considera un valor de $\alpha = 4.6051$ (Murguia y Ruths, 2016b).

Los resultados obtenidos se muestran en las Figuras (8) y (9).

Cuando la salida del sistema está libre de ataques, el valor esperado del estado se mantiene alrededor de cero, tal y como se muestra en la Figura (8).

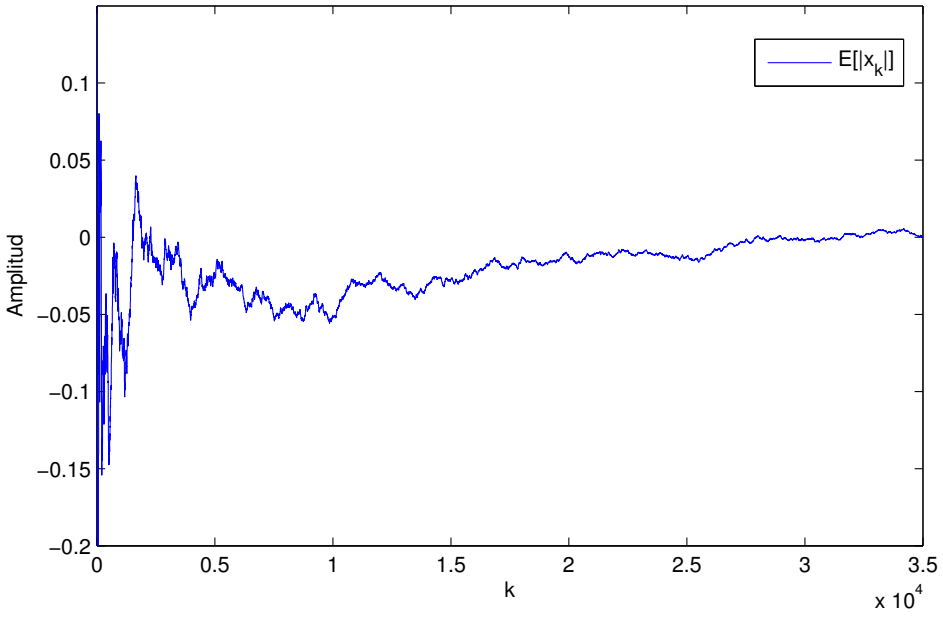

Figura 8. Evolución de la media estado sin ataque.

Sin embargo, cuando se aplica el ataque (28), el cual es iniciado en el instante de muestreo $k = 5 \times 10^3$, el valor esperado del estado se degrada como se muestra en la Figura (9).

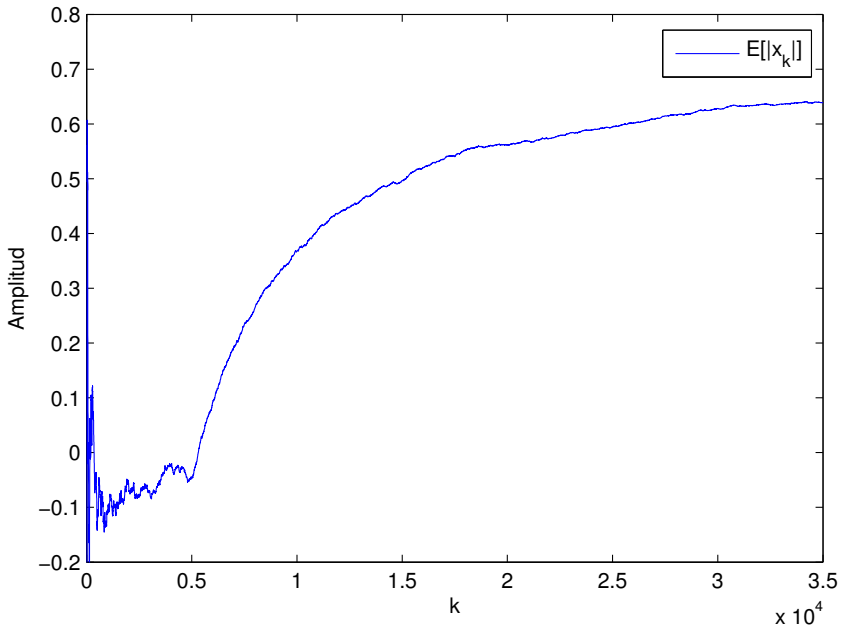

Figura 9. Degradación de la estado estimado debido a un ataque sigiloso.

Capítulo 3. Detector de ataques en CPS lineales

En este capítulo se presenta el diseño de un detector de ataques o anomalías en la salida medida, aplicable a sistemas lineales de segundo orden. El diseño del detector está basado en un observador discontinuo. La efectividad del detector para diferentes ataques y bajo diferentes circunstancias se investiga por medio de simulaciones numéricas.

Además, se discuten dos posibles extensiones del detector aquí propuesto: una de ellas consiste en filtrar la salida del sensor antes de aplicarla al detector y otra posible extensión en la que se usan dos sensores, es decir, para el caso en que se tiene redundancia en la medición.

3.1. Descripción del sistema y planteamiento del problema

Sea el sistema lineal de segundo orden

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{37}$$

$$y(t) = Cx(t), \tag{38}$$

donde $x(t) \in \mathbb{R}^2$ es el estado del sistema, $u(t) \in \mathbb{R}$ es la entrada de control, $y(t) \in \mathbb{R}$ es la salida del sistema y las matrices A, B y C dadas por

$$A = \begin{bmatrix} 0 & 1 \\ -a & -b \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{39}$$

con $a, b \in \mathbb{R}^+$.

Planteamiento del problema

Considere el sistema (37)-(39), y suponga que la salida y es atacada por un intruso o 'hacker' por medio de la señal $\delta(t) \in \mathcal{C}^2$, la cual es dos veces continuamente differenciable, de tal manera que la salida atacada del sistema está dada por

$$\bar{y}(t) = y(t) + \delta(t) = Cx(t) + \delta(t). \tag{40}$$

Entonces surgen las siguientes preguntas, ¿es posible detectar que la salida ha sido atacada? o más aún ¿es posible determinar la naturaleza de $\delta(t)$? Esta tesis aborda estas preguntas y propone posibles soluciones por medio del diseño de un detector para identificar que la salida original del sistema ha sido atacada.

La solución aquí propuesta para identificar el ataque $\delta(t)$ se muestra de manera esquemática en el diagrama de bloques de la Figura 10. En dicha figura se muestra el procedimiento para identificar la presencia del ataque $\delta(t)$. Como primer paso, la salida atacada es aplicada a un observador robusto que contiene un término discontinuo. Después, el término discontinuo del observador es aplicado a un filtro pasabajas, el cual generará una señal de residuo. Finalmente, se usa un algoritmo de detección el cual tiene como entrada la salida del filtro, Este detector es caracterizado y sintonizado de tal manera que se activan alarmas cuando la salida del sistema está sujeta al ataque $\delta(t)$.

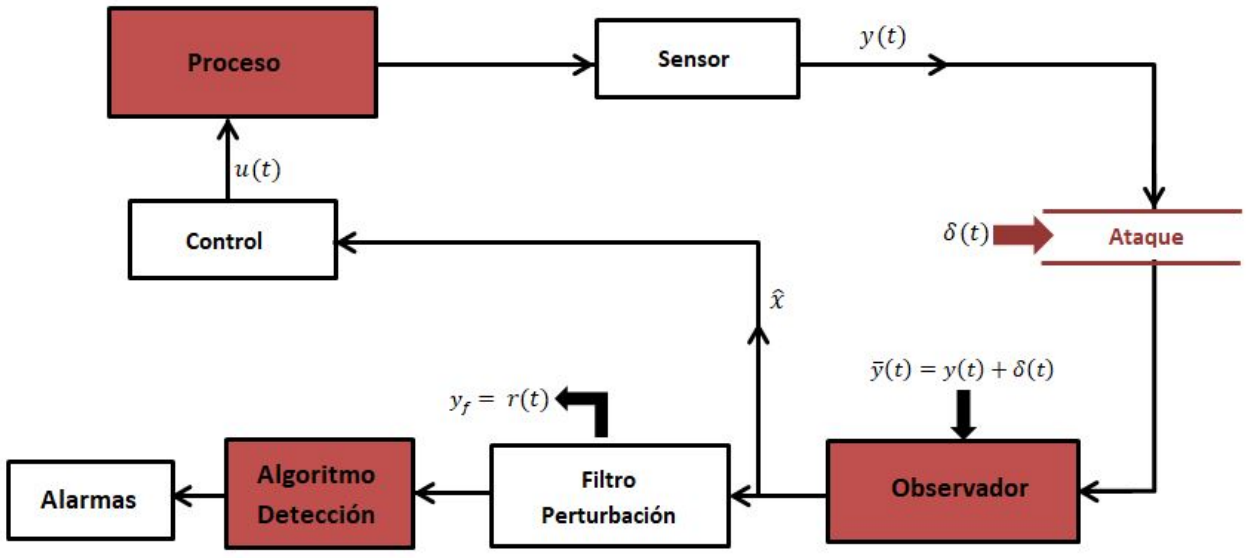

Figura 10. Diagrama a bloques del detector de ataques.

A continuación se describe en detalle el observador, el filtro y el algoritmo de detección propuesto.

3.2. Observador discontinuo

Para reconstruir el estado del sistema (37) y (72), se utiliza el siguiente observador discontinuo, el cual fue propuesto por (Rosas *et al.*, 2007) y originalmente fue diseñado no sólo para reconstruir el vector de estados en sistemas Lagrangianos de segundo orden (x_1:posición, x_2: velocidad) sino también para identificar perturbaciones acopladas en los sistemas. El modelo del observador es

$$\dot{\hat{x}} = A\hat{x} + Bu + \Gamma e_1 + Bc_3 \text{sign}(\bar{y} - \hat{y}), \tag{41}$$

$$\hat{y} = C\hat{x}, \tag{42}$$

donde $\hat{x} = [\hat{x}_1,\ \hat{x}_2]^T$, $e_1 = (\bar{y} - \hat{y}) = x_1 - \hat{x}_1 + \delta$, $\hat{x}_i \in \mathbb{R}$, la matriz A y los vectores B y C están definidos en (39). Por otra parte, la matriz Γ está dada por

$$\Gamma = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \tag{43}$$

y, de acuerdo a (Rosas *et al.*, 2007), las ganancias positivas c_i, $i = 1, 2, 3$ se pueden elegir de la siguiente manera.

$$c_1 > 0, \quad c_2 > 0, \quad c_3 > 2\lambda_{max}(P)\sqrt{\frac{\lambda_{max}(P)}{\lambda_{min}(P)}}\left(\frac{c_2\rho}{\theta}\right), \tag{44}$$

donde $P \in \mathbb{R}^{2\times 2}$ es la solución de la equación de Lyapunov $A^TP + PA = -I$, con $A \in \mathbb{R}^{2\times 2}$ dada en (39), $I \in \mathbb{R}^{2\times 2}$ es la matriz identidad, $\rho \in \mathbb{R}^+$ es el límite superior en las perturbaciones presentes en el sistema y θ es un parámetro positivo que satisface $0 < \theta < 1$.

3.2.1. Análisis de la dinámica del error de observación

Como primer paso, se definen los errores de observación

$$e_1 = \bar{y} - \hat{y}, \tag{45}$$

$$e_2 = x_2 - \hat{x}_2 + \dot{\delta} - c_1 e_1. \tag{46}$$

Por consiguiente, la dinámica del error de observación está descrita por

$$\dot{e}_1 = x_2 + \dot{\delta} - \hat{x}_2 - c_1 e_1, \tag{47}$$

$$\dot{e}_2 = -(a + bc_1 + c_2)e_1 - (b + c_1)e_2 - c_3 \text{sign}(e_1) + \ddot{\delta} + a\delta + b\dot{\delta}. \tag{48}$$

Reemplazando (45)-(46) en (47)-(48) se obtiene

$$\dot{e}_1 = e_2, \tag{49}$$

$$\dot{e}_2 = -k_1 e_1 - k_2 e_2 - c_3 \text{sign}(e_1) + \xi. \tag{50}$$

con

$$k_1 = (a + bc_1 + c_2), \quad k_2 = (b + c_1), \quad \text{y} \quad \xi = \ddot{\delta} + a\delta + b\dot{\delta}. \tag{51}$$

Por lo tanto, se puede concluir que si c_3 se escoge de acuerdo con (44), con $\rho := |\xi|_{max}$ entonces la dinámica del error converge a la superficie de discontinuidad dada por

$$e_1 = 0. \tag{52}$$

Para analizar el comportamiento del sistema en la superficie de discontinuidad, se tiene,

$$\ddot{e}_1 = -k_1 e_1 - k_2 e_2 - c_3 \text{sign}(e_1) + \xi, \tag{53}$$

y en el modo deslizante, $e_1 = \dot{e}_1 = \ddot{e}_1 = 0$. De acuerdo al método de control equivalente,

$$-\overline{c_3 \text{sign}(e_1)} + \xi = 0. \tag{54}$$

donde la barra superior en el término discontinuo significa filtrado.

Dado que en la superficie de discontinuidad la ecuación (52) se satisface, se tiene que (53) se reduce a

$$\overline{c_3 \text{sign}(e_1)} = \xi = \ddot{\delta} + a\delta + b\dot{\delta}. \tag{55}$$

Observación 3.1 *Nótese que si el ataque es cualquier función suave, es decir suficientemente diferenciable, o al menos dos veces diferenciable, entonces al filtrar el término discontinuo del observador, se obtiene 'la perturbación' que el ataque induce en el sistema.*

Además, de la ecuación (55) se tiene que si el ataque $\delta(t)$ es un ataque constante, entonces filtrando el término discontinuo es posible reconstruir el ataque si el parámetro α es conocido.

Observación 3.2 *Cuando los errores de observación convergen a cero, es decir $e_1 = e_2 = 0$, de la ecuación (46) se tiene que*

$$x_2 - \hat{x}_2 = -\dot{\delta}. \tag{56}$$

Este resultado lleva a las siguientes observaciones para el caso en que la salida del sensor no está influenciada por ruido Gaussiano, es decir para el caso en que $\eta(t) = 0$, en la ecuación (38):

1. *Si el ataque $\delta(t)$ es variante en el tiempo el observador puede reconstruir el estado x_1 pero no puede reconstruir el estado x_2, ya que como se muestra en (56), habrá un error de observación para el estado x_2.*

2. *Si el ataque $\delta(t)$ es constante entonces el observador puede reconstruir el vector de estado completo.*

3. *De la ecuación (56), se deduce que si el ataque es variante en el tiempo, entonces podemos reconstruir el ataque mediante la integración de e_2, es decir*

$$\int_0^t (x_2 - \hat{x}_2) d\tau = -\delta(t) + \delta(0). \tag{57}$$

   Sin embargo, cabe mencionar que, en la práctica, dicho método de reconstrucción puede no converger, debido al ruido de medición y además, si no se elige la condición inicial en el integrador adecuadamente, entonces habrá un offset entre el ataque real y el ataque reconstruído.

3.3. Filtro pasa bajas

En la sección anteriorse mostró que mediante la filtración del término discontinuo $c_3\text{sign}(e_1)$ es posible obtener información de las anomalías o ataques presentes en el sistema, véase la ecuación (55).

Por lo tanto, en esta sección, se propone el uso de un filtro Butterworth de segundo orden para filtrar el término discontinuo presente en el observador. La estructura del filtro está dada de la siguiente manera

$$\begin{aligned} \dot{x}_f &= A_f x_f + B_f c_3 \text{sign}(e_1), \\ y_f &= C_f x_f, \end{aligned} \tag{58}$$

donde $x_f = [x_{f1}\ x_{f2}]^T$, $x_{f1}, x_{f2} \in \mathbb{R}$ y

$$A_f = \begin{bmatrix} 0 & 1 \\ -\omega_c^2 & -1.4142\omega_c \end{bmatrix}, \quad B_f = \begin{bmatrix} 0 \\ \omega_c^2 \end{bmatrix}, \tag{59}$$

$$C_f = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{60}$$

donde ω_c es la frequencia de corte del filtro. Si ω_c se escoge de manera adecuada, es posible minimizar el retraso de fase tal que

$$\lim_{t\to\infty} y_f \approx \ddot{\delta} + a\delta + b\dot{\delta}. \tag{61}$$

3.4. Algoritmo de detección

Finalmente, con la información obtenida del observador discontinuo y del filtro, se diseña el siguiente algoritmo de detección.

$$\begin{cases} |y_f(t)| \leq \alpha_f \longrightarrow \text{no alarma} \quad \longrightarrow \quad alarma = 0, \\ \\ |y_f(t)| > \alpha_f \quad \longrightarrow \text{alarma} \quad \longrightarrow \quad alarma = 1. \end{cases} \tag{62}$$

Nótese que idealmente, la salida del filtro debe ser igual a cero cuando no hay

ataques y por consiguiente α_f debe escogerse como $\alpha_f = 0$, en el detector (62). Sin embargo, debido a la alta frecuencia en el término discontinuo, la salida del filtro no es exactamente cero, por lo que, para evitar falsas alarmas, el valor de α_f debe ser diferente de cero aún si no hay ataque en la salida del sensor.

Por ahora, no se cuenta con un procedimiento formal para determinar el valor del umbral superior α_f en el detector (62). Sin embargo, se tiene el siguiente procedimiento empírico:

1. Se asume que inicialmente el sistema no está atacado, es decir $\delta(t) = 0$ en (38).
2. A continuación se mide la salida del filtro y_f, después de la respuesta transitoria del sistema, durante un tiempo 'suficientemente largo'.
3. El valor de α_f es calculado a partir de la serie de tiempo obtenida en el paso 2, de la siguiente manera:

$$\alpha_f = max(|y_f(t)|). \tag{63}$$

3.5. Resultados numéricos

En esta sección se presenta un ejemplo numérico con el que se ilustra la capacidad del detector (62) para detectar ataques en la salida de un sistema lineal de segundo orden.

Como caso particular, se considera la dinámica de un sistema mecánico masa-resorte-amortiguador, cuyo comportamiento dinámico está gobernado por las ecuaciones

$$\dot{x}_1 = x_2, \tag{64}$$
$$\dot{x}_2 = (-kx_1 - bx_2 + u)\frac{1}{m},$$
$$y = x_1 \tag{65}$$

donde $k = 175$ [N/m], $b = 2.1$ [Ns/m], y $m = 0.74$ [kg]. Además, se considera el caso

en que el sistema está excitado por la entrada

$$u(t) = Asin(\omega t), \tag{66}$$

con $A, \omega \in \mathbb{R}_+$.

Para reconstruir la salida $y(t)$ dada en (65), se usa el observador anteriormente descrito en (41), el cual, para el ejemplo bajo consideración toma la forma

$$\dot{\hat{x}}_1 = \hat{x}_2 + c_1 e_1, \tag{67}$$

$$\dot{\hat{x}}_2 = (u - b\hat{x}_2 - k\hat{x}_1)\frac{1}{m} + c_2 e_1 + c_3 \text{sign}(e_1),$$

$$\hat{y} = \hat{x}_1, \tag{68}$$

donde los errores de observación e_1 y e_2, están definidos como $e_1 = y - \hat{y}$ y $e_2 = x_2 - \hat{x}_2 + \dot{\delta} - c_1 e_1$. Los parámetros del observador son iguales a los del sistema, con ganancias $c_1 = 20$, $c_2 = 20$,y $c_3 = 12$.

En un primer estudio numérico se caracteriza el detector (62) para el caso en que no hay ataques en la salida. Para esto, el sistema (64)-(67) se integra numéricamente con los valores paramétricos antes mencionados y con condiciones iniciales $x_1(0) = \hat{x}_1(0) = 0.1$ y $x_2(0) = \hat{x}_2(0) = 0$. La integración numérica se realizó utilizando la rutina ode45 de Matlab, con un tamaño de paso de 1×10^{-3}.

Los resultados obtenidos se muestran en la Figura 11 de la cual es claro ver que la salida observada converge a la salida del sistema. Por otra parte, los errores de observación convergen después de la respuesta transitoria del observador como se muestra en la Figura 12.

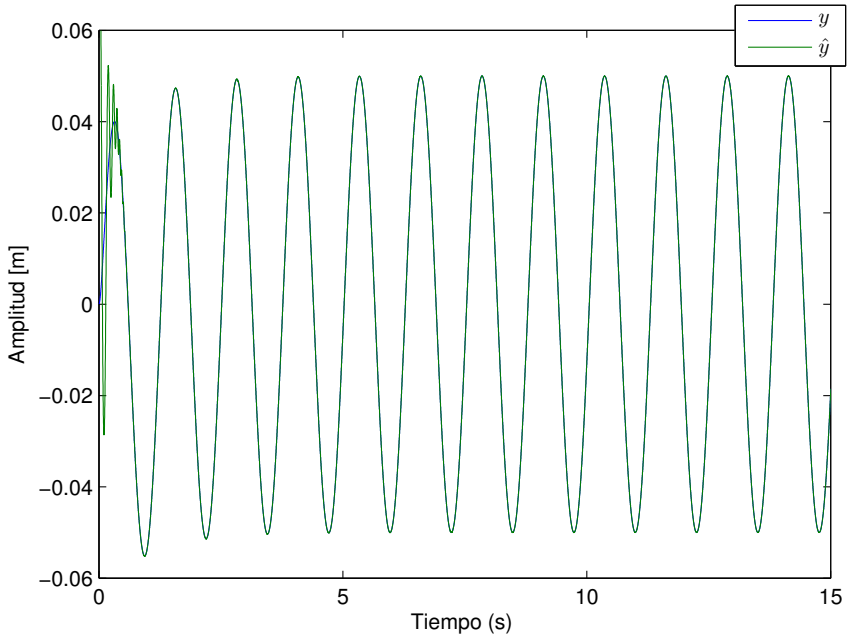

Figura 11. Respuesta del observador con respecto a la planta.

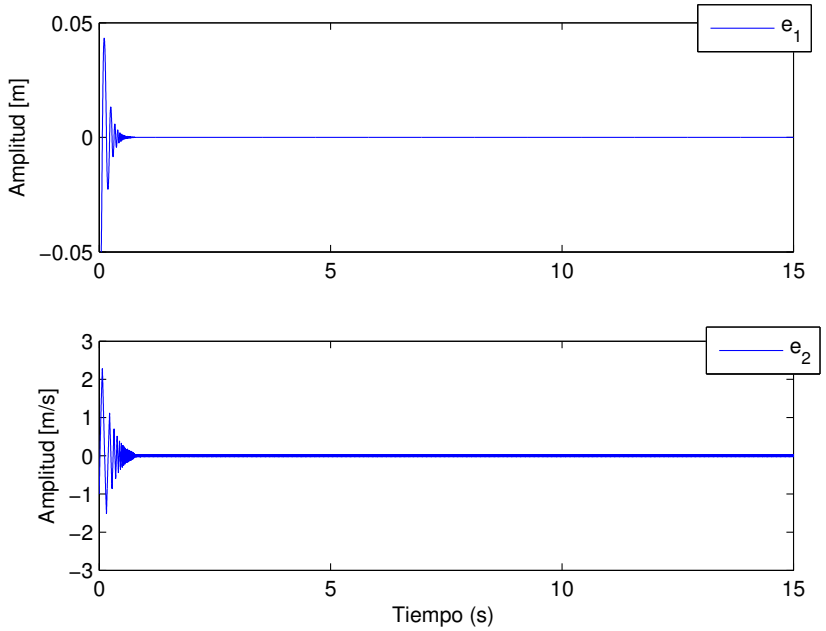

Figura 12. Errores entre los estados de la planta con los estados observados.

A continuación, el término discontinuo $c_3\text{sign}(e_1)$ en el observador (67) se filtra utilizando el filtro de Butterworth descrito en la ecuación (58)-(60), con frecuencia $\omega_c = 15$ [rad]. En la Figura 13 se muestra el comportamiento de $|y_f|$, también el valor de $\alpha_f = 0.0126$ el cual es diferente de cero debido a la presencia del fenómeno Chattering

generado por el término discontinuo.

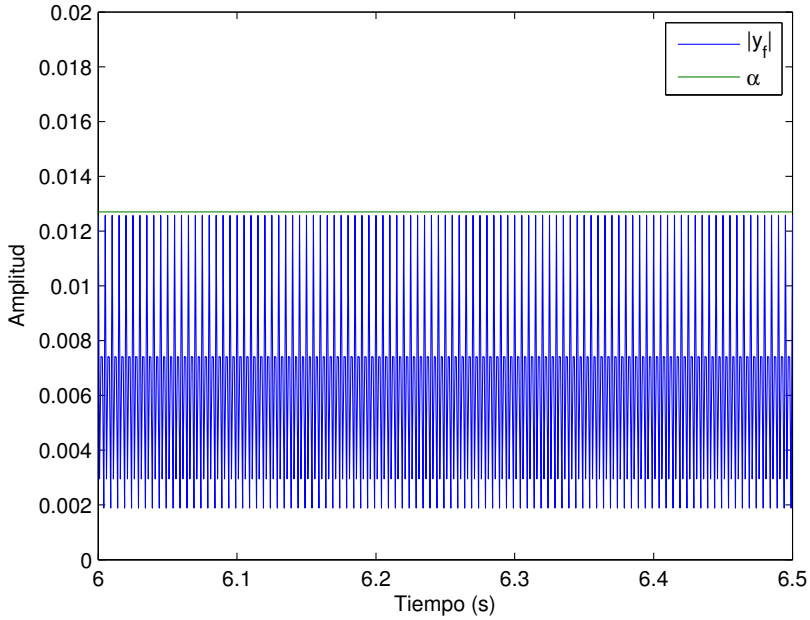

Figura 13. Salida del filtro (en valor absoluto) y valor del umbral α_f para el caso en que la salida del sistema no está atacada.

Para verificar que el detector no da falsas alarmas con el valor obtenido de α_f, se corrieron 10 simulaciones usando los mismos parámetros y condiciones iniciales, sin embargo en ninguna de la simulaciones se activó la alarma.

3.5.1. Ataque a la salida del sistema

En la sección anterior se mostró que el detector ha quedado caracterizado para el caso en que no hay ataques en la salida del sistema. Ahora, en esta sección, se verifica la efectividad de este detector en el caso en que la salida del sistema ha sido corrompida con un ataque $\delta(t)$, de tal manera que la salida (65) se convierte en

$$\bar{y} = y + \delta = x_1 + \delta. \tag{69}$$

En particular, se considera el siguiente ataque variante en el tiempo

$$\delta(t) = \begin{cases} 0, & \text{si} \quad 0 \leq t < 10, \\ 0.004sin(5t), & \text{de otra manera.} \end{cases} \tag{70}$$

La Figura 14 muestra el efecto que el ataque tiene en la salida. De esta figura se puede ver que el ataque (70) puede considerarse como un ataque sigiloso ya que la degradación de la señal es muy ‘pequeña’.

Por otra parte, la Figura 15 muestra los errores de observación. Se puede apreciar que la salida observada $\hat{y}$ converge a la salida atacada $\bar{y}$.

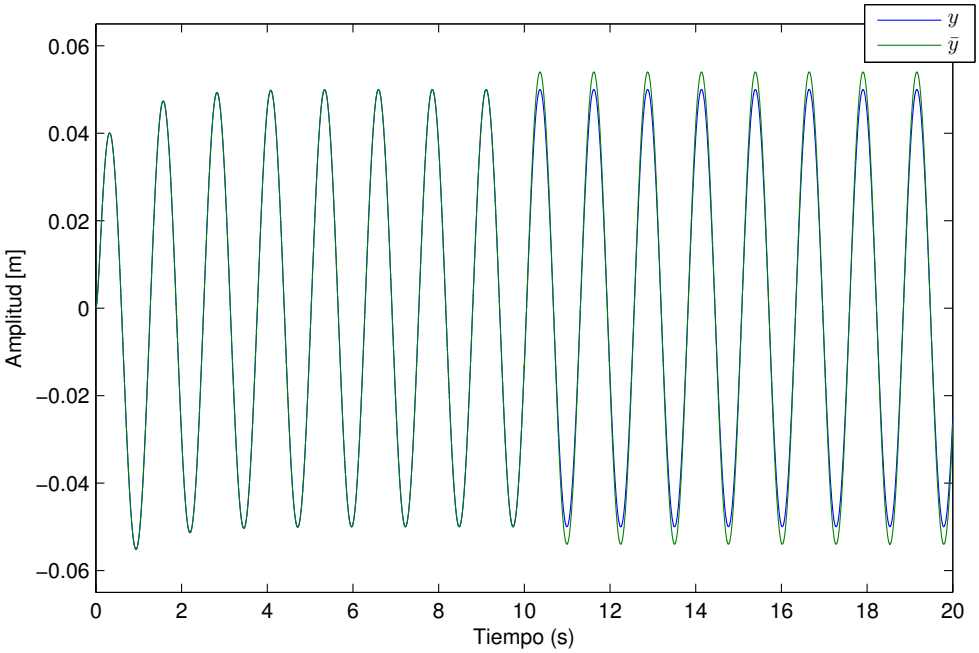

Figura 14. Respuesta de la salida nominal (línea azul) y la salida atacada (línea verde).

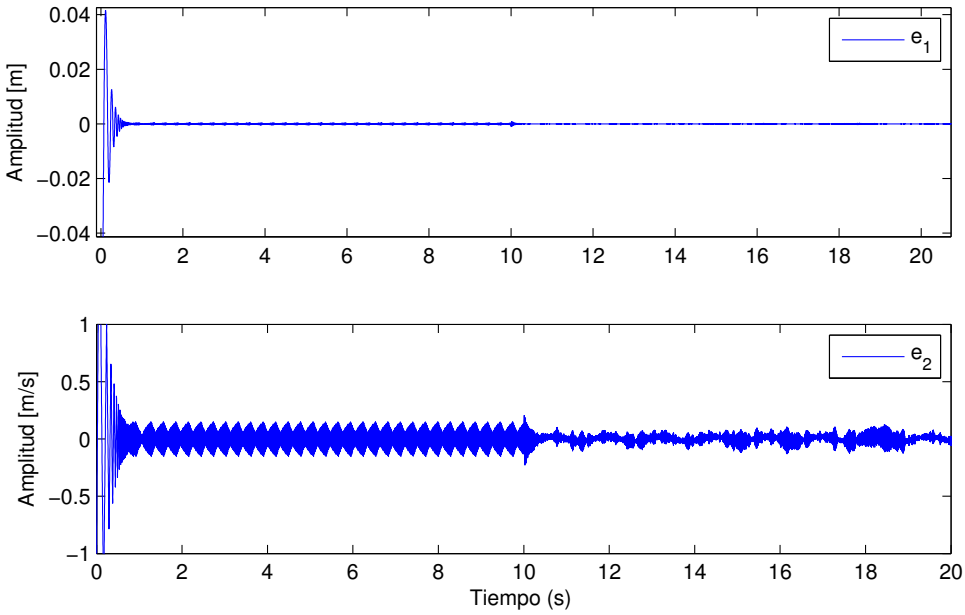

Figura 15. Errores de observación.

A pesar de que la magnitud del ataque (70) es relativamente pequeña, de solo el 5 % de la magnitud de la señal, el detector es capaz de identificar este ataque tal y como se muestra en la Figura 16. En esta figura es claro ver que cuando el ataque es aplicado al instante de tiempo $t \geq 10$, la salida del filtro (línea azul) sobrepasa el valor de umbral α_f calculado anteriormente (línea verde). Como consecuencia, el detector (62) activa las alarmas (línea roja).

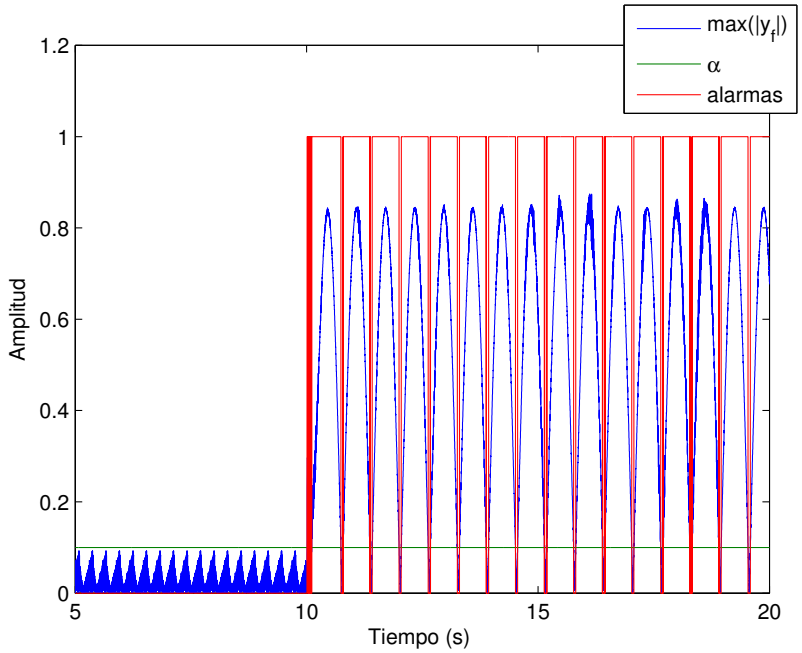

Figura 16. Detección del ataque. Cuando el ataque (70) es aplicado al instante $t > 10$, la salida del filtro (señal azul) excede el valor de umbral α_f (línea verde). Como consecuencia, el detector (62) enciede las alarmas (señal roja) para indicar la presencia del ataque.

3.5.2. Identificación del control equivalente

En la Sección 3.2 se mostró que el control equivalente, es decir, la salida y_f del filtro the Butterworth dado en (58) está dado por la ecuación (55). En esta sección se verifica esta ecuación de manera numérica para el ejemplo presentado en la subsección anterior.

En particular, se presenta una comparación entre el valor predicho por la ecuación (55) y el valor obtenido en la simulación numérica considerando el ataque (70). El resultado de dicha comparación se muestra en la Figura 17 para dos valores diferentes de la frecuencia de corte ω_c del filtro (58). El panel superior de la Figura 17 muestra la comparación para el caso en que la frecuencia de corte del filtro es $\omega_c = 15$ y el panel inferior de la figura muestra la comparación para el caso en que la frecuencia de corte se incrementa a $\omega_c = 45$. Nótese que al utilizar una frecuencia de corte mayor la señal no presenta desfasamiento y se tiene una mejor coincidencia entre los resultados analíticos y numéricos. Sin embargo, el precio a pagar es que al incrementar la frecuencia de corte del filtro la señal a la salida del filtro es más ruidosa.

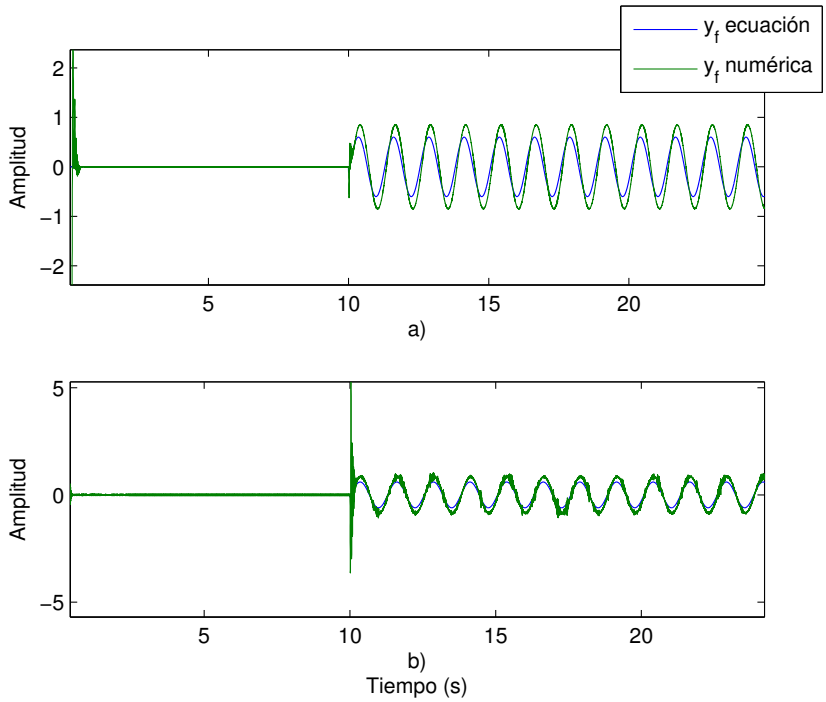

Figura 17. Identificación de la salida del filtro (control equivalente). Línea verde: salida del filtro predicha por la ecuación (55). Línea azul: salida del filtro obtenida en simulación numérica. Panel superior: $\omega_c = 15$. Panel inferior: $\omega_c = 45$. Se obtiene una mejor coincidencia entre los resultados numéricos y analíticos cuando la frecuencia de corte del filtro es mayor.

3.5.3. Reconstrucción del ataque

Una de las ventajas que presentan las simulaciones es que se tiene acceso al vector de estados completo del sistema. Esto no siempre es el caso en aplicaciones prácticas, donde tener sensores para cada variable del sistema puede resultar muy costoso. En esta sección se aprovecha el accesso al vector de estados completo del sistema atacado para ilustrar como el detector propuesto puede ser usado para reconstruir un estimado del ataque $\delta(t)$ presente en la salida del sistema.

Para hacer la reconstrucción del ataque, se utiliza la ecuación (57).

Como primer ejemplo, se reconstruye el ataque (70). El resultado obtenido se muestra en la Figura 18 a), de donde es claro que existe un offset entre el ataque reconstruído y el ataque real. Como se mencionó anteriormente, dicho offset lo introduce el integrador debido a la diferencia en condiciones iniciales. Sin embargo, si este se suprime, entonces el ataque reconstruído converge al ataque real.

Para mostrar la capacidad del detector de identificar ataques de diversa naturaleza,

se considera el ataque

$$\delta(t) = \begin{cases} 0, & \text{si} \quad 0 \leq t < 10, \\ 0.001 sign(sin(5t)), & \text{de otra manera.} \end{cases} \tag{71}$$

La Figura 18b) muestra la reconstrución de este ataque. Una vez más, si se remueve el offset introducido por el integrador, entonces se puede observar que el ataque estimado y el real coinciden.

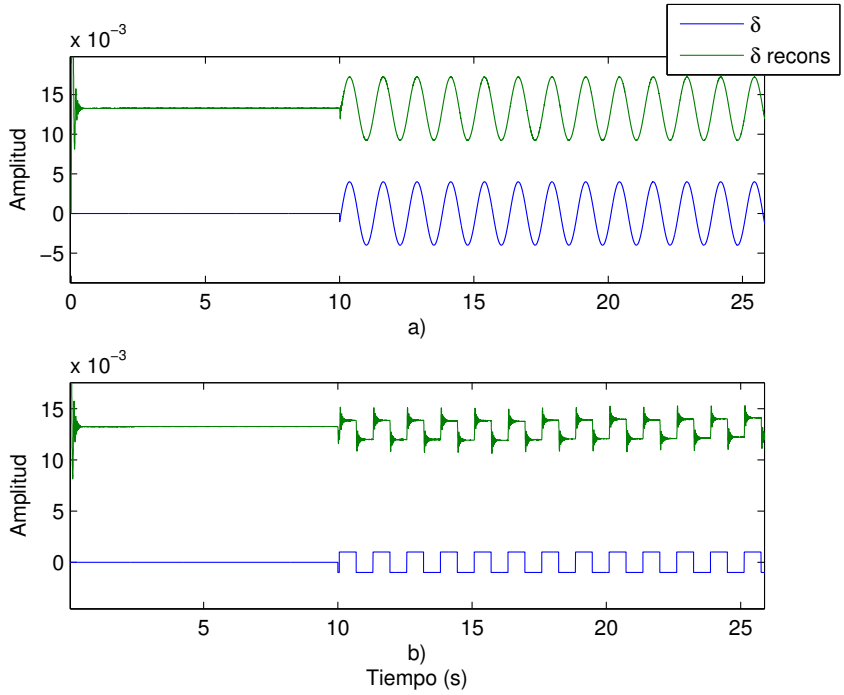

Figura 18. Ataques reconstruídos. a) Ataque (70). b) Ataque (71).

3.6. Detector de ataques con ruido en el sensor

En la sección anterior se consideró un escenario muy ideal en el sentido que la salida medida del sensor estaba libre de ruido. Sin embargo, en aplicaciones prácticas, el ruido de medición casi siempre está presente. Por lo tanto, en esta sección, investigamos el rendimiento del detector (62) para el caso en que la salida del sistema tiene ruido Gaussiano de media cero, es decir, se considera que la salida está dada por

$$\bar{y}(t) = y(t) + \eta(t) + \delta(t) = Cx(t) + \eta(t) + \delta(t), \tag{72}$$

donde $\eta(t)$ es el ruido Gaussiano de media cero. El diagrama a bloques para este caso se muestra en la Figura 19.

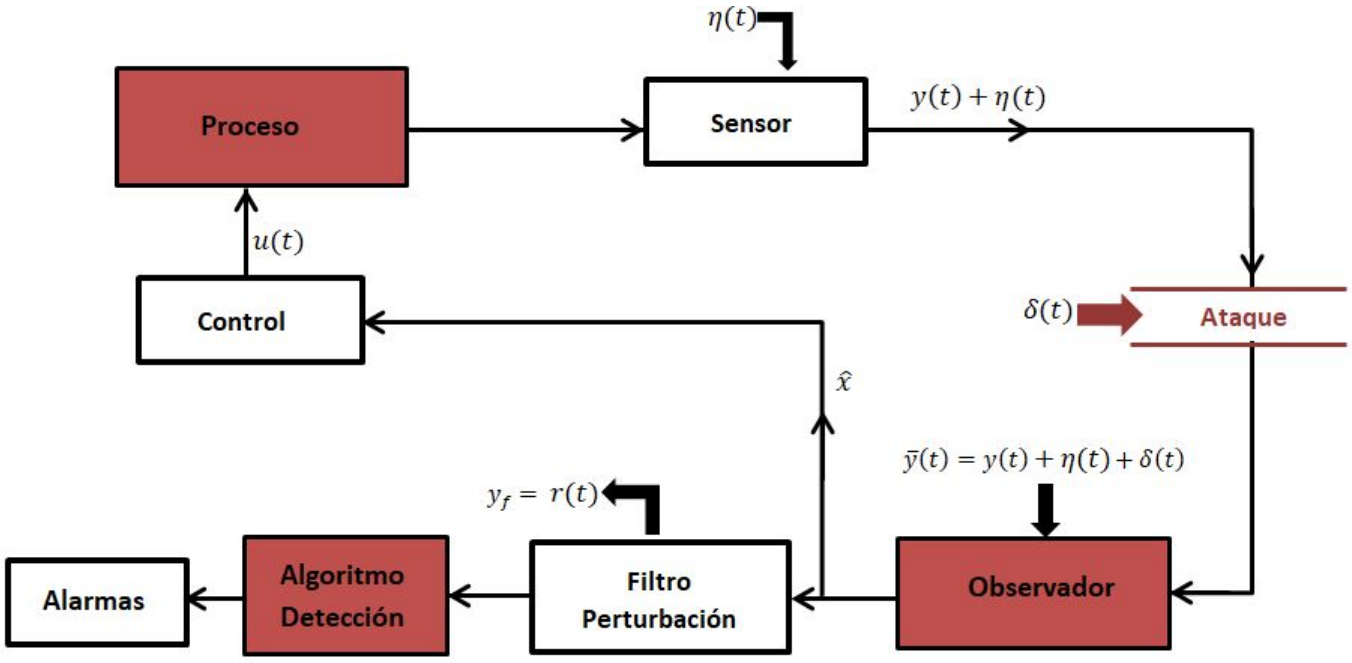

Figura 19. Diagrama a bloques del detector de ataques.

Como ejemplo particular, se considera el sistema mecánico masa-resorte-amortiguador descrito por las ecuaciones (64)-(65), con los mismos parámetros usados en la Sección 3.5. Además, se considera que la salida del sistema está contaminada con ruido Gaussiano $\eta(t)$ con media cero y varianza 0.001. Se utiliza la misma entrada de control y el observador diseñados en la Sección 3.5.

Los resultados obtenidos se muestran en la Figura 20 de la cual es claro ver que la salida observada prácticamente converge a la salida del sistema. Por otra parte, debido al ruido presente en la salida del sistema, los errores de observación no convergen completamente a cero. Sin embargo, dichos errores permanecen acotados y con amplitud ‘pequeña’ como se muestra en la Figura 21.

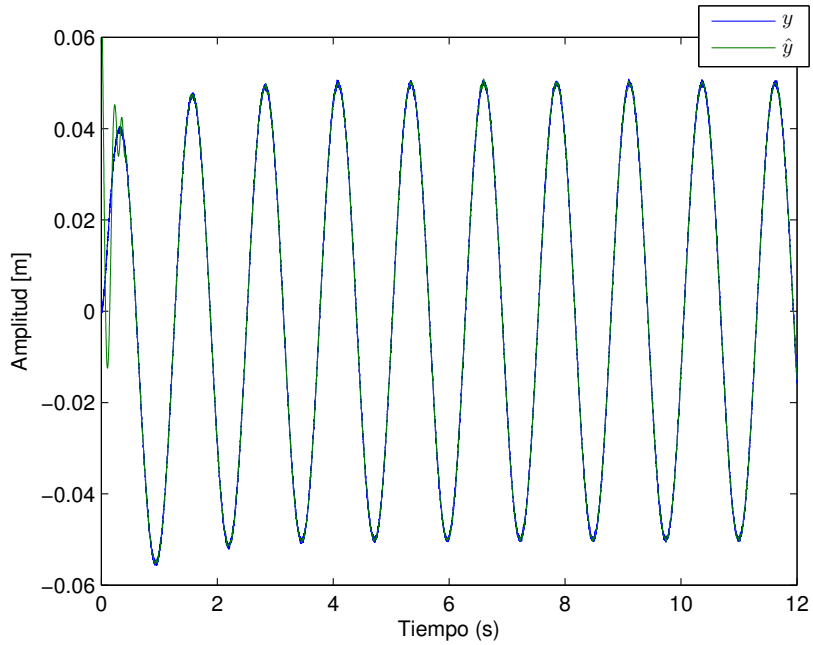

Figura 20. Respuesta del observador con respecto a la planta.

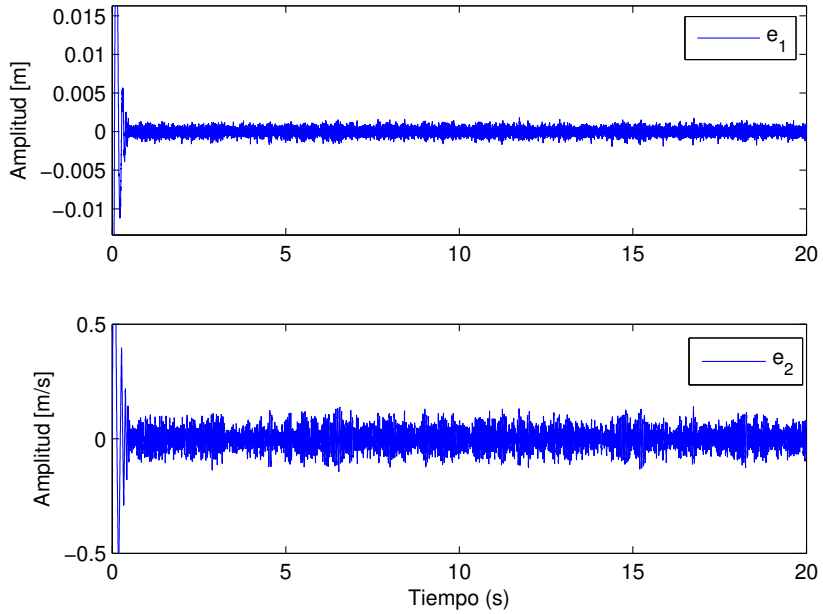

Figura 21. Errores entre los estados de la planta con los estados observados.

Posteriormente, el término discontinuo $c_3\text{sign}(e_1)$ se filtró mediante el filtro pasa bajas descrito en la ecuación (58)-(60), con frecuencia ω_c = 15 [rad]. En la Figura 22 se observa el comportamiento de $|y_f|$, también se muestra el valor del umbral $\alpha = 0.28515$ cuando el sistema está libre de ataques.

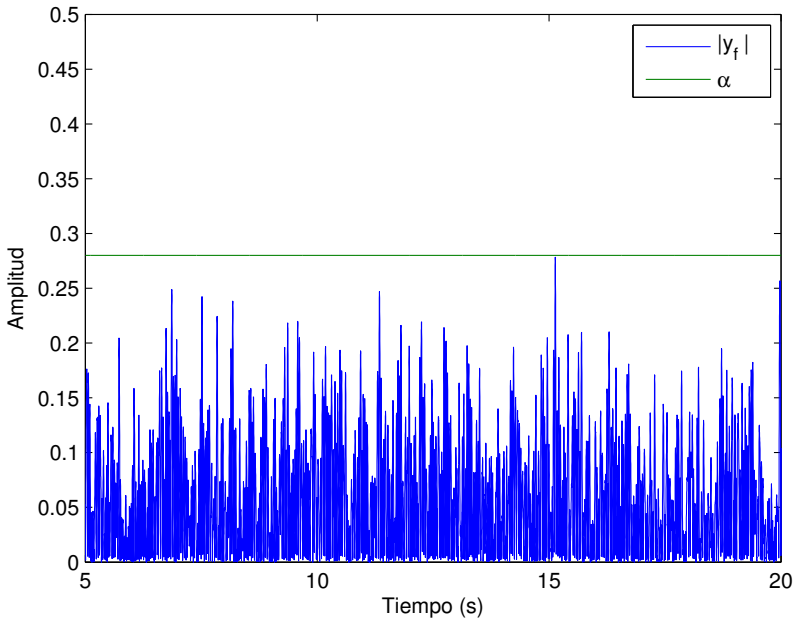

Figura 22. Salida del filtro (en valor absoluto) y valor del umbral α_f para el caso en que la salida del sistema no está atacada.

Para corroborar que no se presentan falsas alarmas en el detector con el valor obtenido de α_f, se realizaron 10 simulaciones usando los mismos parámetros y condiciones iniciales, excepto por el ruido Gaussiano que es generado aleatoriamente, aunque en todas las simulaciones se dejó fija la varianza del ruido al valor inicial de 0.001.

3.6.1. Ataque variante en el tiempo

Para ilustrar el funcionamiento del detector después de sintonizarlo libre de ataque, se considera que la salida del sistema es atacada, por lo tanto la salida (65) se convierte en

$$\bar{y} = y + \eta + \delta = x_1 + \eta + \delta. \tag{73}$$

En este caso se implementa el ataque variante en el tiempo descrito en (70). En la Figura 23 se muestra la degradación de la señal. Por otro lado, se puede observar que a pesar de la influencia del ruido en el sensor el detector funciona correctamente cuando la salida del sistema ha sido atacada como se muestra en la Figura 62.

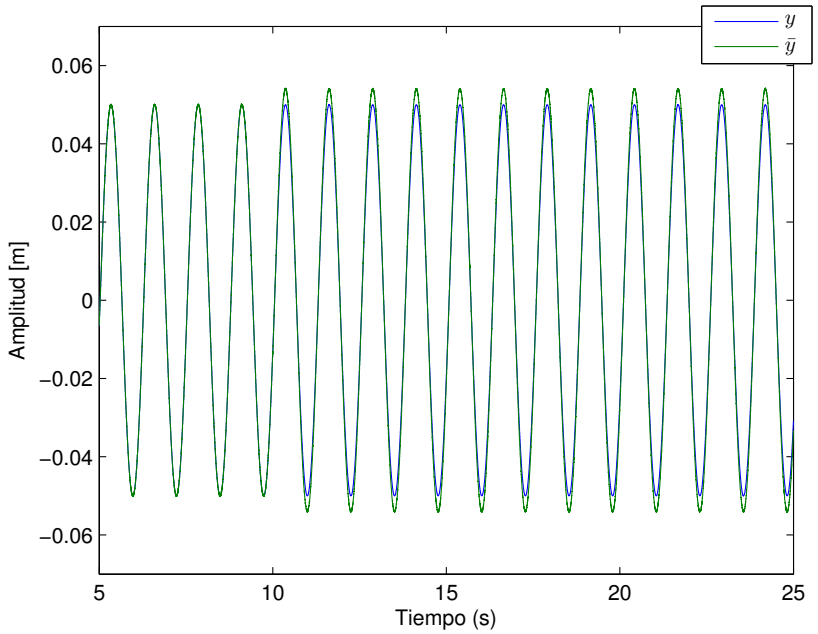

Figura 23. Respuesta de la salida nominal (línea azul) y la salida atacada (línea verde).

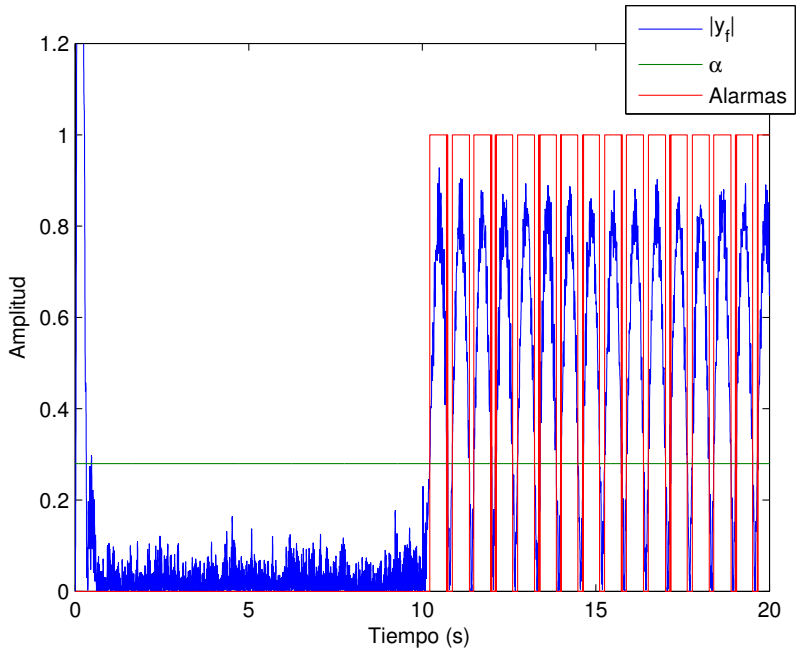

Figura 24. Detección del ataque. Cuando el ataque (70) es aplicado al instante $t \geq 10$, la salida del filtro (señal azul) excede el valor de umbral α_f (línea verde). Como consecuencia, el detector (62) enciede las alarmas (señal roja) para indicar la presencia del ataque.

3.7. Detector de ataques con filtrado del sensor

A continuación se propone una modificación del detector propuesto en la sección anterior. Dicha modificación consiste en filtrar la salida medida del sensor antes de ser aplicada al observador, tal y como se muestra en el diagrama de bloques de la Figura 26. La razón por la cual se ha incluído este filtro es para reducir el efecto del ruido en la señal de salida del sensor y de esta manera aumentar la sensibilidad del detector.

El diseño del detector se realiza siguiendo la misma metodología de la sección anterior, excepto que en el observador robusto se usa la señal *filtrada* de la salida y en lugar de simplemente usar la salida y. El filtro utilizado para filtrar la salida del sensor es un filtro pasabajas Butterworth de segundo orden descrito por (58).

Esta idea se aplicó también a los detectores clásicos CUSUM y Chi-cuadrado mencionados en el Capítulo 2. El resultado obtenido (ver el artículo de conferencia en la sección de anexos de esta tesis) mostró que efectivamente la inclusión de este filtro permite hacer más sensibles a los detectores para identificar ataques o anomalías en la salida del sistema.

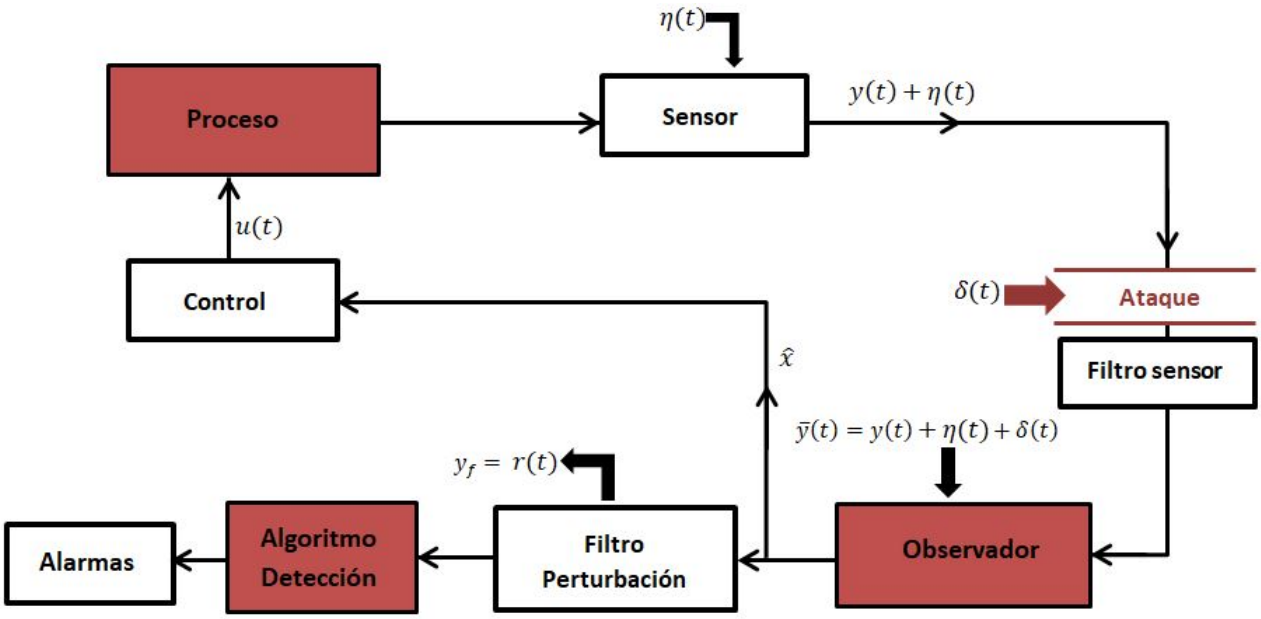

Figura 25. Detector de ataques con filtro en el sensor.

En esta sección, se presenta un ejemplo que ilustra la ventaja de incluir un filtro a la salida del sensor. Considere el ataque

$$\delta = \begin{cases} 0, & \text{si} \quad 0 \leq t < 10, \\ 0.001sin(5t), & \text{de otra manera,} \end{cases} \tag{74}$$

Si se emplea el detector presentado en la sección anterior, el detector no es capaz de identificar los ataques, ya que ocurren falsos negativos en la señal de alarma, tal y como se muestra en la Figura 26a). Sin embargo, cuando se usa el detector con filtro en la salida, el detector es capaz de identificar la presencia del ataque, tal y como se muestra en la Figura 26. En esta comparación, se usaron exactamente los mismos valores paramétricos y condiciones iniciales en ambos casos, excepto que para la Figura 26b) se utilizó la salida filtrada del sensor.

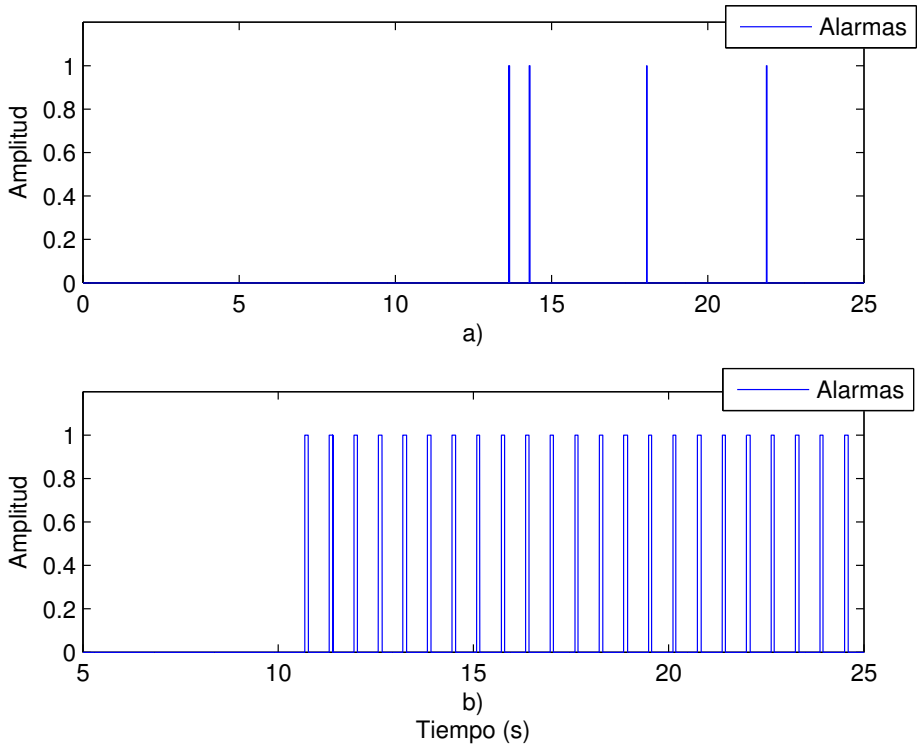

Figura 26. Activación de las alarmas ante un ataque inducido al instante $t = 10$. a) Algoritmo de detección sin filtro en el sensor, b) Algoritmo de detección con filtro en el sensor.

3.8. Detector basado en redundancia

A continuación se presenta otra posible modificación del detector presentado en la Sección 3.1. La modificación consiste en usar redundancia en la medición, es decir que se tienen dos sensores para medir la misma variable, tal y como se muestra en la Figura (28).

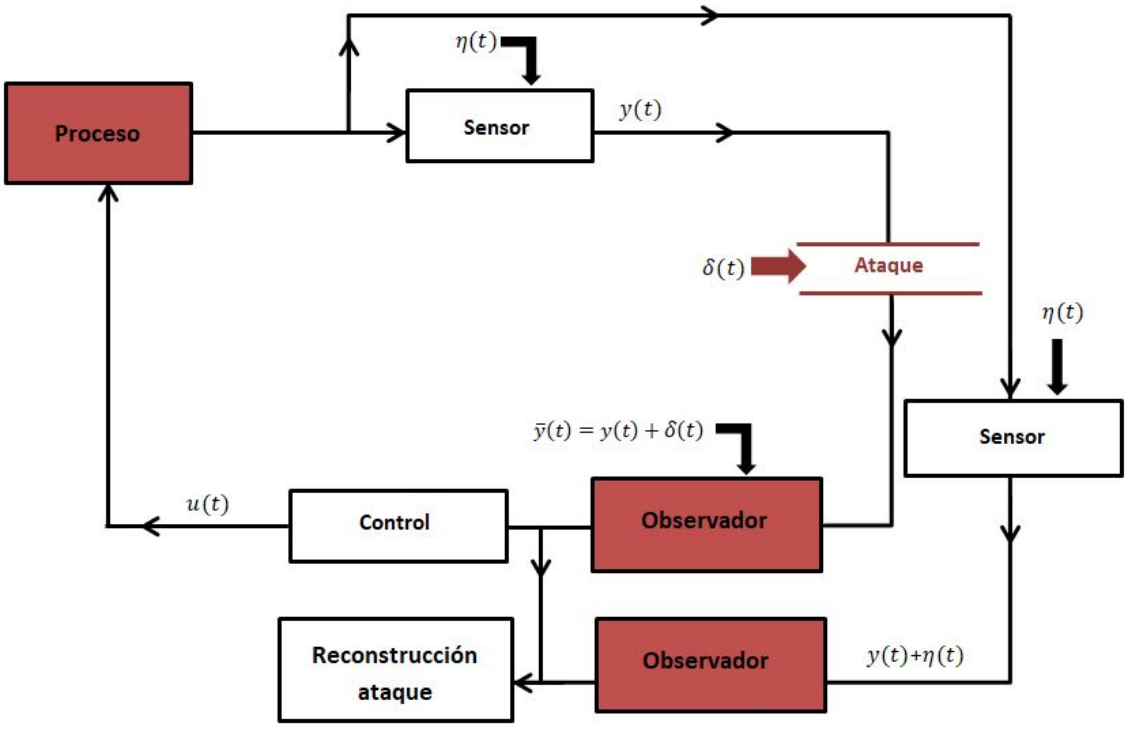

Figura 27. Detector basado en redundancia

Considere el sistema de la forma (37)-(39), suponga que se tienen dos sensores para medir la salida (38) los cuales están libres de ruido, y que solamente uno de los dos sensores es atacado, de tal manera que la salida del sensor 1 está dada por

$$y_1 = x_1 + \delta(t), \tag{75}$$

y la salida del sensor 2 está dada por

$$y_2 = x_1. \tag{76}$$

Nótese que en esta configuración la reconstrucción del ataque es trivial, ya que basta restar las salidas de los sensores para obtener $\delta(t)$.

Sin embargo, si se pretende utilizar un observador para reconstruir el estado del sistema, entonces la estimación del ataque se puede hacer de la siguiente manera

Primero, para reconstruir el estado del sistema usando la medición (75) del sensor

1 se utiliza el observador

$$\dot{\hat{x}}_1 = \hat{x}_2 + c_2\hat{e}_1, \tag{77}$$

$$\dot{\hat{x}}_2 = a\hat{x}_1 + b\hat{x}_2 + u + c_2e_1 + c_3\text{sign}(\hat{e}_1), \tag{78}$$

$$\hat{y} = \hat{x}_1,$$

donde e_1 está dado por

$$\hat{e}_1 = y_1 - \hat{y}. \tag{79}$$

Por otra parte, para reconstruir el estado del sistema usando la medición (76) del sensor 2, la cual está libre de ataque, se utiliza el observador

$$\dot{\tilde{x}}_1 = \tilde{x}_2 + c_2\tilde{e}_1, \tag{80}$$

$$\dot{\tilde{x}}_2 = a\tilde{x}_1 + b\tilde{x}_2 + u + c_2e_1 + c_3\text{sign}(\tilde{e}_1), \tag{81}$$

$$\tilde{y} = \tilde{x}_1,$$

donde

$$\tilde{e}_1 = y_2 - \tilde{y}. \tag{82}$$

Debido a que solamente uno de los sensores se ve afectado por el ataque, es posible reconstruir el ataque usando los errores de observación de cada observador. Específicamente, se tiene que

$$\hat{e}_1 - \tilde{e}_1 = (y_1 - \hat{y}) - (y_2 - \tilde{y}). \tag{83}$$

Cuando los errores de observación $\hat{e}_1$ y $\tilde{e}_1$ y sustituyendo (75) y (76) en (83), se obtiene que

$$\delta = \hat{y} - \tilde{y}. \tag{84}$$

Es decir, el ataque se puede reconstruir usando solamente las salidas estimadas por los observadores (77) y (80).

Para ilustrar el funcionamiento de este detector basado en redundancia, considere el sistema de segundo orden definido por las ecuaciones (37)-(39) con $a = 4$ y $b = 20$ y entrada $u = Asin(\omega t)$, con $A = 8$ y $\omega = 5$. También, considere los observadores (77) y (80) con ganancias $c_1 = c_2 = 25$, $c_3 = 40$. Ahora, suponga que la salida del sistema está sujeta al ataque

$$\delta(t) = \begin{cases} 0, & \text{si} \quad 0 \leq t < 8, \\ 0.1sin(2t), & \text{de otra manera.} \end{cases} \tag{85}$$

El sistema completo, es decir planta más observadores, se simula y el ataque se detecta usando la ecuación (84). El resultado de la reconstrucción del ataque se muestra en la Figura (28), en la cual es claro ver que el ataque reconstruido coincide con el ataque real aplicado a la salida del sistema.

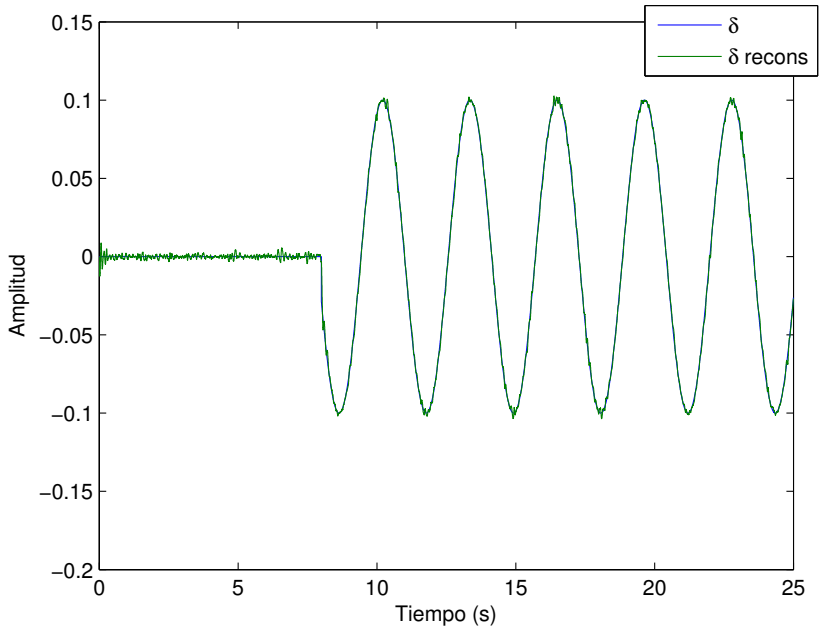

Figura 28. Reconstrucción del ataque

En conclusión este detector de ataques basado en redundancia, permite reconstruir el ataque a la salida del sistema siempre y cuando uno de los sensores no sea atacado. Aunque este detector parece atractivo, puede tener la desventaja de tener un alto costo, debido a que requiere de la implementación de más de un sensor para medir la misma variable.

Capítulo 4. Resultados experimentales

En este capítulo se valida experimentalmente el algoritmo de detección propuesto en el capítulo anterior en dos sistemas CPS cuya capa física consiste de un sistema mecánico masa-resorte-amortiguador, y de un circuito electrónico, el cual emula la dinámica de un sistema de segundo orden.

4.1. Resultados experimentales en un sistema mecánico

En este caso se consideró el sistema mecánico masa-resorte-amortiguador, el cual representa la capa física del CPS la cual se comunica mediante una tarjeta de adquisición de datos a una computadora, siendo esta última la capa virtual. En la capa virtual (PC) se construyó el algoritmo de detección utilizando el software Simulink, mientras que la tarjeta de adquisición de datos realiza el procesamiento de las lecturas de medición del sensor de posición del sistema masa-resorte-amortiguador y envía la señal de control al sistema físico. Ambas acciones se realizan en tiempo real. El modelo del sistema masa-resorte-amortiguador tiene la forma

$$\dot{x}_1 = x_2, \tag{86}$$

$$\dot{x}_2 = (-kx_1 - bx_2 + u)\frac{1}{m}, \tag{87}$$

$$y = x_1, \tag{88}$$

donde k es la constante de elasticidad del resorte medido en $[N/m]$, b es el coeficiente de amortiguamiento viscoso dado en $[Ns/m]$, y m $[kg]$ es la masa.

Figura 29. Sistema masa-resoste-amortiguador disponible en el Laboratorio de Control de CICESE.

4.1.1. Caracterización del detector

Para realizar la caracterización del algoritmo de detección de ataques se utilizó la metodología y las herramientas descritas en la Subsección 3.1 del capítulo anterior. Se consideró el sistema mecánico masa-resorte-amortiguador dado por las ecuaciones (86)-(88), donde $k = 175\ N/m$, $b = 2.1\ Ns/ms$ y $m = 0.74\ kg$. Para estimar el estado de la planta se utilizó el observador discontinuo diseñado anteriormente en el Capítulo 3, descrito por las ecuaciones (67)-(68), las ganancias seleccionadas fueron $c_1 = 20, c_2 = 20, c_3 = 30$. Nótese que es el mismo observador discontinuo implementado en el ejemplo numérico del Capítulo 3. La convergencia del estado observado con la respuesta del estado de la planta se muestra en la Figura 30, y en la Figura 31 se muestra el error de posición entre la medición de la planta y del observador. No se puede comparar el error en velocidad porque el sistema físico no tiene sensor de velocidad.

Nótese que las gráficas de esta sección se muestran a partir de los 10 segundos y hasta los 20 segundos, debido a que la capacidad de almacenamiento de la tarjeta de adquisición de datos utilizada en la validación de este experimento está limitada a una ventana de 10 segundos. Los primeros 10 segundos del experimento se ignoraron porque contienen la respuesta transitoria del sistema, la cual no es de interés en este experimento.

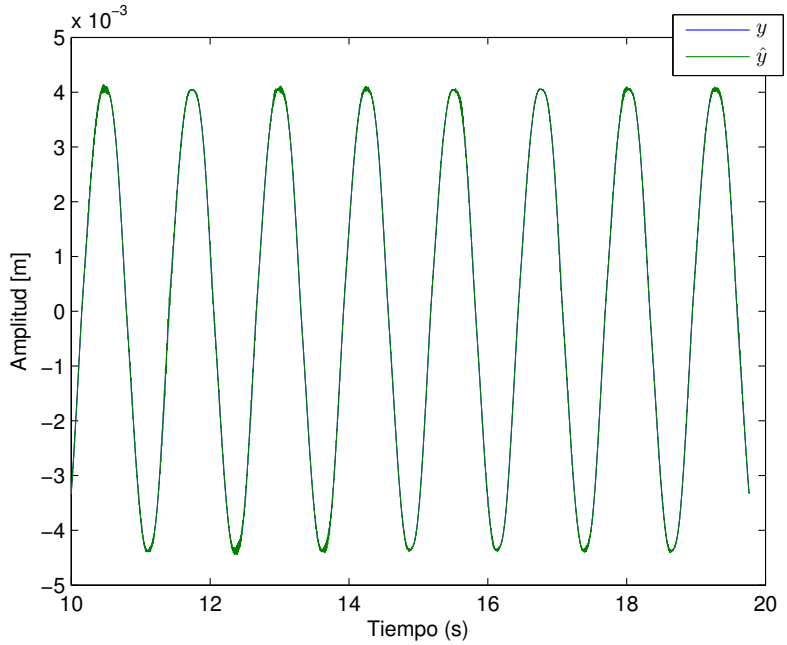

Figura 30. Respuesta de la salida medida y la salida observada.

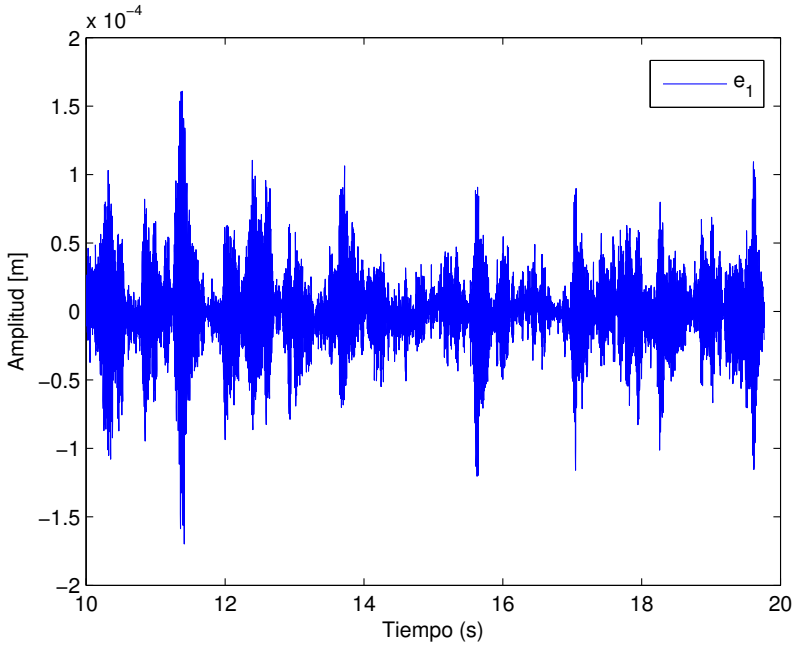

Figura 31. Error de observación en posición.

Para diseñar el filtro se utilizó un bloque de Simulink en el cual se describió el filtro en variables de estado, dando el valor de la frecuencia de corte (ω_c) y el término discontinuo $c_3\text{sign}(e_1)$ del observador como entrada del filtro de segundo orden. Para este caso se usó un valor de $\omega_c = 12$ y el orden del filtro es 2.

Se midió la salida del filtro y_f después de la respuesta transitoria del sistema sin ataque durante un tiempo suficientemente largo para obtener el parámetro de umbral α_f del detector (62). Para esta caracterización del algoritmo de detección, se obtuvo el valor de umbral de $\alpha_f = 2.071$.
La Figura 32 muestra el valor de α_f y el comportamiento de $|y_f|$ cuando la salida del sistema no se encuentra bajo ataque.

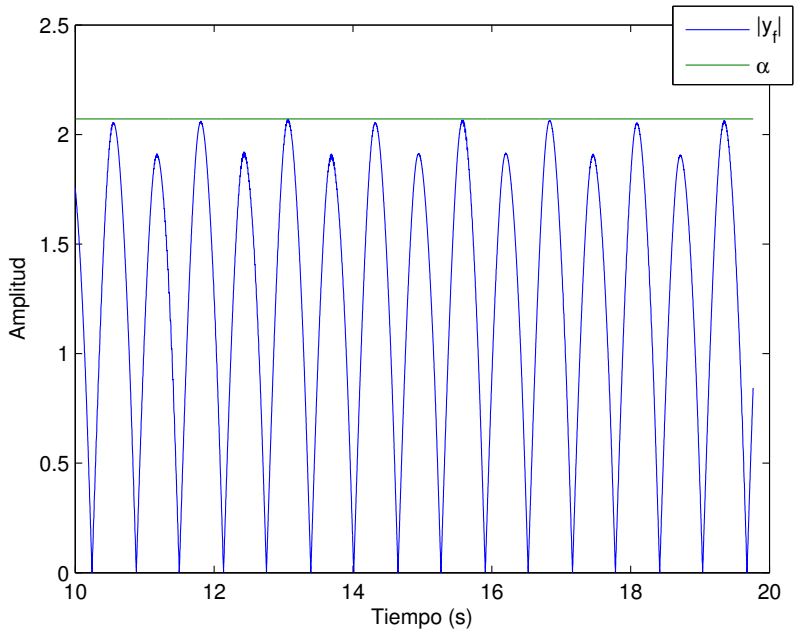

Figura 32. Valor de α

4.1.2. Aplicación de un ataque variante en el tiempo

Para evaluar el desempeño del detector de ataques se implementó el ataque variante en el tiempo dado por $\delta(t) = 0.004\ sin(5t)$, el cual fue activado a los 12 segundos de operación del sistema mecánico.

En la Figura (33) se observa que las salida estimada converge a la salida medida aún en presencia del ataque. En la Figura (34) se muestra el error de observación, el cual al iniciar el ataque se presenta un incremento, pero después muestra convergencia a cero. Finalmente, la Figura (35) presenta la salida obtenida del filtro pasabajas. Se observa que cuando el ataque es aplicado, el valor de $|y_f|$ se incrementa y sobrepasa el valor de umbral α_f, indicado por la línea horizontal de color verde. Como consecuencia, el detector (62) enciende las alarmas, las cuales se indican en color

rojo.

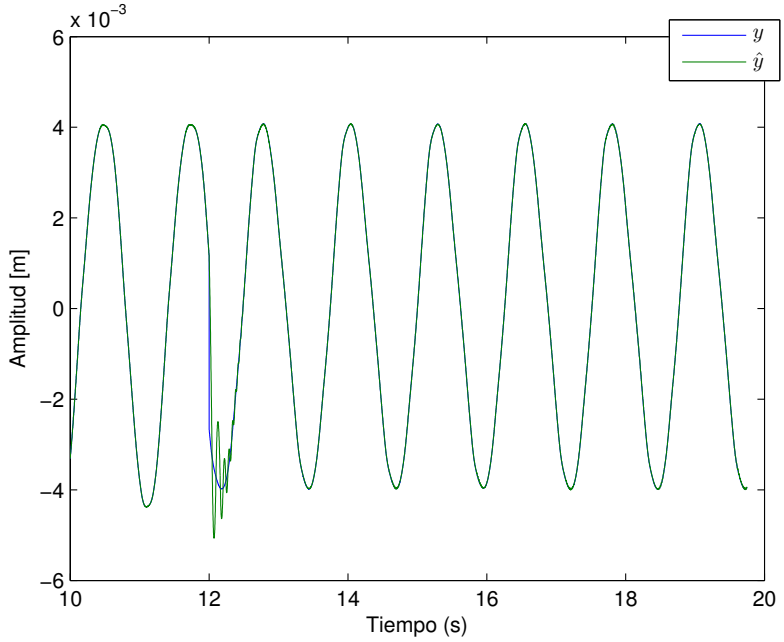

Figura 33. Salida medida de la planta y salida estimada por el observador.

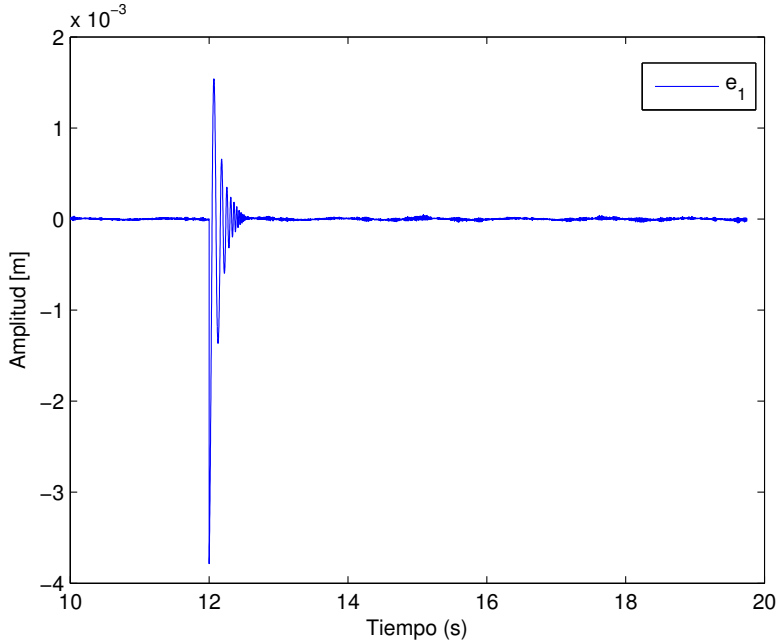

Figura 34. Error de observación de posición.

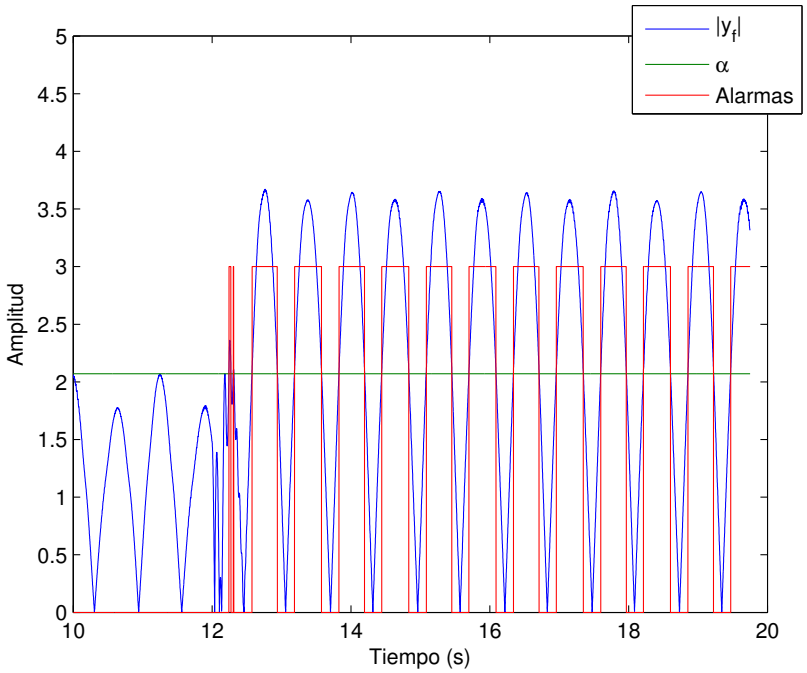

Figura 35. El algoritmo (62) activa las alarmas cuando se ataca la salida al instante de tiempo $t = 12$.

4.2. Resultados experimentales usando un circuito electrónico

En esta sección se analiza la detección de ataques en un sistema ciberfísico, cuya capa física está compuesta de un circuito electrónico construido con amplificadores operacionales y componentes simples como resistencias y capacitores. El sistema emula la dinámica del sistema masa-resorte amortiguador considerado en la sección anterior.

Se sabe que es posible simular o construir la ecuación diferencial que describe a cierto sistema mediante circuitos analógicos, esto específicamente con amplificadores operacionales. Esta manera de obtener o diseñar las soluciones se le conoce como cálculo analógico, cuyo objetivo es construir, mediante una red eléctrica usando amplificadores y componentes electrónicos (resistencias, capacitores, diodos, etc.), la solución de una ecuación diferencial (Roberge, 1975).

A continuación, en la Figura 36 se muestra el diseño del circuito electrónico que simula el modelo del sistema masa-resorte-amortiguador, el cual fue diseñado y validado en el sotfware Multisim.

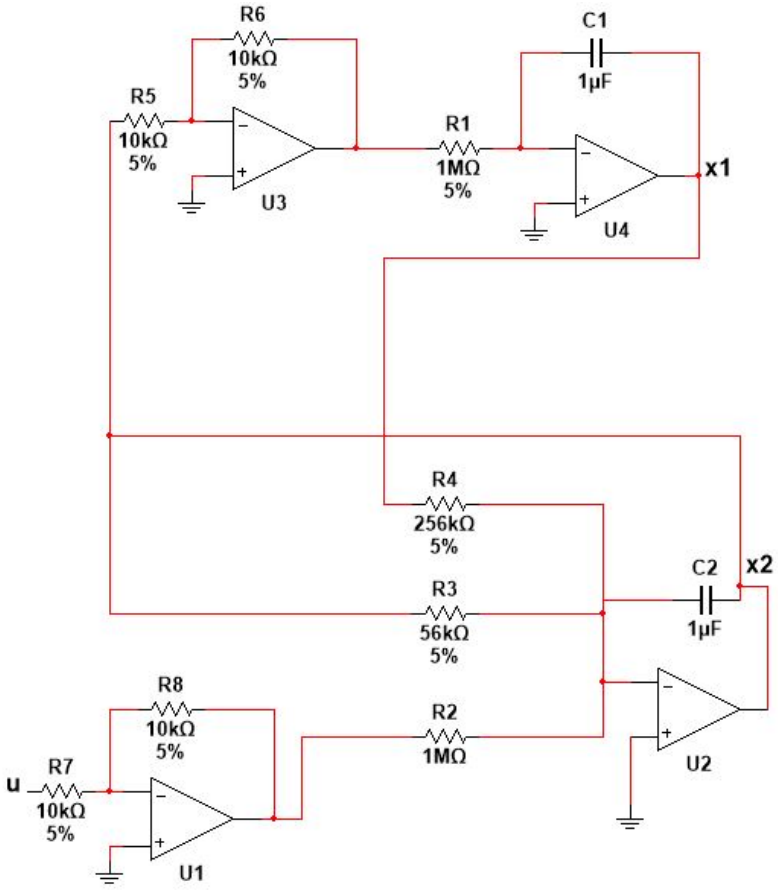

Figura 36. Diseño del circuito analógico.

En la Tabla 1 se tiene una lista de los componentes electrónicos utilizados en la implementación del circuito.

Tabla 1. Lista de componentes utilizados en el cirtuito.

Cantidad	Componente	Referencia diagrama
4	OP AMP, TL074	U1,U2,U3,U4
2	Capacitor, 1uF Poliéster	C1,C2
2	Resistencia, 1 *M*Ω	R1,R2
4	Resistencia, 10 *k*Ω	R5,R6,R7,R8
1	Resistencia, 56 *k*Ω	R3

La medición de las variables del sistema se realizarón con una tarjeta de adquisición de datos Dspace. Mediante esta tarjeta se realiza la comunicación del mundo físico con el mundo virtual y viceversa. La capa virtual de este CPS está compuesta por una computadora en la cual se diseña la ley de control que se envía al sistema físico. También en esta capa virtual se implementa el algoritmo de detección.

4.2.1. Caracterización del algoritmo de detección

El circuito electrónico emula la dinámica del sistema (37) con matrices A, B y C dadas en (39), los parámetros del sistema utilizados son $a = 4$, $b = 20$. El modelo del observador para estimar el estado del sistema tiene la forma (41)-(42). Los errores de observación están definidos como $e_1 = x_1 - \hat{x}_1$ y $e_2 = -x_2 - \hat{x}_2 + \dot{\eta} + \dot{\delta} - c_1 e_1$, y las ganancias implementadas son $c_1 = 20$, $c_2 = 20$, $c_3 = 25$, donde las condiciones iniciales del observador y de la planta son cero.

Se diseñó una entrada de control para que el sistema realice el seguimiento de una trayectoria deseada $x_d = 0.1sin(3t)$.

Los errores de seguimiento están definidos de la siguiente manera,

$$e_1 = x_d - x_1, \tag{89}$$

$$e_2 = \dot{x}_d - \hat{x}_2. \tag{90}$$

Por otra parte, la dinámica del error de seguimiento es

$$\dot{e}_2 = \ddot{x}_d + ax_1 + b\hat{x}_2 + k_p e_1 + k_d - u. \tag{91}$$

Por lo tanto, la entrada de control para lograr el seguimiento de trayectoria es la siguiente

$$u = kx_1 + b\hat{x}_2 + k_p e_1 + k_d e_2 + \ddot{x}_d. \tag{92}$$

La dinámica del error de seguimiento está dada por

$$\dot{e} = \begin{bmatrix} 0 & 1 \\ -k_p & -k_d \end{bmatrix} e. \tag{93}$$

Por lo tanto, se requiere que las ganancias $k_p > 0$ y $k_d > 0$ para conseguir que esta dinámica sea estable. Las ganancias del control PD para este caso son $k_p = 25$ y $k_d = 4$. En la Figura (37) se observa la salida del circuito electrónico, la cual es muy ruidosa, y la salida del estado observado. Por otra parte, los errores de observación se muestran en la Figura (38).

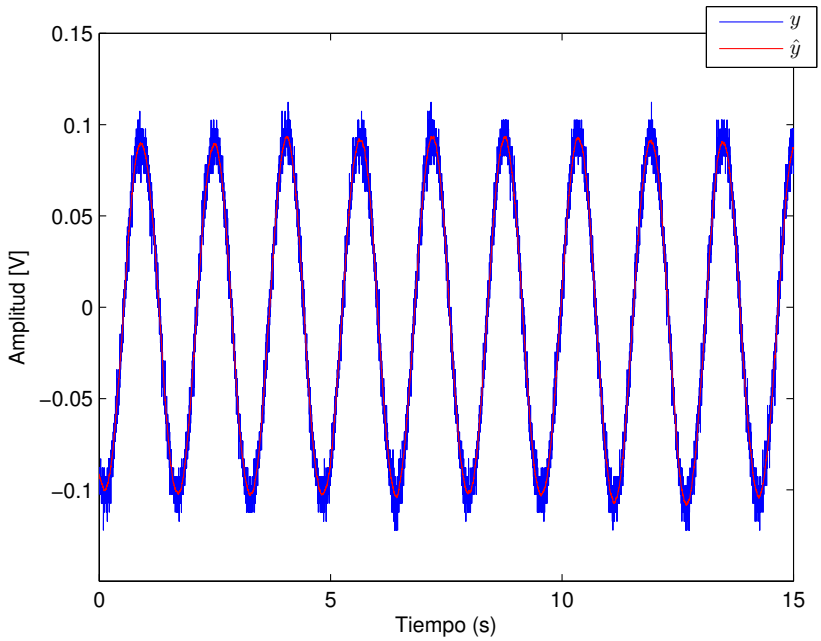

Figura 37. Salida medida y salida estimada de la planta.

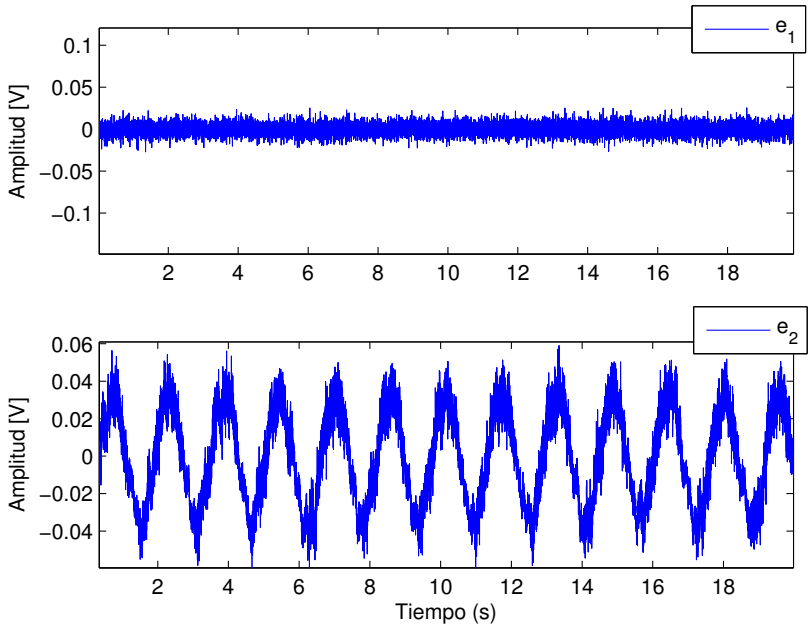

Figura 38. Errores de observación.

Después de estimar el estado de la planta, debido a la filtración del término discontinuo $c_3\text{sign}(e_1)$ del observador, se pueden obtener las variaciones provocadas por la influencia de las anomalías o ataques a la variable medida, de tal manera que se caracterizó el valor del umbral α_f siguiendo la metodología presentada anteriormente

en la Sección (3.4), cuando el sistema presenta un comportamiento nominal ante la ausencia de ataques. La frecuencia de corte que se utilizó en el filtro de Butterworth pasa bajas es $\omega_c = 20$ [rad], obteniendo el valor de $\alpha_f = 0.57$ después de conocer el comportamiento de la señal $|y_f|$. El comportamiento de la salida del filtro $|y_f|$ y el parámetro de umbral α_f se muestran en la Figura 39.

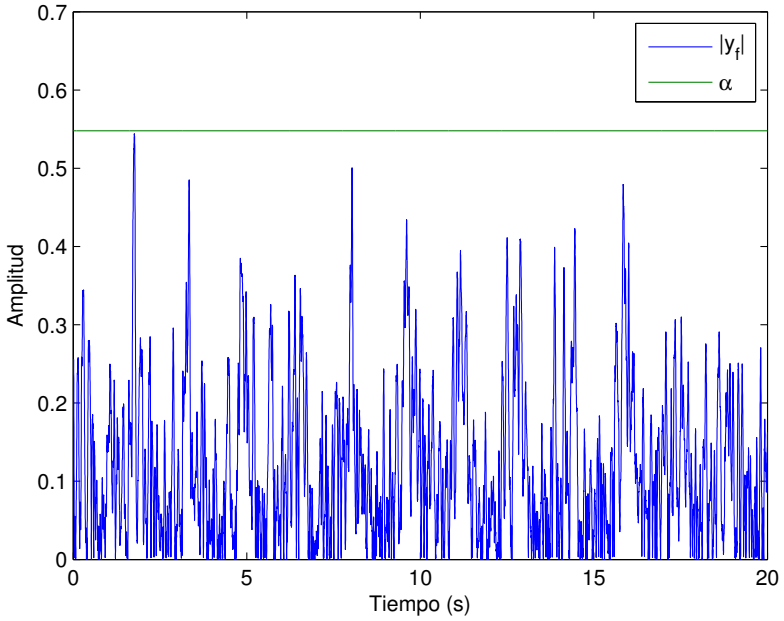

Figura 39. Valor del umbral α_f (línea horizontal verde) y valor absoluto de la salida del filtro $|y_f|$ (línea azul).

4.2.2. Aplicación de un ataque variante en el tiempo

Después de haber realizado la sintonización del detector estando el sistema bajo un comportamiento nominal, sin ser afectado por alguna anomalía se realizó la evaluación del detector. Se consideró un ataque invariante en el tiempo dado por $\delta(t) = 0.05\ sin(3t)$, el cual permaneció activo en el sistema desde el inicio de operación del circuito. En la Figura 40 se puede ver la salida original y y la salida atacada $\bar{y} = y + \delta$. Por otra parte, la Figura 41 muestra la convergencia del observador con el estado atacado.

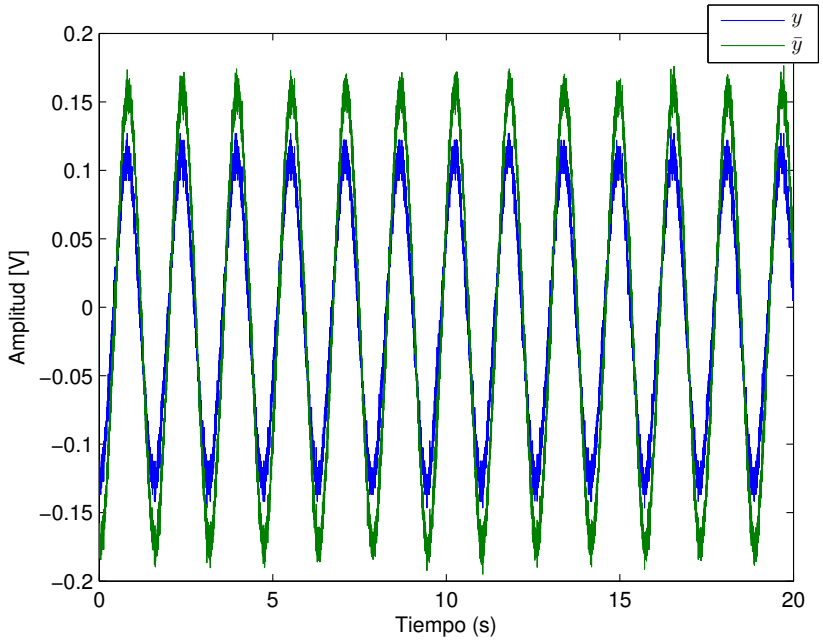

Figura 40. Salida original y y salida atacada $\bar{y}$.

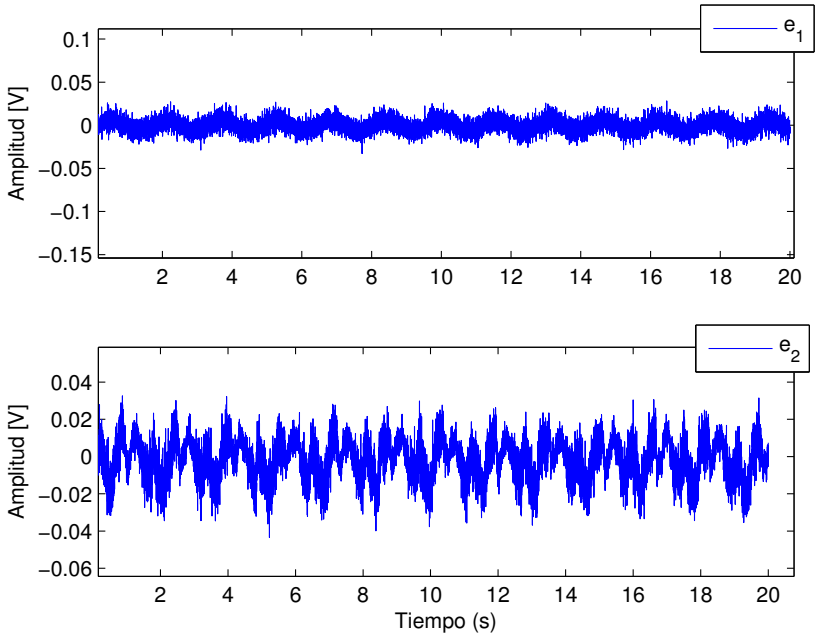

Figura 41. Errores de observación.

En la Figura 42 se puede observar que el algoritmo de detección funciona de manera adecuada, ya que cuando la amplitud de $|y_f|$ incrementa debido al ataque, se excede el valor de umbral α_f y como consecuencia el detector activa las alarmas.

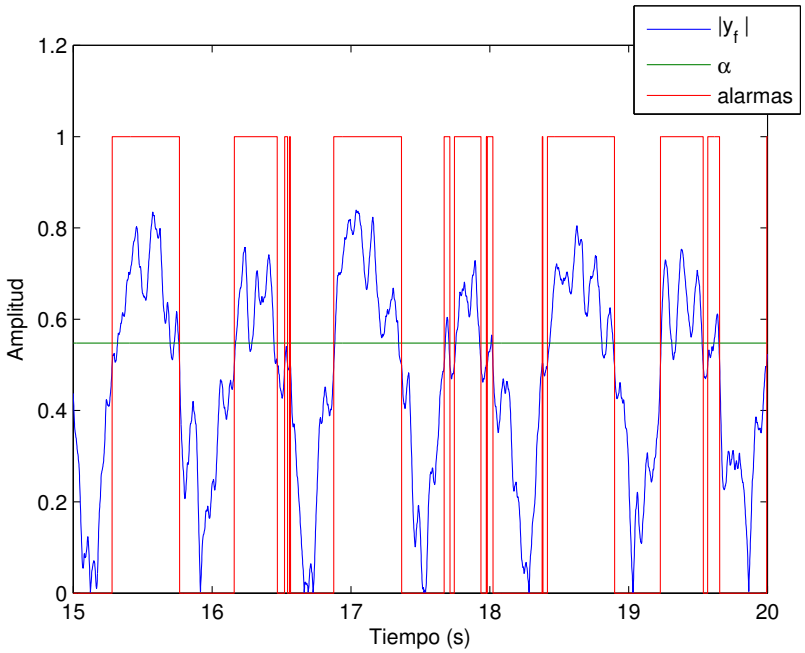

Figura 42. Activación de las alarmas debido a un ataque en la señal de salida.

Capítulo 5. Detección de ataques en sistemas conmutados

Muchos sistemas industriales presentan dinámicas conmutadas. Por ejemplo, en un proceso industrial existen válvulas que se abren y se cierran, sistemas de bombeo que son activados y desactivados, y sistemas de calefacción y enfriamiento que son accionados cada vez que la temperatura del proceso excede ciertos valores de umbral predeterminados, por mencionar solo algunos.

Este tipo de sistemas son suceptibles de ser atacados resultando en conmutaciones a instantes no deseados o planeados. Como ejemplo, podemos considerar un sistema de enfriamiento. El atacante puede corromper el sensor de temperatura de tal manera que la temperatura indicada por el sensor nunca rebase el valor de umbral al cual el sistema de enfriamiento debe activarse, resultando en un sobrecalentamiento o destrucción del sistema a enfriar.

Existen muchos detectores en la literatura, incluidos aquellos presentados en el Capítulo 2 de esta tesis. Sin embargo, estos detectores han sido desarrollados para sistemas con dinámica contínua o discreta pero su aplicación al caso de sistemas conmutados aún no ha sido investigada y no parece evidente que dichos detectores sean directamente aplicables al caso conmutado.

En este capítulo se presenta el diseño de un detector para sistemas conmutados. Al igual que en lo capítulos anteriores, el diseño del detector está basado en un observador discontínuo, el cual reconstruye el estado del sistema conmutado a la vez que permite identificar discrepancias entre la salida medida y la salida observada. Esta información se utiliza en un algoritmo que permite activar alarmas cada vez que las diferencias entre la salida medida y la salida observada excede cierto valor de umbral. Como caso particular de estudio, se presenta un sistema de dos tanques de agua, en el que el flujo de entrada a los tanques conmuta cada vez que el líquido en los tanques alcanza cierto valor. Los resultados obtenidos se ilustran por medio de simulaciones numéricas.

5.1. Descripción del sistema

Sea el sistema conmutado de primer orden

$$\dot{x} = -a_i x + b_i u, \qquad x \in \xi_i, \quad \xi_i \subset \mathbb{R} \tag{94}$$

$$y = x + \eta. \tag{95}$$

donde a_i, y b_i son constantes, $u \in R$, es la entrada de control, η es el ruido Gaussiano de media cero presente en el sensor, y el subíndice $i = (1, 2..., N)$ indica el número de modos de operación que tiene el sistema conmutado. El sistema opera en el modo i siempre y cuando $x \in \xi_i$.

Los sistemas conmutados descritos por (94)-(95) presentan diferentes modos de operación y tienen $(N-1)$ superficies de conmutación.

En este trabajo, el caso de interés es cuando la salida (95) del sistema conmutado (94) se encuentra sometida a un ataque externo $\delta \in R$, de tal manera que la salida atacada está dada por

$$\bar{y} = x + \delta + \eta. \tag{96}$$

Igual que en capítulos anteriores, se supone que la naturaleza de δ es en principio desconocida y cualquier valor de $\delta \neq 0$ será considerado un ataque.

5.2. Diseño del algoritmo de detección

Para detectar la presencia del ataque δ en la salida medida (96), se utilizará la metodología y las herramientas descritas en el Capítulo 4. Como primer paso, se realiza el diseño de un observador discontinuo para el sistema conmutado. Después, se introduce el filtro que será utilizado para identificar la presencia de anomalías en la salida $\bar{y}$, y finalmente, se presenta el algoritmo de detección y se proporciona un método heurístico para sintonizar los parámetros en el algoritmo.

5.2.1. Observador discontinuo

El observador aquí considerado para la reconstrución del estado del sistema conmutado (94) es una modificación del observador presentado en Rosas *et al.* (2007) y está definido de la siguiente manera

$$\begin{aligned} \dot{\hat{x}} &= -a_i\hat{x} + b_i u + c_1 e + c_2 \text{sign}(e) \qquad x \geq \xi_i, \\ \hat{y} &= \hat{x}, \end{aligned} \tag{97}$$

donde $c_1 \in \mathbb{R}_+, c_2 \in \mathbb{R}_+$ son ganancias y e es el error de observación, el cual está dado por

$$e = \bar{y} - \hat{y}. \tag{98}$$

Para el caso en que el sensor está libre de ruido y suponiendo que el ataque δ está acotado y es diferenciable respecto del tiempo, se tiene que la dinámica del error de observación está descrita por

$$\dot{e} = -(a_i + c_1)e + \delta a_i + \dot{\delta} - c_2 sign(e). \tag{99}$$

Por lo tanto, si se escogen las ganancias $c_1 > 0$ y c_2 'suficientemente grande', entonces, siguiendo el desarrollo presentado en Rosas *et al.* (2007) es posible mostrar que el sistema del error (99) es estable.

Cuando el sistema alcanza la superficie de discontinuidad $e = 0$, se tiene que

$$\overline{c_2 sign(e)} = \delta a_i + \dot{\delta}. \tag{100}$$

donde la barra sobre el término discontinuo indica promediación.

Sin embargo, cabe mencionar que dicho análisis no contempla la conmutación del sistema, por lo que se requiere de un estudio de estabilidad más riguroso, el cual está más allá de este trabajo de tesis. Por ahora, el análisis realizado garantiza que el observador es capaz de reconstruir el estado cuando ambos, el sistema y el observador,

se encuentran operando en el modo i.

5.2.2. Filtro pasa bajas

Para poder obtener el promedio del término discontinuo, el cual contiene información sobre el ataque, véase la sección 5.2.1, se utiliza el filtro de Butterworth de segundo orden pasa bajas

$$\dot{x}_f = A_f x_f + B_f u_f, \tag{101}$$

$$y_f = C x_f, \tag{102}$$

donde $x_f = [x_{f1}\ x_{f2}]^T$ y las matrices A_f,B_f y C_f están dadas en (60) y el término discontinuo del observador (98) es la entrada del filtro, es decir

$$u_f = c_2 sign(e). \tag{103}$$

Por otro lado, la salida del filtro coincide (aproximadamente) con el control equivalente dado en la sección 5.2.1

$$y_f \approx \overline{c_2 sign(e)} = \delta a_i + \dot{\delta}. \tag{104}$$

5.2.3. Algoritmo de detección

Finalmente, se presenta el algoritmo de detección, cuyo desarrollo es muy similar al algoritmo presentado en la Sección 3.4. En la implementación del algoritmo se usa la salida y_f, véase la ecuación (104), del filtro.

El detector está descrito de la siguiente manera

$$\begin{cases} |y_f(t)| \leq \alpha_f \longrightarrow \text{no alarma} \quad \longrightarrow \quad alarma = 0, \\ |y_f(t)| > \alpha_f \quad \longrightarrow \text{alarma} \quad \longrightarrow \quad alarma = 1. \end{cases} \tag{105}$$

Por el momento, no se encontró un procedimiento formal para determinar el valor del umbral α_f. En su lugar, el valor de α_f es obtenido de manera heurística:

1. Primero, se supone que inicialmente el sistema no es atacado, es decir, $\delta = 0$.

2. Después, se mide la salida del filtro y_f en estado estacionario (sin transitorio) y por un instante de tiempo suficientemente largo.

3. Finalmente, el valor de α es calculado, a partir de la serie de tiempo obtenida en el paso 2, de la siguiente manera

$$\alpha_f = max(|y_f(t)|). \tag{106}$$

5.3. Caso de estudio

En esta sección se diseña un detector siguiendo la metodología antes mencionada para un sistema conmutado, el cual se muestra en la Figura 43. Dicho sistema consiste de dos tanques con agua, los cuales tienen un flujo de salida constante. En la parte de arriba de los tanques hay una manguera, la cual está dedicada a llenar a uno u otro tanque, pero no a los dos al mismo tiempo. Se supone que la manguera puede cambiar instantáneamente entre los tanques.

Sean $x_i \geq 0$ y $v_i > 0$ el volumen de agua y el flujo de salida constante, respectivamente, para el tanque i, con $i = 1, 2$ y sea $\omega > 0$ el flujo constante de agua que alimenta a los tanques (uno a la vez). El objetivo es mantener el volumen de agua en los tanques por encima de los valores r_1 y r_2, los cuales están indicados con una línea punteada en la Figura 43. Además, se supone que inicialmente, el volumen de agua en los tanques está por encima de r_1 y r_2, respectivamente, es decir, se considera el caso en que $x_1(0) > r_1$ y $x_2(0) > r_2$. Finalmente, se supone que inicialmente, la manguera vierte agua sobre el tanque 1 y cuando $x_2 \leq r_2$ entonces la manguera se cambia instantáneamente al tanque 2. Después, cuando $x_1 \leq r_1$ la manguera se cambia de regreso al tanque 1 y así sucesivamente (Lin y Antsaklis, 2014).

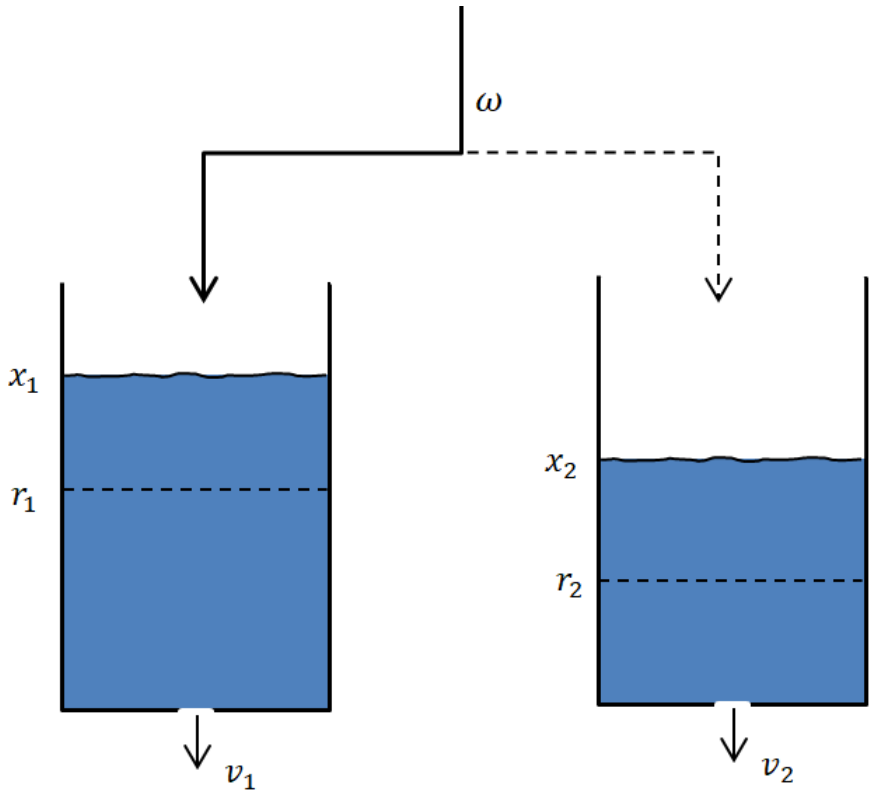

Figura 43. Sistema de dos tanques.

Bajo las consideraciones antes mencionadas, el sistema de tanques mostrado en la Figura 43 puede modelarse como un sistema híbrido con dos modos discretos y dos estados continuos. Los dos modos discretos están dados por q_1: tanque 1 llenándose y q_2: tanque 2 llenándose. En cada uno de estos modos discretos los estados continuos, dados por x_1 y x_2, evolucionan de manera diferente.

Para el caso del **modo q_1**, la evolución temporal del volumen de agua en los tanques está gobernada por

$$\dot{x}_1 = \omega - v_1, \tag{107}$$

$$\dot{x}_2 = -v_2. \tag{108}$$

El sistema permanece en este modo mientras se satisfaga que $x_2 \geq r_2$. Cuando esta condición no se satisface entonces el sistema conmuta al **modo q_2**, en el cual la

evolución temporal está descrita por las ecuaciones

$$\dot{x}_1 = -v_1, \tag{109}$$

$$\dot{x}_2 = \omega - v_2. \tag{110}$$

El sistema permanece en este modo mientras se satisfaga la condición $x_1 \geq r_1$ y cuando esta condición no se cumple entonces el sistema conmuta al modo q_1.

Este modelo, el cual representa un sistema híbrido, puede caracterizarse como el autómata híbrido mostrado en la Figura 44.

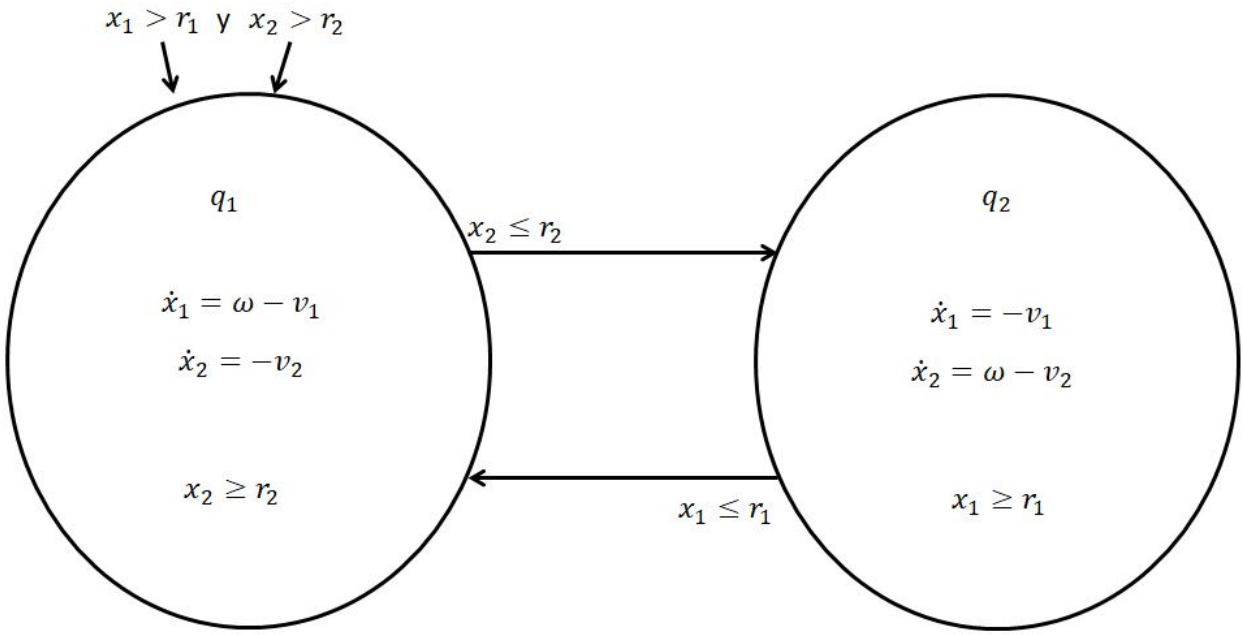

Figura 44. Autómata hibrido para el sistema de dos tanques de la Figura 43.

Finalmente, se considera que la salida medible en el tanque 1 y 2 es el volumen de agua x_1 y x_2, respectivamente. Y, además, se supone que la medición del sensor del tanque j, $j = 1, 2$, está contaminada con ruido Gausiano de media cero η_j, y que dicha medición está sujeta a un ataque δ_j.

En otras palabras, se considera el caso en que las salidas del tanque 1 y 2 están dadas por

$$y_j = x_j + \eta_j + \delta_j, \quad j = 1, 2. \tag{111}$$

5.3.1. Diseño del observador

A continuación, se presenta el diseño de los observadores para reconstruir el estado x_1 y x_2 correspondientes al volumen de agua en los tanques 1 y 2, respectivamente. Los observadores tienen la forma (41) y se tiene un observador por sensor y por modo de operación.

Cuando el sistema de tanques evoluciona en el modo q_1, los observadores están descritos por

$$\textbf{modo } \boldsymbol{q_1} = \begin{cases} \dot{\hat{x}}_1 = \omega - v_1 + c_1 e_1 + c_2 sign(e_1) + y_{f1}, \\ \dot{\hat{x}}_2 = -v_2 + c_3 e_2 + c_4 sign(e_2) + y_{f2}, \end{cases} \tag{112}$$

donde

$$e_j = \bar{y}_j - \hat{y}_j, \quad j = 1, 2, \tag{113}$$

con $\bar{y}_j$, $j = 1, 2$ dada en (111) y $\hat{y}_j = \hat{x}_j$.

Por otra parte, cuando los tanques operan en el modo q_2, los observadores están dados por

$$\textbf{modo } \boldsymbol{q_2} = \begin{cases} \dot{\hat{x}}_1 = -v_1 + c_1 e_1 + c_2 sign(e_1) + y_{f1}, \\ \dot{\hat{x}}_2 = \omega - v_2 + c_3 e_2 + c_4 sign(e_2) + y_{f2}. \end{cases} \tag{114}$$

Nótese que en ambos modos, los observadores contienen el término y_{fj}, $j = 1, 2$, el cual corrresponde a la salida del filtro pasa bajas j, descrito en la ecuación 101.

Finalmente, se muestra el autómata híbrido que describe el modelo de los observadores para los diferente modos de operación.

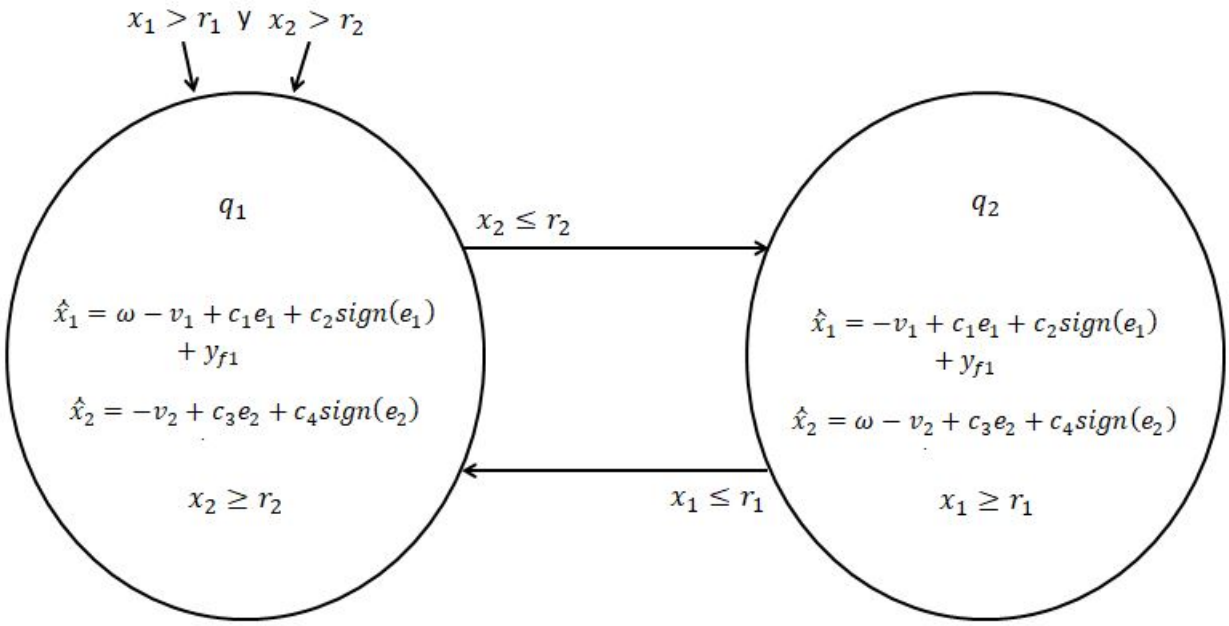

Figura 45. Autómata híbrido del observador.

5.3.2. Evolución del sistema con ruido en el sensor pero sin ataque en la salida medida

A continuación, se caracteriza el comportamiento dinámico del sistema de tanques para el caso en que la salida está libre de ataques, es decir, para el caso en que $\delta_j = 0$ en (111).

En el análisis consideramos el sistema conmutado (107)-(110), con parámetros $\omega =$ 60, $v_1 = 30$, $v_2 = 20$, $r_1 = 25$ y $r_2 = 20$, junto con los observadores (112)-(114) con ganancias c_1,c_2,c_3,c_4 = 70. Además, por cada tanque consideramos un filtro pasa bajas de la forma (101) con frecuencia de corte $\omega_{c1} = 20$.

Por otra parte, se considera que las salidas y_j, $j = 1, 2$ son como se definió en la ecuación (111) con ruido Gausiano η_j de media cero y covarianza 0.001 y $\delta_j = 0$.

Los resultados obtenidos se muestran en la Figura 46, en la cual se observan las salidas medidas y las salidas observadas. De esta figura se infiere que las salidas observadas convergen a las salidas medidas del sistema. Sin embargo, debido a la presencia del ruido Gausiano en la medición, los errores de observación no convergen a cero, tal y como se puede apreciar en la Figura 47.

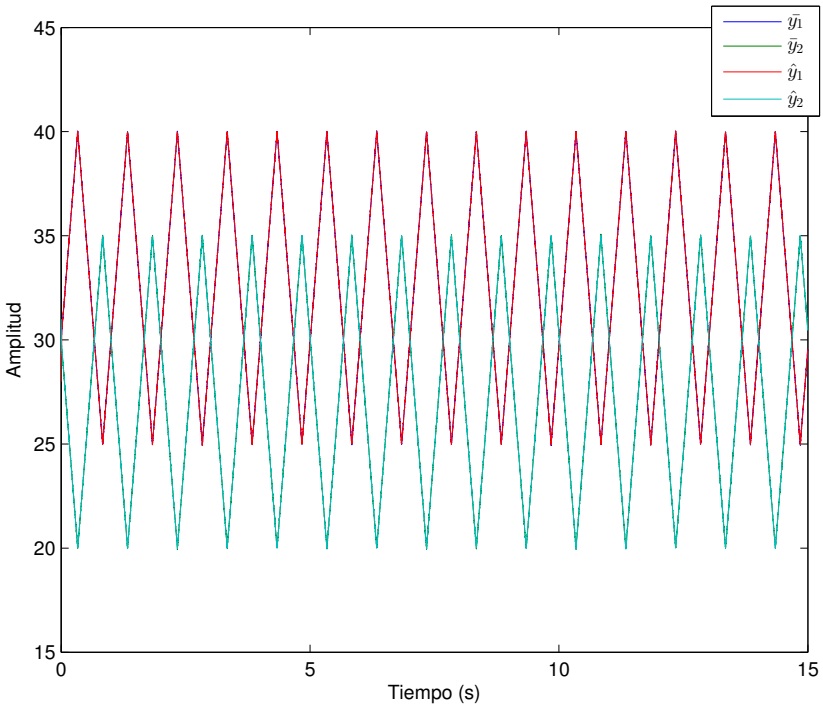

Figura 46. Convergencia de los estados observados con los del sistema.

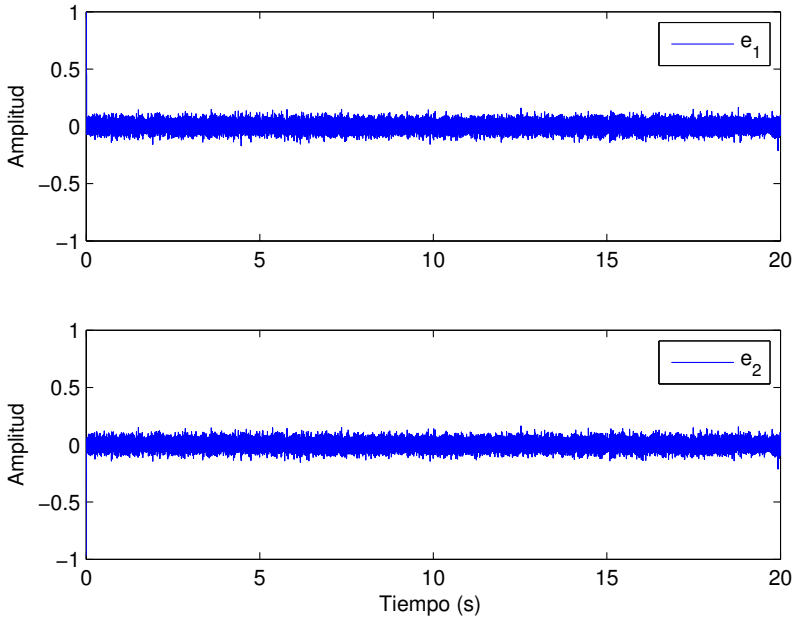

Figura 47. Errores de observación.

Posteriormente, el detector se sintoniza utilizando el algoritmo descrito en la Sección 5.2.3. Como primer paso se requiere obtener la salida (sin comportamiento tran-

sitorio) del filtro pasa bajas (101). Esto se muestra en la Figura 48, en donde la señal azul corresponde al valor absoluto de la salida del filtro pasa bajas en el sensor 1 (panel superior) y el sensor 2 (panel inferior). El máximo de dichas señales está indicado por una línea horizontal roja. En el caso del filtro del sensor 1, el máximo es 0.241 mientras que para la señal del filtro del sensor 2, el máximo es 0.254.

Finalmente, esta información es usada para sintonizar el detector de ataques (105) en el sensor 1 y el sensor 2. En particular, para el detector del sensor 1, el valor de α_f en (105) es $\alpha_f = \alpha_1 = 0.1791$, mientras que para el detector del sensor 2, el valor de α_f en (105) es $\alpha_f = \alpha_2 = 0.254$.

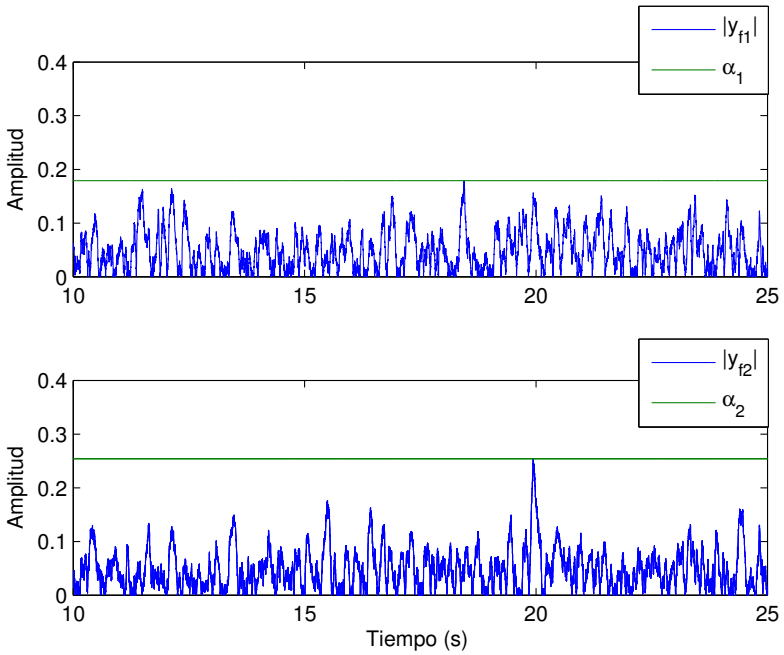

Figura 48. Valor del umbral α_1 y α_2.

Finalmente, para verificar que los detectores (105) han sido correctamente sintonizados, se vuelve a correr una simulación y una vez que el sistema está en estado estacionario se activa el detector. Dado que no se ha introducido ningún ataque, el detector no leventa ninguna alarma. Por consiguiente, el detector ha quedado correctamente sintonizado.

En la siguiente sección se introducirá un ataque para ver si el detector es capaz de identificarlo.

5.3.3. Ataque variante en el tiempo

A continuación, para ilustrar y verificar el funcionamiento de los detectores de ataques aquí propuestos, se considerará el caso en que las salidas (111) del sistema están sujetas a los siguientes ataques

$$\delta_1 = \begin{cases} 0 & \text{si} \quad 0 \leq t < 12, \\ 0.06cos(5t) & \text{de otra manera,} \end{cases} \tag{115}$$

y

$$\delta_2 = \begin{cases} 0 & \text{si} \quad 0 \leq t < 15, \\ 0.06cos(5t) & \text{de otra manera.} \end{cases} \tag{116}$$

En la Figura 49 se muestra como los ataques inducidos al sistema son sigilosos ya que la diferencia entre la salidas con ataque y sin ataque es prácticamente indetectable a simple vista. Sin embargo, los detectores sí son capaces de identificar los ataques, tal y como se muestra en las Figuras 50 y 51.

En particular, la Figura 50 muestra que cuando los ataques son activados, el valor de $|y_{fj}|, j = 1, 2$ excede el valor de umbral α_j y como consecuencia los detectores (105) encienden las alarmas, como se observa en la Figura 51.

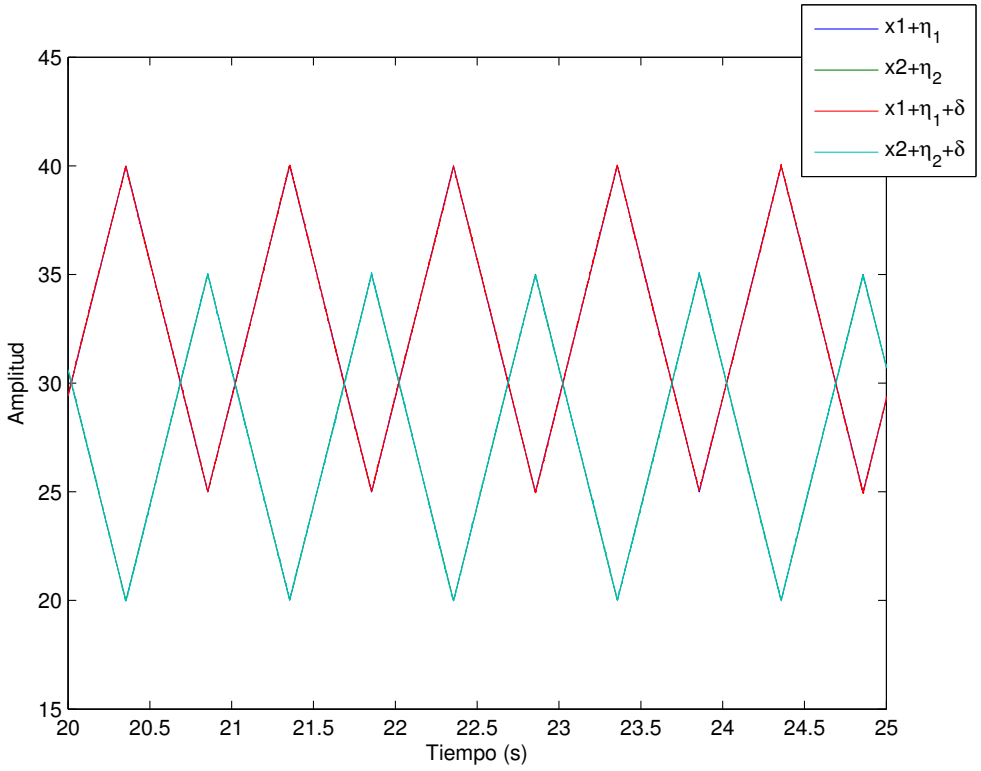

Figura 49. Comparación entre los estados del sistema sin ataque y después de ser atacados.

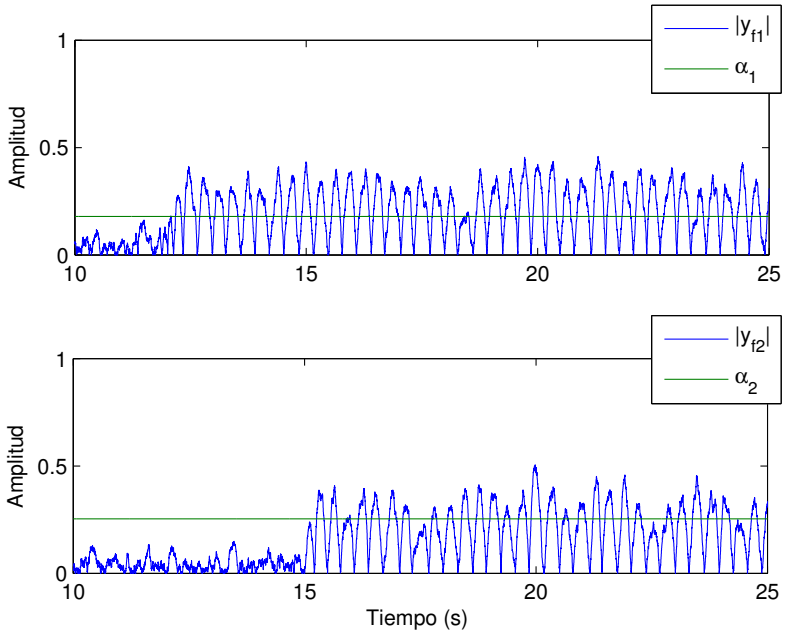

Figura 50. La salida del filtro $|y_{fj}|$, $j = 1, 2$ (línea azul) excede el valor de umbral α_j (línea verde) cuando el sensor es atacado.

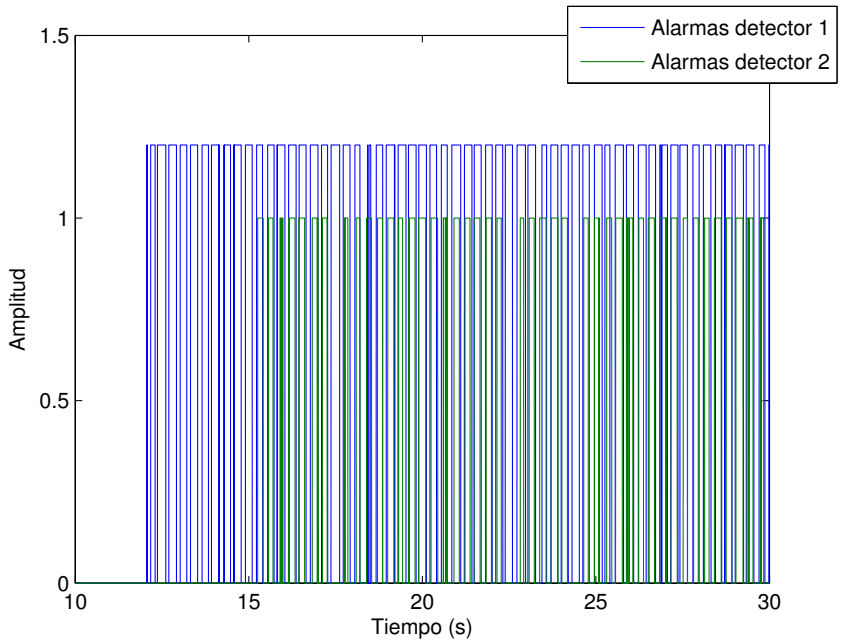

Figura 51. Los detectores activan las alarmas cuando los ataques (115) y (116) son aplicados a los sensores. Línea azul: alarmas cuando el sensor 1 está sujeto al ataque (115). Línea roja: alarmas para el caso en que el sensor 2 es contaminado con el ataque (116).

Capítulo 6. Conclusiones

En este trabajo de tesis se realizó el diseño de distintos detectores de ataques para dos clases de sistemas ciberfísicos, sistemas lineales de segundo orden y sistemas conmutados.

Al implementar y validar el detector de ataques experimentalmente, se observó que efectivamente el detector funciona. Además, los distintos experimentos realizados permitieron concluir que al realizar la caracterización del detector es recomendable sintonizar el observador usando ganancias 'pequeñas', debido a que la convergencia del error de observación se produce de manera más lenta y el detector se vuelve más sensible a pequeñas variaciones entre la salida original y la salida medida, la cual puede estar sujeta a un ataque.

Al trabajar con este detector numérica y experimentalmente se pudo concluir que el ruido Gaussiano juega un papel importante en la caracterización del detector, debido a que si la influencia del ruido en el sensor es muy grande, habrá ataques para los cuales el detector no será capaz de identificarlos, produciendo así falsos negativos en el algoritmo de detección. Este problema se puede reducir o eliminar mediante el filtrado de la salida medida del sistema antes de ser aplicada al observador, tal y como se observó en la Sección 3.7, ya que al filtrar la salida se puede obtener mayor sensibilidad en el detector.

Bajo ciertas condiciones, el detector aquí propuesto permite reconstruir el ataque, lo que representa una gran ventaja en comparación con los detectores clásicos disponibles en la literatura, los cuales no pueden reconstruir el ataque sino solamente identificar su presencia.

Como trabajo futuro se propone realizar la implementación del detector basado en un observador discontinuo, donde la comunicación entre el capa física y virtual sea de manera inalámbrica mediante algún protocolo de comunicación, por ejemplo, Wifi, Bluetooth, Ethernet, entre otros. Por otra parte, falta validar experimentalmente el detector para sistemas conmutados utilizando la teoría desarrollada en este proyecto de tesis. Finalmente, otra posible extensión sería encontrar el valor del umbral α_f en el detector (62) de manera analítica.

Literatura citada

Alvarez, J., Rosas, D., y Peña, J. (2009). Analog implementation of a robust control strategy for mechanical systems. *IEEE Transactions on Industrial Electronics*, **56**(9): 3377–3385.

Bangemann, T., Riedl, M., Thron, M., y Diedrich, C. (2016). Integration of classical components into industrial cyber–physical systems. *Proceedings of the IEEE*, **104**(5): 947–959.

Brook, D. y Evans, D. A. (1972). An approach to the probability distribution of cusum run length. *Biometrika*, **59**(3): 539–549.

Cárdenas, A. A., Amin, S., y Sastry, S. (2008). Secure control: Towards survivable cyber-physical systems. *Proceedings - International Conference on Distributed Computing Systems*, pp. 495–500.

Collins, S. y McCombie, S. (2012). Stuxnet: the emergence of a new cyber weapon and its implications. *Journal of Policing, Intelligence and Counter Terrorism*, **7**(1): 80–91.

Dibaji, S. M., Pirani, M., Flamholz, D. B., Annaswamy, A. M., Johansson, K. H., y Chakrabortty, A. (2019). A systems and control perspective of cps security. *Annual Reviews in Control*, **47**: 394 – 411.

Ding, D., Han, Q.-L., Xiang, Y., Ge, X., y Zhang, X.-M. (2018). A survey on security control and attack detection for industrial cyber-physical systems. *Neurocomputing*, **275**: 1674 – 1683.

Dunaka, P. R. y McMillin, B. (2017). Cyber-physical security of a chemical plant. *Proceedings of IEEE International Symposium on High Assurance Systems Engineering*, (60): 33–40.

Fawzi, H., Tabuada, P., y Diggavi, S. (2011). Secure state-estimation for dynamical systems under active adversaries. *2011 49th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2011*, pp. 337–344.

Fink, G., Edgar, T., Rice, T., MacDonald, D., y Crawford, C. (2017). Chapter 9 - security and privacy in cyber-physical systems. En: H. Song, D. B. Rawat, S. Jeschke, y C. Brecher (eds.), *Cyber-Physical Systems*. Academic Press, Boston, Intelligent Data-Centric Systems, pp. 129 – 141.

Gao, Z., Cecati, C., y Ding, S. X. (2015). A survey of fault diagnosis and fault-tolerant techniques-part i: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, **62**(6): 3757–3767.

Goh, J., Adepu, S., Tan, M., y Lee, Z. S. (2017). Anomaly detection in cyber physical systems using recurrent neural networks. En: *2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)*. pp. 140–145.

Karnouskos, S. (2011). Stuxnet worm impact on industrial cyber-physical system security. *IECON Proceedings (Industrial Electronics Conference)*, pp. 4490–4494.

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

Koren, I. (2018). Detecting and counteracting benign faults and malicious attacks in cyber physical systems. En: *2018 7th Mediterranean Conference on Embedded Computing (MECO)*. pp. 2–2.

Labinaz, G., Bayoumi, M., y Rudie, K. (1996). Modeling and Control of Hybrid Systems: A Survey. *IFAC Proceedings Volumes*, **29**(1): 4718–4729.

Lin, H. y Antsaklis, P. J. (2014). Hybrid dynamical systems: An introduction to control and verification. *Foundations and Trends in Systems and Control*, **1**(1): 1–172.

Loukas, G. (2015). A cyber-physical world. En: G. Loukas (ed.), *Cyber-Physical Attacks*. Butterworth-Heinemann, Boston, pp. 1 – 19.

Mitchell, R. y Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Computing Surveys (CSUR)*, **46**.

Mo, Y. y Sinopoli, B. (2009). Secure control against replay attacks. *2009 47th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2009*, pp. 911–918.

Mo, Y. y Sinopoli, B. (2016). On the Performance Degradation of Cyber-Physical Systems under Stealthy Integrity Attacks. *IEEE Transactions on Automatic Control*, **61**(9): 2618–2624.

Murguia, C. y Ruths, J. (2016a). CUSUM and chi-squared attack detection of compromised sensors. pp. 474–480.

Murguia, C. y Ruths, J. (2016b). Characterization of a CUSUM model-based sensor attack detector. *2016 IEEE 55th Conference on Decision and Control, CDC 2016*, (Cdc): 1303–1309.

Pasqualetti, F., Dorfler, F., y Bullo, F. (2013). Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, **58**(11): 2715–2729.

Roberge, J. K. (1975). *Operational amplifiers: theory and practice*, Vol. 197. Wiley New York.

Rosas, D., Alvarez, J., y Fridman, L. (2007). Robust observation and identification of ndof lagrangian systems. *International Journal of Robust and Nonlinear Control*, **17**: 842 – 861.

Serpanos, D. (2018). The cyber-physical systems revolution. *Computer*, **51**(3): 70–73.

Shakarian, P., Shakarian, J., y Ruef, A. (2013). Chapter 13 - attacking iranian nuclear facilities: Stuxnet. En: P. Shakarian, J. Shakarian, y A. Ruef (eds.), *Introduction to Cyber-Warfare*. Syngress, Boston, pp. 223 – 239.

Sun, S. F., Gu, D., y Huang, Z. (2014). Fully Secure Wicked Identity-Based Encryption Against Key Leakage Attacks. *Computer Journal*, **58**(10): 2520–2536.

Wang, Y., Gu, D., Peng, D., Chen, S., y Yang, H. (2012). Stuxnet vulnerabilities analysis of scada systems. En: J. Lei, F. L. Wang, M. Li, y Y. Luo (eds.), *Network Computing and Information Security*, Berlin, Heidelberg. Springer Berlin Heidelberg, pp. 640–646.

Zhang, T., Wang, Y., Liang, X., Zhuang, Z., y Xu, W. (2017). Cyber attacks in cyber-physical power systems: A case study with gprs-based scada systems. *Proceedings of the 29th Chinese Control and Decision Conference, CCDC 2017*, pp. 6847–6852.

Zhong, H., Du, D., Li, C., y Li, X. (2018). A novel sparse false data injection attack method in smart grids with incomplete power network information. *Complexity*, **2018**.

Anexo

Se presenta la versión final del artículo ya aceptado en el congreso The 58th IEEE Conference on Decision and Control (2019), a celebrarse del 11 al 13 diciembre, en Niza Francia, bajo el título de Filtering Approaches for Dealing with Noise in Anomaly Detection.

Filtering Approaches for Dealing with Noise in Anomaly Detection

Navid Hashemi[1], Eduardo Verdugo German[2], Jonatan Pena Ramirez[2], and Justin Ruths[1]

***Abstract*— The leading workhorse of anomaly (and attack) detection in the literature has been residual-based detectors, where the residual is the discrepancy between the observed output provided by the sensors (inclusive of any tampering along the way) and the estimated output provided by an observer. These techniques calculate some statistic of the residual and apply a threshold to determine whether or not to raise an alarm. To date, these methods have not leveraged the frequency content of the residual signal in making the detection problem easier, specifically dealing with the case of (e.g., measurement) noise. Here we demonstrate some opportunities to combine filtering to enhance the performance of residual-based detectors. We also demonstrate how filtering can provide a compelling alternative to residual-based methods when paired with a robust observer. In this process, we consider the class of attacks that are stealthy, or undetectable, by such filtered detection methods and the impact they can have on the system.**

I. INTRODUCTION

In anomaly detection, we seek to differentiate normal behavior from anomalous behavior - essentially anything that deviates away from the nominal model. This task is easy if our confidence is high in the nominal model but as uncertainty rises, distinguishing normal from anomalous becomes more challenging. From a control theory perspective, one of the most fundamental sources of uncertainty is measurement noise and in this work we present an intuitive, but to-date-unexplored, idea of combining traditional model-based detection schemes with low-pass filtering to reduce the impact of noise on our ability to detect anomalies in control systems. Beyond demonstrating that this approach enables better detection, the novelty lies in characterizing this performance boost analytically, in characterizing how this effects the impact an attacker can have on the system, and in understanding what new kind of attacks are possible when filtering is involved.

There is a deep literature from Fault Detection that leverages model-based detectors to identify the occurrence of faults [1]. More recently these tools have been reused in identifying the presence of attacks in control systems [2]-[11]. Attacks, in particular their strategic and exploitative nature, offer new challenges to both control theory and anomaly detection. While many research groups have rallied behind the banner of analyzing attacks in control systems and the challenges they raise, in what follows we describe the main representative contributions of these groups that relate to characterizing the tuning and performance of model-based detectors.

While all papers [2]-[11] investigate the use of model-base detectors, and [2]-[10] consider various types of attacks, it has only been recently that the impact of these attacks has been evaluated [6]-[10]. Of these [4], [8], [10] use the reachable set to characterize attack impact and [6], [7], [9] use various norms of system state or the estimation error covariance [5]. To a large extent this work on attacker impact has been facilitated by analytic methods to tune model-based detectors, as provided in [4], [7], [8], [9], [10], [11] and for work to define various types of worst-case attacks [3], [5]-[10]. In all of these cases Gaussian noise(s) are assumed on the measurements of the system (and possibly the system state), except in [11] which tunes a detector for generalized noise distributions.

While much has been done with model-based detectors, no study has integrated a low-pass filter to attenuate the effect of the noise on the attacker's capabilities to change the system state. The recent work to to tune classical detectors to desired levels of performance (desired false alarm rate) and to characterized the attacker's impact on the system state (given the attacker's desire to remain stealthy to classical detectors) positions us well to now add another layer, i.e., filtering, to the detection scheme. In Section III, we characterize the improvement gained by adding the filter (in terms of sensitivity to attacks and reduction in attacker capabilities) traded off with the new attack definitions that are stealthy to the introduced *low-pass chi-squared detector*.

This notion of filtering is not exclusive to retrofit conventional residual-based detectors, but generalizes in compelling ways to other types of model-based observers. In particular, in Section IV, we review how the discontinuous observer, an estimator that uses a sliding mode to yield finite-time convergence, produces a discontinuous term that can be filtered to produce an estimate of the disturbance that enters a system [12]. In this case, we reframe the disturbance estimation problem as a method to approximate the anomaly (e.g., failure and/or attack), which provides an attractive alternative to residual-based detectors. In this case the disturbance estimate not only provides a way to detect the presence of an anomaly, but also find the form of the anomaly so that advanced mitigation strategies can be employed - such as using the estimated anomaly as feedback to avoid the estimate to deviate away from the true state value.

In Section II we introduce the system and our notation as well as the attack context; in Section V we provide an

*This work is supported by the ConTex grant award 2018-56A and 2018-56B .

[1]These authors are with the Departments of Mechanical and Systems Engineering at the University of Texas at Dallas, Richardson, Texas, USA navid.hashemi@, jruths@utdallas.edu

[2]These authors are with the Department of Electronics and Telecommunications at the Scientific Research and Advanced Studies Center of Ensenada, Mexico (CICESE) jverdugo@cicese.edu.mx, jpena@cicese.mx

example that combines and compares both of these tools.

II. SYSTEM & CONTEXT

We consider a continuous, linear, time-invariant control system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + \eta(t), \end{cases} \tag{1}$$

in which $x(t) \in \mathbb{R}^n$ is the system state, $u(t) \in \mathbb{R}^m$ is the input, A$\in \mathbb{R}^{n\times n}$ is the system matrix, $B \in \mathbb{R}^{n\times m}$ is the input matrix, $y \in \mathbb{R}^p$ is the system output, and $\eta(t) \in \mathbb{R}^p$ is zero mean Gaussian additive measurement noise with covariance R, i.e., $\eta(t) \sim \mathcal{N}(0, R)$. The pair (A, C) is detectable and the pair (A, B) is stablizable.

We use the output y and an estimator to produce an estimate of the state, $\hat{x}(t)$. In this work we consider two different choices of estimators. We use this estimate for to construct the feedback law,

$$u(t) = K\hat{x}(t). \tag{2}$$

The process measurements are potentially corrupted by an attacker, who is able to read and arbitrarily change the sensor measurements. We model this as an additive attack $\delta(t) \in \mathbb{R}^p$ such that $y(t)$ is changed to $\bar{y}(t)$,

$$\bar{y}(t) = y(t) + \delta(t) = Cx(t) + \eta(t) + \delta(t). \tag{3}$$

A. Attacker Capabilities

In this work we assume that the attacker has access to all system information including, e.g., dynamics, states, estimator states, and detector parameters. In particular, the attacker can view (disclosure) and edit (disruption) sensor measurements.

III. RESIDUAL-BASED ANOMALY DETECTION

The residual quantifies the difference between what we receive from the sensor, $\bar{y}(t)$ and what we expect based on an estimator, $\hat{y}(t) = C\hat{x}(t)$,

$$r(t) = \bar{y}(t) - C\hat{x}(t), \tag{4}$$

Anomaly detectors use this quantity to make real-time choices in the state of the control system they monitor.

The Kalman filter is the standard choice of estimator used when employing residual based detectors. Here we specifically use the steady state Kalman filter has been chosen with observer gain $L \in \mathbb{R}^{n\times p}$.

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L(\bar{y}(t) - C\hat{x}(t)). \tag{5}$$

Armed with this estimator, we can express the difference between the state and expected state as the estimation error $e(t) = x(t) - \hat{x}(t)$, which leads to the following coupled closed-loop equations,

$$\begin{cases} \dot{x}(t) = \big(A + BK\big)x(t) - BKe(t), \\ \dot{e}(t) = \big(A - LC\big)e(t) - L\eta(t) - L\delta(t), \\ r(t) = Ce(t) + \eta(t) + \delta(t). \end{cases} \tag{6}$$

The steady state covariance of the estimation error can be calculated from the stochastic dynamics of $e(t)$ using the following Riccati equation,

$$\dot{P} = 0 = (A - LC)P + P(A - LC)' + LRL'. \tag{7}$$

Because the detection process is inherently a sampled approach we discretize (6) with uniform step size τ such that $\xi(k\tau) = \xi_k$,

$$\begin{cases} x_{k+1} = Fx_k + Ge_k, \\ e_{k+1} = He_k - L_d(\eta_k + \delta_k), \\ r_k = Ce_k + \eta_k + \delta_k. \end{cases} \tag{8}$$

Residual-based detection relies on quantifying the distribution of the residual under nominal operation, i.e., without attacks/anomalies. These statistics form a one-sided hypothesis test that is either accepted (no anomalies) or rejected (alarm raised). The discrete covariance of η_k is $\Sigma_\eta = \tau R$ and the discrete covariance of the estimation error is $\Sigma_e = \tau P$. In the absence of attack the distribution of the residual follows a zero-mean Gaussian distribution with the covariance [13],

$$\Sigma_r = E[r_k r_k'] = C\Sigma_e C' + \Sigma_\eta. \tag{9}$$

In the following subsections we will review the conventional chi-squared detector and subsequently extend this detector to enhance its performance.

A. Chi-Squared Detector

The role of the detector is to create a non-negative scalar valued random variable from the residual that can be easily compared with a threshold. One of the most widely used approaches to do this is given by the chi-squared detector which introduces a distance measure

$$z(t) = r'(t)\Sigma_r^{-1} r(t). \tag{10}$$

Since the residual is normally distributed, the distance measure, as a sum of squared Gaussian random variables, is chi-squared distributed (with p degrees of freedom, where p is the number of the sensors). By normalizing by the residual covariance, we eliminate system dependence. The detection procedure for the chi-squared detector is summarized as follows. For given a threshold $\alpha \in \mathbb{R}_{>0}$ and the distance measure in (10):

$$\begin{cases} z(t) \leq \alpha \quad \longrightarrow \quad \text{no alarm}, \\ z(t) > \alpha \quad \longrightarrow \quad \text{alarm}, \end{cases} \tag{11}$$

Since the chi-squared distribution has p degrees of freedom the distance measure has mean $E[z(t)] = p$ and covariance $E[z^2] = 2p$ during nominal (no anomaly) operation. In the presence of attacks/anomalies the distance measure will, in general, not fall according to a chi-squared distribution with mean value p and covariance $2p$. As indicated in (11), the detector is required to make a decision to raise or not raise an alarm at every time instant. Therefore, it is impractical to build a sample distribution of "current behavior" to decide if the distance measure was still distributed in chi-squared fashion. Instead, the alarms, and more specifically, the rate of

alarms can be used as a metric for how deviated the system is from nominal behavior.

Here *alarm rate* is defined as the rate of generation of alarm by the detector, which empirically is the fraction of time instants in which an alarm was raised. We expect alarms to be raised regularly, even under normal operation, because of the infinite support of the measurement noise. We can, however, predict the false alarm rate $\mathcal{A}$ - the alarm rate under normal operation - from the distance measure distribution, $\mathbf{Pr}(z \geq \alpha) = \mathcal{A}$. More importantly, we can set the false alarm rate to a desired value $\mathcal{A}^*$ by selecting the threshold α appropriately.

Lemma 1: [13]. Assume that there are no attacks to the system and consider chi-squared detector, with threshold $\alpha \in \mathbb{R}_{>0}$, $r_k \sim N(0, \Sigma)$. Let $\alpha = \alpha^* := 2P^{-1}(1-\mathcal{A}^*, \frac{p}{2})$, where $P^{-1}(\cdot,\cdot)$ denotes the inverse regularized lower incomplete gamma function, then $\mathcal{A} = \mathcal{A}^*$.

If the attacker wants to remain stealthy (undetected) to the detector, the threshold of the chi-squared detector puts a limit on the distance measure. It, therefore, limits the capabilities of stealthy attacks to change the system behavior. We derive worst-case stealthy attacks by assuming a powerful, knowledgeable attacker that knows the noisy measurement $Cx + \eta$ and also knows the state estimate $\hat{x}$, the attack can compensate for the terms in the residual

$$\delta_k = -\bar{y}_k + C\hat{x}_k + \Sigma_r^{\frac{1}{2}}\bar{\delta}_k, \tag{12}$$

such that

$$z_k = \bar{\delta}_k'\bar{\delta}_k. \tag{13}$$

By increasing the norm of $\bar{\delta}_k$ the attacker can increase the impact of the attack up to the point where the attack is identified by the detector.

In *zero alarm attacks* the attacker takes the perspective that raising no alarms is the way to avoid detection, essentially keeping the distance measure below the threshold,

$$z_k = \bar{\delta}_k'\bar{\delta}_k \leq \alpha, \tag{14}$$

and hence $\|\bar{\delta}_k\| \leq \sqrt{\alpha}$.

In *hidden attacks*, the attacker realizes that under normal operation alarms are generated at the rate of $\mathcal{A}^*$, hence, it is reasonable to generate an attack sequence that mimics the false alarm rate,

$$\Pr(z_k > \alpha) = \Pr\big(\|\bar{\delta}_k\| > \sqrt{\alpha}\big) = \mathcal{A}^*. \tag{15}$$

Hidden attacks are inherently more potent than zero-alarm attacks, especially when keeping in mind that the norm of the attack can be arbitrarily large at the time instants in which alarms are raised.

For more detail on these attacks and ways in which to evaluate the impact of these attacks through state deviation or the induced reachable set see [9], [10].

B. Filtered Chi-Squared Detector

In this section we introduce a modification to the chi-squared detector to leverage apriori knowledge that the noise typically exists in substantially higher frequencies than the disturbances as well as many anomalies. By filtering the residual signal, we aim to reduce the covariance of the statistic that forms the distance measure so that attacks are more readily apparent. Exploiting frequency domain information has not been combined with attack detection and here we introduce the approach and characterize its performance.

We filter the residual element-wise using a bank of p identical Butterworth filters (of second order) such that $\xi_i(t) = [\xi_{1i}(t), \xi_{2i}(t)]'$ and

$$\begin{cases} \dot{\xi}_i(t) = \Phi\xi_i(t) + \Psi r_i(t), \\ \rho_i(t) = [1\ 0]\xi_i(t) \end{cases} \tag{16}$$

for $i = 1, \ldots, p$ and where

$$\Phi = \begin{bmatrix} 0 & 1 \\ -\omega_c^2 & -\sqrt{2}\omega_c \end{bmatrix}, \quad \Psi = \begin{bmatrix} 0 \\ \omega_c^2 \end{bmatrix}. \tag{17}$$

and we use the filtered output $\rho(t)$ to construct a new distance measure. To characterize the distribution of this distance measure, first characterize the covariance of $\rho_i(t)$.

Theorem 1: Given a residual signal $r_k \sim \mathcal{N}(0, \Sigma_r)$ filtered through the Butterworth filter in (16)-(17) with bandwidth ω_c, the discrete-time filter with sampling rate τ output ρ_k is zero mean Gaussian with covariance,

$$\Sigma_\rho = \frac{\tau\omega_c}{2\sqrt{2}}\Sigma_r. \tag{18}$$

Proof: We first discretize the filter with time step τ

$$\begin{cases} \xi_k^{(i)} = \Phi_d\xi_k^{(i)} + \Psi_d r_k^{(i)}, \\ \rho_k^{(i)} = [1\ 0]\xi_k^{(i)} \end{cases} \tag{19}$$

where $\zeta_k^{(i)} = \zeta_i(k\tau)$ and for small enough τ, $\Phi_d = I + \Phi\tau$ and $\Psi_d = \Psi\tau$. To find the covariance of filtered residue vector ρ_k we collect all filters together,

$$\dot{\xi}_k = \widetilde{\Phi}\xi_k + \widetilde{\Psi}r_k, \tag{20}$$

with $\widetilde{\Phi} = \Phi_d \otimes I_p$ and $\widetilde{\Psi} = \Psi_d \otimes I_p$ and where $\xi_k \in \mathbb{R}^{2p}$, $\widetilde{\Phi} \in \mathbb{R}^{2p\times 2p}$ and $\widetilde{\Psi} \in \mathbb{R}^{2p\times p}$. We can define the covariance of ξ_k as $\mathcal{P} = [\mathcal{P}_{i,j}]$, $i = 1, \ldots, p$ and $j = 1, \ldots, p$, where $\mathcal{P}_{i,j} \in \mathbb{R}^{2\times 2}$. This total covariance is supplied by the discrete-time Riccati equation (in steady state),

$$0 = \widetilde{\Phi}\mathcal{P}\widetilde{\Phi}' - \mathcal{P} + \widetilde{\Psi}\Sigma_r\widetilde{\Psi}', \tag{21}$$

and following some algebra manipulation reveals that we can split this into Riccati equations on each subblock

$$\Phi_d\mathcal{P}_{i,j}\Phi_d' - \mathcal{P}_{i,j} + \Psi_d[\Sigma_r]_{i,j}\Psi_d' = 0. \tag{22}$$

From (22) we see that $\mathcal{P}_{i,j} = \mathcal{P}_{j,i}$ and in addition we know that $\mathcal{P}_{i,j} = \mathcal{P}_{j,i}'$, therefore matrix $\mathcal{P}_{i,j}$ is a symmetric square matrix. The covariance of $\rho(t)$ can be computed element-wise using the structure of Φ_d and Ψ_d leaving,

$$\begin{aligned} [\Sigma_\rho]_{i,j} &= [\mathcal{P}_{i,j}]_{11} \\ &= -\tau\omega_c\frac{(\tau\omega_c)^2 - \sqrt{2}\tau\omega_c + 2}{(\tau\omega_c)^3 - 3\sqrt{2}(\tau\omega_c)^2 + 8\tau\omega_c - 4\sqrt{2}}\Sigma_r \end{aligned} \tag{23}$$

If the sampling time is taken small enough the higher order terms vanish leaving

$$[\Sigma_\rho]_{i,j} = \frac{\tau\omega_c}{2\sqrt{2}}[\Sigma_r]_{i,j}, \tag{24}$$

which leads directly to (18). ■

Remark 1: Note that τ is the sampling time of the system, which highlights that the filtered covariance is expected to be significantly smaller than the covariance of the original residual. By reducing the covariance of the nominal behavior using this filtering technique the aim is to make it easier to distinguish attacks and anomalies (especially those with low frequency components).

Using this covariance, we can create a new normalized distance measure

$$z_k = \rho_k' \Sigma_\rho^{-1} \rho_k = \frac{2\sqrt{2}}{\tau\omega_c} \rho_k' \Sigma_r^{-1} \rho_k. \tag{25}$$

Because ρ_k is a Gaussian random variable, $z(t)$ is also chi-squared. In fact, because of the normalization by the covariance, this functions exactly like the conventional chi-squared detector (including using Lemma 1 to select the threshold). The major difference lies in the scale of the normalizing covariance; here being smaller provides an advantage to identify smaller attacks. To execute these attacks, attackers will also need to know the cut-off frequency of the filter. Analogous to (12), we define

$$\delta_k = -y_k + C\hat{x}_k + \left(\frac{\tau\omega_c}{2\sqrt{2}}\Sigma_r\right)^{\frac{1}{2}} \bar{\delta}_k, \tag{26}$$

where $\bar{\delta}_k$ is an independent random variable that attackers may select to define different classes of stealthy attacks.

As before the *zero alarm attack* is generated in a way so that no alarm is generated by the detector. The precise characterization of the sequence $\bar{\delta}_k$ is a topic for future work, however, intuitively the fact that the attack signal is also filtered by the low-pass filter, it follows that the frequency content (as quantified by a Fourier Transform) and not just the amplitude (norm) of the attack vector plays a role in quantifying zero alarm attacks. To start, note that zero alarm attacks of the conventional chi-squared detector will also be zero alarm attacks of the filtered chi-squared detector. On top of this, there is an opportunity to inject high frequency content into the attack, which would be attenuated by the action of filter. If the filtered chi-squared used an *ideal* low pass filter, it would be possible to inject any signal that had frequency content higher than the cut-off frequency of the filter. The second order filter requires some adjustment to balance the small contributions from the stopband frequencies with that of the passband frequencies.

Similar to the unfiltered case, we employ the zero alarm attack constraint in a probabilistic fashion for the corresponding *hidden attack*. In this case $\bar{\delta}_k$ is selected such that the Fourier frequency content is now bounded in probability. Again the full characterization of these attacks are for future work.

IV. DISCONTINUOUS OBSERVER

In this section, we present an anomaly detector based on a robust observer and a low pass filter for a particular class of second order systems. Although we consider a second order system here, any physical plant that can be partitioned into a collection of second order systems can be captured by an extended version of the presented methods [12]. The observer, which has a discontinuous term, is used to reconstruct the state of the plant/system and by filtering the discontinuous term with a low-pass filter it is possible to identify disturbances/anomalies. The design of this detector is explained in the following lines.

Consider system (1) with state $x = [x_1\ x_2]^T$, $x_i \in \mathbb{R}$ and

$$A = \begin{bmatrix} 0 & 1 \\ -a & -b \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{27}$$

and assume that the output of the system is the noisy and attacked output $\bar{y}$, see (3), with

$$C = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{28}$$

In order to reconstruct the state of system (1) and (27), we use the following robust observer, which has been proposed in [12]

$$\dot{\hat{x}} = A\hat{x} + Bu + \Gamma e + Bc_3 \text{sign}(\bar{y} - \hat{y}), \tag{29}$$

$$\hat{y} = C\hat{x}, \tag{30}$$

where $\hat{x} = [\hat{x}_1\ \hat{x}_2]^T$, $e = [(\bar{y} - \hat{y})\ (x_2 - \hat{x}_2)]^T$, $\hat{x}_i \in \mathbb{R}$, matrix A and vectors B, C as defined in (27)-(28) and

$$\Gamma_1 = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}. \tag{31}$$

According to [12], the positive gains $c_i\ i = 1, \ldots, 3$ can be chosen as follows

$$c_1 > 0, \quad c_2 > 0, \quad c_3 > 2\lambda_{max}(P)\sqrt{\frac{\lambda_{max}(P)}{\lambda_{min}(P)}}\left(\frac{c_2\rho}{\theta}\right), \tag{32}$$

where $P \in \mathbb{R}^{2\times 2}$ is the solution of the Lyapunov equation $A^TP + PA = -I$, with $A \in \mathbb{R}^{2\times 2}$ as given in (27), $I \in \mathbb{R}^{2\times 2}$ is the identity matrix, $\rho \in \mathbb{R}_+$ is an upper bound on the disturbances present on the system and θ is a positive parameter satisfying $0 < \theta < 1$.

Next, define the observation errors

$$e_1 = \bar{y} - \hat{y}, \quad \text{and} \quad e_2 = x_2 - \hat{x}_2. \tag{33}$$

Then, the corresponding observation error dynamics are described by

$$\dot{e}_1 = e_2 - c_1 e_1 + \dot{\eta} + \dot{\delta}, \tag{34}$$

$$\dot{e}_2 = -(a + c_2)e_1 - be_2 - c_3 \text{sign}(e_1) + a(\eta + \delta). \tag{35}$$

when the system reaches the discontinuous surface, it follows that $e_1 = \dot{e}_1 = 0$. Replacing this into (34) it follows that on the discontinuous surface, the observation error e_2 satisfies

$$e_2 = -\dot{\eta} - \dot{\delta}, \tag{36}$$

and consequently, it also holds that

$$\dot{e}_2 = -\ddot{\eta} - \dddot{\delta}. \tag{37}$$

Next, by using the equivalent control method [14], it follows from (35)-(37) that, on the discontinuous surface, the filtered version of the discontinuous term $c_3\text{sign}(e_1)$ satisfies

$$\overline{c_3\text{sign}(e_1)} = \ddot{\eta} + \dddot{\delta} + a(\eta + \delta) + b(\dot{\eta} + \dot{\delta}), \tag{38}$$

where the over bar on the discontinuous term denotes filtering.

In order to obtain $\overline{c_3\text{sign}(e_1)}$, we use the following low-pass Butterworth filter, written in state-space form

$$\dot{x}_f = A_f x_f + B_f c_3\text{sign}(e_1), \tag{39}$$

$$y_f = C_f x_f, \tag{40}$$

where $x_f = [x_{f1}\ x_{f2}]^T$ and

$$A_f = \begin{bmatrix} 0 & 1 \\ -\omega_c^2 & -1.4142\omega_c \end{bmatrix}, \quad B_f = \begin{bmatrix} 0 \\ \omega_c^2 \end{bmatrix}, \tag{41}$$

$$C_f = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{42}$$

where ω_c is the cut-off frequency of the filter. Then, by choosing a ω_c that minimizes the phase delay, it is possible to assume that

$$y_f \approx \overline{c_3\text{sign}(e_1)} = \ddot{\eta} + \dddot{\delta} + a(\eta + \delta) + b(\dot{\eta} + \dot{\delta}). \tag{43}$$

A. Anomaly output detector with robust observer and filter

In order to design the anomaly output detector, it is worth nothing that the output y_f of the filter, see (43), can indeed be seen as a *residual* because it contains information about the noise and the attack/anomaly in the output of the plant.

Then, similar to (11), we define the following detector

$$\begin{cases} y_f(t) \leq \alpha_f & \longrightarrow \quad \text{no alarm}, \quad \longrightarrow \quad alarm = 0, \\ y_f(t) > \alpha_f & \longrightarrow \quad \text{alarm} \quad \longrightarrow \quad alarm = 1. \end{cases} \tag{44}$$

For the time being, we do not have a formal procedure for determining the upper threshold α_f. Instead, the value of α_f is empirically determined as follows. First, note that for the free-attack case, the output of the filter is approximately given by

$$y_{free-attack} \approx \ddot{\eta} + a\eta + b\dot{\eta}. \tag{45}$$

Hence we can obtain the value of α_f by directly measuring $y_{free-attack}$ at the output of the filter for sufficiently long time and then the value of α_f will correspond to the L_∞-norm computed from the obtained measurement.

Remark 2: Note that for the free-noise case, i.e. $\eta = 0$, and considering that the attack is constant, i.e. $\delta := \delta_0$, with $\delta_0 \in \mathbb{R}$, then the output of the filter, see (43), satisfies

$$y_f \approx a\eta. \tag{46}$$

Hence, in this case, it is possible not only to identify that there is an attack in the output but also it is possible to reconstruct the attack. Note that this fact opens the possibility of using the estimated attack in the control in order to make the system/plant immune to constant attacks by preventing the drift of the estimation error. Furthermore, in this case there exist, however, a drawback: to estimate the attack δ_0 it is necessary to have perfect knowledge of parameter a, which indeed is a parameter of the plant.

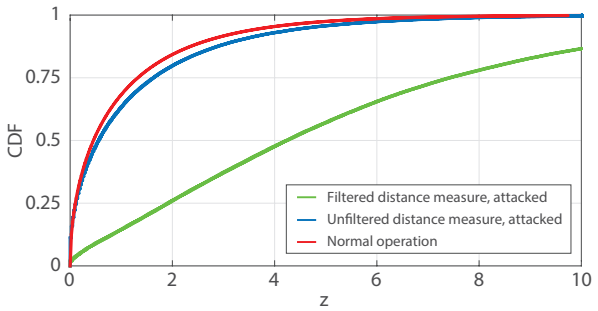

Fig. 1. This figure shows the performance of the proposed filter-based detector. Using a threshould of $\alpha = 3.84$ both detectors have been tuned to provide a $\mathcal{A}^* = 0.05$ (5%) false alarm rate. The green curve plots the cumulative distribution function of the distance measure under attack when computed from the filtered residual ρ_k and exhibits a dramatic difference from what the distribution looks like under normal operation (red). This is quantified by a higher than expected false alarm rate $\mathcal{A} = 0.55$ (55%). In contrast, the blue curve plots the cumulative distribution function of the distance measure under attack when computed from the residual without filtering. Not only does the distribution stay quite similar under attack, but the alarm rate is 7% which is only 2% bigger than the false alarm rate and would not supply as robust a detection.

V. EXAMPLE

This section presents a numerical example, which illustrates the performance of the detectors presented in Sections III and IV. In particular, we consider a second order system of the form (1) with

$$A = \begin{bmatrix} 0 & 1 \\ -4 & -20 \end{bmatrix}, \ B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \ C = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{47}$$

Furthermore it is assumed that the output is influenced by zero-mean Gaussian noise $\eta(t) \sim \mathcal{N}(0, R)$.

A. Residual-Based Methods

We now demonstrate these tools and provide a comparison between the filtered and unfiltered version of the chi-squared detector. The system matrices are provided above; here we use a control gain $K = [1, 1]$ and an observer gain $L = [0, 2]'$. The sampling time is $T = 0.001$ and the corresponding measurement noise covariance is $\Sigma_\eta = 2$.

The idea behind this comparison is that filtration makes the covariance of the filtered residual smaller and removes the high frequency content of the measurement noise, therefore, the attack will be more identifiable when filter is applied. This ease of detecting the attack is apparent when the alarm rate corresponding to the filtered detector deviates more significantly from the desired/tuned false alarm rate, indicating the detector's sensitivity to the attack is higher. We consider a simple attack to demonstrate this point,

$$\delta(t) = 1. \tag{48}$$

Figure 1 shows the cumulative distribution of distance measure calculated from r_k under attack (blue) compared

with the cumulative distribution of the distance measure under no attacks (red). Because the distributions are closely similar, the conventional chi-squared detector is not able to easily distinguish the attack. This figure also compares the cumulative distribution of the distance measure calculated from the filtered residual ρ_k (green). In this case the small attack significantly changes the distribution of the distance measure and detection is easy. This is captured quantitatively by comparing the false alarm rates. Here the detectors were tuned to provide $\mathcal{A}^* = 0.05$ (5%) false alarms (i.e., under no attacks) by selecting the threshold as $\alpha = 3.84$. Under attack the conventional detector's alarm rate was 7% whereas the filtered detector's alarm rate was 55%, demonstrating the dramatic difference in sensitivity provided by the additional filtration.

Because the residual and filtered residual are not themselves normalized (as opposed to the distance measure which is normalized), the role that covariance plays is quite clear in Fig. 2. Here the reduced covariance of the filtered residual clearly distinguishes the attack scenario from the normal operation. This distinction is less obvious in the unfiltered residual.

B. Discontinuous Observer

For the case of the observer-based detector discussed in Section IV, we consider observer (29)-(30) with A and B as" given in (47) and the gains (32) are chosen as follows: $c_1 = c_2 = 5$, and $c_3 = 12$. For this choice of c_i, $i = 1, \ldots, 3$, the observation error is globally asymptotically stable for the case of free-noise ($\eta = 0$) and free-attack ($\delta = 0$), see [12].

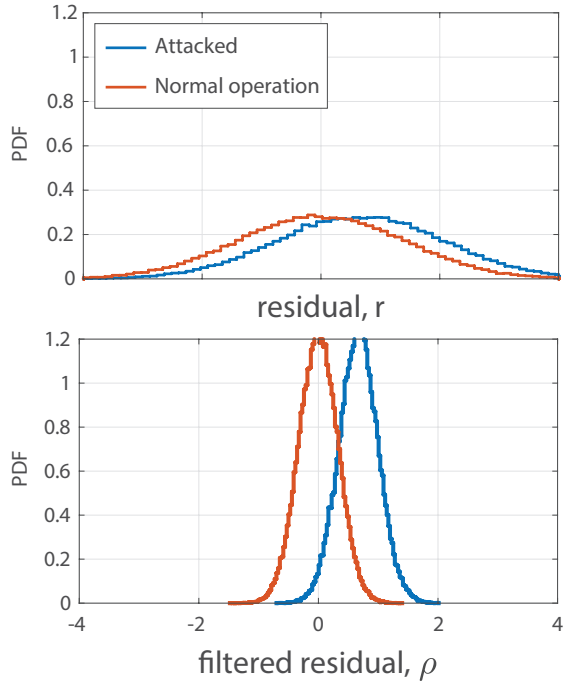

Fig. 2. The probability density functions, since they are not normalized, aptly demonstrate the role of shrinking covariance in the detection problem. The difference between the attacked and nominal behavior is more dramatic in the filtered residual ρ_k (lower panel) than in the residual r_k (upper panel).

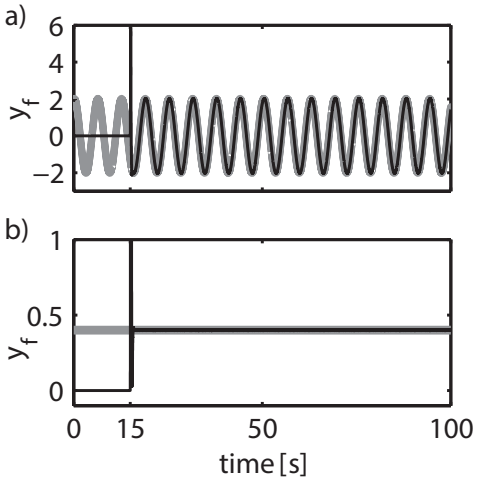

Fig. 3. Residual signal for the free-noise case. a) Residual when the output is subject to the time varying attack (49). Black line: measured residual. Gray line: predicted residual from (43). b) Residual for the constant attack (50). Black line: measured residual. Gray line: predicted residual from Eq. (46).

Next, in order to obtain the residual signal y_f, see (43), we use the second order low-pass filter (39)-(42) with cut-off frequency $\omega_c = 12$.

After that, we obtain the threshold value α_f, see (44). This requires numerically integrating the equations describing the system, the observer, and the filter with parameter values as described above. The numerical integration is performed by using Simulink (MATLAB) and the solver Runge-Kutta with fixed step size of 0.001. Further it is assumed that the output of the system contains zero-mean Gaussian noise η with covariance $R = 0.001$. Finally, by using the empirical procedure mentioned in Subsection IV-A, we obtain that the threshold value in the detector (44) is $\alpha_f = 1.55$.

Now that we have characterized the detector (44), we proceed to evaluate the performance of this detector by considering two different attacks, namely a time varying attack and a constant attack.

The time varying attack is assumed to be described by

$$\delta(t) = \begin{cases} 0 & \text{if} \quad t \leq 15, \\ 0.1 \sin t & \text{otherwise}, \end{cases} \tag{49}$$

whereas the constant attack is simply given by

$$\delta(t) = \begin{cases} 0 & \text{if} \quad t \leq 15, \\ 0.1 & \text{otherwise}. \end{cases} \tag{50}$$

The identified residual y_f for each attack is shown in Fig. 3. For the case of the time varying attack (49) the measured output of the filter for the free-noise case is shown by the black signal in Fig. 3a). Note that the measured output indeed corresponds to the expected output computed from (43) setting $\eta = 0$. On the other hand, the residual for the case of the constant attack (50) is shown in Fig. 3b), see the black signal. Also for this case, the measured residual corresponds to the predicted residual computed from (46), which is denoted by the gray signal.

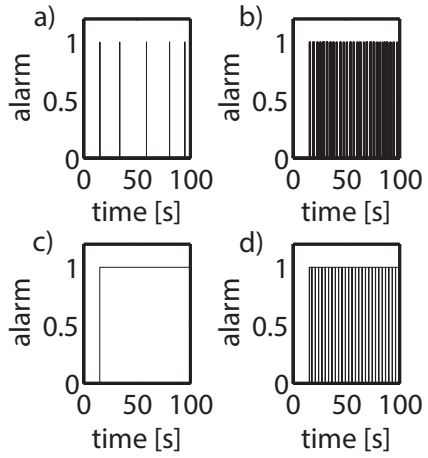

Fig. 4. Output of the detection algorithm (44) for diffrent attacks δ. a) $\delta = 0.1$, b) $\delta = 0.1 \sin t$, c) $\delta = 1.1$, d) $\delta = 1.1 \sin t$. In all cases the attack is applied for $t \geq 15$.

Finally, the performance of the detector when the output is subjected to noise and to attack, i.e $\eta \neq 0$, $\delta \neq 0$, is illustrated in Fig. 4. In particular, Figs. 4a) and 4c) show the output of the detector, i.e., variable $alarm$ in (44) for the case of a constant attack, which is applied for $t \geq 15$. The result shown in Fig. 4a) corresponds to the 'small' attack $\delta = 0.1$, whereas Fig. 4c) shows the alarms obtained for the 'large' attack $\delta = 1.1$. From these figures it is easy to see that when the attack is small, there are instants at which the attack is hidden from the detector and therefore there are instants at which no alarms are raised even though the output is being attacked, see Fig. 4a). On the other hand, when the attack is large ($\delta = 1.1$), the detector keeps the alarm raised during all the time that the output is being attacked, as clearly shown in Fig. 4c).

A similar situation occurs for the cases that the output is subjected to a sinusoidal attack, see Figs. 4b) and 4d). In this case, however, the alarm is always switched-off whenever the sinusoidal attack crosses the zero axis, which is expected since at that instants the attack is zero.

VI. CONCLUSIONS

In this paper we've introduced two techniques to exploit the a priori knowledge that the noise injects high frequency disturbance to aid in making the attack/anomaly detection task easier. In the first case, we retrofit the conventional chi-squared detector with a filter to reduce the covariance of the detector's distance measure statistic. In the second case, we adopted a robust observer which has a discontinuous term that reveals the presence of an attack or anomaly in the system. As fault and anomaly detection tools are poised to bridge to industry, such practical techniques are compelling ways to boost performance and constraint potential attackers even further.

REFERENCES

[1] J. Chen and R. J. Patton, *Robust Model-based Fault Diagnosis for Dynamic Systems*. Norwell, MA, USA: Kluwer Academic Publishers, 1999.

[2] A. Cárdenas, S. Amin, Z. Lin, Y. Huang, C. Huang, and S. Sastry, "Attacks against process control systems: Risk assessment, detection, and response," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, 2011, pp. 355–366.

[3] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

[4] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Transactions on Automatic Control*, vol. 61, pp. 2618–2624, 2016.

[5] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal Linear Cyber-Attack on Remote State Estimation," *IEEE Transactions on Control of Network Systems*, vol. PP, no. 99, pp. 1–10, 2016.

[6] J. Milošević, D. Umsonst, H. Sandberg, and K. H. Johansson, "Quantifying the impact of cyber-attack strategies for control systems equipped with an anomaly detector," in *2018 European Control Conference (ECC)*. IEEE, 2018, pp. 331–337.

[7] D. Umsonst and H. Sandberg, "Anomaly detector metrics for sensor data attacks in control systems," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 153–158.

[8] C. Murguia and J. Ruths, "On reachable sets of hidden cps sensor attacks," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 178–184.

[9] R. Tunga, C. Murguia, and J. Ruths, "Tuning windowed chi-squared detectors for sensor attacks," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 1752–1757.

[10] N. Hashemi, C. Murguia, and J. Ruths, "A comparison of stealthy sensor attacks on control systems," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 973–979.

[11] N. Hashemi and J. Ruths, "Generalized chi-squared detector for lti systems with non-gaussian noise," in *2019 Annual American Control Conference (ACC)*. IEEE, 2019.

[12] D. I. Rosas Almeida, J. Alvarez, and L. Fridman, "Robust observation and identification of ndof lagrangian systems," *International Journal of Robust and Nonlinear Control*, vol. 17, no. 9, pp. 842–861. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rnc.1156

[13] C. Murguia and J. Ruths, "Cusum and chi-squared attack detection of compromised sensors," in *proceedings of the IEEE Multi-Conference on Systems and Control (MSC)*, 2016.

[14] V. I. Utkin, *Sliding Modes in Control and Optimization*. Springer-Verlag Berlin Heidelberg, 1992.